HAV815V

Practical Genomic data analysis in R

Module 6/8 : Enrichment Analysis / Single Cells.

Villemin Jean-Philippe, PhD - Bioinformatician
Jean-Philippe.villemin@inserm.fr

Institute of Cancer Research, Montpellier

UNIVERSITÉ DE MONTPELLIER

Gene Ontology Enrichment vs Gene Set Enrichment Analysis

Gene Ontology -GO- **Enrichment** performs a hypergeometric test comparing the set of "significant" genes against the "universe" (or background) genes.

Gene Set Enrichment Analysis GSEA() is a Komolgorov-Smirnov test on the whole gene list, testing if some category (e.g., a specific pathway) is more abundant at the top of the list than expected by chance. (two modes available **Standard** or **PreRanked**)

- Input are generally normalized counts from DESEQ2
- For standard mode you need to provide a file with phenotype labeling (class definition) for all samples.(Control vs Disease) .
- If you have fewer than 7 samples per group you would need to switch the permutation method from "phenotype" to "genes_set" .
- GSEA Preranked, because it doesn't have access to the sample level information has to run in gene_set permutation mode.
- FDR of 25% indicates that the result is likely to be valid 3 out of 4 times

While GO Enrichment require a list of input genes only, GSEA asks for an expression profile of all genes as its input file. So, a key difference is that GSEA does not require a cutoff - you use all your genes.

Gene Ontology Enrichment :
http://bioinformatics.sdstate.edu/go/

https://david.ncifcrf.gov/summary.jsp (Functional Annotation Chart)

Gene Set Enrichment Analysis :
https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/FAQ

https://cloud.genepattern.org/gp/pages/login.jsf

https://guangchuangyu.github.io/software/clusterProfiler/

Online Youtube tutorials :

https://liulab-dfci.github.io/bioinfo-combio/de.html

Venn Diagram :
https://bioinformatics.psb.ugent.be/webtools/Venn/

Database sources for annotation

GO Terms :

- Biological Process
- Molecular Function
- Cellular Component

KEGG Pathways

Reactome...

The Molecular Signatures Database (MSigDB) :

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.	**C5** **ontology gene sets** consist of genes annotated by the same ontology term.
C1 **positional gene sets** corresponding to human chromosome cytogenetic bands.	**C6** **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.
C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.	**C7** **immunologic signature gene sets** represent cell states and perturbations within the immune system.
C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.	**C8** **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.
C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.	MSigDB Molecular Signatures Database

https://www.gsea-msigdb.org/gsea/msigdb/

David

https://david.ncifcrf.gov/summary.jsp
(Functional Annotation Chart output treated with custom R script)

Functional Annotation Chart

Current Gene List: demolist2
Current Background: Homo sapiens
394 DAVID IDs

Options

Rerun Using Options | Create Sublist

Parameter Panel

Enriched Functional Annotations

Enrichment P-values

Sublist	Category	Term	RT	Genes	Count	%	P-Value
	GOTERM_MF_4	transcription factor activity	RT		45	11.4	4.0E-6
	GOTERM_MF_4	kinase activity	RT		42	10.7	3.1E-5
	GOTERM_MF_4	cytokine activity	RT		16	4.1	3.6E-4
	GOTERM_MF_4	sequence-specific DNA binding	RT		24	6.1	1.8E-3
	GOTERM_MF_4	phosphotransferase activity, alcohol group as acceptor	RT		30	7.6	3.3E-3
	GOTERM_MF_4	growth factor activity	RT		11	2.8	9.0E-3
	GOTERM_MF_4	damaged DNA binding	RT		5	1.3	1.4E-2
	GOTERM_MF_4	G-protein-coupled receptor binding	RT		6	1.5	2.5E-2
	GOTERM_MF_4	phorbol ester receptor activity	RT		3	0.8	3.8E-2
	GOTERM_MF_4	Ras guanyl-nucleotide exchange factor activity	RT		6	1.5	4.3E-2

