Practical Course on Big data analysis: transcriptomics, metabolomics, CyTOF and flow cytometry

Milena Sokolowska
Domingo Barber
Jozef Janda

EAACI
EUROPEAN ACADEMY OF ALLERGY AND CLINICAL IMMUNOLOGY

WS 2020 Chamonix, France

Understanding next generation sequencing through gene ontologies and user-friendly platforms

Milena Sokolowska MD, PhD

Head Immune Metabolism
Swiss Institute of Allergy and Asthma Research
University of Zurich

WS 2020 Chamonix, France

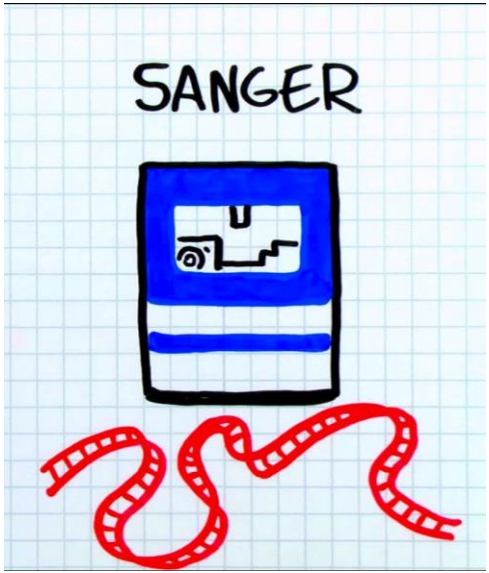
Disclosure

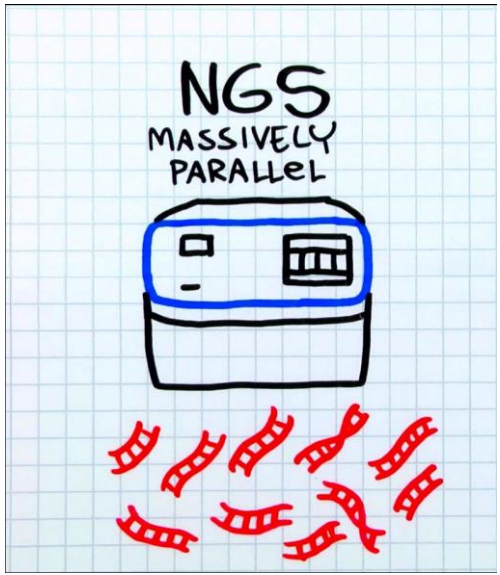
In relation to this presentation, I declare the following, real or perceived conflicts of interest:

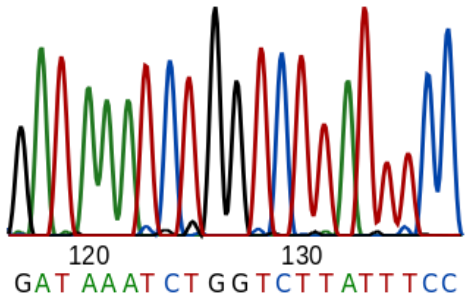
Type	Company
Employment full time / part time	n/a
Research Grant (P.I., collaborator or consultant; pending and received grants)	SNSF grnat, Allergopharma grant, GSK grant
Other research support	n/a
Speakers Bureau / Honoraria	n/a
Ownership interest (stock, stock-options, patent or intellectual property	n/a
Consultant / advisory board	n/a

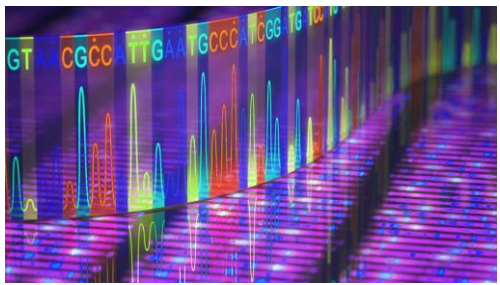
A conflict of interest is any situation in which a speaker or immediate family members have interests, and those may cause a conflict with the current presentation.
Conflicts of interest do not preclude the delivery of the talk, but should be explicitly declared. These may include financial interests (eg. owning stocks of a related company, having received honoraria, consultancy fees), research interests (research support by grants or otherwise), organisational interests and gifts.

