Institute for Research and Innovation in Biomedicine

UNIVERSITÉ DE ROUEN NORMANDIE

Normandie Université

UE 5 Bioinformatique en sciences omiques 2

matière: Analyse et annotation en génomique et transcriptomique 2 – HD

Enrichment Analysis (EA) : toward the comprehensive functional analysis of large gene lists

Hélène Dauchel

Computer science, information processing and systems laboratory EA 4108

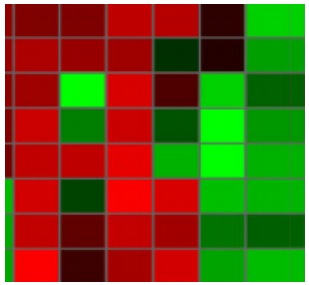
litis

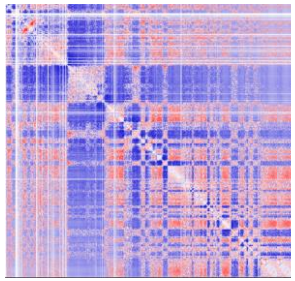
Team **TIBS**

Information Processing in Health and Biology

❑ **Differential analysis in high throughput experiments** : a common procedure to go insight molecular profile associated to two phenotypes

Normalization
Models and
differential analysis
(Cf UE4)

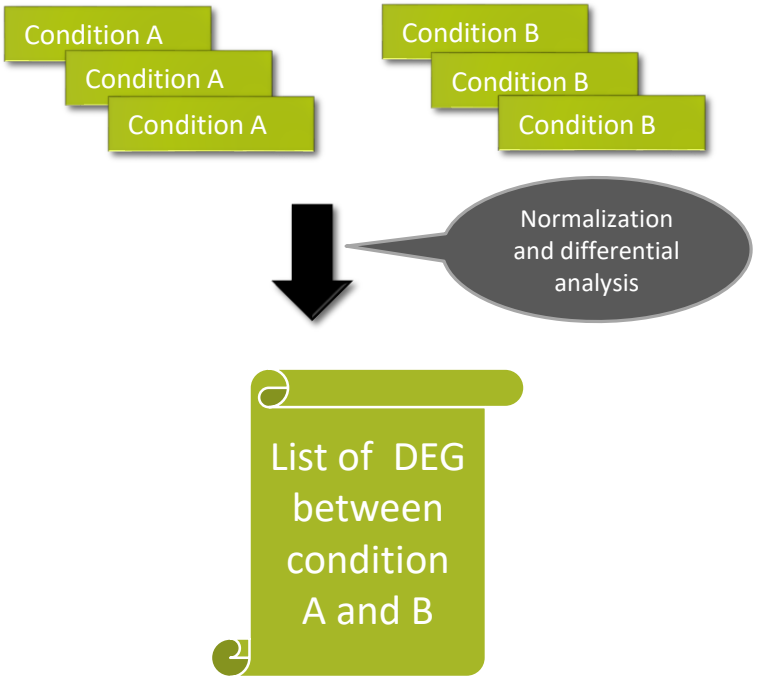




- ✓ RNA-seq : Differential Expression of Gene (DEG)
- ✓ ChIP seq and Hi-C : Peak calling and interaction matrix
- ✓ Comparative Metagenomics
- ✓ But also other omics…Proteomics…Interactomics

❑ **Molecular profile data** : The final stage of many genetic, proteomic, or metabolic analyses is the production of a list of 'interesting' biomolecules.

- ✓ lists of genes ranked by differential or co-expression investigated in transcriptomics experiments
- ✓ lists of single nucleotide polymorphism (SNP)-containing genes ranked by *p*-values determined by genetic association to a phenotype of interest through a genome-wide association study (GWAS)
- ✓ lists of putative transcription factor or miRNA targets ordered by probability.
- ✓ Lists of proteins
- ✓ Etc.