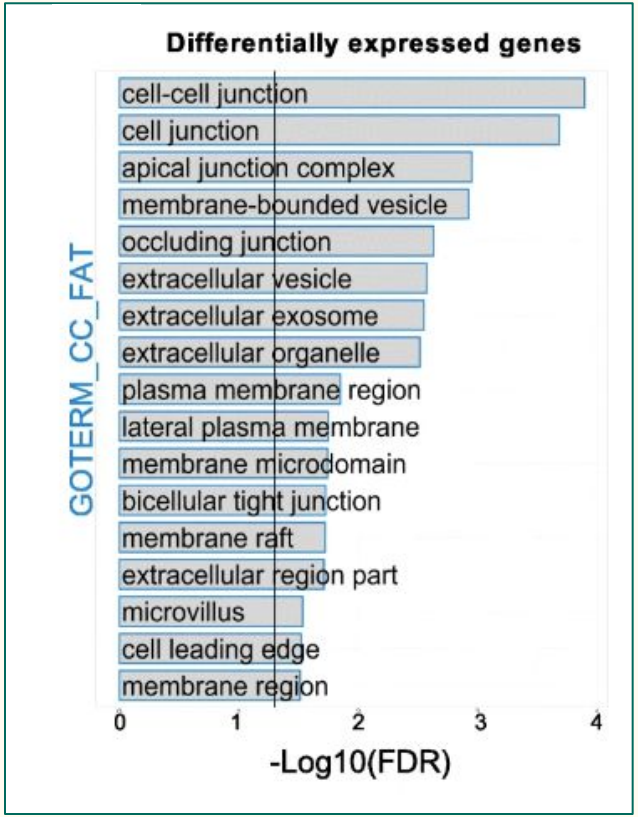
David :Functional Annotation Chart Output