Sequencing of any generation









Types of NGS experiments

a. Genomics

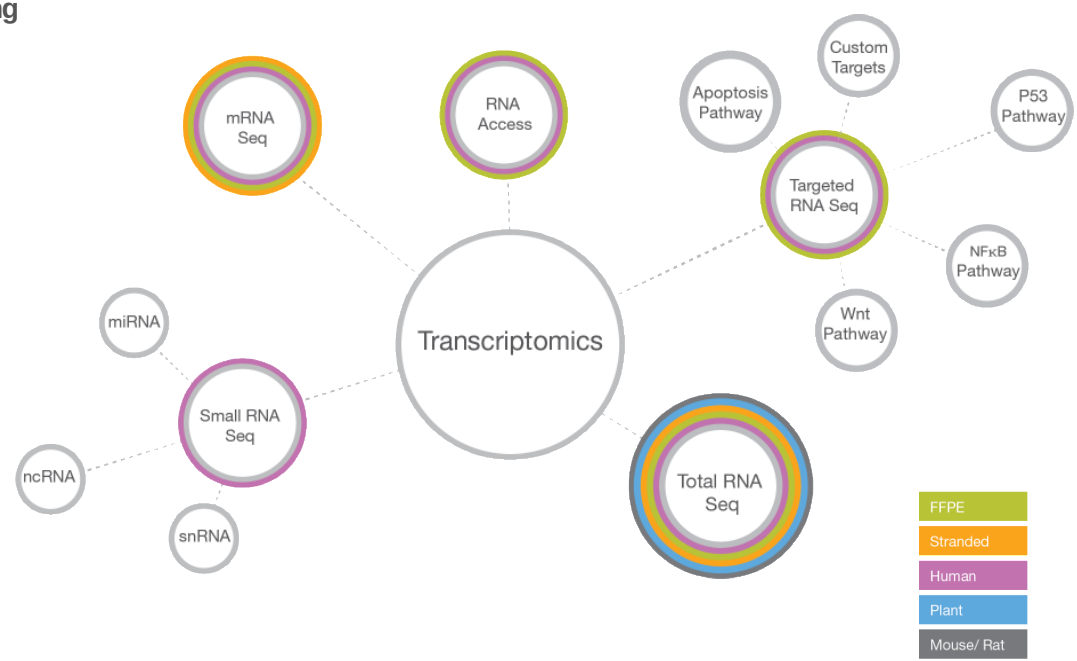
Whole-Genome Sequencing

Exome Sequencing

***De novo* Sequencing**

Targeted Sequencing

b. Transcriptomics



c. Epigenomics

Methylation Sequencing

ChIP Sequencing

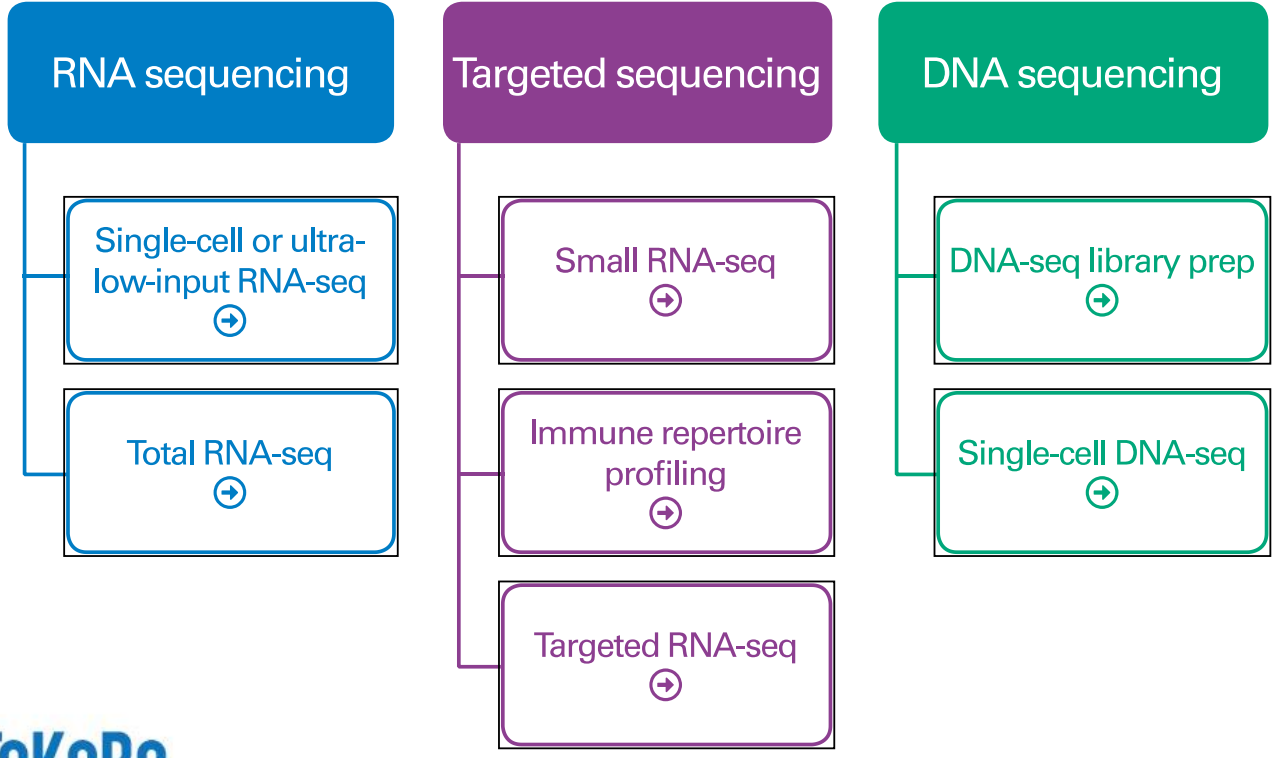
Ribosome Profiling

ATAC sequencing

illumina®

NGS workflow

Choosing the source and type of nucleic acid

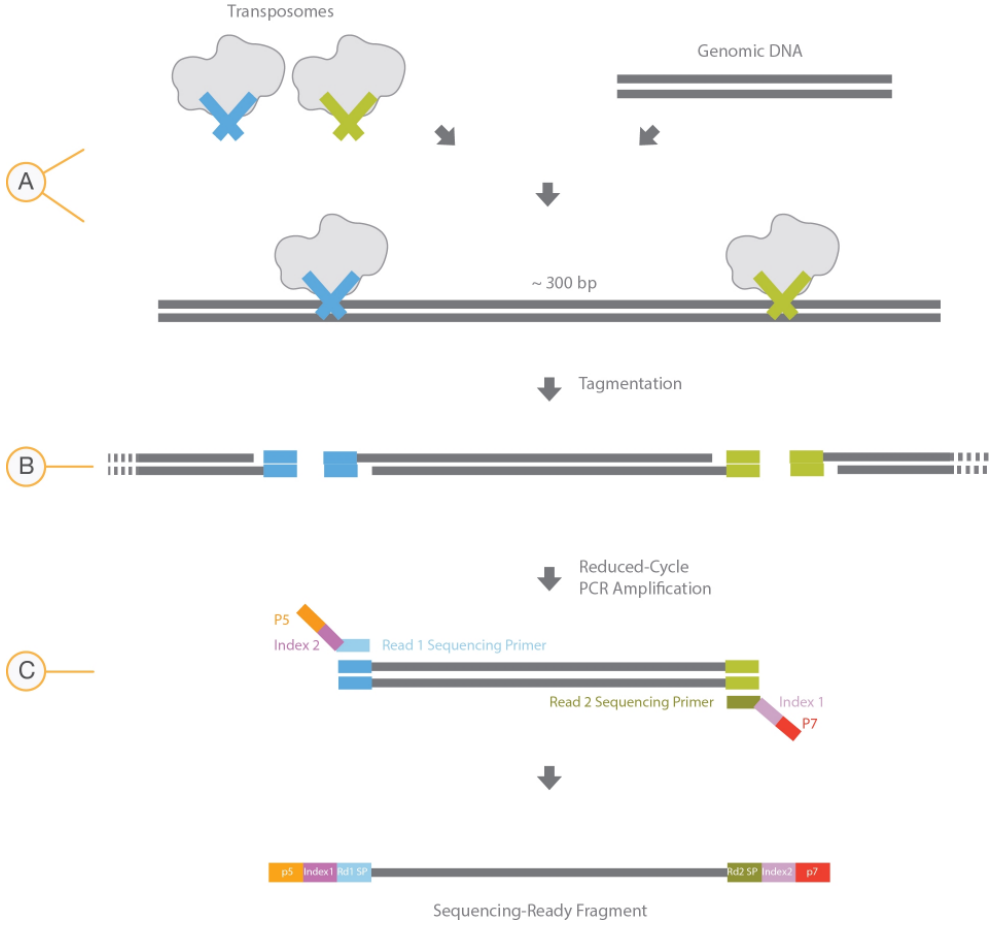


Library preparation methods

- **Whole genome libraries**
  - *de novo* or resequencing
- **RNA-seq libraries**
  - mRNA-seq, total RNA-seq, small RNA-seq
- **Shotgun metagenomics**
  - Sequencing multiple genomes or transcriptomes from the same sample
- **Bisulfite libraries**
  - Discover sites of DNA methylation
- **Customized libraries**

illumina®

Library preparation



illumina®

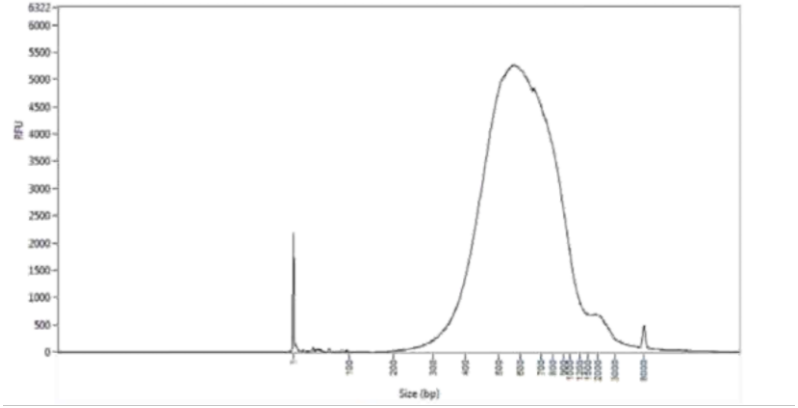
Library preparation

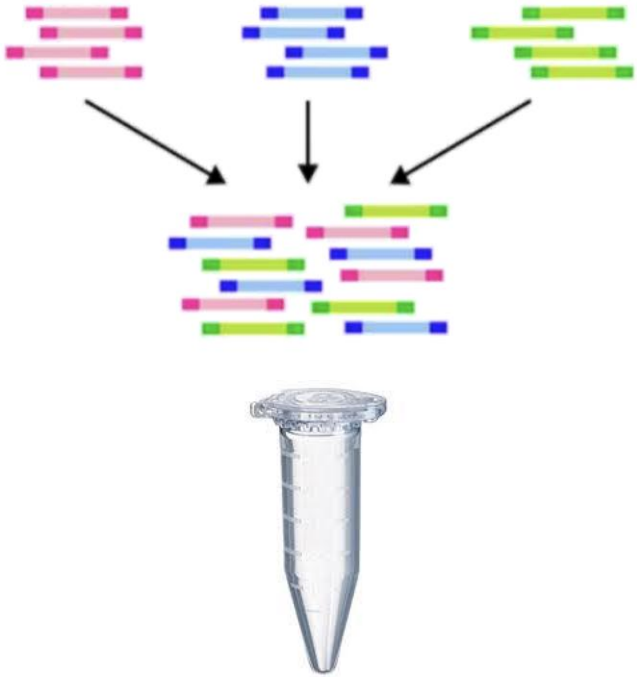
For clustering:
Libraries must have P5 and P7 binding regions on either end of a library

For sequencing:
Libraries must have sequencing primer binding regions

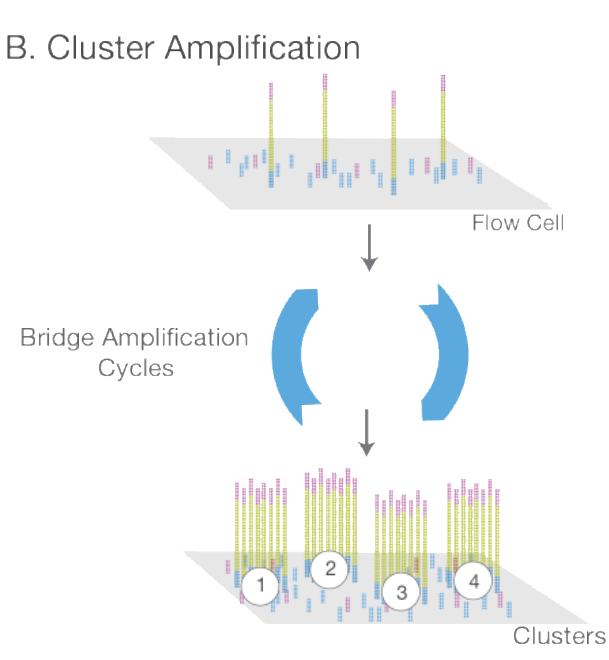
For pooling samples:
Libraries must have a unique index or barcode sequence