The case of RNA-seq : often thousands of genes which are used for the analysis

1. Context : Biological interpretation of differential data

The case of RNA-seq

Condition A Condition B

Whole transcriptome sequencing

Mapping on genome

Counting RNA reads corresponding to all genes of a genome

Comparing transcriptomes by statistical analysis

➢ **Normalization of counts** (different models)

➢ **Differential analysis**

Calcul of a Fold-Change for each gene

$$\frac{\text{Count Condition A}}{\text{Count Condition B}}$$

Expression in **Log 2**

Log 2 Fold Change

Statistical tests: calcul of a Probability P = **Test of the significativity** of the observed Fold-Change then Pval adj : FDR, FWER

List of DEG between condition A and B associated with a p- value and pval adj

Optional step

Filtering : Genes having a padj -value inferior to a threehold (usually 0,01 or 0,05) can be only retained.

List of significantly differentially expressed genes for a given padj value

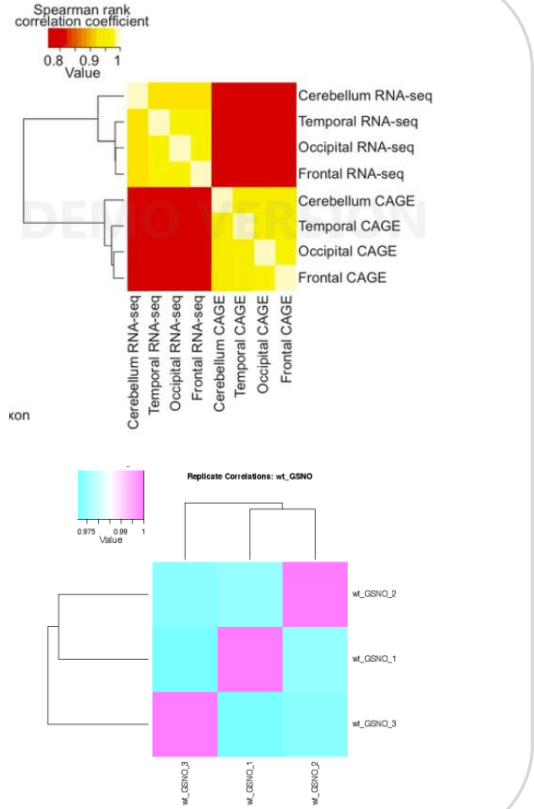
1. Context : Biological interpretation of differential data