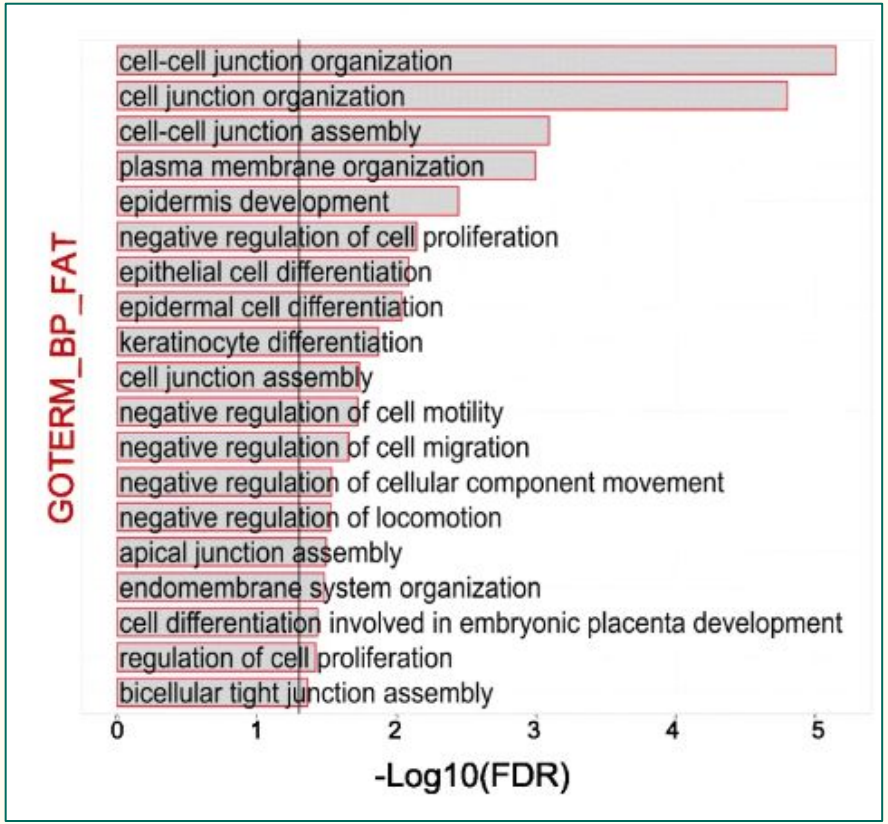
https://david.ncifcrf.gov/summary.jsp

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_CC_DIRECT	GO:0005886~plasma membrane	962	34.67916366258111	1.1094925937807518E-46	SPINT2, CLDN1, PREX1, ENDOU, CAPNS2, C3AR1, CX CLDN8, CLDN7, TTYH3, COL13A1, FPR1, IL20RA, IL20RB, LYPD3, SLCO5A1, SLC5A1, THY1, TAPBPL, IQGAP2, LYPD5, SLC5A3, LYPD6, IL22RA2, KCNMB4, RHBDL1, JAG2, NGFR, C LYPD6B, FAT2, NOX4, FXYD6, FAT3, NOX5, HRG, FERMT2, HLA-DRB1, ACHE, LRRC4, TMEM47, JPH2, TUBA1A, FLVCR2, HLA-DPA1, FGFBP1, GPR37, SPHK1, EMP1, EMP3, SLC7A4, S RAPGEF5, SLC29A1, GPRC5C, HCST, SLC29A2, CRB1, SERPINB12, PTAFR, KCNA7, SLC7A2, ADAM28, GPA33, ADAM23, HAS3, HAS2, CSMD3, FLNC, STX2, KCNB1, WNT7B, DAB2IP, DC HEPHL1, ADORA3, ADORA1, SLC16A6, PIM1, SLC16A7, ROS1, PPP1R16B, TICAM2, SDCBP2, SLC9A5, FRMD6, SLC9A7, SLC9A9, LY6D, PKP2, SLC27A6, PKP1, S100A9, SLC27A2, DSC CORO1A, CYP2C18, BMF, NPTX1, SLC38A5, GRIA3, SLC38A4, TNFRSF9, HTR1D, MCC, TNFRSF10D, PTPRB, HLA-DRA, NFE2L2, KCNK5, CNTNAP1, FLT1, ITGAM, ITGB5, LRRK2, ITGB3 ITGB7, DSP, OVCH2, ITGA1, HTR3A, NRG2, TNFRSF1B, SPTB, ADORA2A, ADORA2B, NRG4, OR7A5, DSG1, ITGA5, DSG3, LCP1, SLC25A4, TNFRSF21, SLC47A2, ATP10B, LRP3, TSPAN MARCKSL1, KCNJ6, SLC12A5, PCDH9, TIE1, PCDH7, MERTK, EFR3B, SORL1, SYT17, BAIAP3, SYT12, USH1G, SYT11, TSPAN18, IL2RB, PIK3AP1, RECK, EPHB6, ERRFI1, CPNE5, CP EPHB3, PAQR8, EPHA5, SLC34A2, ARL11, ENTPD2, ARL14, ENTPD3, ACTN1, SLC34A1, ENTPD8, ANK3, EREG, BTC, STX1B, SFRP1, FNDC4, KCNQ1, KCNQ3, CD226, TRIM16, EPHA1, PCDH12, CD1D, PCDH19, NKAIN1, RRAS, MICAL1, TSPAN7, TSPAN5, B4GALNT1, TSPAN2, CD14, TSPAN1, SLC10A4, NTNG2, SLC10A6, IFITM10, TNFSF15, TNFSF12, IL36G, NBEAL2, PTGDR2, DIRAS1, PTPRN, CTSZ, PTPRO, PTPRJ, GRIK2, SLC7A11, PTPRK, PTPRH, C1ORF210, AKAP12, KCNT1, MUC12, MUC15, CD36, LYNX1, FCER1G, MME, SYK, MMD, ANO4, SYN2 HSD17B8, STX11, REM1, PERP, GPC3, CD59, GPC4, PAK3, CD74, GABBR2, CD70, RFTN1, PLEKHA4, PCDHB12, KITLG, EPGN, XK, GPAM, KCNS3, PCDHB16, TACC2, CD68, IL18R1, C