Library quantification and pooling

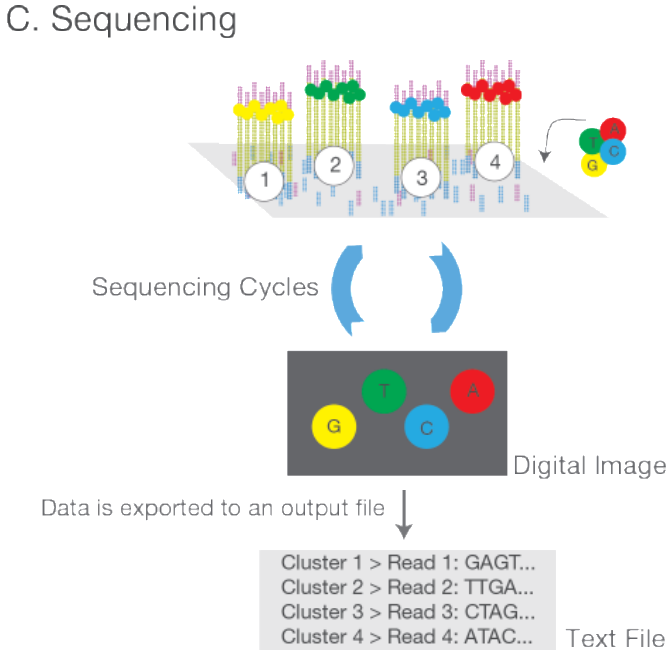




Sequencing



Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

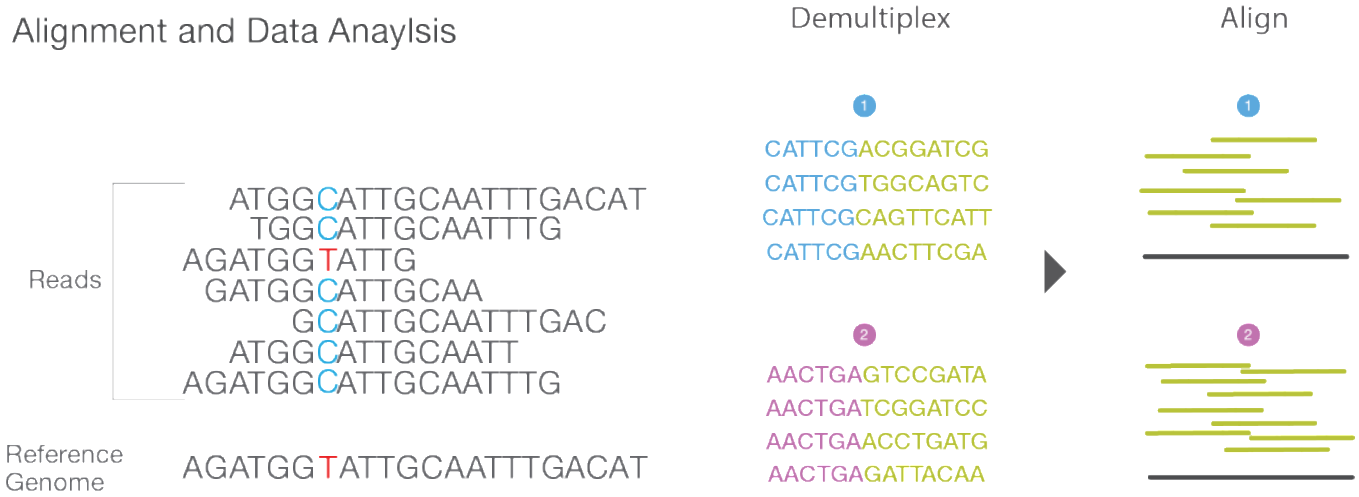


Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

illumina®

Data analysis

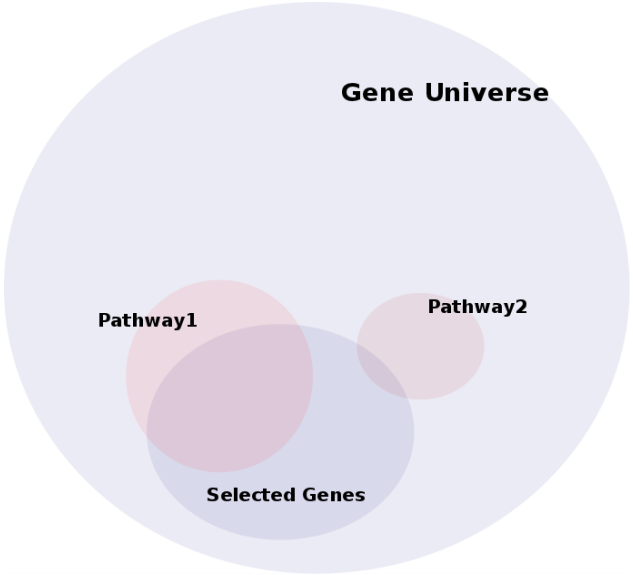
D. Alignment and Data Anaylsis



Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

E. Statistical analysis

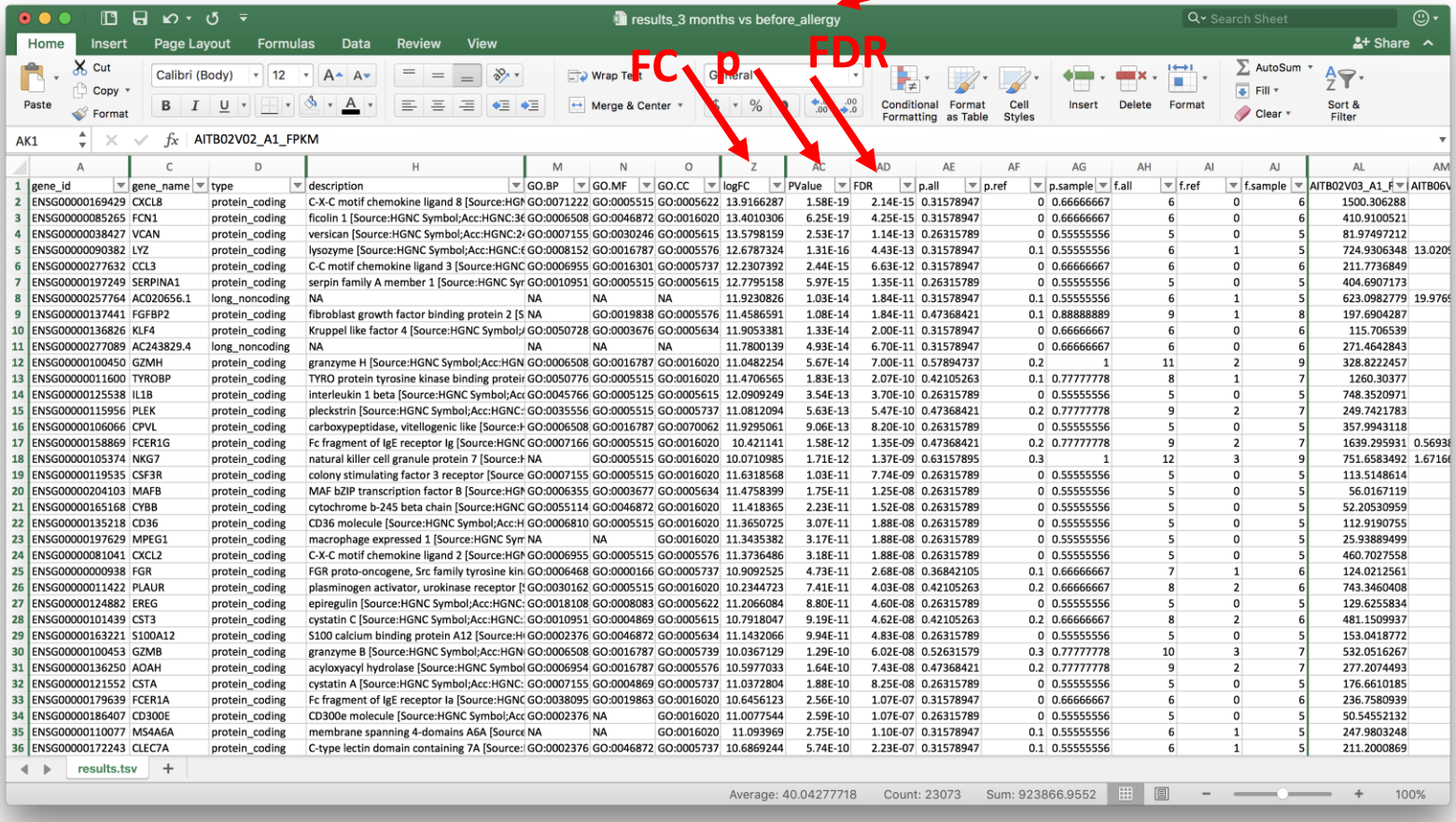
Explorative Functional Analysis



- NGS data/results
- Functional databases/gene ontologies
- Enrichment of gene sets
- Tools

Step 1: What are your data about?

Which comparison?