id	baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj
F	12.1890439437(	23.7377255562253	0.640362331186405	0.0269765664646	-5.21214945378257	0.000352602716413288	0.00873102805488948
F	13.4562757526(	26.1859644092797	0.726587095931825	0.0277471963444(	-5.17151418495594	0.000365853472796627	0.00899722118645353
F	120.7327260214	205.053458307232	36.4119937356143	0.177573175484112	-2.49351443278781	0.000368138897676588	0.00903202241155467
F	1087.704212226	2015.32624783994	160.0821766133	0.0794323880735(	-3.65412881230948	0.000370057264029156	0.00905767520305326
F	30.2547976154'	55.0291204595349	5.48047477141162	0.0995922654341	-3.32782248616755	0.000372464357077757	0.00909514140647757
F	253.525048423	54.623538440241	452.426558405952	8.2826300039297'	3.05008894298517	0.000381894942802932	0.00930353454556063
F	24.64373901715	45.8800750971259	3.40740293726723	0.0742675972097(	-3.75112328642914	0.000383221124129706	0.00931397851573322
F	1868.64553374(	3495.38682105273	241.904246437232	0.0692067169734(	-3.8529441219421	0.000391858924802309	0.00948679362370035
F	47.91835812697	15.8389190159735	79.9977972379775	5.0507106676472	2.33648639866749	0.000392159805797596	0.00948679362370035
F	502.1345963153	889.784608306999	114.484584323675	0.128665503150815	-2.95830279427682	0.000405404290715818	0.00975558052828336
F	94.52312609765	32.651240112349	156.39501208304	4.78986438324866	2.25998480911711	0.000404492662578144	0.00975558052828336
F	102.2028884533	174.399265074172	30.0065118322796	0.172056412161587	-2.5390464357602	0.000406090844885181	0.00975558052828336
F	477.9738199104	804.777215184255	151.170424636643	0.187841332712226	-2.41241354553835	0.000409409466376248	0.00981258993545659
F	354.6652143723	596.030609185225	113.299819559504	0.190090605773335	-2.39524085874006	0.000411867322171289	0.00984875361634478
F	1395.741389473	2606.9853227829			04	0.00041346012957271	0.00986411316024273
F	1476.565717936	193.046083660297			4	0.000414493483606259	0.009866085717582
F	484.676376330	868.712364723123			28	0.000416409891934221	0.00988902027115183
F	118.4885137784	210.129556526221			78	0.000422059737354687	0.0100003103978697
F	147.2398418053	33.1080878275052			6	0.000430246905943441	0.0101710760589545
F	80.6746048924	143.769232485127			78	0.000432059871147548	0.0101907212335665
F	472.6168309916	855.613986319722	89.6196756639412	0.104743116752246	-3.25507265529385	0.000441555909427073	0.0103910821497373
F	77.4250043924	140.019865701284	14.8301430835683	0.10591456440335	-3.23902710567285	0.00044478257067235	0.0104433337521214
F	16.9430677180(	31.4434276899411	2.44270774623912	0.0776857971823(	-3.6862053261162	0.000447001472507941	0.0104585227345027
F	453.443307838	785.891890799193	120.994724877205	0.153958484934886	-2.69338671517785	0.000447444988833996	0.0104585227345027
F	76.5823507628	18.9451464148729	134.219555110785	7.0846406869368'	2.82469468556116	0.000454112351967445	0.0105727331117153
F	44.3768868394	8.45756213782498	80.2962115410742	9.49401378701828	3.24701814482195	0.000454368757739931	0.0105727331117153
F	744.6242011482	138.947961092578	1350.30044120388	9.71802990548531	3.28066387209596	0.000459003234557312	0.01065667912357
F	100.3048054246	24.3565454295428	176.253065419693	7.23637372670717	2.85526691870319	0.000460207072570688	0.0106607790159344
F	11633.31584648	22261.7938096746	1004.83788328807	0.045137327740922	-4.46953518039521	0.000466266834240085	0.0107531493460969
F	192.850365427'	328.326366713301	57.3743641422887	0.17474796409631	-2.51665244732822	0.000466180116755962	0.0107531493460969
F	733.0432176618	1317.95964958839	148.126785735325	0.112390986917985	-3.15340175016565	0.000479914387007424	0.011043351459785
F	2820.501806946	5354.14087193725	286.862741944265	0.053577735215715	-4.22222259146343	0.000483548807911145	0.011102366213506