DUOXA2, CCRL2, ARHGEF40, IL6R, TGM2, KIRREL3, PDGFRB, PDGFRA, CD96, CD93, SCARA5, SLC30A2, CYBB, PRLR, OSCAR, ADAM19, SCNN1G, GPRIN2, SCN8A, SCNN1B, SCNN1A, D CACNA1D, CACNA1C, PLD1, CACNA1E, CACNA1G, HCAR2, ENHO, HCAR3, GRK5, LPXN, WNT3, WNT4, PACSIN1, MFAP3L, GGT5, MGAM, SYT1, DENND4C, SLC52A1, SYT9, SYT8, SYT7, G SIRPB2, LDB2, RND2, LTB4R, RND1, GRM4, CA2, CA9, HS3ST3B1, PLA2G4E, MMP2, PLA2G4C, ARHGEF18, EPN3, CDHR1, CAT, CDHR4, SAMHD1, CLCN1, IL1RL1, GNG2, CLMP, MARVE MX1, ATP2B4, MAPK10, OPN1SW, TEC, CD7, CD9, GPD1L, TEK, VIM, CNTFR, GLDC, FRMPD1, SERPINE1, SLC4A3, ABCA12, ENO2, EPS8, GPR173, GJA1, GPR176, ALCAM, EVA1A, GJ MAP1B, GPR160, TYRO3, PRKD1, GJC1, FAM171A1, FLRT2, S100A12, PLEK2, SLC19A2, PROM2, GSDMC, ST14, AFAP1L2, GSDMA, FZD5, MCAM, FZD8, CGN, SLC16A14, EHD3, GJB2, LPCAT1, FMN2, SLC6A2, ABCC11, ASGR1, PHEX, SLC6A4, FCRLA, GPR132, C1QTNF1, AIFM2, PLEKHN1, SERPINB2, SLC2A10, FCRL6, IFNGR1, NFAM1, SLC2A12, PRSS12, CD200R1, TRIM16L, ADCY4, ADCY7, ADCY5, GPNMB, EPCAM, GPR156, LAPTM5, GPR4, GPR3, HIPK3, SULF2, GPR141, GPR143, VEPH1, SLC28A2, PLCD4, PLCD1, DGKG, RAB3B, ATP8A2, CHRM1 BEST1, RAB44, CDH4, CDH3, CDH2, CDH1, LAMP3, SLCO2B1, GRAP2, CLEC5A, GRB10, SLC39A8, SLC26A11, STK32A, TECTA, SEMA6B, SLC15A2, RALGAPA2, CACNB1, DISP2, C1ORF1 NOTCH1, CD151, NOTCH4, ILDR1, RHOBTB1, GOLGA7B, ADAP2, TBXA2R, PLXNA2, CLCA2, DRD2, KIAA0319, CLCA4, PLXNA4, DRD4, STRA6, ATP8B4, ATP8B2, ATP8B1, ADAP1, DAB2, PIK3C2G, ADRA1B, TREM1, FCAR, PIK3C2B, PSTPIP2, SGCG, AMN, SLC13A4, UGT1A1, SGIP1, COBL, VSTM2L, RTN4RL2, KCNAB1, ADRA2A, RTN4RL1, LCK, PECAM1, SMPDL3B, LCT, P2RY2, P2RY1, SEMA4A, ANGPT1, TRPA1, KCNJ12, SEMA4B, CLCA3P, SEMA4C, KCNJ15, APOBR, PTPN13, CPPED1, P2RX7, P2RX6, NFASC, P2RX5, FRAS1, CROCC, RAB19, RAB17, FO AQP3, RAPGEFL1, EDNRA, VSIG2, INSC, SYNPO, GLUL, CDON, ARRDC4, ACSL1, IL1R1, ARRDC2, IL1R2, ARRDC3, SLC6A14, KRT1, MTUS1, SLC6A13, ACSL5, SLC6A11, PSG4, LLGL2 