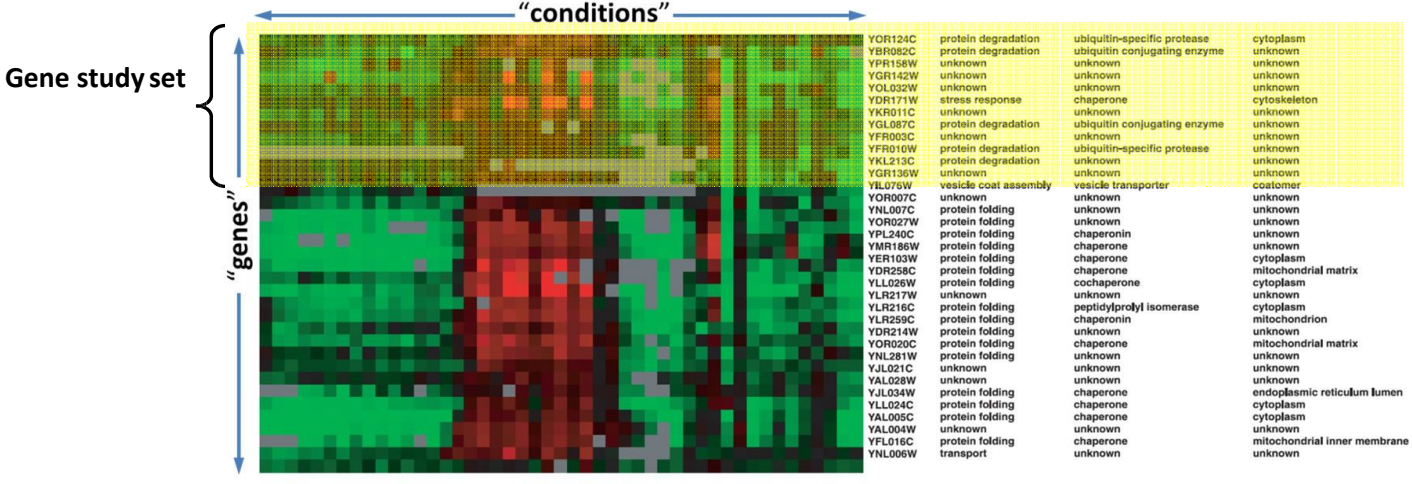


	A	C	D	H	M	N	O	Z	AC	AD	AE	AF	AG	AH	AI	AJ	AL
1	gene_id	gene_name	type	description	GO.BP	GO.MF	GO.CC	logFC	PValue	FDR	p.all	p.ref	p.sample	f.all	f.ref	f.sample	AITB02V03_A1_F
2	ENSG00000169429	CXCL8	protein_coding	C-X-C motif chemokine ligand 8 [Source:HGN	GO:0071222	GO:0005515	GO:0005622	13.9166287	1.58E-19	2.14E-15	0.31578947	0	0.66666667	6	0	6	1500.306288
3	ENSG00000085265	FCN1	protein_coding	ficolin 1 [Source:HGNC Symbol;Acc:HGNC:3	GO:0006508	GO:0046872	GO:0016020	13.4010306	6.25E-19	4.25E-15	0.31578947	0	0.66666667	6	0	6	410.9100521
4	ENSG00000038427	VCAN	protein_coding	versican [Source:HGNC Symbol;Acc:HGNC:2	GO:0007155	GO:0030246	GO:0005615	13.5798159	2.53E-17	1.14E-13	0.26315789	0	0.55555556	5	0	5	81.97497212
5	ENSG00000090382	LYZ	protein_coding	lysozyme [Source:HGNC Symbol;Acc:HGNC:6	GO:0008152	GO:0016787	GO:0005576	12.6787324	1.31E-16	4.43E-13	0.31578947	0.1	0.55555556	6	1	5	724.9306348
6	ENSG00000277632	CCL3	protein_coding	C-C motif chemokine ligand 3 [Source:HGNC	GO:0006955	GO:0016301	GO:0005737	12.2307392	2.44E-15	6.63E-12	0.31578947	0	0.66666667	6	0	6	211.7736849
7	ENSG00000197249	SERPINA1	protein_coding	serpin family A member 1 [Source:HGNC Syr	GO:0010951	GO:0005515	GO:0005615	12.7795158	5.97E-15	1.35E-11	0.26315789	0	0.55555556	5	0	5	404.6907173
8	ENSG00000257764	AC020656.1	long_noncoding	NA	NA	NA	NA	11.9230826	1.03E-14	1.84E-11	0.31578947	0.1	0.55555556	6	1	5	623.0982779
9	ENSG00000137441	FGFBP2	protein_coding	fibroblast growth factor binding protein 2 [S	NA	GO:0019838	GO:0005576	11.4586591	1.08E-14	1.84E-11	0.47368421	0.1	0.88888889	9	1	8	197.6904287
10	ENSG00000136826	KLF4	protein_coding	Kruppel like factor 4 [Source:HGNC Symbol;A	GO:0050728	GO:0003676	GO:0005634	11.9053381	1.33E-14	2.00E-11	0.31578947	0	0.66666667	6	0	6	115.706539
11	ENSG00000277089	AC243829.4	long_noncoding	NA	NA	NA	NA	11.7800139	4.93E-14	6.70E-11	0.31578947	0	0.66666667	6	0	6	271.4642843
12	ENSG00000100450	GZMH	protein_coding	granzyme H [Source:HGNC Symbol;Acc:HGN	GO:0006508	GO:0016787	GO:0016020	11.0482254	5.67E-14	7.00E-11	0.57894737	0.2	1	11	2	9	328.8222457
13	ENSG00000011600	TYROBP	protein_coding	TYRO protein tyrosine kinase binding protei	GO:0050776	GO:0005515	GO:0016020	11.4706565	1.83E-13	2.07E-10	0.42105263	0.1	0.77777778	8	1	7	1260.30377
14	ENSG00000125538	IL1B	protein_coding	interleukin 1 beta [Source:HGNC Symbol;Acc	GO:0045766	GO:0005125	GO:0005615	12.0909249	3.54E-13	3.70E-10	0.26315789	0	0.55555556	5	0	5	748.3520971
15	ENSG00000115956	PLEK	protein_coding	pleckstrin [Source:HGNC Symbol;Acc:HGNC:	GO:0035556	GO:0005515	GO:0005737	11.0812094	5.63E-13	5.47E-10	0.47368421	0.2	0.77777778	9	2	7	249.7421783
16	ENSG00000106066	CPVL	protein_coding	carboxypeptidase, vitellogenic like [Source:H	GO:0006508	GO:0016787	GO:0070062	11.9295061	9.06E-13	8.20E-10	0.26315789	0	0.55555556	5	0	5	357.9943118
17	ENSG00000158869	FCER1G	protein_coding	Fc fragment of IgE receptor Ig [Source:HGNC	GO:0007166	GO:0005515	GO:0016020	10.421141	1.58E-12	1.35E-09	0.47368421	0.2	0.77777778	9	2	7	1639.295931
18	ENSG00000105374	NKG7	protein_coding	natural killer cell granule protein 7 [Source:H	NA	GO:0005515	GO:0016020	10.0710985	1.71E-12	1.37E-09	0.63157895	0.3	1	12	3	9	751.6583492
19	ENSG00000119535	CSF3R	protein_coding	colony stimulating factor 3 receptor [Source	GO:0007155	GO:0005515	GO:0016020	11.6318568	1.03E-11	7.74E-09	0.26315789	0	0.55555556	5	0	5	113.5148614
20	ENSG00000204103	MAFB	protein_coding	MAF bZIP transcription factor B [Source:HGN	GO:0006355	GO:0003677	GO:0005634	11.4758399	1.75E-11	1.25E-08	0.26315789	0	0.55555556	5	0	5	56.0167119
21	ENSG00000165168	CYBB	protein_coding	cytochrome b-245 beta chain [Source:HGNC	GO:0055114	GO:0046872	GO:0016020	11.418365	2.23E-11	1.52E-08	0.26315789	0	0.55555556	5	0	5	52.20530959
22	ENSG00000135218	CD36	protein_coding	CD36 molecule [Source:HGNC Symbol;Acc:H	GO:0006810	GO:0005515	GO:0016020	11.3650725	3.07E-11	1.88E-08	0.26315789	0	0.55555556	5	0	5	112.9190755
23	ENSG00000197629	MPEG1	protein_coding	macrophage expressed 1 [Source:HGNC Syn	NA	NA	GO:0016020	11.3435382	3.17E-11	1.88E-08	0.26315789	0	0.55555556	5	0	5	25.93889499
24	ENSG00000081041	CXCL2	protein_coding	C-X-C motif chemokine ligand 2 [Source:HGN	GO:0006955	GO:0005515	GO:0005576	11.3736486	3.18E-11	1.88E-08	0.26315789	0	0.55555556	5	0	5	460.7027558
25	ENSG00000000938	FGR	protein_coding	FGR proto-oncogene, Src family tyrosine kin	GO:0006468	GO:0000166	GO:0005737	10.9092525	4.73E-11	2.68E-08	0.36842105	0.1	0.66666667	7	1	6	124.0212561
26	ENSG00000011422	PLAUR	protein_coding	plasminogen activator, urokinase receptor [	GO:0030162	GO:0005515	GO:0016020	10.2344723	7.41E-11	4.03E-08	0.42105263	0.2	0.66666667	8	2	6	743.3460408
27	ENSG00000124882	EREG	protein_coding	epiregulin [Source:HGNC Symbol;Acc:HGNC:	GO:0018108	GO:0008083	GO:0005622	11.2066084	8.80E-11	4.60E-08	0.26315789	0	0.55555556	5	0	5	129.6255834
28	ENSG00000101439	CST3	protein_coding	cystatin C [Source:HGNC Symbol;Acc:HGNC:	GO:0010951	GO:0004869	GO:0005615	10.7918047	9.19E-11	4.62E-08	0.42105263	0.2	0.66666667	8	2	6	481.1509937
29	ENSG00000163221	S100A12	protein_coding	S100 calcium binding protein A12 [Source:H	GO:0002376	GO:0046872	GO:0005634	11.1432066	9.94E-11	4.83E-08	0.26315789	0	0.55555556	5	0	5	153.0418772
30	ENSG00000100453	GZMB	protein_coding	granzyme B [Source:HGNC Symbol;Acc:HGN	GO:0006508	GO:0016787	GO:0005739	10.0367129	1.29E-10	6.02E-08	0.52631579	0.3	0.77777778	10	3	7	532.0516267
31	ENSG00000136250	AOAH	protein_coding	acyloxyacyl hydrolase [Source:HGNC Symbol	GO:0006954	GO:0016787	GO:0005576	10.5977033	1.64E-10	7.43E-08	0.47368421	0.2	0.77777778	9	2	7	277.2074493
32	ENSG00000121552	CSTA	protein_coding	cystatin A [Source:HGNC Symbol;Acc:HGNC:	GO:0007155	GO:0004869	GO:0005737	11.0372804	1.88E-10	8.25E-08	0.26315789	0	0.55555556	5	0	5	176.6610185
33	ENSG00000179639	FCER1A	protein_coding	Fc fragment of IgE receptor Ia [Source:HGNC	GO:0038095	GO:0019863	GO:0016020	10.6456123	2.56E-10	1.07E-07	0.31578947	0	0.66666667	6	0	6	236.7580939
34	ENSG00000186407	CD300E	protein_coding	CD300e molecule [Source:HGNC Symbol;Acc	GO:0002376	NA	GO:0016020	11.0077544	2.59E-10	1.07E-07	0.26315789	0	0.55555556	5	0	5	50.54552132
35	ENSG00000110077	MS4A6A	protein_coding	membrane spanning 4-domains A6A [Source	NA	NA	GO:0016020	11.093969	2.75E-10	1.10E-07	0.31578947	0.1	0.55555556	6	1	5	247.9803248
36	ENSG00000172243	CLEC7A	protein_coding	C-type lectin domain containing 7A [Source:	GO:0002376	GO:0046872	GO:0005737	10.6869244	5.74E-10	2.23E-07	0.31578947	0.1	0.55555556	6	1	5	211.2000869

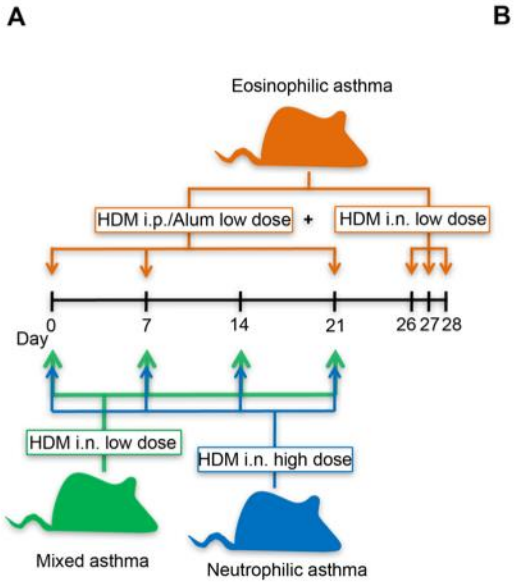
GOAL: Identify: key molecules (TFs, miRNAs, master regulators), Enriched Biological Processes/Pathways, Networks (links across candidates)