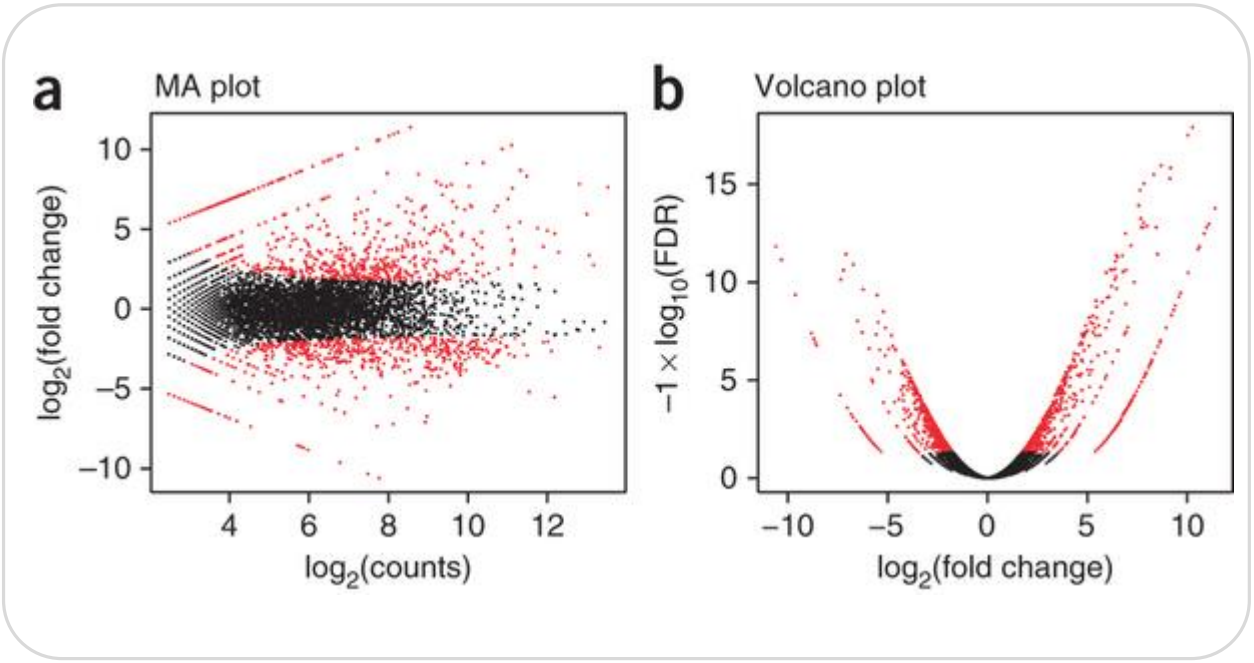
ID

DEG results : list of gene IDs ranked by *p*-values
(marked threehold at 0,01)

➢ Gene expression profiling : whole data inspection, replicats, representation

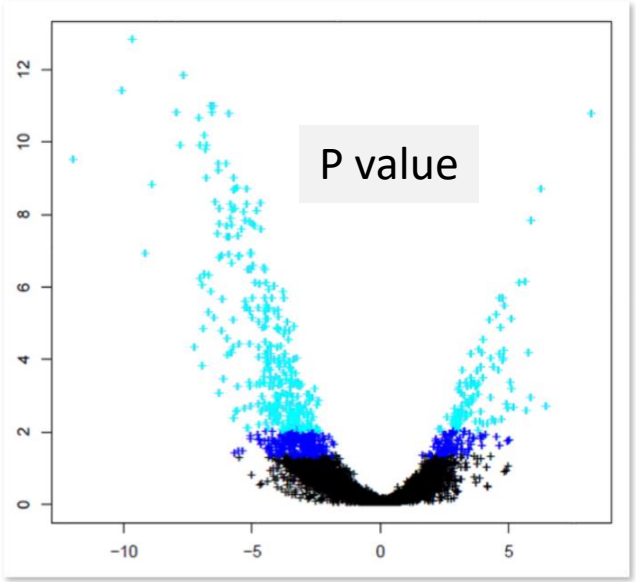


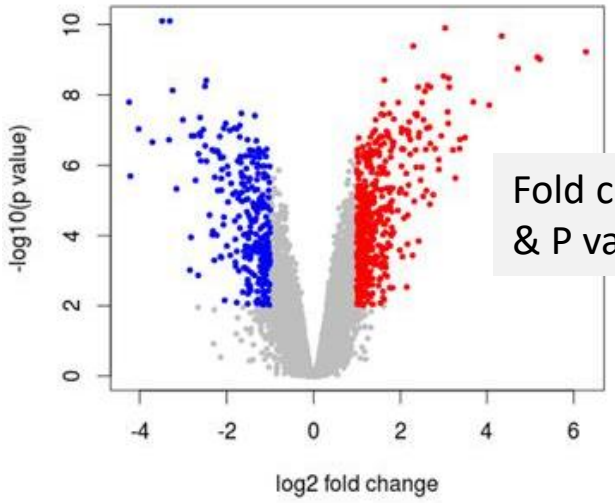
Pearson correlation matrix for all replicates



