Canbio

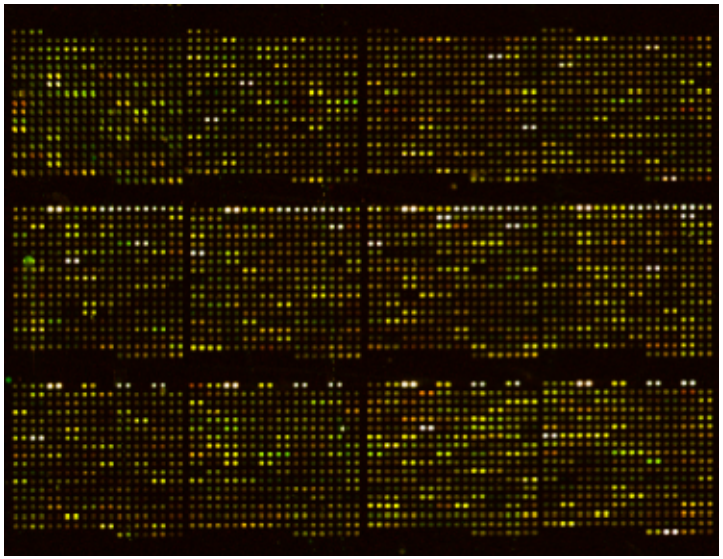
Differential expression: what after?

LSRU | LIFE SCIENCES RESEARCH UNIT

UNIVERSITÉ DU LUXEMBOURG

High-throughput data analysis

» Type: arrays, mass spec proteomics, sequencing



Statistical analysis

SYMBOL	Amean	foldchange	pvalue
ADORA3	7.12792056	0.44480534	0.00846262
GHRL	8.24047829	1.07076669	0.85660397
C21orf45	13.0753666	0.92330364	0.89128031
TMEM49	11.7636167	1.01655947	0.99883589
EIF2C2	6.99395709	1.04511855	0.96441451
TMEM161B	11.3165458	1.05526798	0.99883589
MED10	11.2615344	1.06442834	0.98588405
PHC3	7.06910377	1.0365252	0.99883589
SLC26A1	7.47149457	0.95917318	0.99883589
FOXA1	6.78519517	0.95598414	0.99883589
QKI	8.11889009	1.10117399	0.85651347
HSPA12A	6.92080912	1.00100565	0.99981939
PAIP2	13.2987896	1.04998854	0.98652252
PEX12	9.745034	1.03036326	0.99883589
KIF19	6.80986665	1.00783966	0.99883589
FKBP5	10.6355378	0.96295372	0.99883589
KCNJ2	7.04040671	0.97833506	0.99883589
PTCD1	12.7651266	1.0460091	0.99883589
RIOK2	11.3041597	1.00441562	0.99883589
EGR3	6.91669737	0.46712639	0.01098933
OSTM1	12.420112	1.09315862	0.85811565
CNTN6	7.03514811	1.041474	0.99883589
PSG11	6.74578333	0.98670973	0.99883589
SFRS17A	6.89251423	0.99662773	0.99945858
DSE	9.49447969	0.99466745	0.99883589
NLGN4Y	7.29439202	0.77455199	0.95652323
FAM83B	6.96152742	0.50994257	0.00919348
PHTF1	9.51073102	1.0173747	0.99883589
HCG4P6	7.23686146	0.87449529	0.94561853
JAM2	6.89375188	1.04486647	0.99883589
HSPA13	8.4752902	1.01480614	0.99883589
LOC157860	7.09086032	1.01008759	0.99883589
MPHOSPH8	10.6631937	1.00403233	0.99883589
SLC7A5	11.0708419	0.8505054	0.72905641
SULT1E1	6.43897318	1.06202335	0.99883589

What to do?

- Geneset analyses:
  - zoom out to see the big pattern
- Transcription factors:
  - Do they change?
  - What regulate my genes?
- MiRNAs
  - What regulate my genes?
  - Which target could be regulated
- Public data
  - Why –– where – what ?

scheme

- Aim
- Way to understand the results
- I , O
- Tools available
- Example usage

Geneset analyses

What is going on?

LSRU | LIFE SCIENCES RESEARCH UNIT

UNIVERSITÉ DU LUXEMBOURG

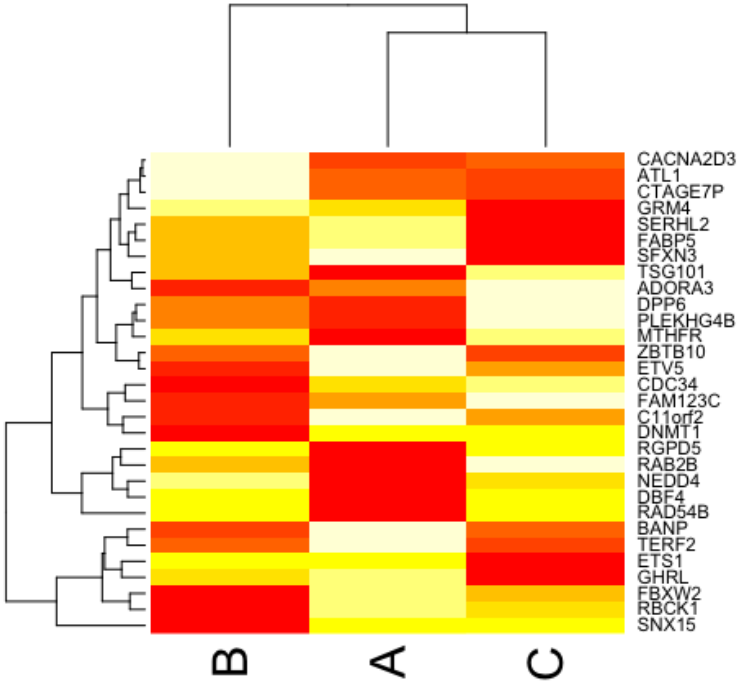
Summary

- Geneset analyses
  - General principle: robust/more general
  - Sources of genesets
  - Hypergeometric
    - What means the test
    - Importance of the background
    - Tools
  - GSEA
    - Difference in input > meaning
    - Results
    - Sample vs gene resampling
    - Tools

Summarize results

SYMBOL	Amean	foldchange	pvalue
ADORA3	7.12792056	0.44480534	0.00846262
GHRL	8.24047829	1.07076669	0.85660397
C21orf45	13.0753666	0.92330364	0.89128031
TMEM49	11.7636167	1.01655947	0.99883589
EIF2C2	6.99395709	1.04511855	0.96441451
TMEM161B	11.3165458	1.05526798	0.99883589
MED10	11.2615344	1.06442834	0.98588405
PHC3	7.06910377	1.0365252	0.99883589
SLC26A1	7.47149457	0.95917318	0.99883589
FOXA1	6.78519517	0.95598414	0.99883589
QKI	8.11889009	1.10117399	0.85651347
HSPA12A	6.92080912	1.00100565	0.99981939
PAIP2	13.2987896	1.04998854	0.98652252
PEX12	9.745034	1.03036326	0.99883589
KIF19	6.80986665	1.00783966	0.99883589
FKBP5	10.6355378	0.96295372	0.99883589
KCNJ2	7.04040671	0.97833506	0.99883589
PTCD1	12.7651266	1.0460091	0.99883589
RIOK2	11.3041597	1.00441562	0.99883589
EGR3	6.91669737	0.46712639	0.01098933
OSTM1	12.420112	1.09315862	0.85811565
CNTN6	7.03514811	1.041474	0.99883589
PSG11	6.74578333	0.98670973	0.99883589
SFRS17A	6.89251423	0.99662773	0.99945858
DSE	9.49447969	0.99466745	0.99883589
NLGN4Y	7.29439202	0.77455199	0.95652323
FAM83B	6.96152742	0.50994257	0.00919348
PHTF1	9.51073102	1.0173747	0.99883589
HCG4P6	7.23686146	0.87449529	0.94561853
JAM2	6.89375188	1.04486647	0.99883589
HSPA13	8.4752902	1.01480614	0.99883589
LOC157860	7.09086032	1.01008759	0.99883589
MPHOSPH8	10.6631937	1.00403233	0.99883589
SLC7A5	11.0708419	0.8505054	0.72905641
SULT1E1	6.43897318	1.06202335	0.99883589

Clustering: group genes according to their expression profile



Aim

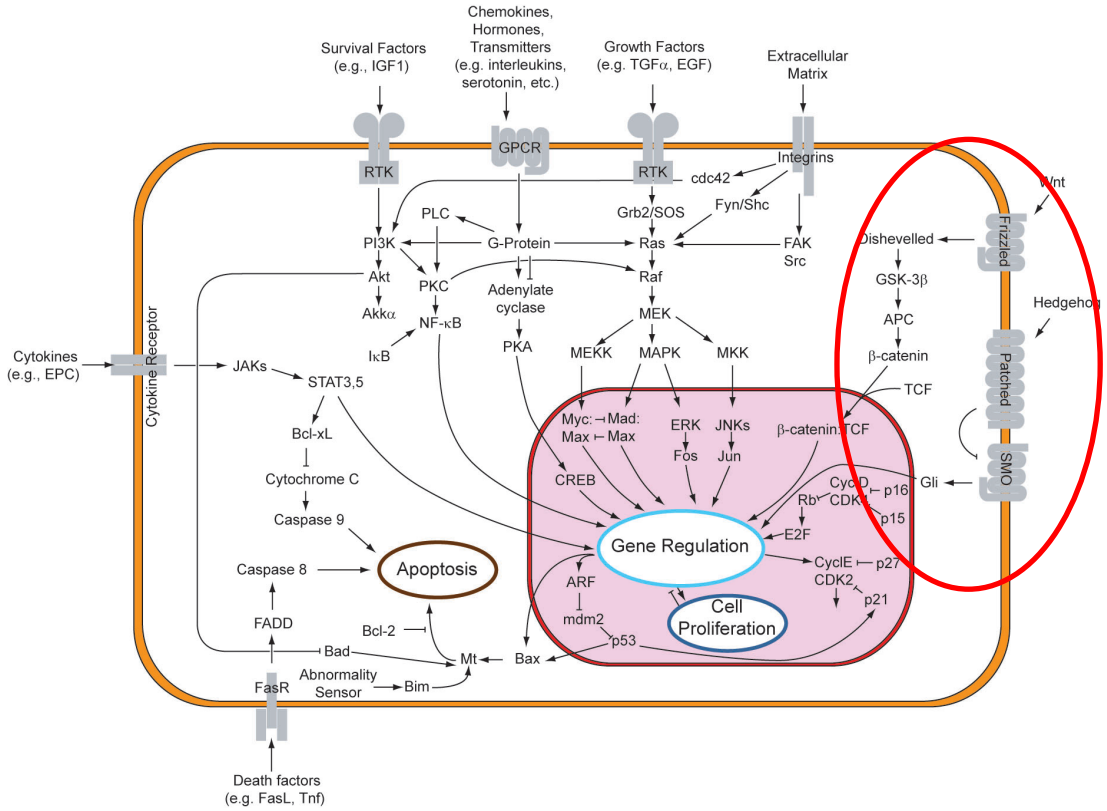
» Interpretation :

See at a glance what's going on in experiment by summarizing list of genes in a biological context

» Get more *robust* and more *subtle* results

Ideal world

SYMBOL	Amean	foldchange	pvalue
ADORA3	7.12792056	0.44480534	0.00846262
GHRL	8.24047829	1.07076669	0.85660397
C21orf45	13.0753666	0.92330364	0.89128031
TMEM49	11.7636167	1.01655947	0.99883589
EIF2C2	6.99395709	1.04511855	0.96441451
TMEM161B	11.3165458	1.05526798	0.99883589
MED10	11.2615344	1.06442834	0.98588405
PHC3	7.06910377	1.0365252	0.99883589
SLC26A1	7.47149457	0.95917318	0.99883589
FOXA1	6.78519517	0.95598414	0.99883589
QKI	8.11889009	1.10117399	0.85651347
HSPA12A	6.92080912	1.00100565	0.99981939
PAIP2	13.2987896	1.04998854	0.98652252
PEX12	9.745034	1.03036326	0.99883589
KIF19	6.80986665	1.00783966	0.99883589
FKBP5	10.6355378	0.96295372	0.99883589
KCNJ2	7.04040671	0.97833506	0.99883589
PTCD1	12.7651266	1.0460091	0.99883589
RIOK2	11.3041597	1.00441562	0.99883589
EGR3	6.91669737	0.46712639	0.01098933
OSTM1	12.420112	1.09315862	0.85811565
CNTN6	7.03514811	1.041474	0.99883589
PSG11	6.74578333	0.98670973	0.99883589
SFRS17A	6.89251423	0.99662773	0.99945858
DSE	9.49447969	0.99466745	0.99883589
NLGN4Y	7.29439202	0.77455199	0.95652323
FAM83B	6.96152742	0.50994257	0.00919348
PHTF1	9.51073102	1.0173747	0.99883589
HCG4P6	7.23686146	0.87449529	0.94561853
JAM2	6.89375188	1.04486647	0.99883589
HSPA13	8.4752902	1.01480614	0.99883589
LOC157860	7.09086032	1.01008759	0.99883589
MPHOSPH8	10.6631937	1.00403233	0.99883589
SLC7A5	11.0708419	0.8505054	0.72905641
SULT1E1	6.43897318	1.06202335	0.99883589



How: what you need

- Results from DE

- Boxes of genes (Genesets)
  - Pathways
  - GO ontologies
  - Signatures from papers