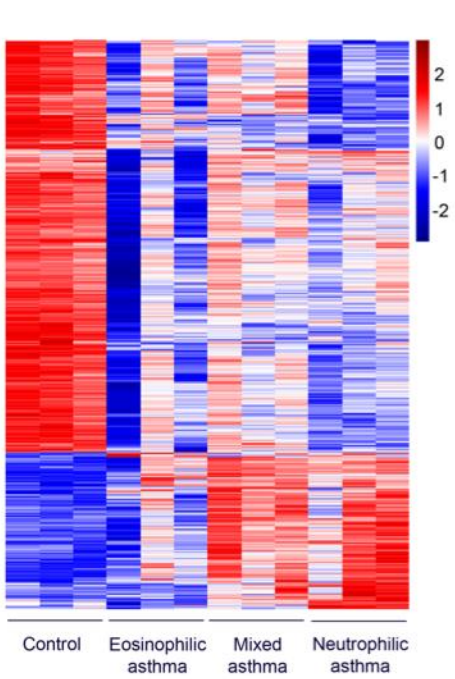
Step 2: Picking “relevant” genes-filtering results

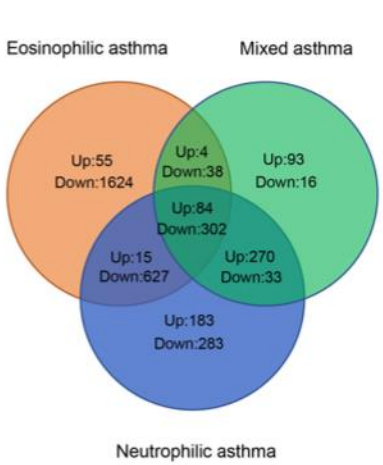
- Fold change cutoff (e.g., > two fold change)
- Fold change rank (e.g., top 10%)
- FDR (e.g, <0.05)
- Combinations of the above



Examples: data presentation-unbiased approach





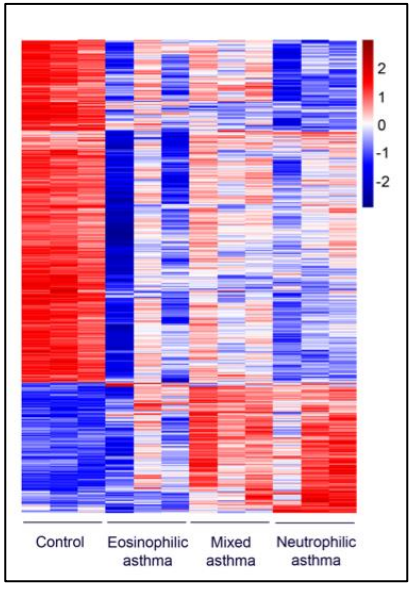


Heatmap

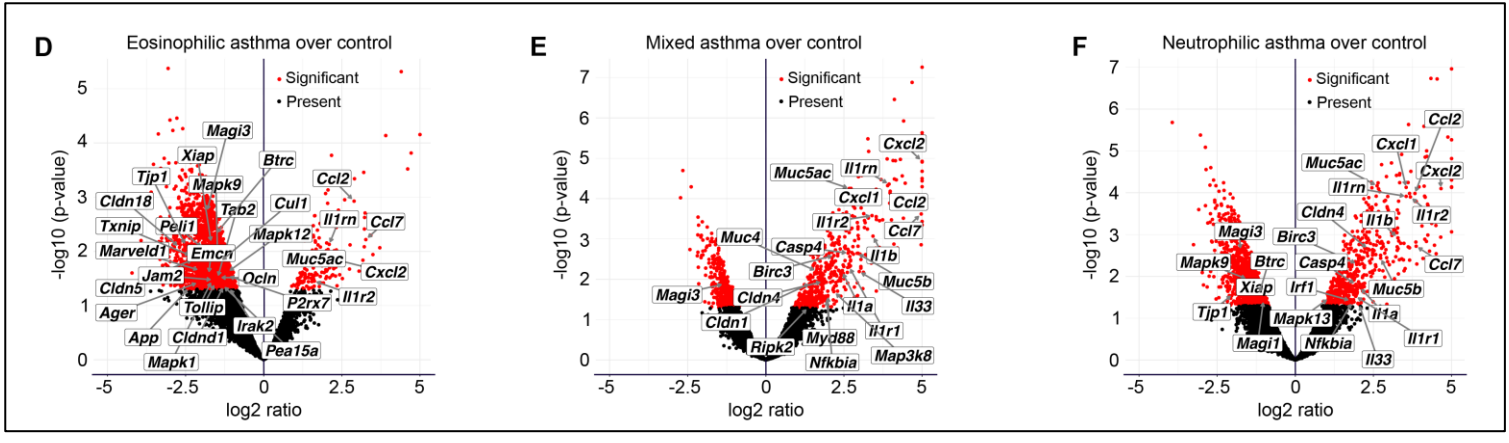
Venn diagram

Tan HT & Hagner S. et al *Allergy* 2019

Are upper and lower panels are showing the same information?



a) Yes
b) No
c) Yes, with some additional information on the volcano plots

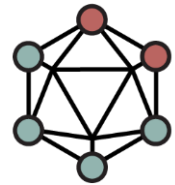


Step 3: Which functional databases/gene ontologies/gene annotations can be interrogated?

- **Gene Ontology**

- **Pathways**

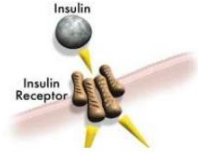
- **Protein class**

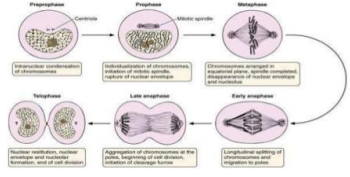


These databases are typically constructed based on **protein-protein interaction experiments, signaling pathway disruption experiment, literature screening** (and combinations of the above)

GO domains

- Three ontology domains:
  1. **Molecular function:** basic activity or task
     e.g. catalytic activity, calcium ion binding
  2. **Biological process:** broad objective or goal
     e.g. signal transduction, immune response
  3. **Cellular component:** location or complex
     e.g. nucleus, mitochondrion
- Genes can have multiple annotations



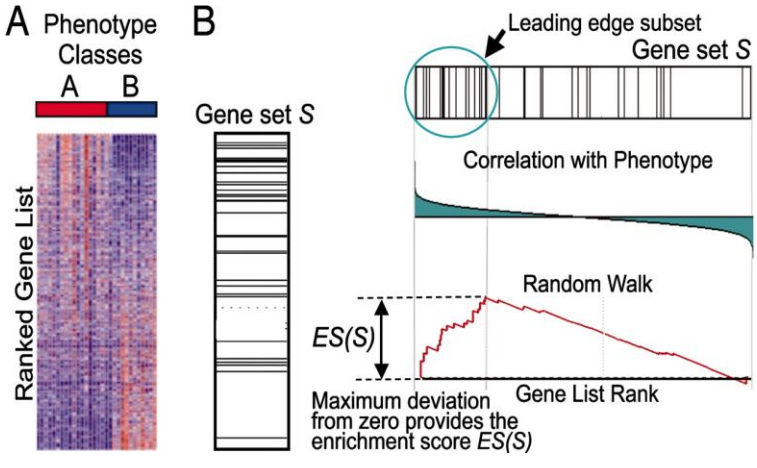


MSigDB
Molecular Signatures Database

IMMPORT
BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

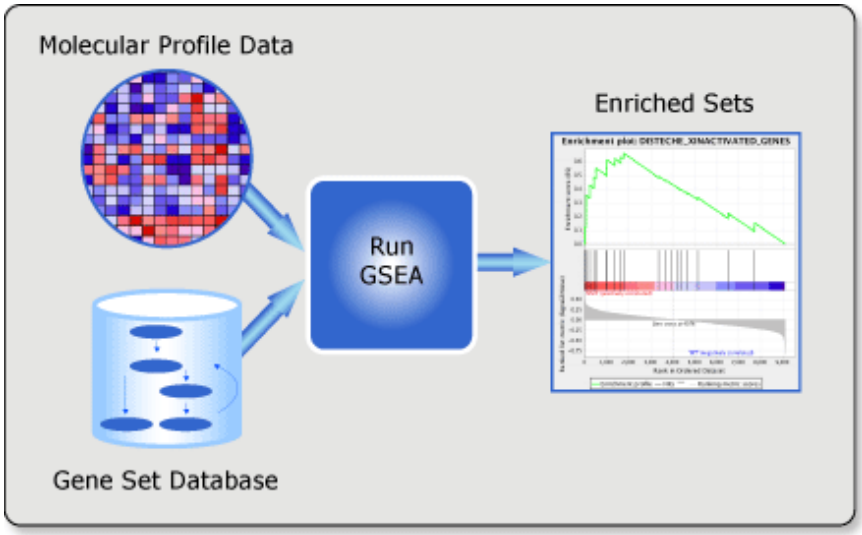
Step 4: Gene set enrichment analysis (GSEA)

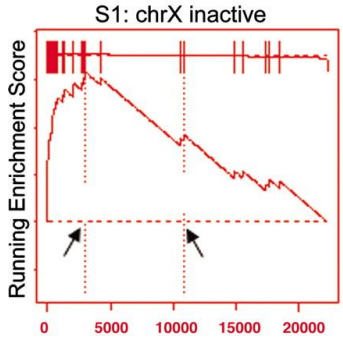
- Input: genes ordered in a ranked list L, according to their differential expression between the classes
- The goal of GSEA is to determine whether members of a gene set S are randomly distributed throughout the list L or tend to occur toward the top (or bottom) of L
  - Enrichment score (ES) reflects the **degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L.**
  - calculated by walking down the list L, increasing a running-sum when we encounter a gene in S and decreasing it when we encounter genes not in S.
  - ES is the max deviation from zero encountered in the random walk (Kolmogorov–Smirnov test)

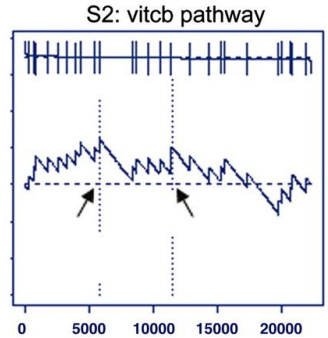


http://software.broadinstitute.org/gsea/index.jsp

Gene set enrichment analysis (GSEA)







MSigDB
Molecular Signatures Database

Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. Epub 2005 Sep 30.

Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.

Subramanian A[1], Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.

Step 5: Tools for functional enrichment analysis

- **Free** tools:
  - GSEA http://software.broadinstitute.org/gsea/index.jsp
  - WebGestalt http://www.webgestalt.org/
  - Panther http://www.pantherdb.org/
  - DAVID https://david.ncifcrf.gov/
  - STRING https://string-db.org
  - Cytoscape https://cytoscape.org/
  - R-Packages: topGO, GSEABase, clusterProfiler,…

- **Commercial** tools :
  - MetaCore/GeneGo - https://portal.genego.com/
  - Ingenuity Pathway Analysis (IPA) https://apps.ingenuity.com/ingsso/login

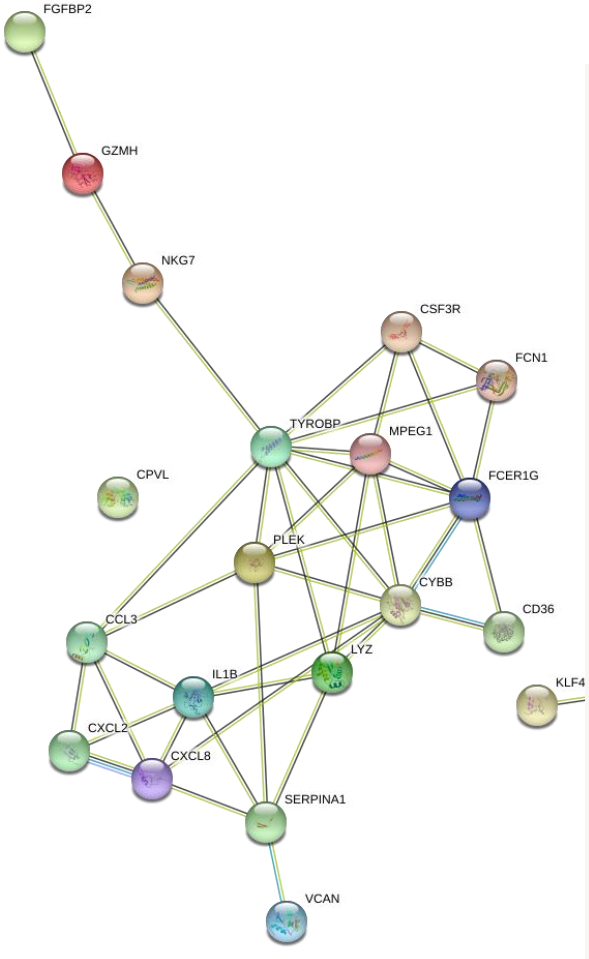
GO Enrichment - Examples

GO-Term	P-Value	Ratio
immune response	2.77e-14	50/289
. . . .T-helper 17 cell differentiati	5.45e-05	3/3
. . . . .positive regulation of natural	4.24e-06	7/16
.regulation of immune response	7.26e-07	16/81
inflammatory response	3.11e-09	39/252
cellular defense response	6.71e-07	13/49
G-protein coupled receptor sig	7.12e-07	28/161
cytokine-mediated signaling pa	2.39e-06	23/196
.chemokine-mediated signaling p	1.66e-05	8/22
negative regulation of viral g	3.33e-06	8/30
positive regulation of natural	2.58e-05	4/5
response to virus	5.88e-05	13/96
.defense response to virus	4.86e-08	20/120
cytolysis	8.07e-05	5/13
positive regulation of cell ad	9.08e-05	8/29

289 Genes in the Gene Universe are annotated with 'Immune Response'

50 Genes in Candidates List are belonging to 'Immune Response'

STRING enrichment



Functional enrichments in your network

Biological Process (GO)

GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0036230	granulocyte activation	9 of 502	7.41e-07
GO:0006952	defense response	12 of 1234	7.41e-07
GO:0002376	immune system process	15 of 2370	7.41e-07
GO:0030593	neutrophil chemotaxis	5 of 59	1.89e-06
GO:0045321	leukocyte activation	10 of 894	2.35e-06
			(more ...)

Molecular Function (GO)

GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0008009	chemokine activity	3 of 48	0.0036
GO:0005515	protein binding	16 of 6605	0.0045
GO:0005125	cytokine activity	4 of 216	0.0045
GO:0045236	CXCR chemokine receptor binding	2 of 16	0.0056
GO:0008329	signaling pattern recognition receptor activity	2 of 17	0.0056
			(more ...)

Cellular Component (GO)

GO-term	*description*	*count in gene set*	*false discovery rate*
GO:0030141	secretory granule	9 of 828	1.27e-05
GO:0005615	extracellular space	7 of 1134	0.0055
GO:0005576	extracellular region	10 of 2505	0.0055
GO:0030667	secretory granule membrane	4 of 298	0.0079
GO:0044433	cytoplasmic vesicle part	7 of 1447	0.0085
			(more ...)

Reference publications

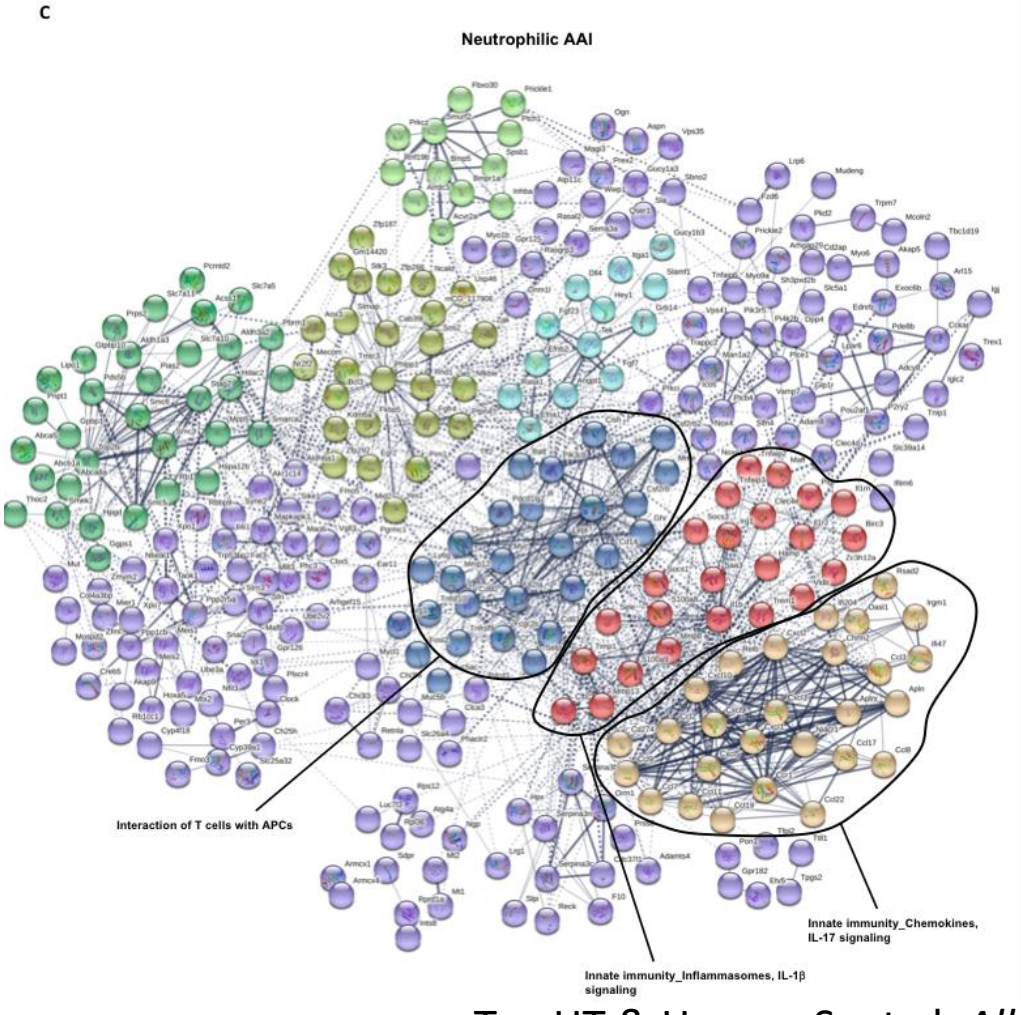
publication	*(year) title*	*count in gene set*	*false discovery rate*
PMID:28573109	(2017) Persistence of Innate Immune Pathways in Late Sta...	8 of 107	3.76e-08
PMID:16552065	(2006) Functional genomics of innate host defense molecul...	5 of 10	2.25e-07
PMID:25828472	(2015) Stimulation of hepatocarcinogenesis by neutrophils ...	5 of 18	1.44e-06
PMID:25767696	(2015) Enhancement of COPD biological networks using a ...	6 of 51	1.44e-06
PMID:20525249	(2010) Response of the mouse lung transcriptome to weldi...	6 of 57	1.62e-06
			(more ...)