APCDD1, PPL, IGSF9, CRHR2, TRPM2, INPP5D, ATP6V0A4, TRPM6, TRPM3, GPM6B, CCR10, SPTBN2, SIGLEC15, PTPRN2, PDE2A, RAB39B, PARP14, ALOX15B, GNG11, VASN, OCLN, V SH3KBP1, GABRR2, EPB41L4B, GLIPR1, SLC22A14, UCHL1, SLC22A17, SLC22A18, DNER, CLDN23, VAV3, IZUMO1, INSRR, RHOH, BTN3A3, GP1BA, PDCD1LG2, IL17RE, RAB33A, RHOB RHOU, CLDN16, RHOV, TLR5, TLR4, TLR3, RAPSN, TLR2, PTGER4, PTGER1, AMIGO2, ZDHHC22, GDPD5, NKD1, DUOX1, NT5E, CXCR1, PDPN, CXCR2, PTCHD3, APBB1, LY6G6C, DUOX2 TJP3, CDK5R2, CDK5R1, F2RL3, STEAP4, PVR, AMOT, IL18RAP, NRCAM, CHRNB4, WNT5B, DSCAM, APLP1, WNT5A, KCTD7, ALDH3A1, CERCAM, STEAP1, ABCG4, STEAP2, C17ORF99, C GAS2L2, RAP1GAP2, TGFB1, OPN3, HSPA5, SMURF2, KLRC2, TGFB3, WNT3A, KLRC3, KLRC4, ESR1, POU2F3, NFKBIA, SLC2A9, EFNA3, TRPV6, BAMBI, DLG4, TRPV4, TGFBI, FGFR4, SLC2A1, HSPB1, SLC2A3, SLC2A4, SLC2A5, SLC2A6, SPRED3, TMEM100, SDR16C5, MFSD6, BDKRB2, CASP1, BLNK, BDKRB1, PDE4A, GPSM3, PTGIR, PCDHGA5, MRGPRX3, ARAP3, LSR LTA, CDH13, LTB, CDH16, DOCK2, CAMK1G, RGS18, RGS17, ATP1A1, TFPI, DLL1, RASD2, PRRG4, DLL4, MUC1, RASD1, ERBB3, PRRG2, LRIG1, STOM, NCAM1, S1PR3, PDE6A, S1PR SPRY1, HCN2	2543	5310	20580	1.4661539042591716	9.153313898691203E-44	9.153313898691203E-44	8.299004601480024E-44				
GOTERM_CC_DIRECT	GO:0005576~extracellular region	460	16.58255227108868	4.439422000602715E-37	PGLYRP4, CDA, PGLYRP3, SERPINE2, GMFG, COL12A1 SERPINF1, CEL, HSPG2, DKK1, UNC13D, BCAN, ACE2, RBP4, BCAM, SPINT1, PADI2, CFD, CFH, COL13A1, CFI, PDGFB, LYPD3, A1BG, THY1, LYPD5, LYPD6, IL22RA2, ADAMTS16, FN1, RNASE13, COL1A1, TMEM98, LYPD6B, REN, HRG, SERPINA3, ACHE, LAD1, SERPINA1, ELN, SERPINA6, ASGR1, C1QTNF1, ADAMTSL4, TMSB4X, TIMP2, TIMP3, GAST, SERPINB3, ANGPTL4, GAS6, HBEGF, CRB1, CTF1, COL11A1, FSTL1, FSTL4, FSTL3, NTF4, ADAM28, APOL6, FRZB, OLFML3, ABI3BP, ADAM23, RNASE7, AGR2, SPP1, METTL7A, APOL1, APOL3, CSF2, CSF1, TNC, DEFB1, OLFML2A, CLU, FGF2, TNF, CXCL16, FGF5, EFEMP2, CDH1, TNR, TECTA, COL27A1, IGFBP4, FST, IGFBP3, IGFBP2, VASH2, PGF, LPAL2, LY6D, S100A4							