Geneset databases

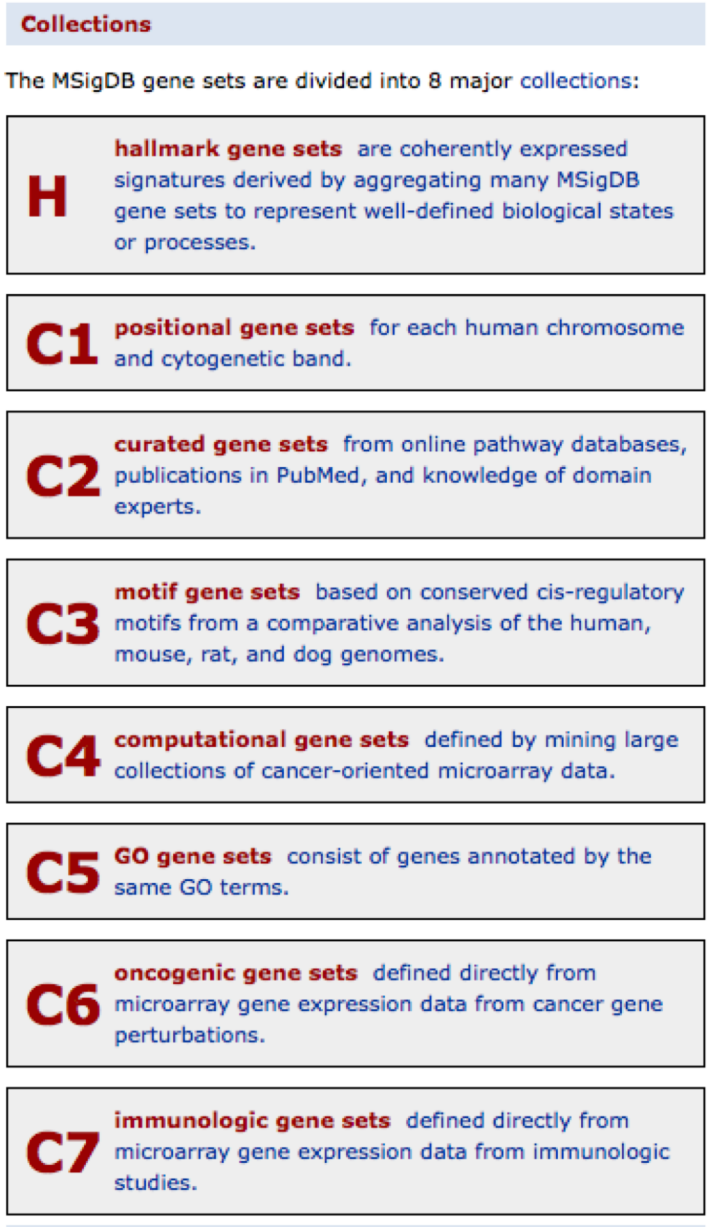
- Curation
- Multiplicity
- Redundancy

Curation, update rate

- GO:
  Different qualities of annotation
  (experimental >> electronically inferred)
- Manually curated collections (Genego, IPA)
- KEGG, Reactome, Biocarta etc:
  updated - but indirect sources can be frozen
- MsigDB
  collections from the Broad institute

MsigDB

- Collections to help focusing
- Maintained and updated
- GO: tried to reduce redundancy



http://software.broadinstitute.org/gsea/msigdb/collections.jsp

Geneset databases

» Multiplicity:
  – pathways, ontologies, protein interaction, miRNA predicted targets, …
  – E.g. pathways: KEGG, Biocarta, Reactome, …

Geneset databases

- Multiplicity
- Redundancy

How to handle redundancy

» Group overlapping categories in clusters (DAVID)

Clusterise results

477 Cluster(s) — Download File

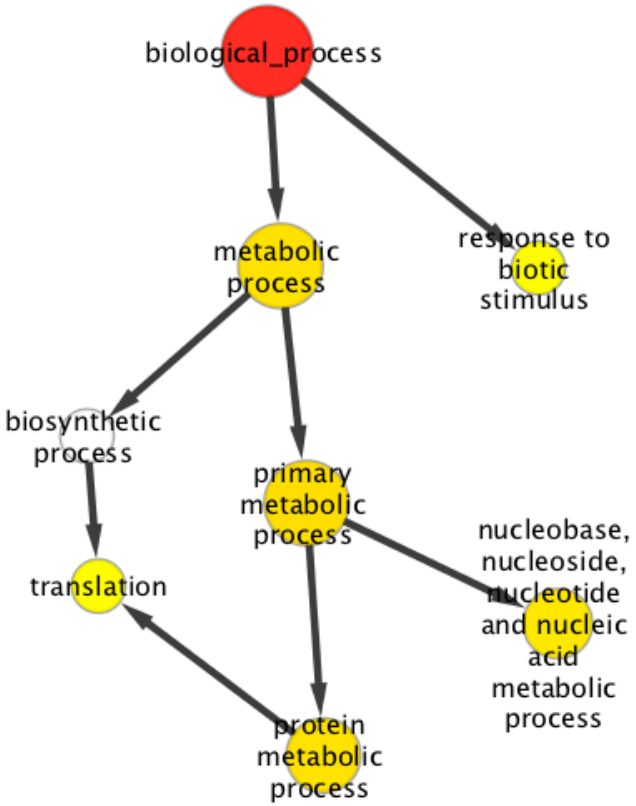
	Annotation Cluster 1	Enrichment Score: 15.55	G		Count	P_Value	Benjamini
	GOTERM_BP_FAT	response to organic substance	RT		388	6.2E-24	5.2E-20
	GOTERM_BP_FAT	cellular response to chemical stimulus	RT		355	8.7E-20	3.7E-16
	GOTERM_BP_FAT	cellular response to organic substance	RT		301	3.3E-18	5.6E-15
	GOTERM_BP_FAT	response to oxygen-containing compound	RT		211	2.5E-14	2.1E-11
	GOTERM_BP_FAT	response to endogenous stimulus	RT		178	4.1E-5	1.6E-3
	Annotation Cluster 2	**Enrichment Score: 8.45**	G		**Count**	**P_Value**	**Benjamini**
	GOTERM_BP_FAT	response to interferon-gamma	RT		42	1.6E-10	4.1E-8
	GOTERM_BP_FAT	cellular response to interferon-gamma	RT		35	7.9E-9	1.2E-6
	GOTERM_BP_FAT	interferon-gamma-mediated signaling pathway	RT		25	3.4E-8	4.3E-6
	Annotation Cluster 3	**Enrichment Score: 7.98**	G		**Count**	**P_Value**	**Benjamini**
	GOTERM_BP_FAT	cell death	RT		250	4.6E-11	1.5E-8
	GOTERM_BP_FAT	programmed cell death	RT		239	4.9E-11	1.5E-8
	GOTERM_BP_FAT	regulation of apoptotic process	RT		189	2.1E-10	5.1E-8
	GOTERM_BP_FAT	apoptotic process	RT		225	2.5E-10	5.7E-8
	GOTERM_BP_FAT	regulation of programmed cell death	RT		190	2.6E-10	5.8E-8
	GOTERM_BP_FAT	regulation of cell death	RT		199	5.0E-10	9.9E-8
	GOTERM_BP_FAT	apoptotic signaling pathway	RT		93	7.8E-9	1.2E-6
	GOTERM_BP_FAT	positive regulation of programmed cell death	RT		87	3.2E-7	2.9E-5
	GOTERM_BP_FAT	positive regulation of cell death	RT		90	4.1E-7	3.5E-5
	GOTERM_BP_FAT	positive regulation of apoptotic process	RT		86	4.3E-7	3.6E-5
	GOTERM_BP_FAT	negative regulation of apoptotic process	RT		113	4.8E-7	4.0E-5
	GOTERM_BP_FAT	negative regulation of programmed cell death	RT		113	9.4E-7	6.9E-5
	GOTERM_BP_FAT	negative regulation of cell death	RT		117	5.2E-6	2.9E-4
	Annotation Cluster 4	**Enrichment Score: 7.93**	G		**Count**	**P_Value**	**Benjamini**
	GOTERM_BP_FAT	defense response	RT		231	6.8E-19	1.9E-15
	GOTERM_BP_FAT	response to cytokine	RT		145	3.3E-18	6.9E-15
	GOTERM_BP_FAT	cytokine-mediated signaling pathway	RT		111	2.7E-17	3.7E-14
	GOTERM_BP_FAT	immune response	RT		228	1.8E-16	2.3E-13

How to handle redundancy

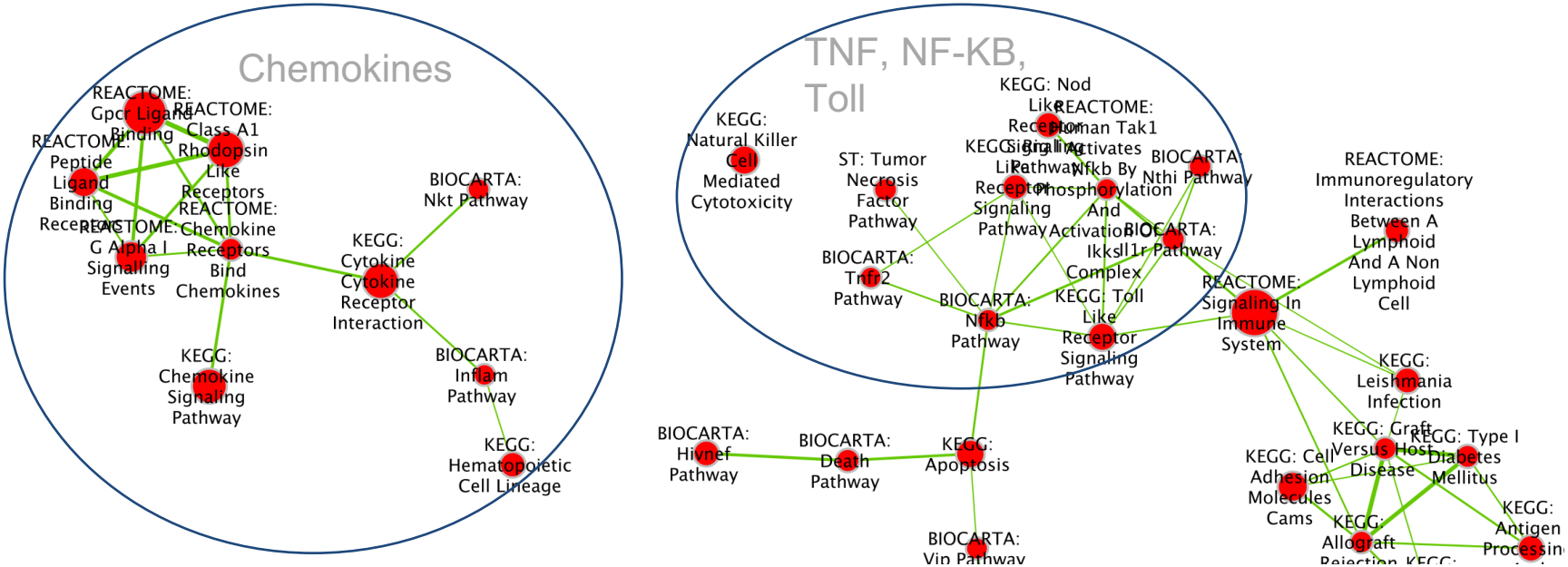
- Group overlapping categories in clusters (DAVID)
- Visualize redundancy among gene sets (BinGO, Enrichment map)

Visualize redundancy

Bingo



Pathways impacted by TNF



How to handle redundancy

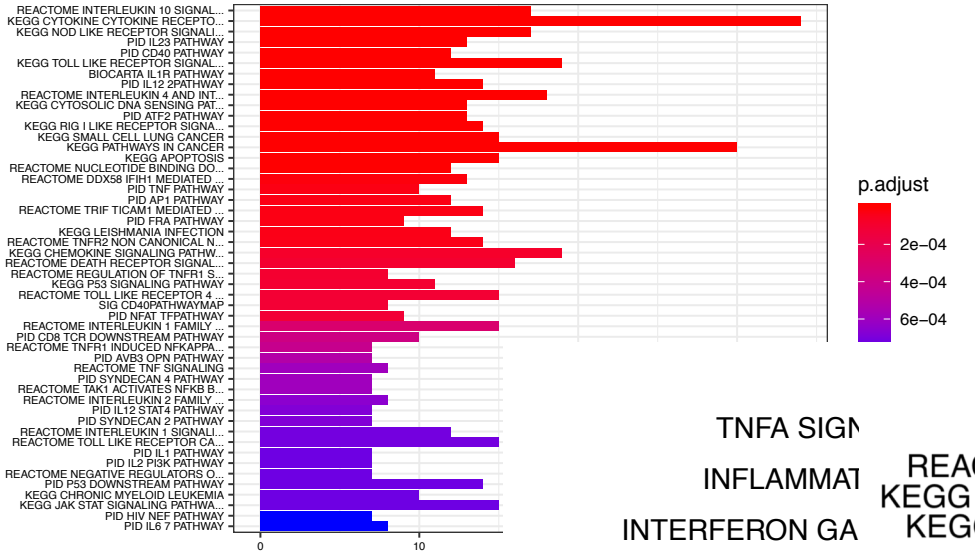
- Group overlapping categories in clusters (DAVID)
- Visualize redundancy among gene sets (BinGO, Enrichment map)
- Reduce : hallmarks: list of 50 comprehensive genesets

Hallmarks

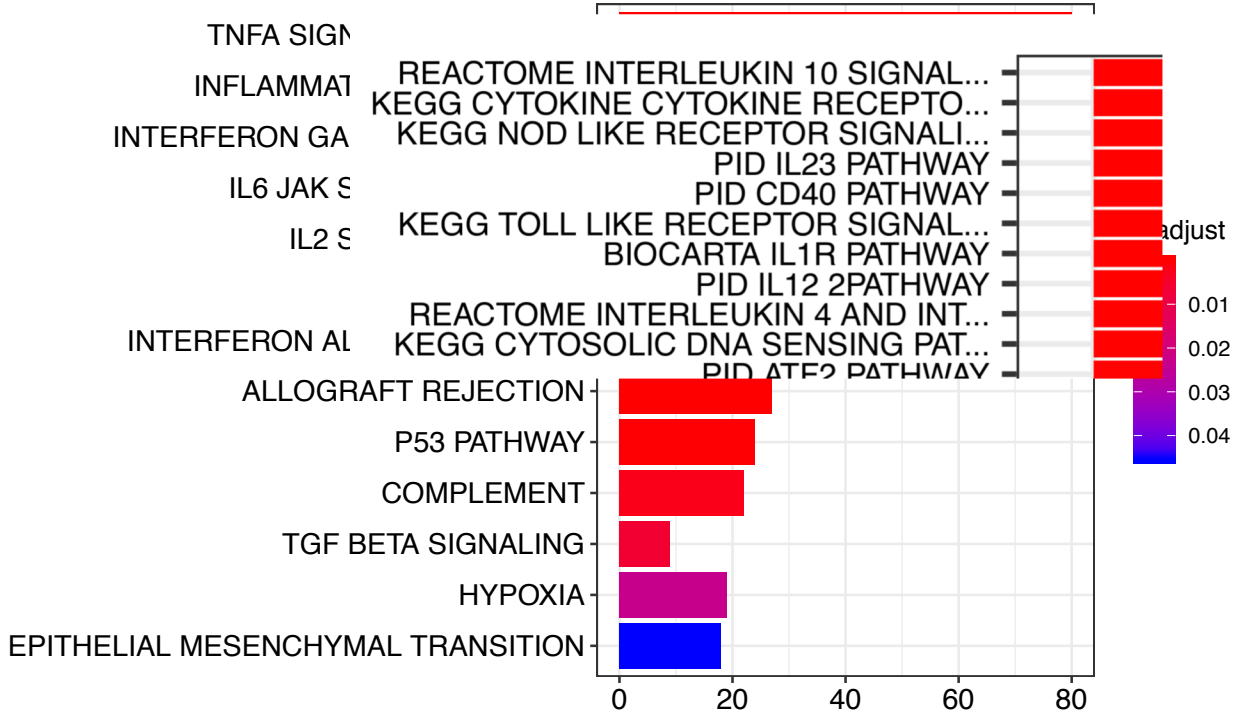
- Combine computational and manual curation
- Reduce
  - redundancy between and
  - variability (of expression) within genesets
- List of 50 clear biological themes