KEGG Pathways

pathway	*description*	*count in gene set*	*false discovery rate*
hsa05132	Salmonella infection	4 of 84	0.00015
hsa04060	Cytokine-cytokine receptor interaction	5 of 263	0.00027

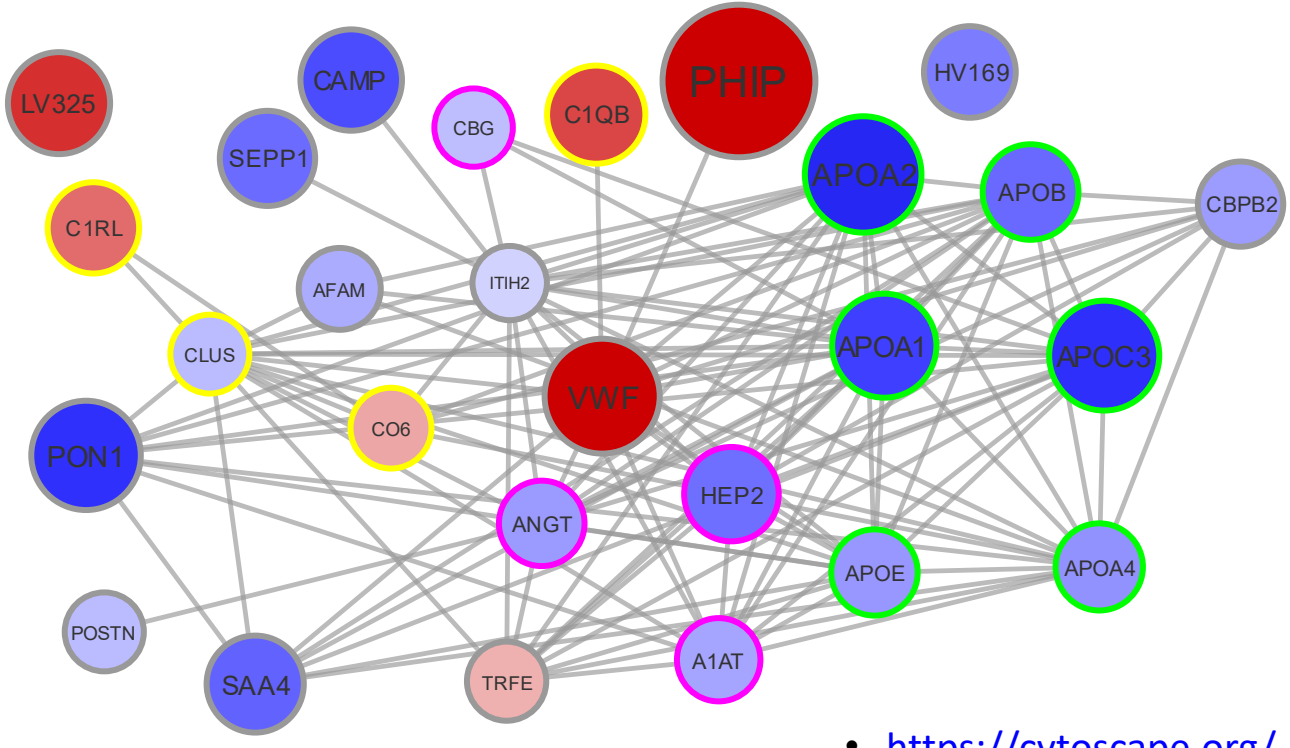
- https://string-db.org

STRING clustering



Tan HT & Hagner S. et al *Allergy* 2019

Cytoscape



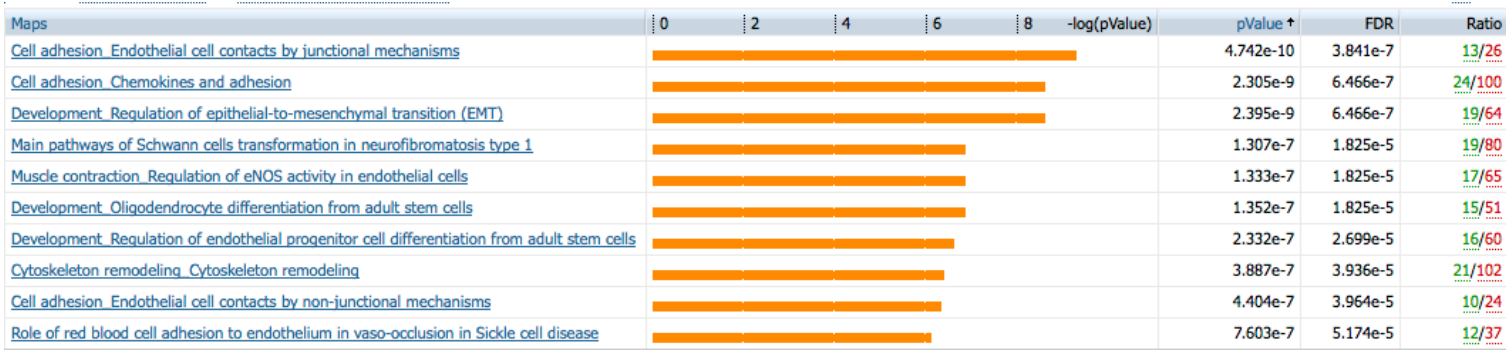
- https://cytoscape.org/

Pathway Enrichment - Example Graphs produced with Metacore

https://portal.genego.com

commercial

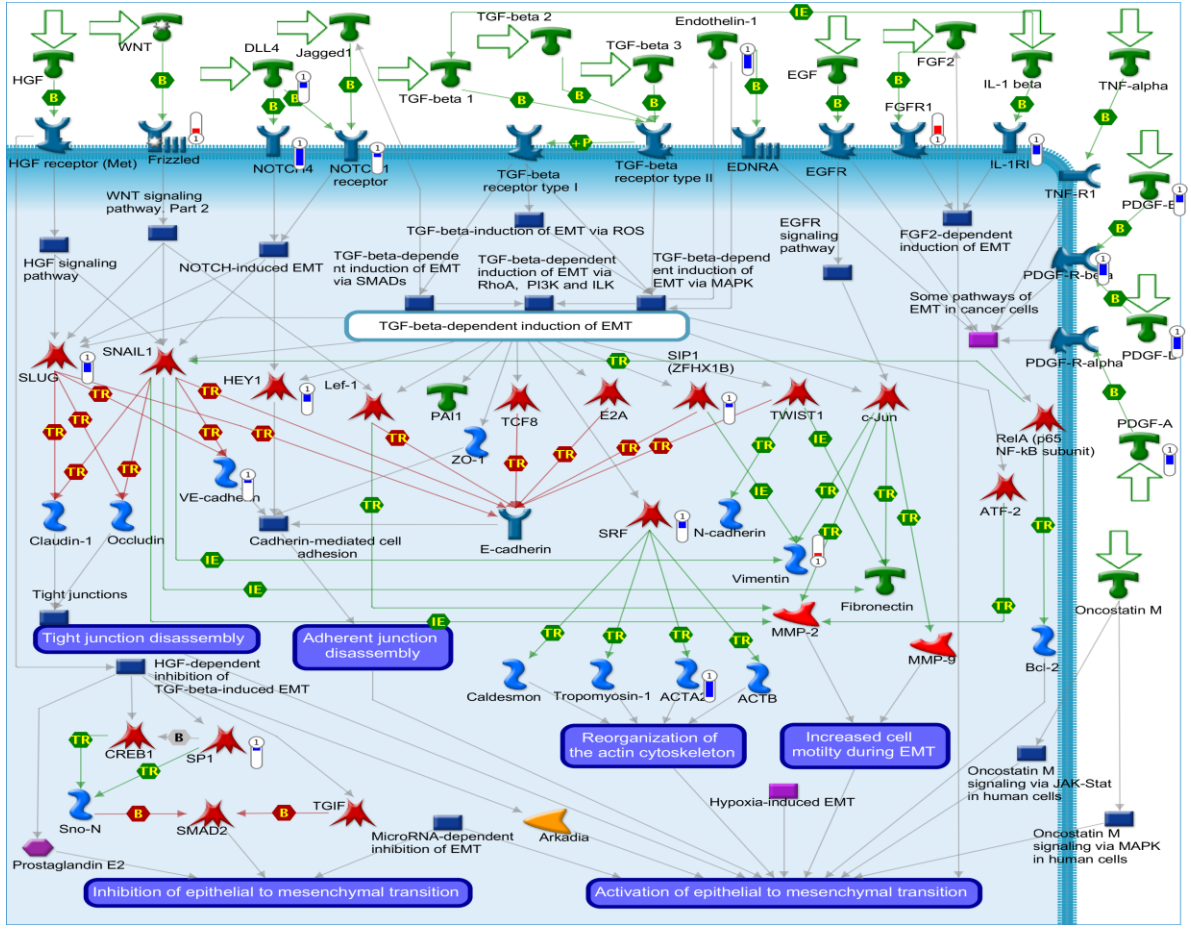
- upload a file with list of genes of interest (eg. differentially expressed genes)
- one click enrichment analysis (eg. pathway enrichment analysis)



Maps	0 2 4 6 8 -log(pValue)	pValue	FDR	Ratio
Cell adhesion_Endothelial cell contacts by junctional mechanisms		4.742e-10	3.841e-7	13/26
Cell adhesion_Chemokines and adhesion		2.305e-9	6.466e-7	24/100
Development_Regulation of epithelial-to-mesenchymal transition (EMT)		2.395e-9	6.466e-7	19/64
Main pathways of Schwann cells transformation in neurofibromatosis type 1		1.307e-7	1.825e-5	19/80
Muscle contraction_Regulation of eNOS activity in endothelial cells		1.333e-7	1.825e-5	17/65
Development_Oligodendrocyte differentiation from adult stem cells		1.352e-7	1.825e-5	15/51
Development_Regulation of endothelial progenitor cell differentiation from adult stem cells		2.332e-7	2.699e-5	16/60
Cytoskeleton remodeling_Cytoskeleton remodeling		3.887e-7	3.936e-5	21/102
Cell adhesion_Endothelial cell contacts by non-junctional mechanisms		4.404e-7	3.964e-5	10/24
Role of red blood cell adhesion to endothelium in vaso-occlusion in Sickle cell disease		7.603e-7	5.174e-5	12/37