Gene Ontology Enrichment

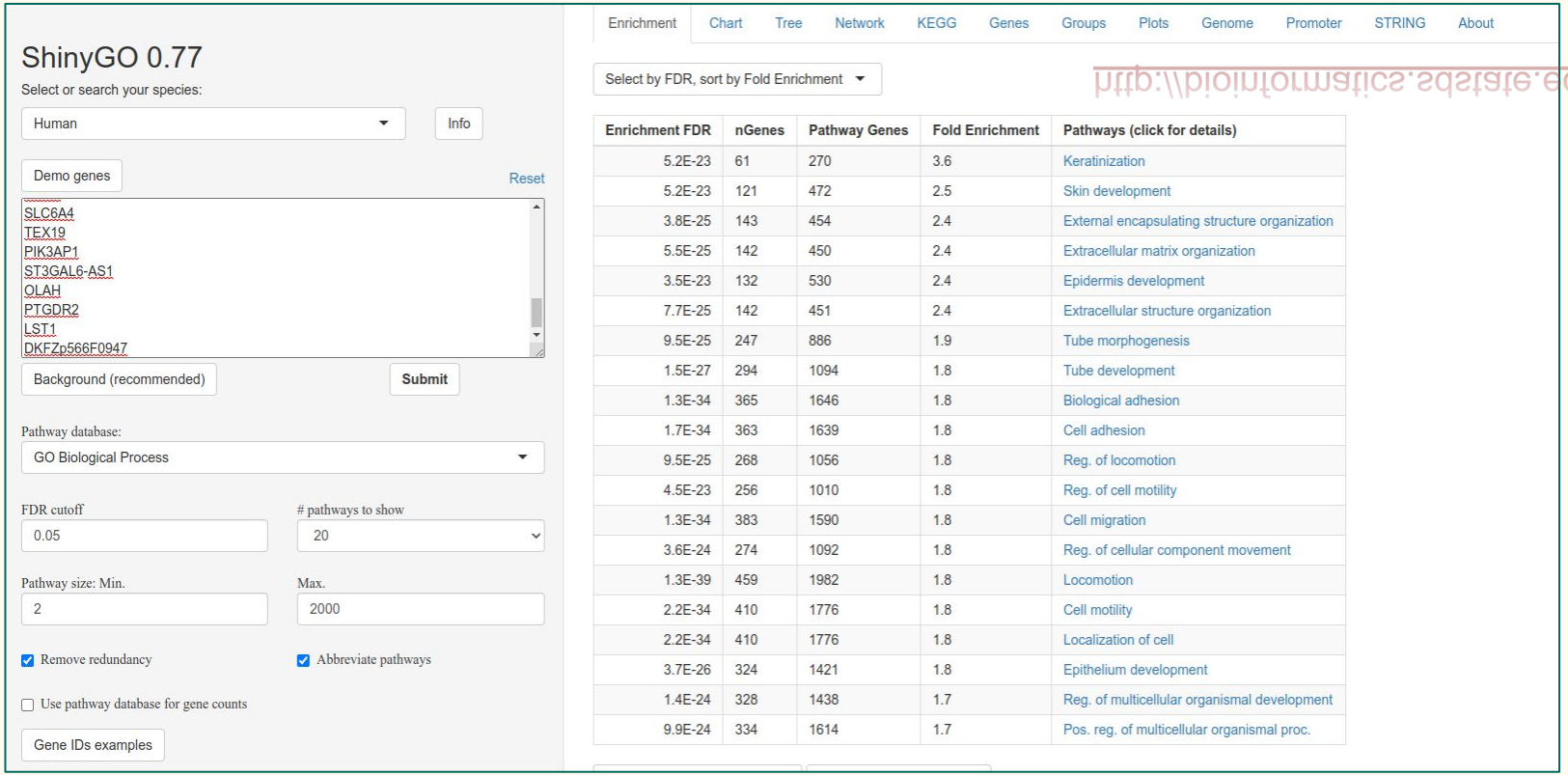
https://david.ncifcrf.gov/summary.jsp
(Functional Annotation Chart output treated with custom R script)





For the lazy ones (or the pragmatists)

http://bioinformatics.sdstate.edu/go/



Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathways (click for details)
5.2E-23	61	270	3.6	Keratinization
5.2E-23	121	472	2.5	Skin development
3.8E-25	143	454	2.4	External encapsulating structure organization
5.5E-25	142	450	2.4	Extracellular matrix organization
3.5E-23	132	530	2.4	Epidermis development
7.7E-25	142	451	2.4	Extracellular structure organization
9.5E-25	247	886	1.9	Tube morphogenesis
1.5E-27	294	1094	1.8	Tube development
1.3E-34	365	1646	1.8	Biological adhesion
1.7E-34	363	1639	1.8	Cell adhesion
9.5E-25	268	1056	1.8	Reg. of locomotion
4.5E-23	256	1010	1.8	Reg. of cell motility
1.3E-34	383	1590	1.8	Cell migration
3.6E-24	274	1092	1.8	Reg. of cellular component movement
1.3E-39	459	1982	1.8	Locomotion
2.2E-34	410	1776	1.8	Cell motility
2.2E-34	410	1776	1.8	Localization of cell
3.7E-26	324	1421	1.8	Epithelium development
1.4E-24	328	1438	1.7	Reg. of multicellular organismal development
9.9E-24	334	1614	1.7	Pos. reg. of multicellular organismal proc.

Gene Ontology Enrichment vs **Gene Set Enrichment Analysis**

Gene Ontology -GO- **Enrichment** performs a hypergeometric test comparing the set of "significant" genes against the "universe" (or background) genes.