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707969/

Hallmarks



Pathways collections



Hallmarks

How: what you need

- Boxes of genes (Genesets)
- Results from DE

How: 2 main methods types

Gene list-based

- Use only significant gene list
- Only based on the names, not on the change direction or amplitude
- Idea is to look for enrichment

Keyword: over-representation

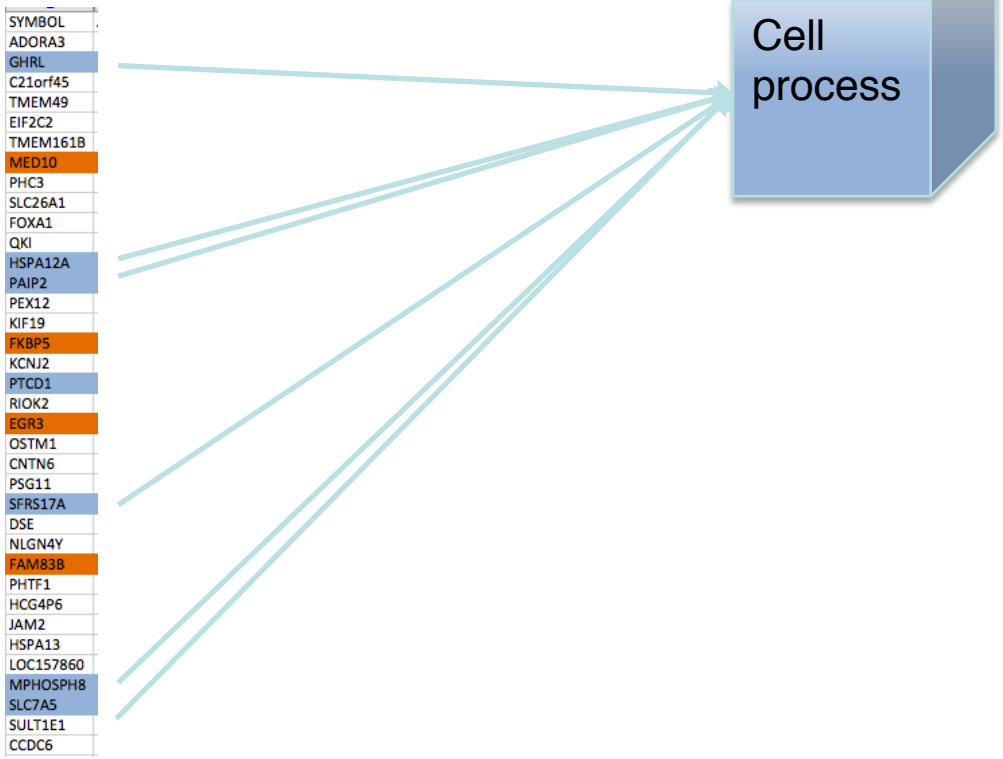
Whole-results

- Use results from whole array
- Use expression change and direction values
- Idea is to test for a “metagene”

Keyword: GSEA

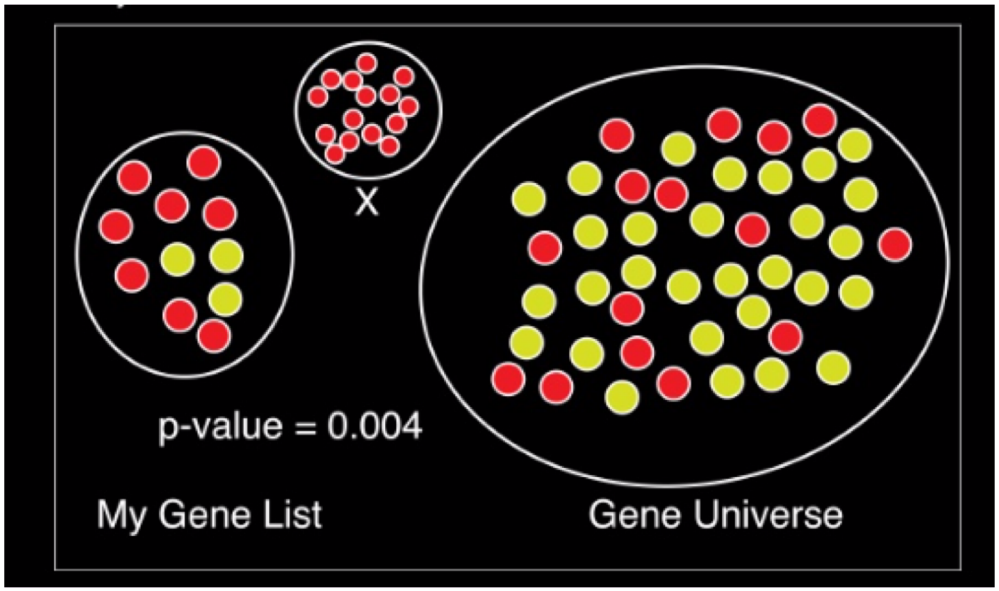
Gene list based methods

gene list



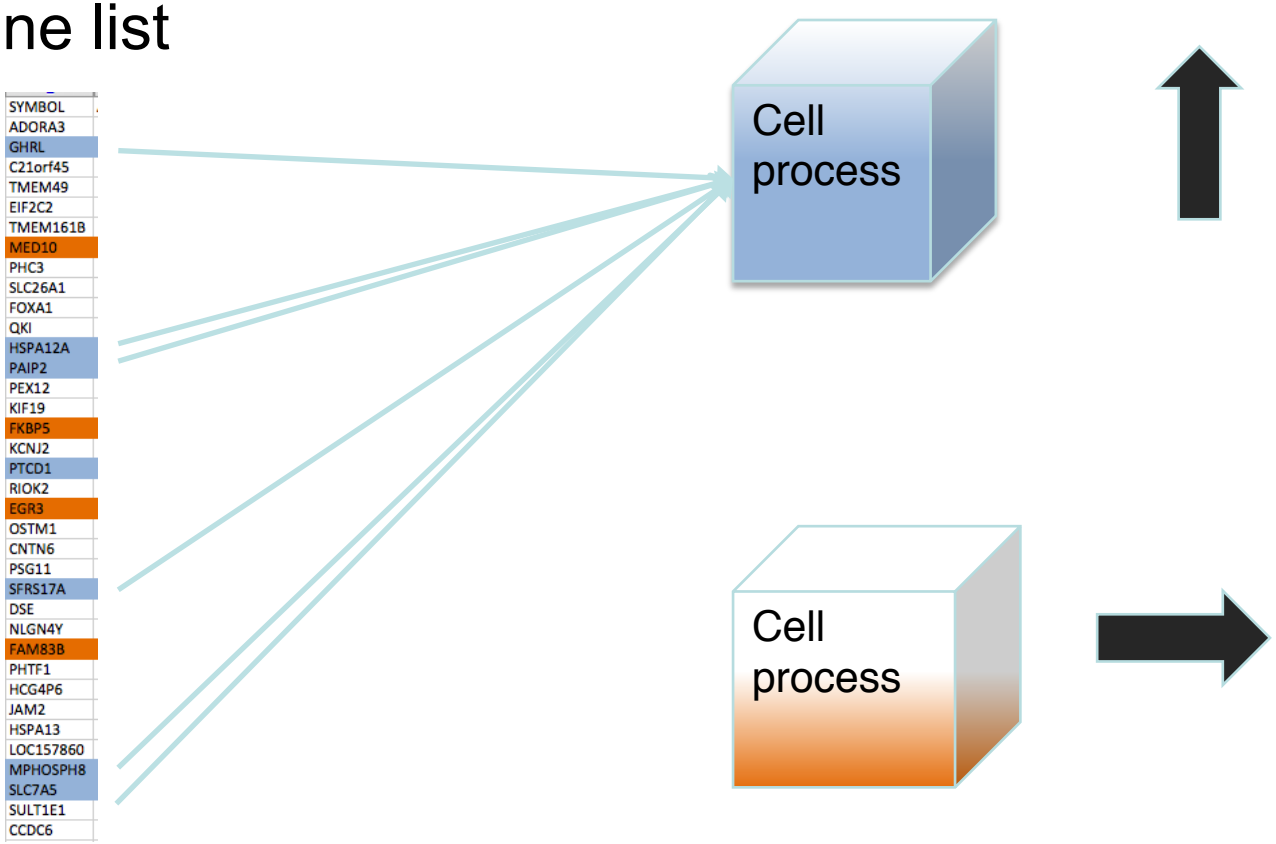
hypergeometric test

Hypergeometric test



Gene list based methods

gene list



Gene list-based enrichment

- (+) Highlight top genes functions
- (-) Small gene list give little results
- (-) Groups can be consistently regulated but their small change doesn't pass the thresholds

Gene list based - I/O

- Input:
  - Gene names from differential expression
  - Background of all genes
  - Genesets : groups of genes
- Output:
  - Enriched genesets
  - (adjusted)p.value
  - number and name of DE genes included

Tools

» Online:
  – DAVID https://david.ncifcrf.gov

» Dedicated applications: Ingenuity, Metacore

» R/Rstudio: clusterprofiler, fry() from edgeR

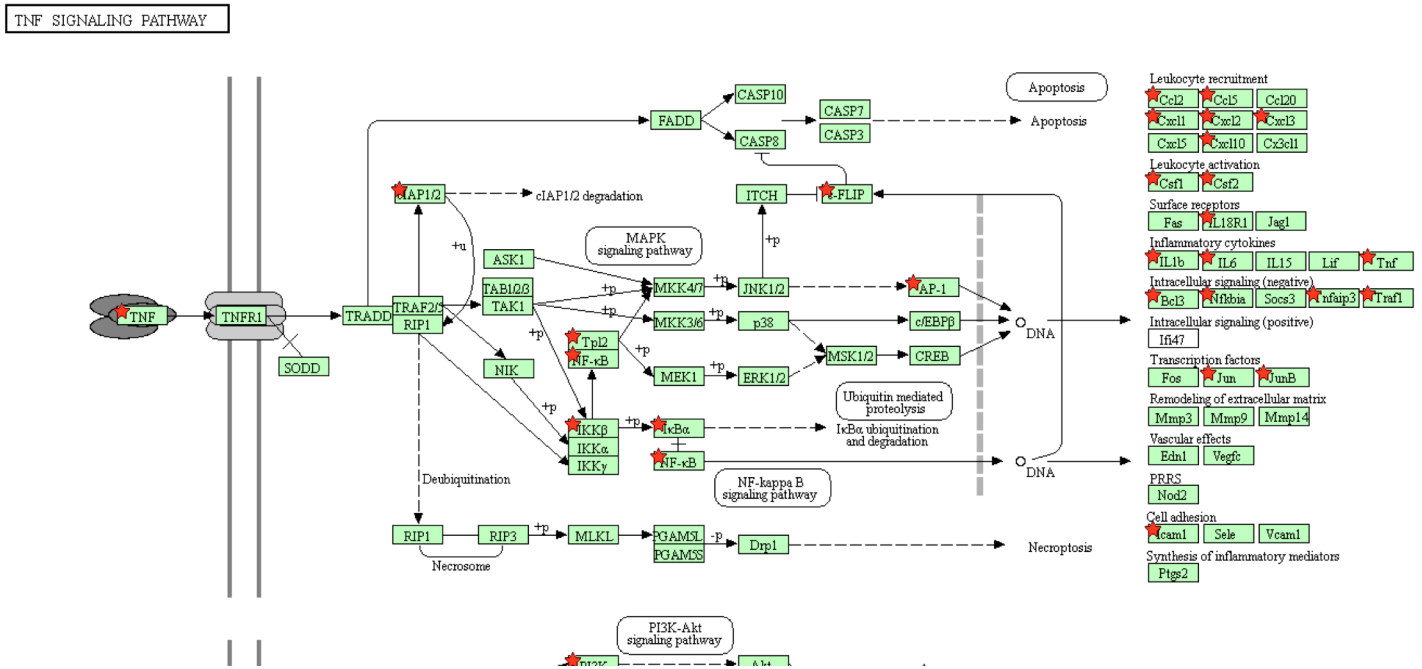
DAVID

- Online web tool, free
- Steps
  - Import gene list
  - set species
  - Define background
  - Select the genesets to test

DAVID

160 chart records

Sublist	Category	Term	RT	Genes
☐	KEGG_PATHWAY	TNF signaling pathway	RT	2
☐	KEGG_PATHWAY	NF-kappa B signaling pathway	RT	2
☐	KEGG_PATHWAY	Measles	RT	2
☐	KEGG_PATHWAY	Influenza A	RT	2
☐	KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT	3
☐	KEGG_PATHWAY	NOD-like receptor signaling pathway	RT	1
☐	KEGG_PATHWAY	HTLV-I infection	RT	3
☐	KEGG_PATHWAY	Toll-like receptor signaling pathway	RT	2



DAVID

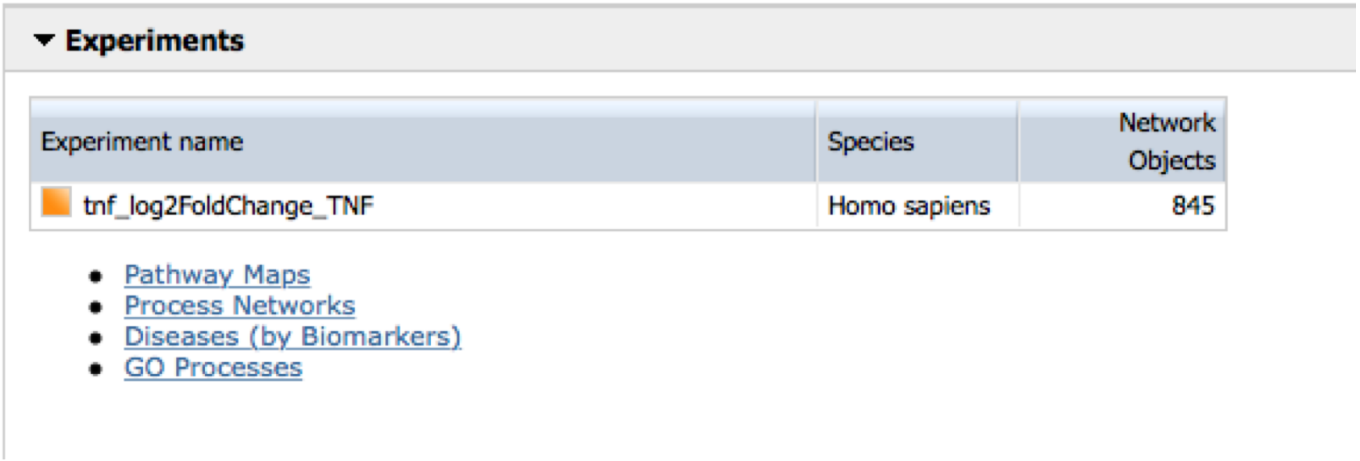
- (+) Easy to use
- (+) Lot of available public genesets (KEGG, GO, PID, PFAM)
- (+) Output: (clusterized) significant categories, gene info, pathways maps
- (-) Not much more visualization