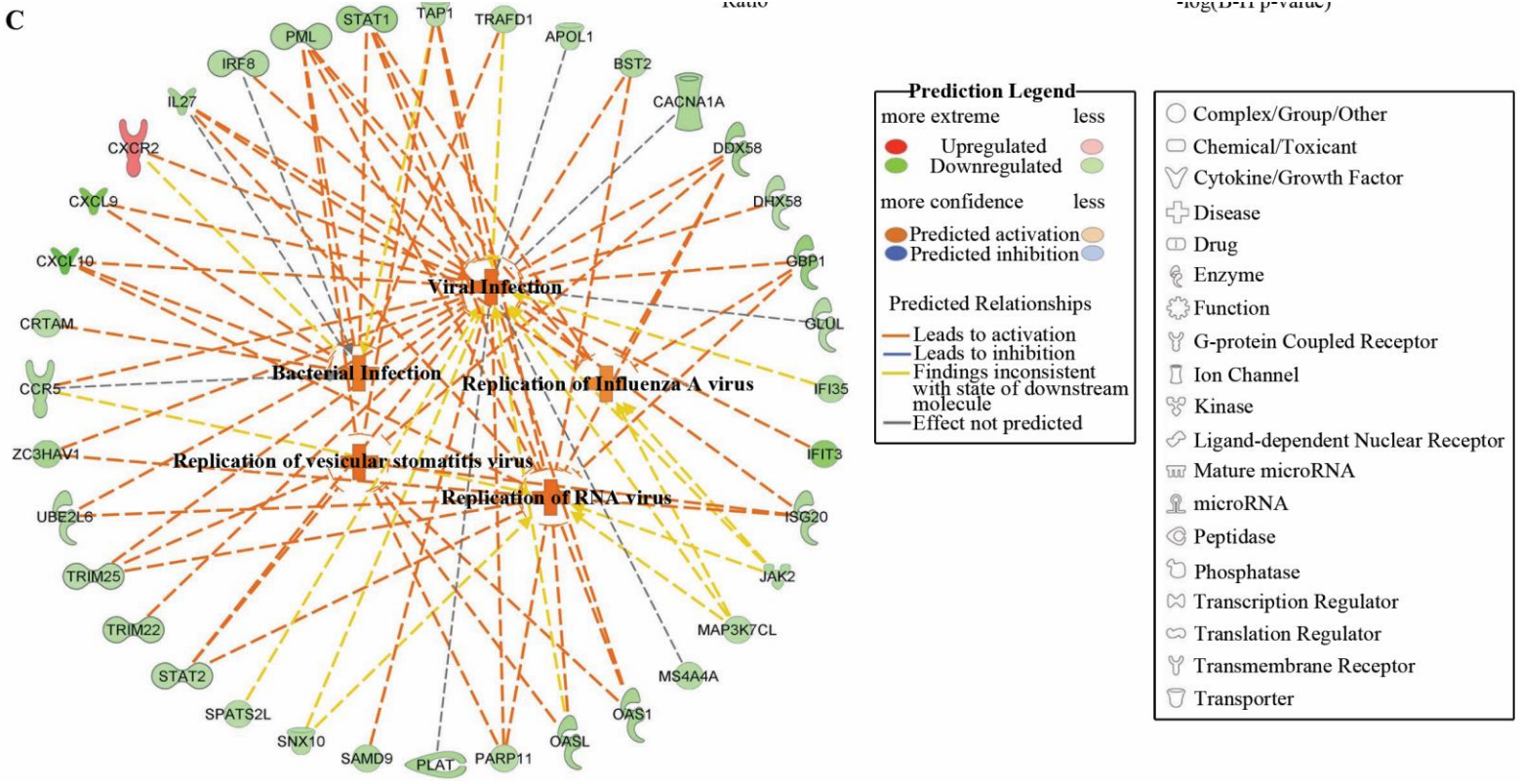
- by clicking on the pathway name, one can get a full picture of the genes involved in that pathway, with genes from the uploaded list specifically marked (example on the next slide: Development regulation of EMT)

Pathway Enrichment - Example Graphs produced with MetaCore
https://portal.genego.com



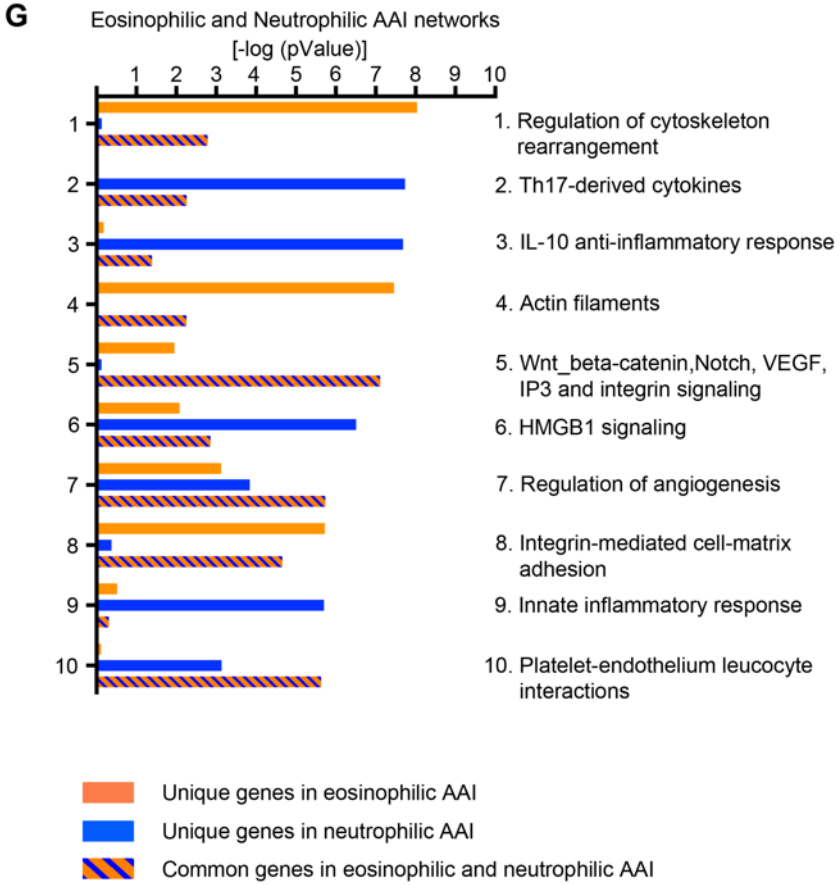
Pathway Enrichment - Example Graphs produced with IPA

commercial



Sokolowska M et al. *J Allergy Clin Immunol* 2017

Question 2: Is innate inflammatory response significantly enriched in eosinophilic AAI?

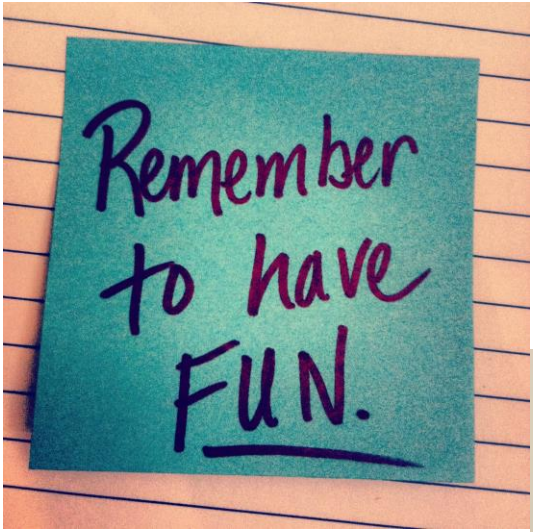


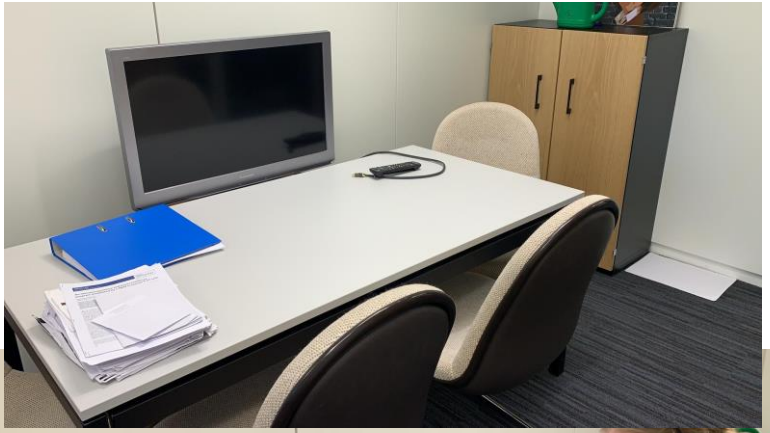
a) Yes
b) No
c) It is significant in both neutrophilic and eosinophilic AAI

Tan HT & Hagner S. et al *Allergy* 2019

Conclusions

- Functional annotation is reliable only for a handful of organisms (notably human and mouse)
- Bear in mind that there are many categories (>1000)
- **Be critical and inspect carefully your enrichment results (e.g. check with different tools)**
- Rank based methods generally are more robust and versatile
- Try to combine thresholds on p-values and fold-change to define the set of differentially expressed genes


Remember
to have
FUN.




...BORING

milena.sokolowska@siaf.uzh.ch