Gene Set Enrichment Analysis GSEA() is a Komolgorov-Smirnov test on the whole gene list, testing if some category (e.g., a specific pathway) is more abundant at the top of the list than expected by chance. (two modes available **Standard** or **PreRanked**)

- Input are generally normalized counts from DESEQ2
- For standard mode you need to provide a file with phenotype labeling (class definition) for all samples.(Control vs Disease) .
- If you have fewer than 7 samples per group you would need to switch the permutation method from "phenotype" to "genes_set" .
- GSEA Preranked, because it doesn't have access to the sample level information has to run in gene_set permutation mode.
- FDR of 25% indicates that the result is likely to be valid 3 out of 4 times

While GO Enrichment require a list of input genes only, GSEA asks for an expression profile of all genes as its input file. So, a key difference is that GSEA does not require a cutoff - you use all your genes.

Gene Ontology Enrichment :
http://bioinformatics.sdstate.edu/go/

https://david.ncifcrf.gov/summary.jsp (Functional Annotation Chart)

Gene Set Enrichment Analysis :
https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/FAQ

https://cloud.genepattern.org/gp/pages/login.jsf

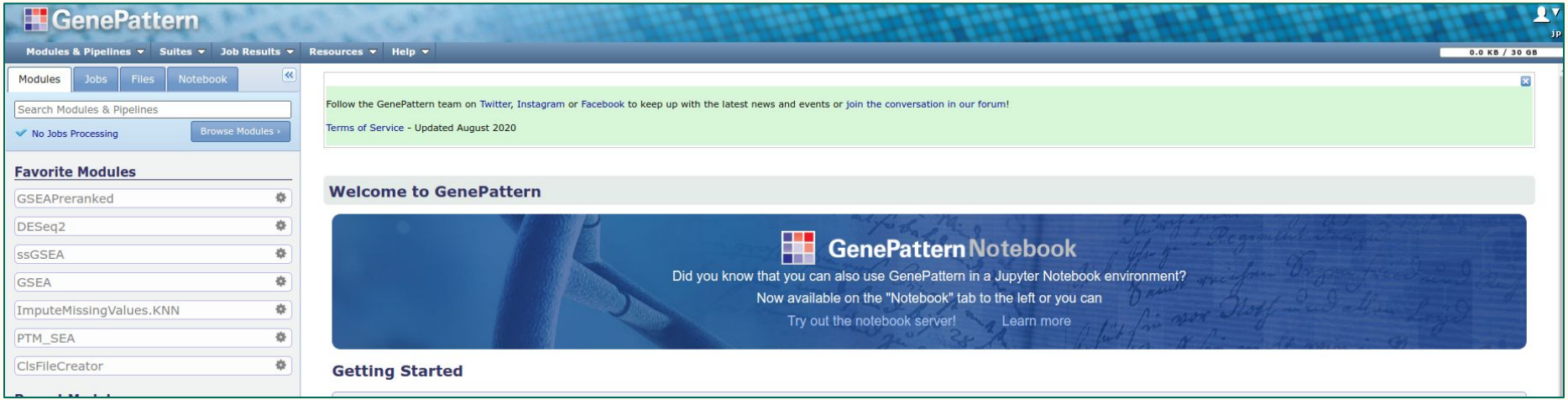
https://guangchuangyu.github.io/software/clusterProfiler/

Online Youtube tutorials :