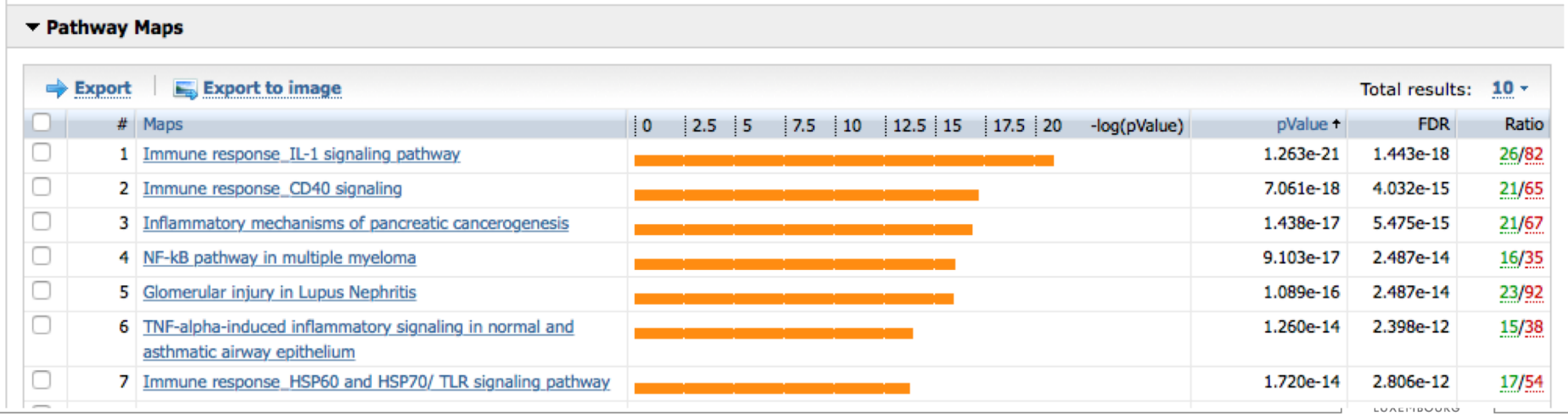
Genego / IPA

- Dedicated tool
- App or online web access
- Not free
- Steps:
  - Import table of DE results
  - set species
  - set thresholds

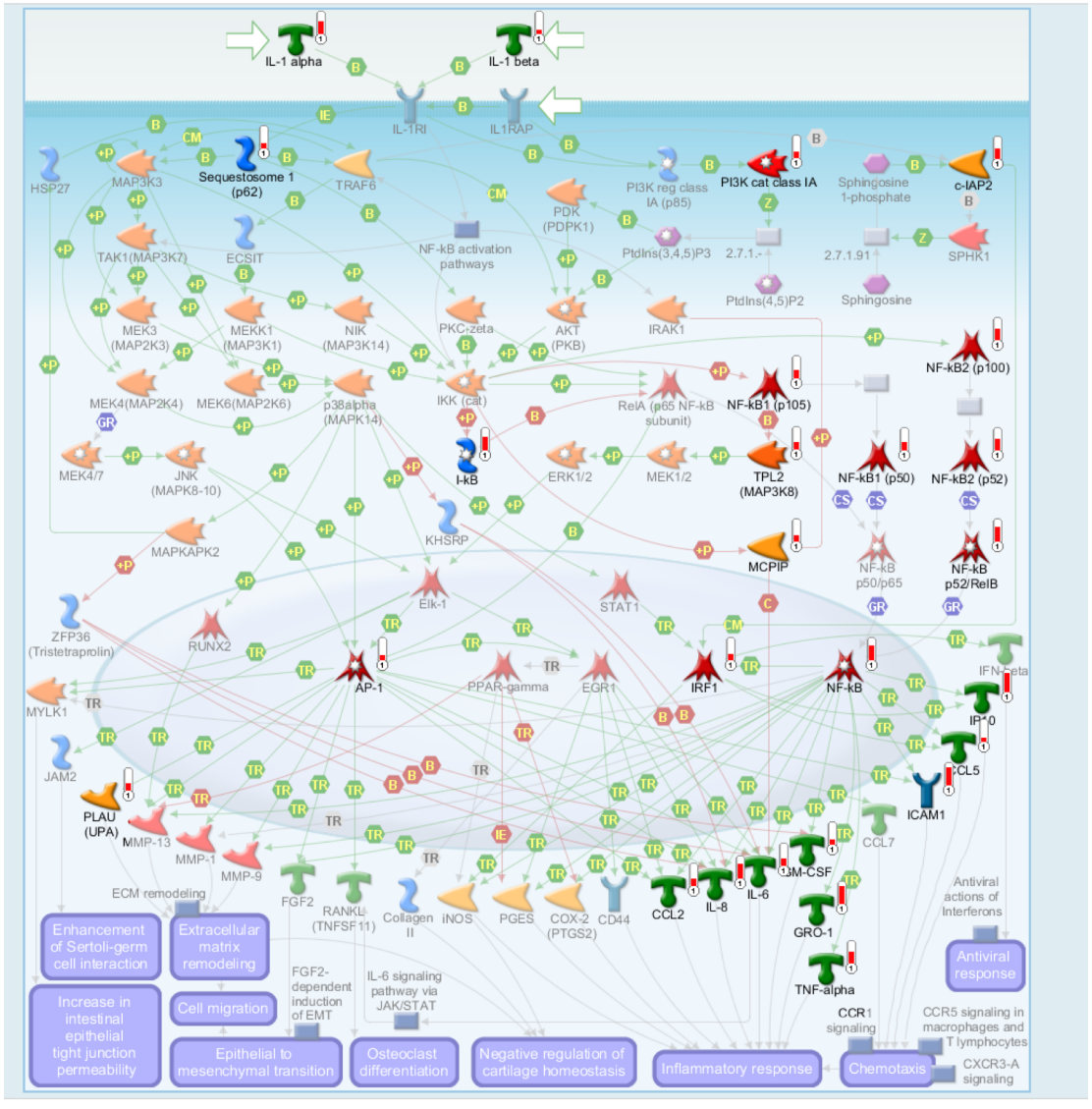
Genego / IPA

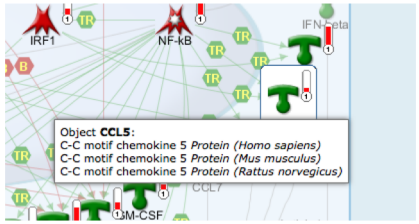
» results





Genego / IPA





LSRU | LIFE SCIENCES RESEARCH UNIT

UNIVERSITÉ DU LUXEMBOURG

Genego / IPA

- (+) Lots of all in one analysis
- (+) Curated collection of pathways
- (-) Not free
- (-) Pathways but not much visualization

Clusterprofiler

- R /R studio
- Can use several sources of geneset, including MsigDB
- Both GSEA / genelist based

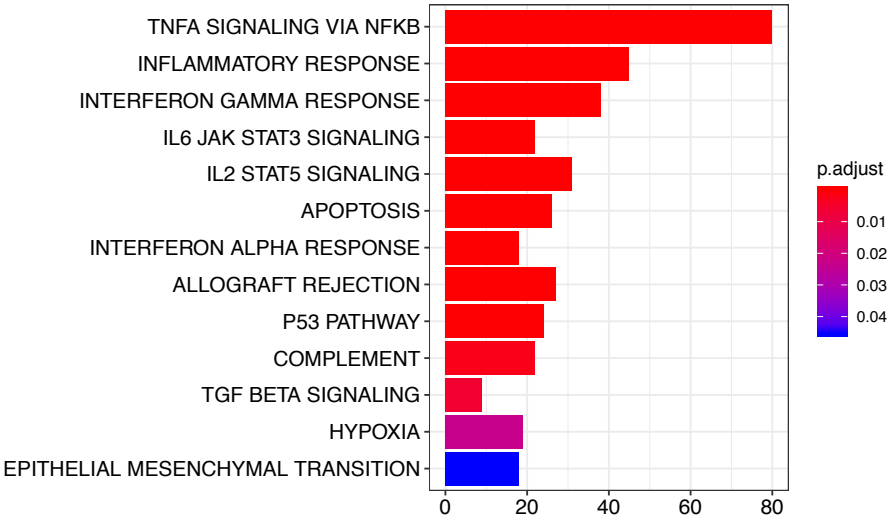
Example with clusterprofiler

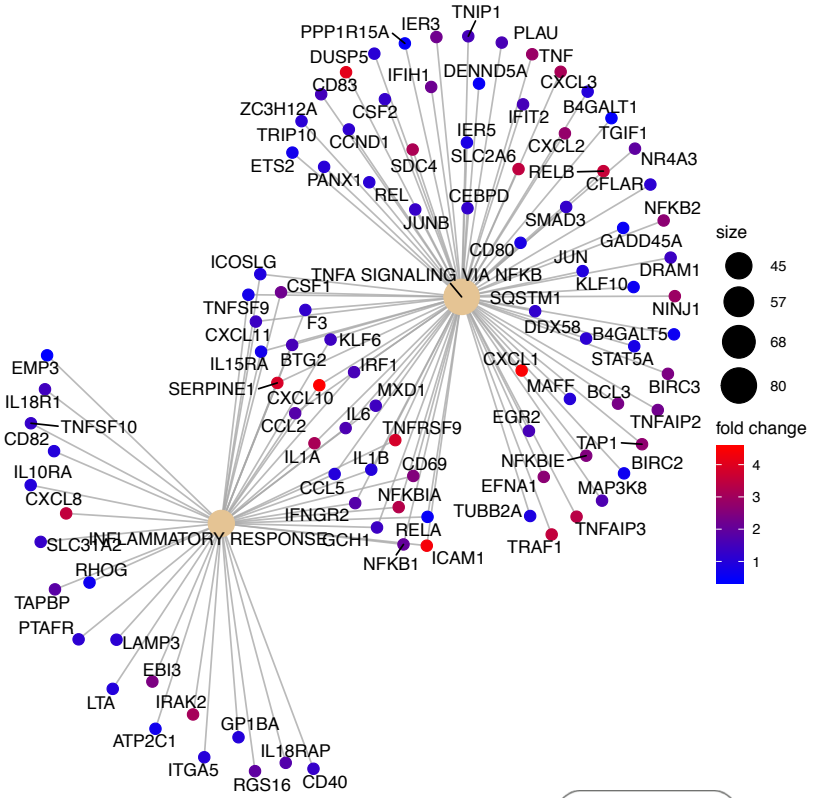
» Steps:
  – list of up-regulated genes from DE object
  – Select your geneset source
  – Select your method (over-representation test)
  – Parameters: min and max size

```
```{r}
tup <- enricher(mydf %>% filter(source=="TNF_up") %>% pull(gene),
 universe = unique(hallmarks$gene),
  minGSSize = 30,
  TERM2GENE = hallmarks,
  pvalueCutoff = 0.05, pAdjustMethod = "BH",
  qvalueCutoff = 1)
tup #13 terms enriched
```
```

Example with clusterprofiler

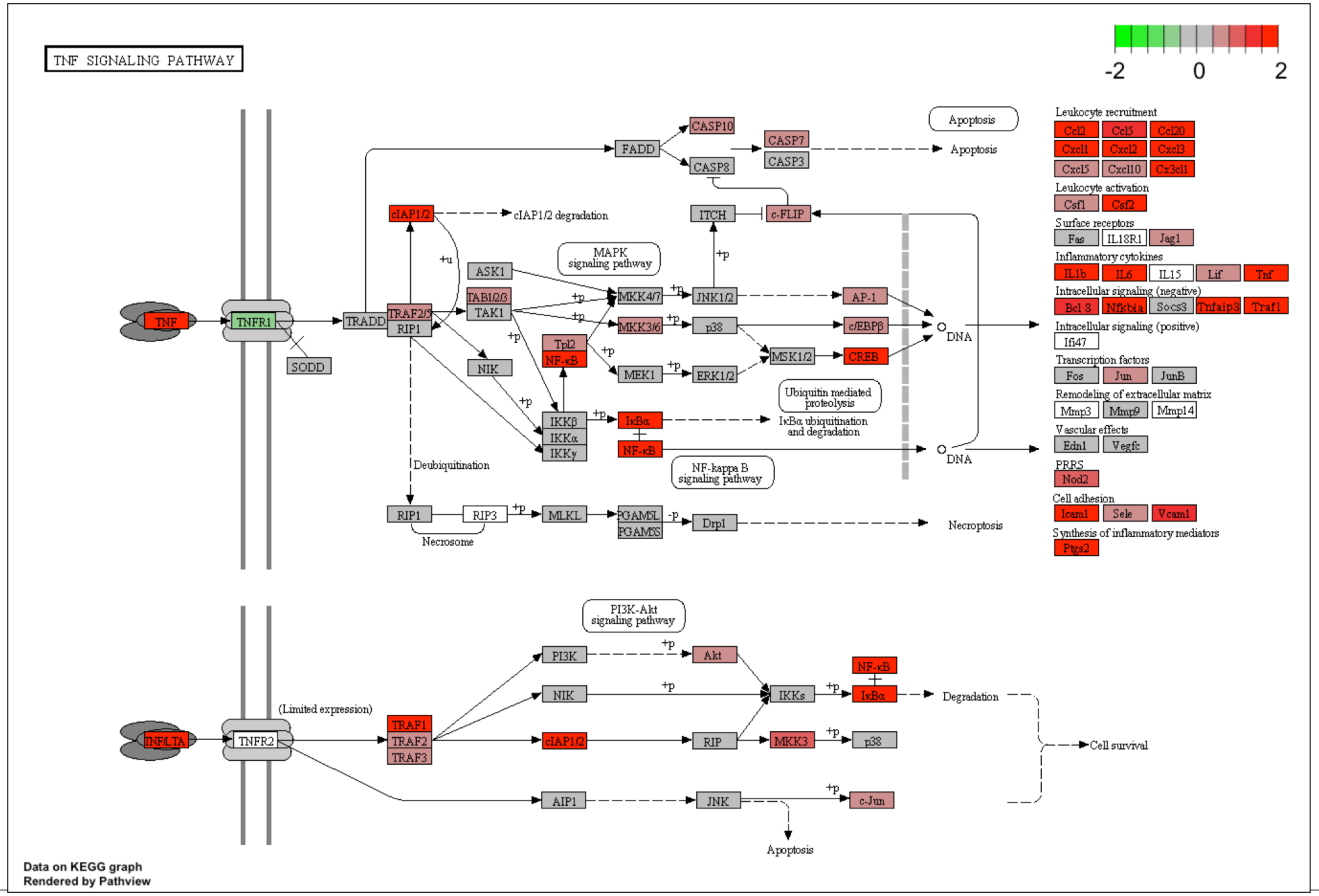
» Results





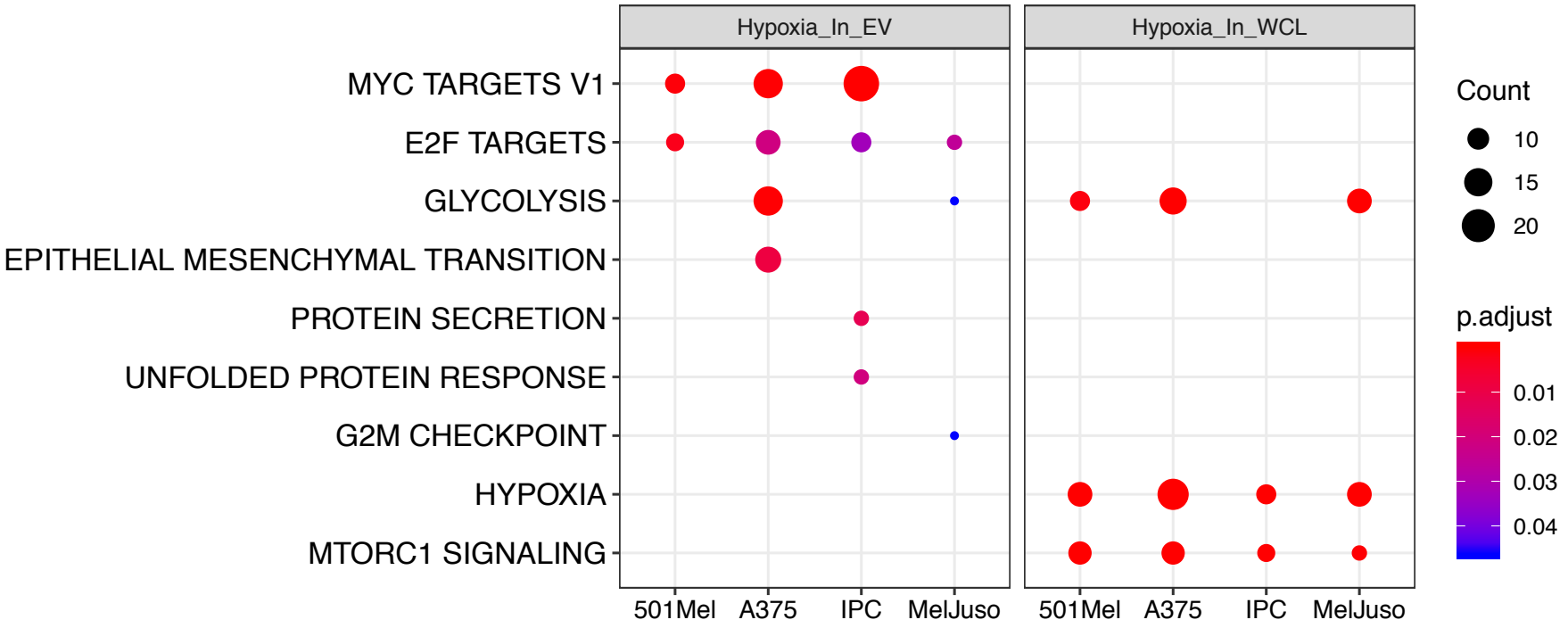
Example with clusterprofiler

» View pathway (KEGG)



Example with clusterprofiler

» Compare results of several differences



clusterprofiler

- (+,-) line coding
- (+) upstream and downstream integration
- (+) Can use several sources of geneset, including MsigDB
- (+) Both GSEA / genelist based
- (+) Several visualisations including networks
- (-) only KEGG pathways can be drawn

How: 2 main methods types

Gene list-based

- Use only significant gene list
- Only based on the names, not on the change direction or amplitude
- Idea is to look for enrichement

Keyword: over-representation

Whole-results

- Use results from whole array
- Use expression change and direction values
- Idea is to test for a “metagene”

Keyword: GSEA

Whole results methods

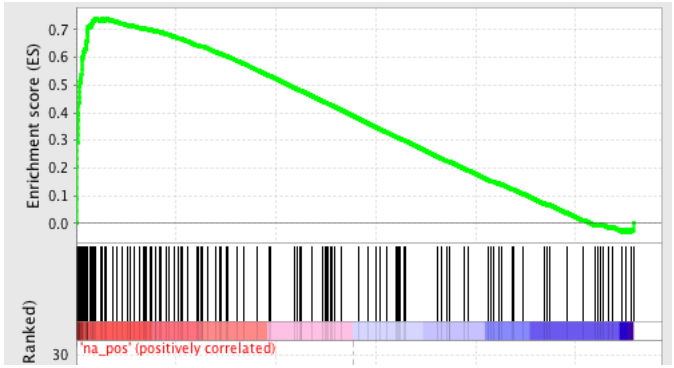
All genes

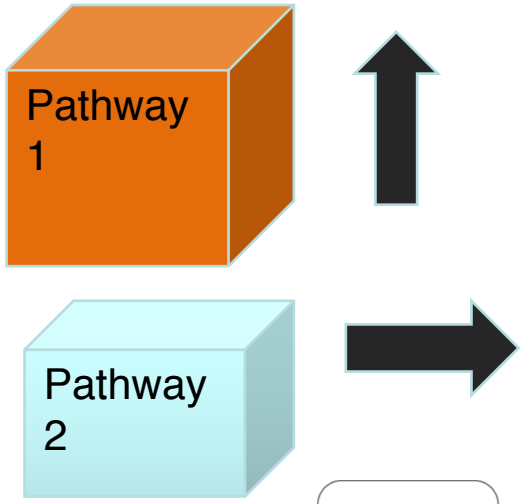
SYMBOL	pvalue
ADORA3	0.00846262
GHRL	0.85660397
C21orf45	0.89128031
TMEM49	0.99883589
EIF2C2	0.96441451
TMEM161B	0.99883589
MED10	0.98588405
PHC3	0.99883589
SLC26A1	0.99883589
FOXA1	0.99883589
QKI	0.85651347
HSPA12A	0.99981939
PAIP2	0.98652252
PEX12	0.99883589
KIF19	0.99883589
FKBP5	0.99883589
KCNJ2	0.99883589
PTCD1	0.99883589
RIOK2	0.99883589
EGR3	0.01098933
OSTM1	0.85811565
CNTN6	0.99883589
PSG11	0.99883589
SFRS17A	0.99945858
DSE	0.99883589
NLGN4Y	0.95652323
FAM83B	0.00919348
PHTF1	0.99883589
HCG4P6	0.94561853

Rank genes

SYMBOL	pvalue
ADORA3	0.00846262
C21orf45	0.89128031
CNTN6	0.99883589
DSE	0.99883589
EGR3	0.01098933
EIF2C2	0.96441451
FAM83B	0.00919348
FKBP5	0.99883589
FOXA1	0.99883589
GHRL	0.85660397
HCG4P6	0.94561853
HSPA12A	0.99981939
JAM2	0.99883589
KCNJ2	0.99883589
KIF19	0.99883589
MED10	0.98588405
NLGN4Y	0.95652323
OSTM1	0.85811565
PAIP2	0.98652252
PEX12	0.99883589
PHC3	0.99883589
PHTF1	0.99883589
PSG11	0.99883589
PTCD1	0.99883589
QKI	0.85651347
RIOK2	0.99883589
SFRS17A	0.99945858
SLC26A1	0.99883589
TMEM161B	0.99883589
TMEM49	0.99883589

Score Genesets





LSRU | LIFE SCIENCES RESEARCH UNIT

UNIVERSITÉ DU LUXEMBOURG

Whole-results methods

- (+) Can detect smaller but consistent changes in expression
- (-, +) Do not focus on top genes
- (-) Results are prone to variability due to permutations procedure

Self contained / competitive

- P.values estimated by permutation
- Can shuffle either genes or samples
  - Shuffle genes: estimate the value of your geneset w.r.t. random genesets. (Scale)
  - Shuffle samples: estimate the value of your geneset w.r.t if there was no phenotype.

Whole results methods - I/O

- Input:
  - Genesets
  - All genes with name and way to rank them / or preranked list
  - No need of background
- Output:
  - Enriched genesets
  - (adjusted)p.value
  - Enrichment score

Tools

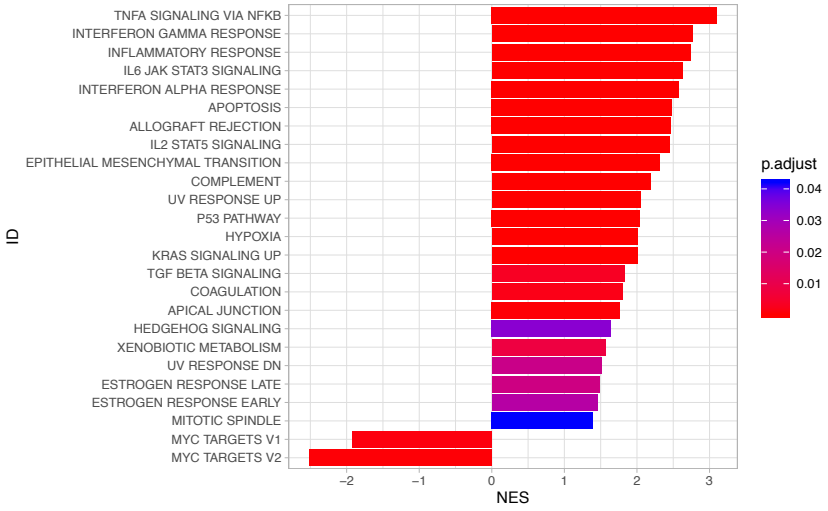
- Online: ?
- GSEA (stand alone)
- R/Rstudio: several packages :
  - R-GSEA (official, old implementation),
  - GSA,
  - clusterprofiler,
  - ROAST (in limma),
  - camera (in limma)

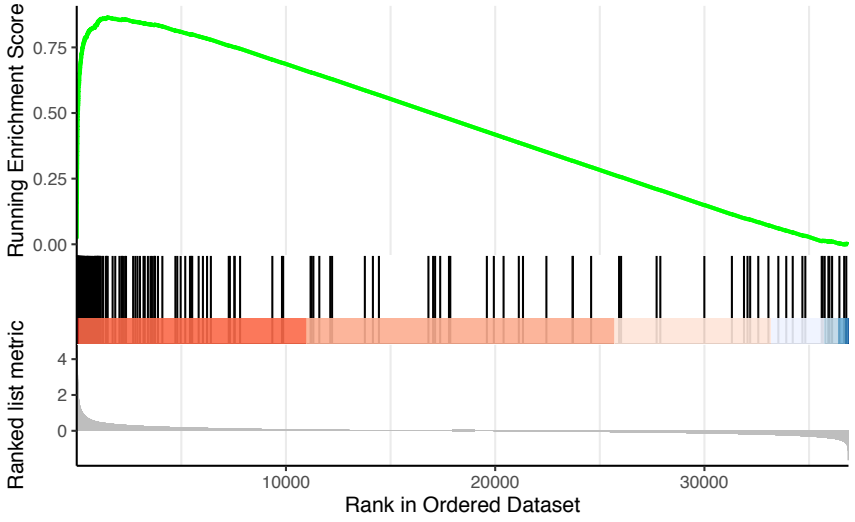
GSEA with clusterprofiler

- TNF effect, MsigDb
- Steps:
  - build list of ranked FC for all genes
  - Define your genesets

GSEA with clusterprofiler

	setSize <int>	enrichmentScore <dbl>	NES <dbl>
TNFA SIGNALING VIA NFKB	196	0.8648216	3.100454
P53 PATHWAY	195	0.5696201	2.039947
EPITHELIAL MESENCHYMAL TRANSITION	193	0.6459036	2.309556
IL2 STAT5 SIGNALING	193	0.6828830	2.441783
INTERFERON GAMMA RESPONSE	193	0.7710122	2.756906
HYPOXIA	190	0.5612657	2.001189





LSRU | LIFE SCIENCES RESEARCH UNIT

UNIVERSITÉ DU LUXEMBOURG

Summary genesets / Points to consider

- Similar results both
- Pathway analysis by hypergeometric: importance of the background
- GSEA like: hypothesis /permutation

Points to consider

- Importance of a clear upstream hypothesis helps downstream interpretation
- Combine if necessary
- Gene expression is not protein function (genesets contains both)
- Annotation mainly for Coding genes

Transcription factor

What can regulate our genes?