https://liulab-dfci.github.io/bioinfo-combio/de.html

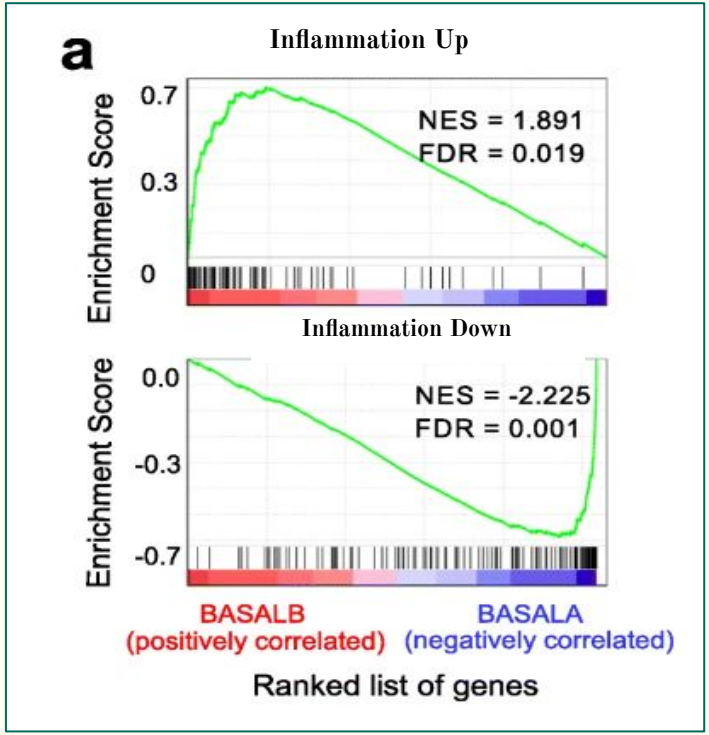
Venn Diagram :
https://bioinformatics.psb.ugent.be/webtools/Venn/

Gene Pattern (another online tool)



https://cloud.genepattern.org/gp/pages/index.jsf

Gene Set Enrichment (Over-representation) Analysis :



ES (enrichment score): reflects the degree to which a gene-set is overrepresented at the top or bottom of a ranked list of genes.

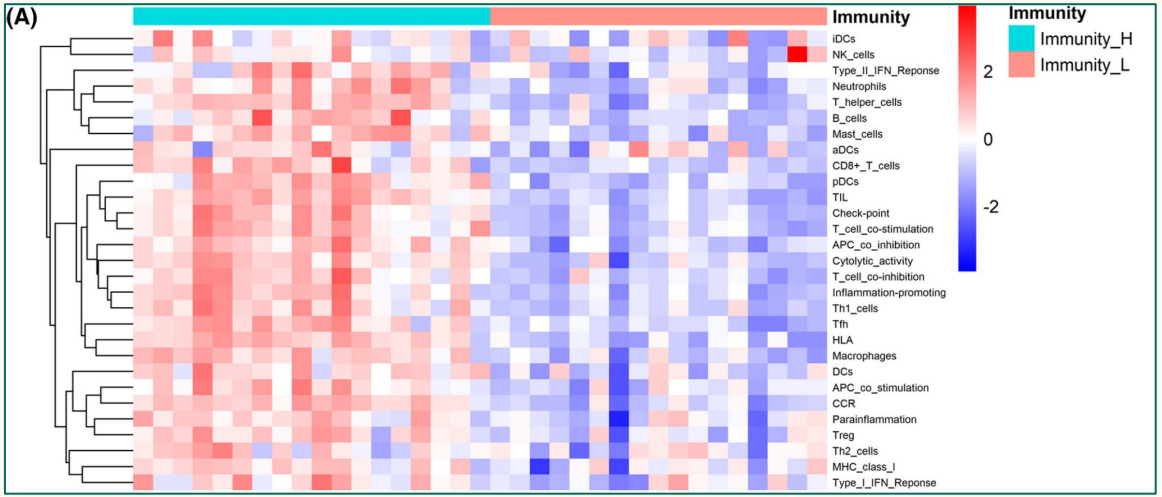
NES (normalized enrichment score): NES corrects for differences in ES between gene-sets due to differences in gene-set sizes.

Single Sample Gene Set Enrichment Analysis (ssGSEA)

Single-sample Gene Set Enrichment Analysis (ssGSEA) is an variation of the GSEA algorithm that instead of calculating enrichment scores for groups of samples (i.e Control vs Disease) and sets of genes (i.e pathways), it provides a single score for each each sample and gene set pair.

Advantages :

Independent of phenotype labeling. In this manner, ssGSEA transforms a single sample's gene expression profile to a *gene set* enrichment profile/score.
No need of all samples to be computed. Only one singe sample (**ss**GSEA)



Example : TGFb-Induced Program In Primary Airway Epithelial Cells (GSE61220)

Here, transforming growth factor-b (TGFb) activates gene expression programs to induce stem cell-like properties, inhibit expression of differentiated epithelial adhesion proteins and express mesenchymal contractile proteins. This process is known as epithelial mesenchymal transition (EMT);