LSRU | LIFE SCIENCES RESEARCH UNIT

UNIVERSITÉ DU LUXEMBOURG

Summary

- TF differently expressed in the data
- TF whose targets are enriched in DE results
- Intersect: differently expressed TF, whose targets are enriched in DE results

Expression of TF present in the data

- Get a TF list: annotation ressources
- Slice data and represent expression

Annotation ressources

» Lists of TFs: most sources very large and include co-factors
  – GO ontology terms : need filtering (evidences)
  – TFcheckpoint: filters, curation (DNA binding and exp. evidence)
» Lists of interactions
  – Encode: Based on Chip-seq data, into cell lines
  – RegNetwork: several DBs, filters
  – TRRUST based on manual literature curation
» Matrices (TFBS enrichment)
  – JASPAR
  – TRANSFAC
  – PAZAR

- Low overlap
- Check updates /maintenance

Expression of TF present in the data

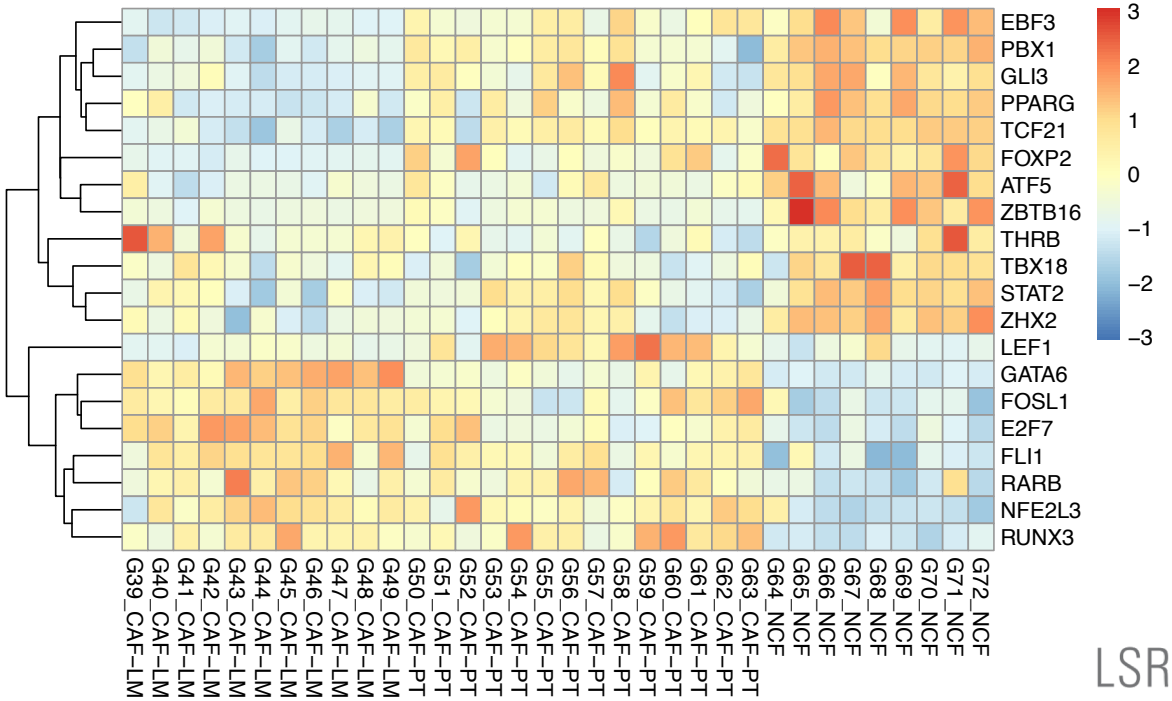
» Get list

```
```{r}
tf_checkpoint <-  read_tsv('http://www.tfcheckpoint.org/data/TFCheckpoint_download_180515.txt')
```
We filter
```{r}
TF <- tf_checkpoint %>%
  filter(DbTF=='yes',evidence.type=='Experimental') %>%
  select(gene_symbol, gene_name)
```
```

Expression of TF present in the data

- Filter results
- Graph

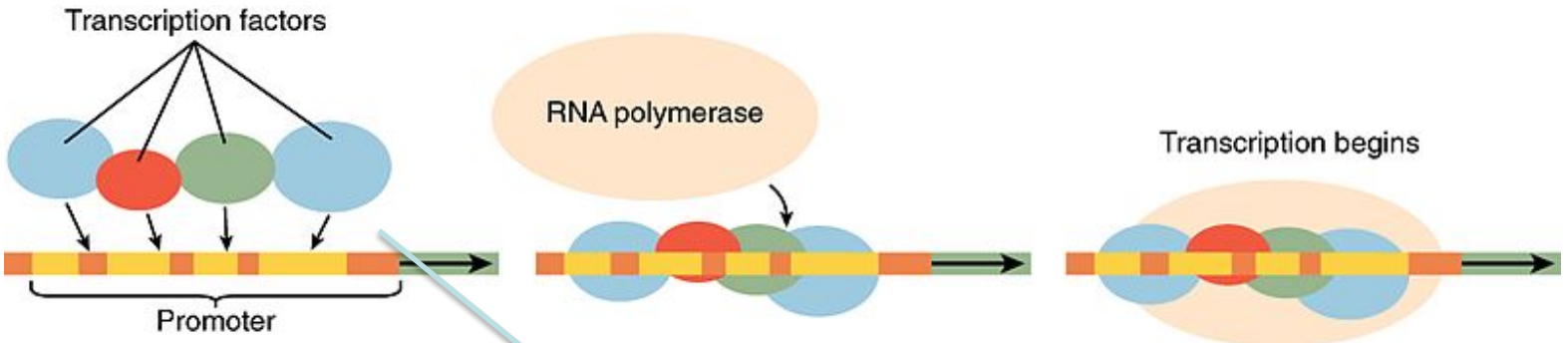
```
p <- all_results %>%
  right_join(TF %>% select(Symbol=gene_symbol)) %>%
  filter(signif != 0) %>%
  select(Symbol, starts_with("G")) %>%
 column_to_rownames("Symbol") %>%
  as.matrix() %>%
  pheatmap::pheatmap(.,scale="row", cluster_cols = F)
```



TF whose targets are enriched in DE results

- What can regulate our genes?
- Principle of binding matrices
- Based again on hypergeometric test – why

Principle of TFBS enrichment



Position weight matrix

GATA1

https://pediaa.com/how-do-transcription-factors-bind-to-dna/
http://jaspar.genereg.net/matrix/MA0035.1/

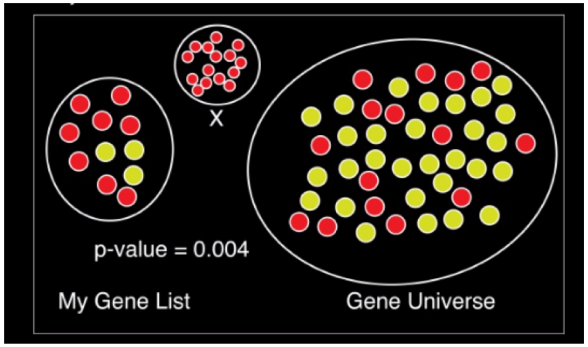
Principle of TFBS enrichment

Matrix list

For each matrix

Gene list

Hypergeometric



TF with over-represented TFBS in this gene list

- TF1
- TF2
- TF3

Ref paper: opossum https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1142402/

input

- List of DE genes (Symbol, EntrezID, RefSeq)
- Annotation: matrices
  - Easier when already inside tool

Tools

- Webtools:
  - Pscan
  - Lasagna
  - Opposum
- Cytoscape: iRegulon
- R/Rstudio: RcisTarget (iRegulon team)

Example Pscan

» Input: list of DE genes: need RefSeq

» ID translation : biomaRt

````
```{r}
library(biomaRt)
ensembl = useDataset("hsapiens_gene_ensembl",mart=useMart("ensembl"))
# searchAttributes(mart = ensembl, pattern = "refseq")
# searchAttributes(mart = ensembl, pattern = "symbol")
```

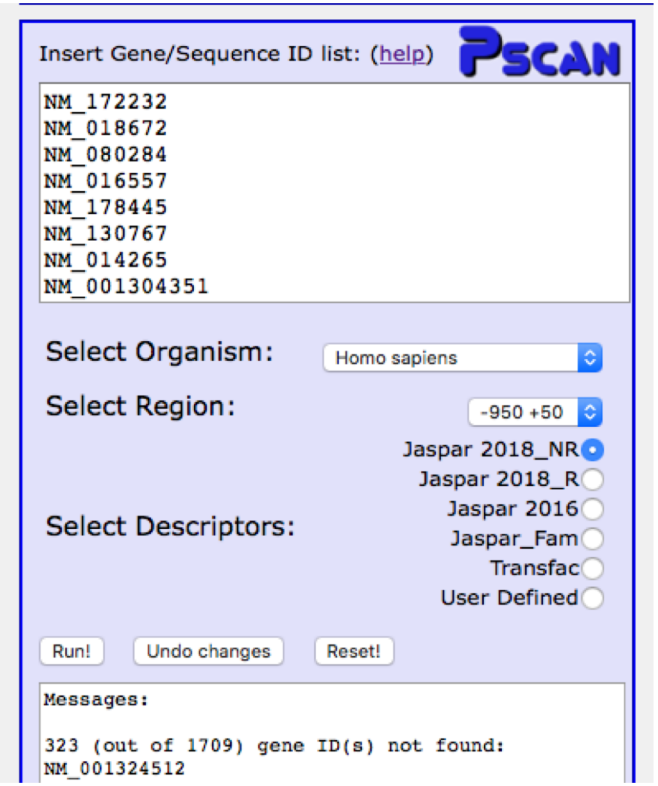
prepare list of significant genes and Get refseq from biomart
```{r}
pscan_input <- all_results %>%
  filter(signif != 0, !is.na(Symbol)) %>%
  dplyr::select(Symbol)

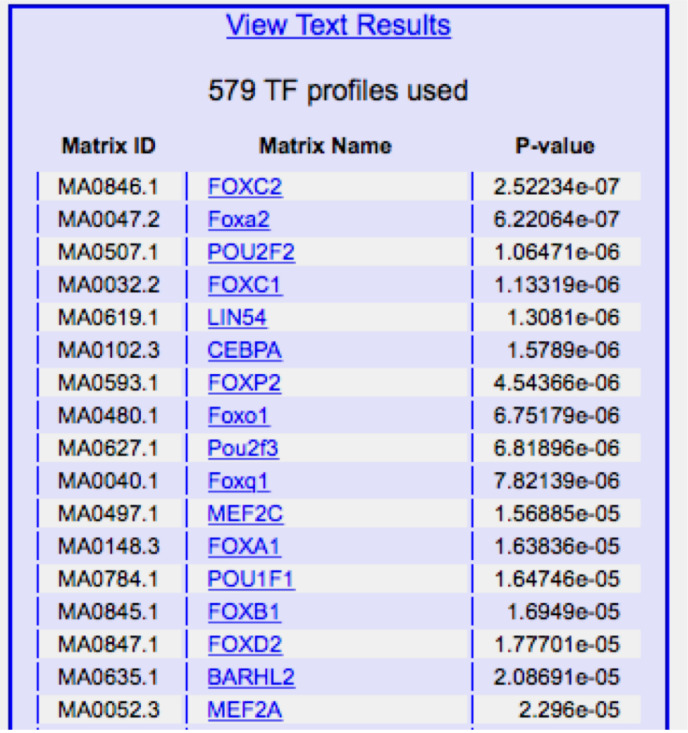
pscan_input <- getBM(attributes=c('hgnc_symbol', 'refseq_mrna'),
      filters = 'hgnc_symbol',
      values = pscan_input$Symbol,
      mart = ensembl) %>%
  filter(refseq_mrna != "")
```
````

Example Pscan

» Load IDs and parameters

» Matrices included





Matrix ID	Matrix Name	P-value
MA0846.1	FOXC2	2.52234e-07
MA0047.2	Foxa2	6.22064e-07
MA0507.1	POU2F2	1.06471e-06
MA0032.2	FOXC1	1.13319e-06
MA0619.1	LIN54	1.3081e-06
MA0102.3	CEBPA	1.5789e-06
MA0593.1	FOXP2	4.54366e-06
MA0480.1	Foxo1	6.75179e-06
MA0627.1	Pou2f3	6.81896e-06
MA0040.1	Foxq1	7.82139e-06
MA0497.1	MEF2C	1.56885e-05
MA0148.3	FOXA1	1.63836e-05
MA0784.1	POU1F1	1.64746e-05
MA0845.1	FOXB1	1.6949e-05
MA0847.1	FOXD2	1.77701e-05
MA0635.1	BARHL2	2.08691e-05
MA0052.3	MEF2A	2.296e-05

Intersect : differently expressed TF, whose targets are enriched in DE results

» Load Pscan results

» Load DE results

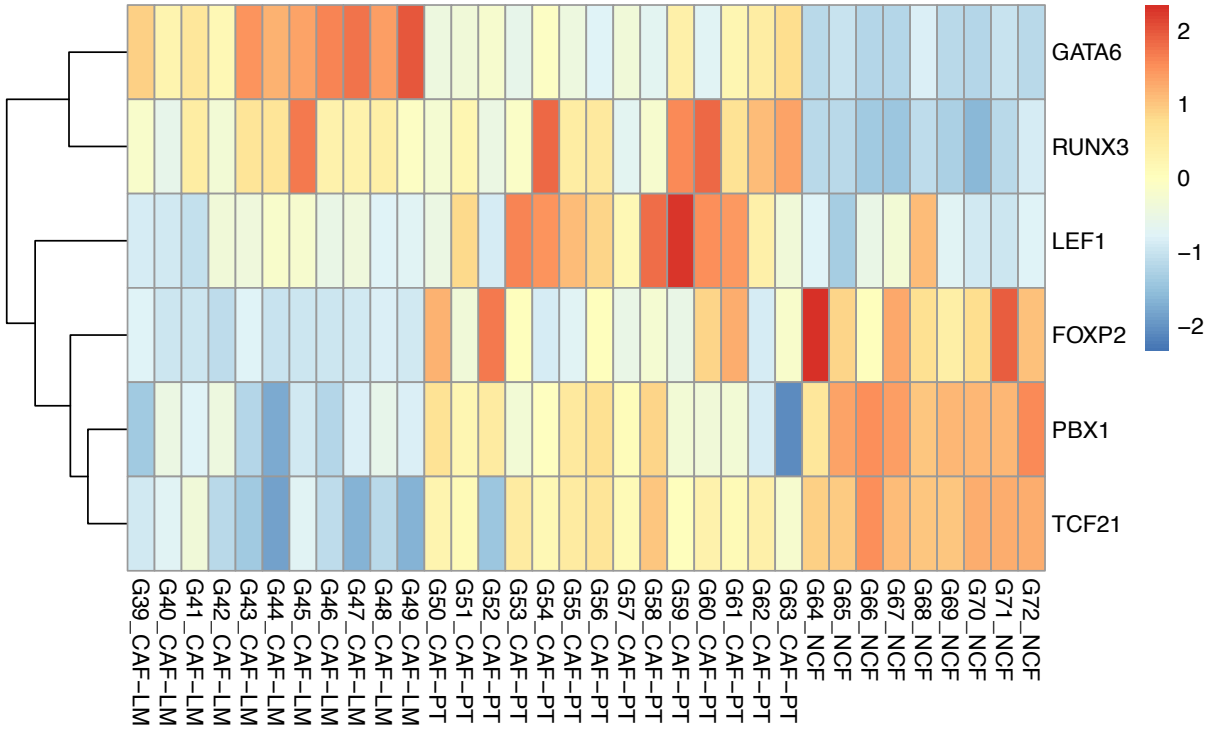
» Filter

```r
Pscan_output <- read_tsv("Pscanresult950_50.txt", ) %>%
  mutate(Symbol=toupper(TF_NAME)) %>%
  filter(P_VALUE<0.05) # keep significant only
```

```r
inter <- all_results %>%
  right_join(TF %>% dplyr::select(Symbol=gene_symbol)) %>%  #select TF in DE results
  filter(signif != 0) %>%  # keep those significant
  inner_join(Pscan_output) # intersect with enriched TFs
```

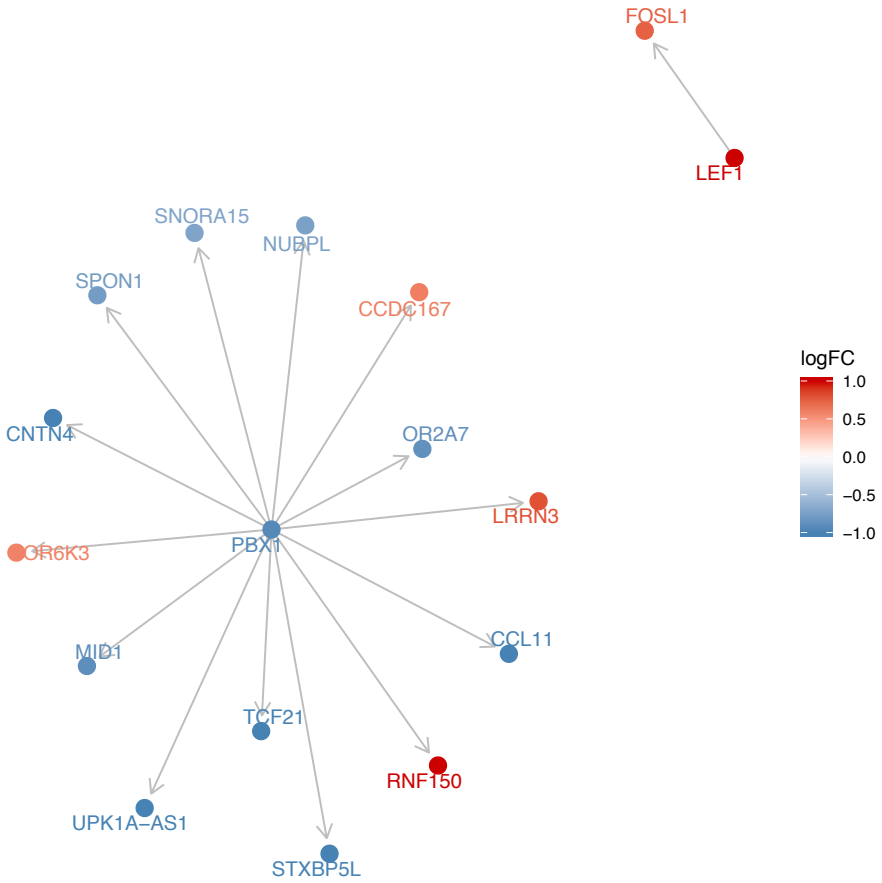
Intersect both

» Graph refined result



Intersect both

» TF > genes: can be represented as a network TF-target:



Points to consider

» Keep in mind that TF can be regulated at protein level > no mandatory mRNA difference to have an effect

miRNAs

What can regulate our genes?

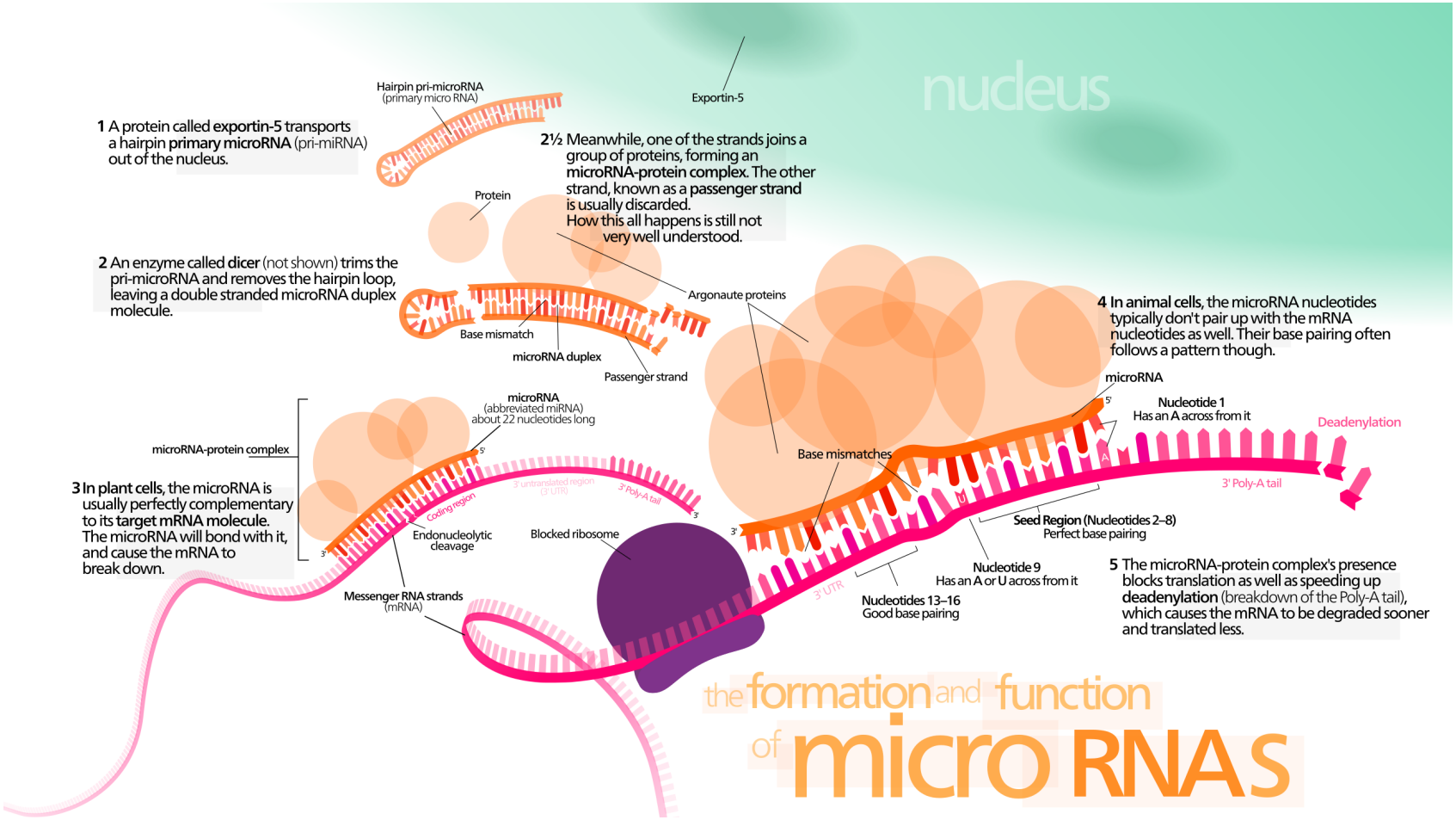
LSRU | LIFE SCIENCES RESEARCH UNIT

UNIVERSITÉ DU LUXEMBOURG

Summary

- Mode of action
- miRNAs :
  - Are the result of DE:
    - look for targets
  - Could target genes of interests:
    - look for regulators

MicroRNAs



https://en.wikipedia.org/wiki/MicroRNA

Target prediction

- Based on seed matching on miRNA
- Often based on 3'UTR on mRNA
- lower free energy > higher binding likelihood
- Considers mRNA 2^{ndy} structure
- Considers conservation status

Ref prediction tools : https://doi.org/10.3390/ijms17121987

Input

- Need list of genes / miRNAS
- Annotations: table of interactions
  - Prediction: TargetScan
  - Experimental: miRTarBase, TarBase (Diana tools)
  - Combining
- Annotation again: quality, updates

Tools

- Webtools: for single gene/miRNA look up
  Targetscan, Diana, miRTarBase sites
- R: SpidermiR, miRNAtap (combine)
- Cytoscape: CytargetLinker (combine)
  - Contains relations from several DBs
  - updated

Cytoscape

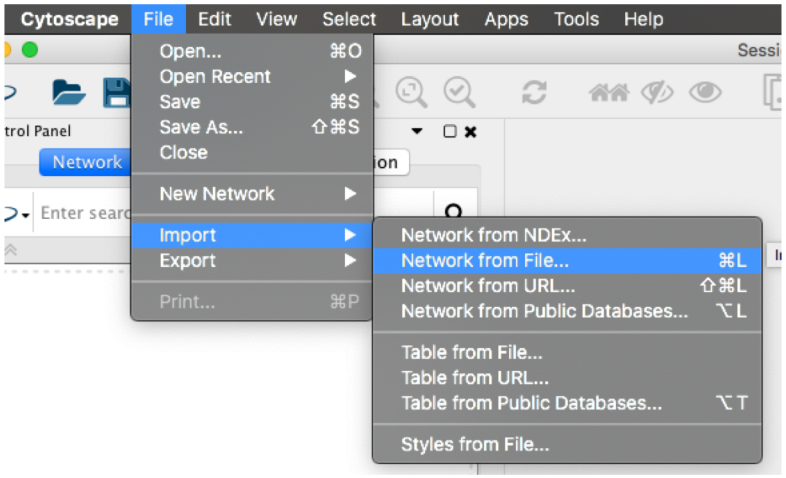
- Stand alone application
- point ‘n’ click
- Network based
- Apps to add functionality:
  - iregulon (TF)
  - cytargetlinker (miRNAs)
  - bingo (GO)
  - geneMania, String (annotations)

https://cytoscape.org

Example Cytoscape

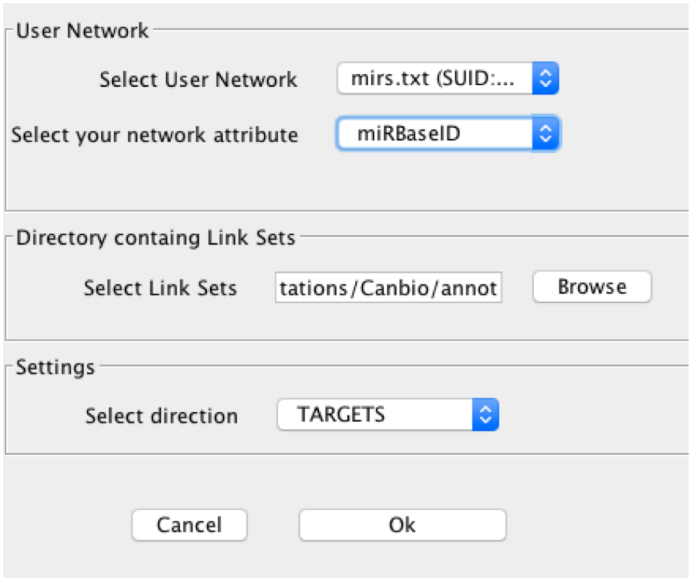
» Download interaction files from https://cytargetlinker.github.io/pages/linksets/mirtarbase

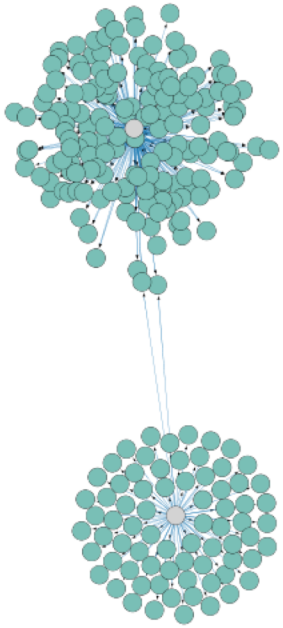
» Build network



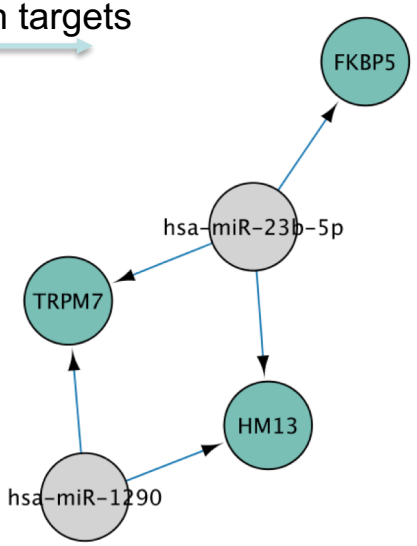
Example Cytoscape

» App > cytargetlinker > Extend network



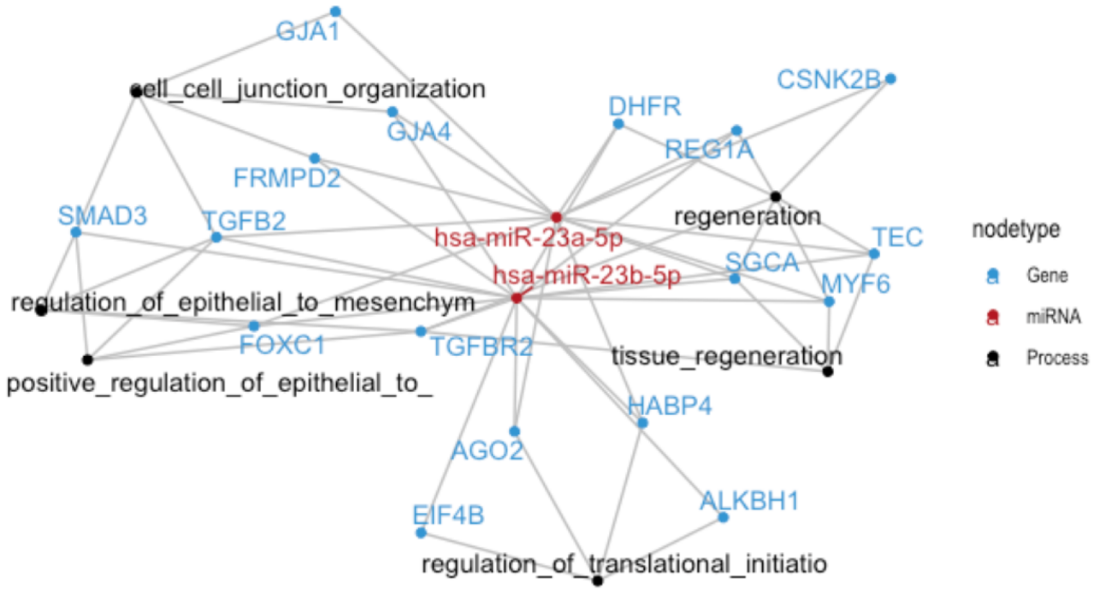


Filter
common targets



Example R

» Retrieve miRNA targets (SpidermiR, miRNAtap)
» Functional analysis of targets
» Network building



Points to consider

» Prediction: filter by score

» Poor overlap between predicted / exp.

Public data

Why -– where – what ?

LSRU | LIFE SCIENCES RESEARCH UNIT

UNIVERSITÉ DU LUXEMBOURG

Public datasets

- Why look for public data?
- Where to find data?
- What data?

Why

- To get preliminary data
  - Is drug effect consistent in other cell types?
  - Is a gene expressed in a specific cell type /cell line?
  - Which panel of cell lines should I use to cover different characteristics?
  - What genes are modified by a treatment but not by another?
- To check results
  - Is my treatment also affecting those genes in other datasets?

Using what have been done elsewhere
can save your time

Why

» To access to patient data
  – Is expression of my protein different between cancer grades?
  – Is survival better when autophagy is lower?
  – Is therapy outcome better for patients with BRCA1 mutation?

Using what have been done elsewhere

give access to patient data

Where - Main Sources

- TCGA: characterize cancer

  http://www.cbioportal.org
- GEO: dataset repository

  http://www.ncbi.nlm.nih.gov/gds
- Genomics of Drug Sensitivity in Cancer (GDSC) database: http://www.cancerrxgene.org
- The protein atlas: http://www.proteinatlas.org/
- CCLE: characterize cancer cell lines

  http://www.broadinstitute.org/ccle

What - TCGA

- Type: DNA copy number, mRNA expression, mutation, protein expression
- Samples: *many* patient samples by cancer type
- Clinical data provided, e.g:
  - Classification (Grade, Stage, FAB, Histological subtype)
  - Age, race, sex
  - Survival data
- Typical use:
  - survival curves for different conditions
  - different expression in different grades
  - co-expression

TCGA - example

Is survival of colorectal cancer patients associated with autophagy level?

» Colorectal cancer dataset (489 patient samples)

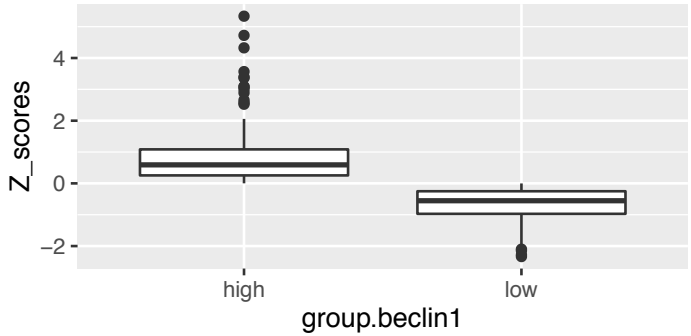
» Protein data: beclin1, p62

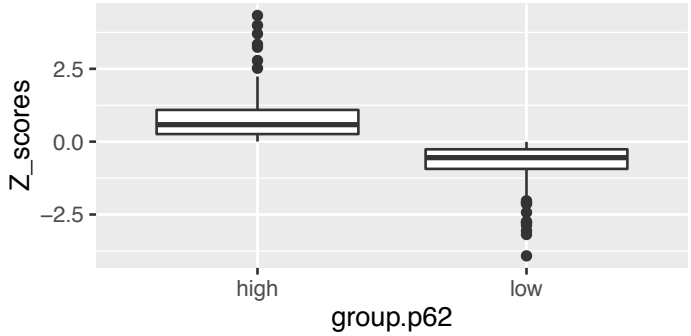
» Separate samples in groups:

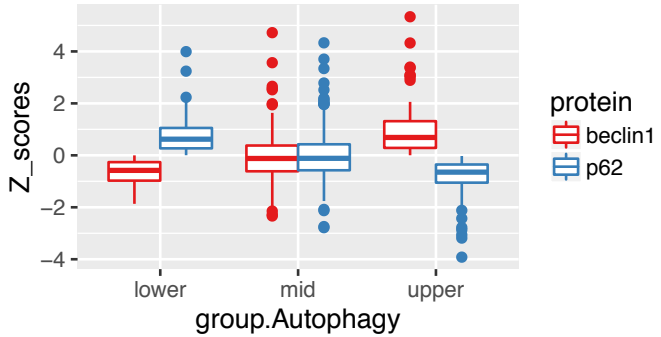
– Low/high beclin expression

– Low/high p62 expression

– “Autophagy“ status





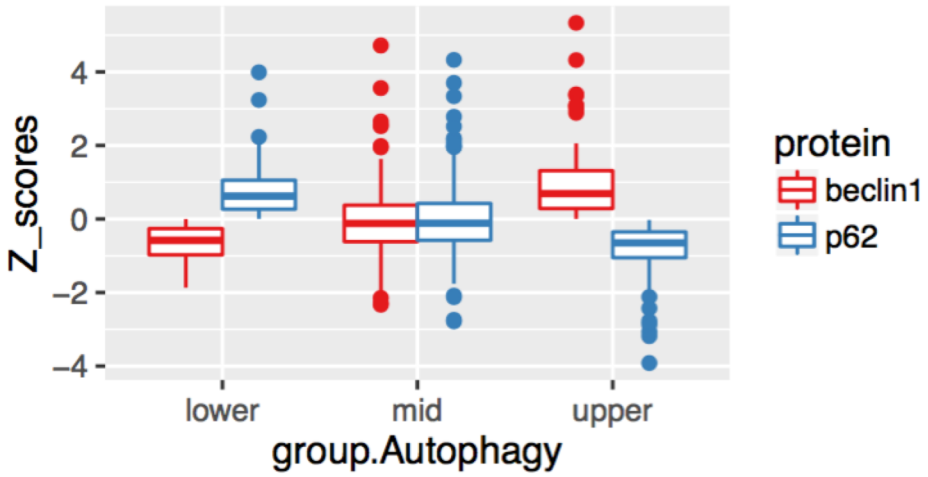


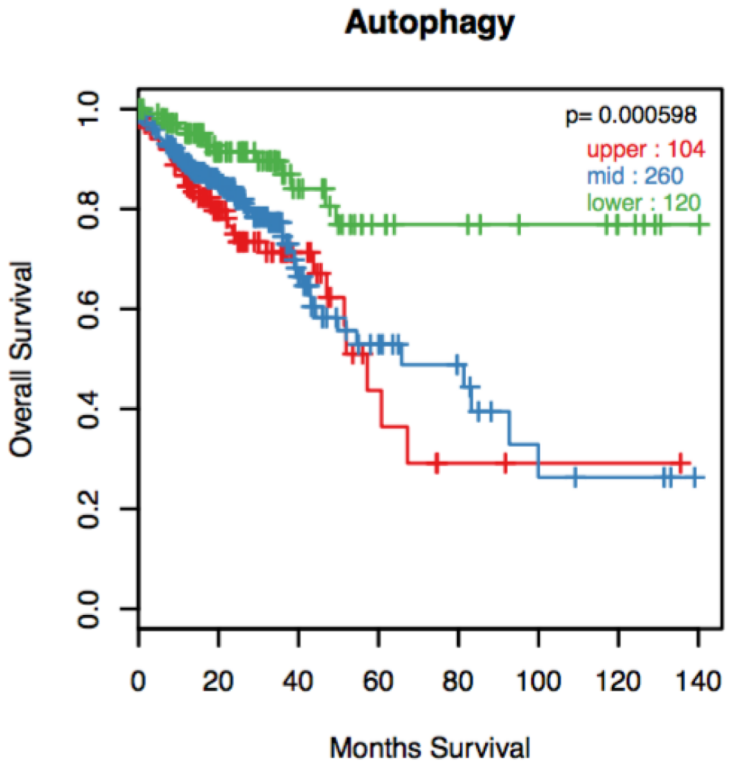
TCGA - example

Compare survival in each autophagy group:

low autophagy (low beclin , high p62)

=> better survival





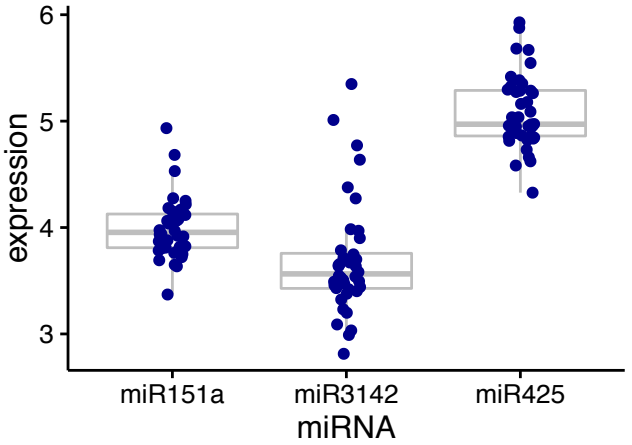
What - GEO

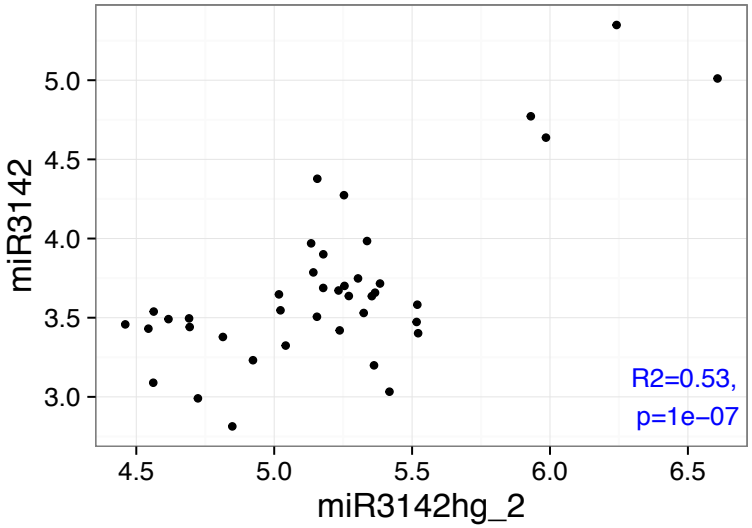
- Type: mRNA expression, Chip seq etc
- Samples: anything - repository of published studies
  - Patients, cell lines
  - Basal expression, control/treated studies
- Typical use:
  - Corroborate our results (e.g. treatment effect in other cell line)
  - Hints for future work: find genes affected by a treatment
  - Co-expression of 2 genes in a large dataset
- ! Limitation: need to find a suitable dataset

GEO - example

LncRNA project: RNA-PPTG1-11 (aka mir3142HG) affected by TNF is host gene of a miRNA:

- Is this miRNA expressed somewhere?
- Is it co-expressed with mir3142HG ?
- Expression change with TNF treatment?



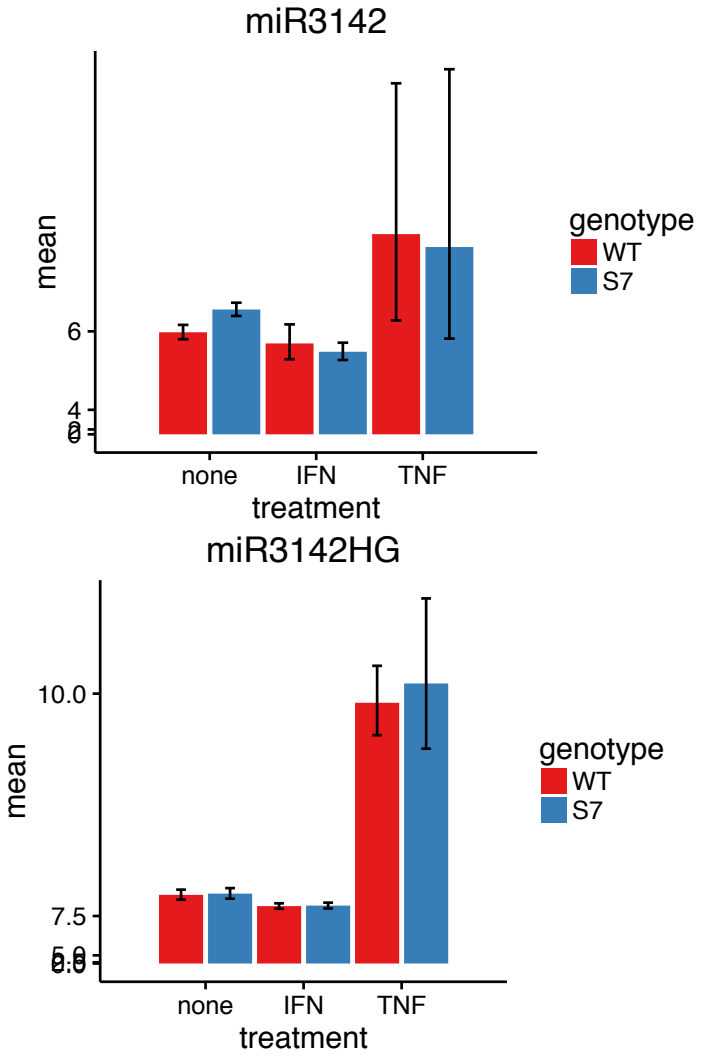


GSE77191 (CML patients), GSE76250 (breast cancer patients), GSE61723 (breast cancer patients), GSE57017 (fibroblast cells)

GEO - example

LncRNA study: RNA-PPTG1-11 is host gene of miR3142:

- Is this miRNA expressed somewhere?
- Is it co-expressed with mir3142HG ?
- Expression change with TNF treatment?



GSE57017 (fibroblast cells)

What - CCLE

- Type: DNA copy number, mRNA expression, mutation
- Samples: 1000 cancer cell lines.
- Typical use:
  - quick look at basal mRNA expression level in different cell lines -> select cell lines matching a specific pattern to perform experiment with

What – Protein Atlas

- Type: mRNA, protein expression
- Samples:
  - By tissues (semi-quantitative), cancer types (semi-quantitative)
  - 46 cell lines (quantitative)
  - 5 leukemia (2 samples each, quantitative)
- Typical use:
  - quick expression overview (subcellular localization, tissue)
  - Antibody information

What – Genomics of Drug Sensitivity in Cancer (GDSC)

» Type: IC50, key ‘features’ linked to sensitivity

- Features analysed: mutation status of cancer genes, chromosomal rearrangements, copy number data from genes causally implicated in cancer, genome-wide transcriptional profiles, tissue type

» Samples: Compounds (140, ongoing), genes, cell lines (>600)

» Typical use:

  – Find typical IC50 range for a drug, value for your cell line
  – Find features associated with sensitivity

Thank you

LSRU | LIFE SCIENCES RESEARCH UNIT

UNIVERSITÉ DU LUXEMBOURG