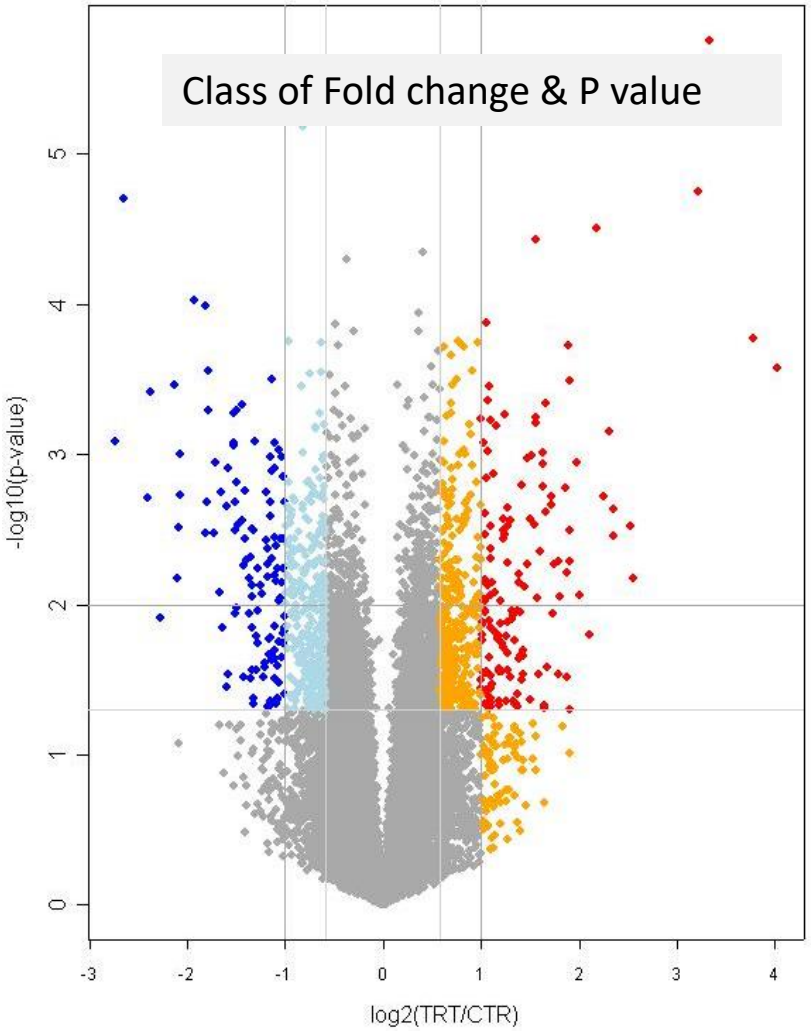
Dispersion estimation

Different threeholds to reduce data before functional analysis



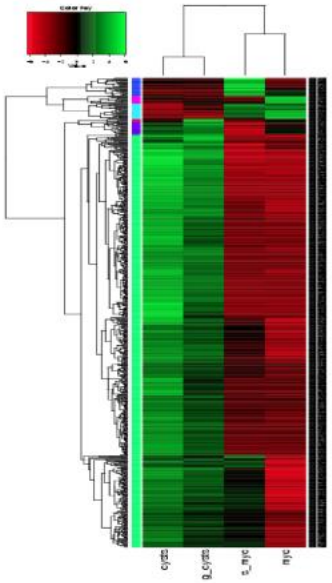


Fold change
& P value



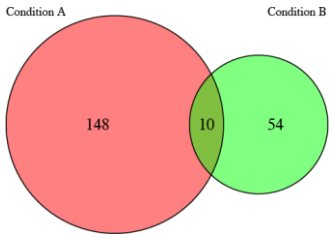
➢ Gene expression profiling : one list or many lists

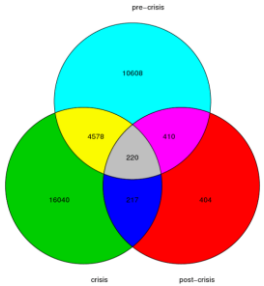
- ✓ **Without classification :** leads to only one list of genes > threeholds or top genes
- ✓ **With classification of genes , representation with heatmap :** leads to many list of genes

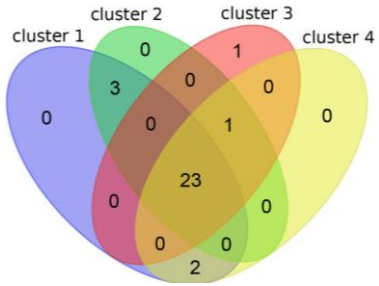


- Clustering performed across samples (raw) and genes (rows)
- Numerical expression value for each significantly and differentially gene are replaced by a color-coded expression value. (threeholds or top genes)
- Genes are classified on the basis of their expression profil : **group of co-regulated gene**

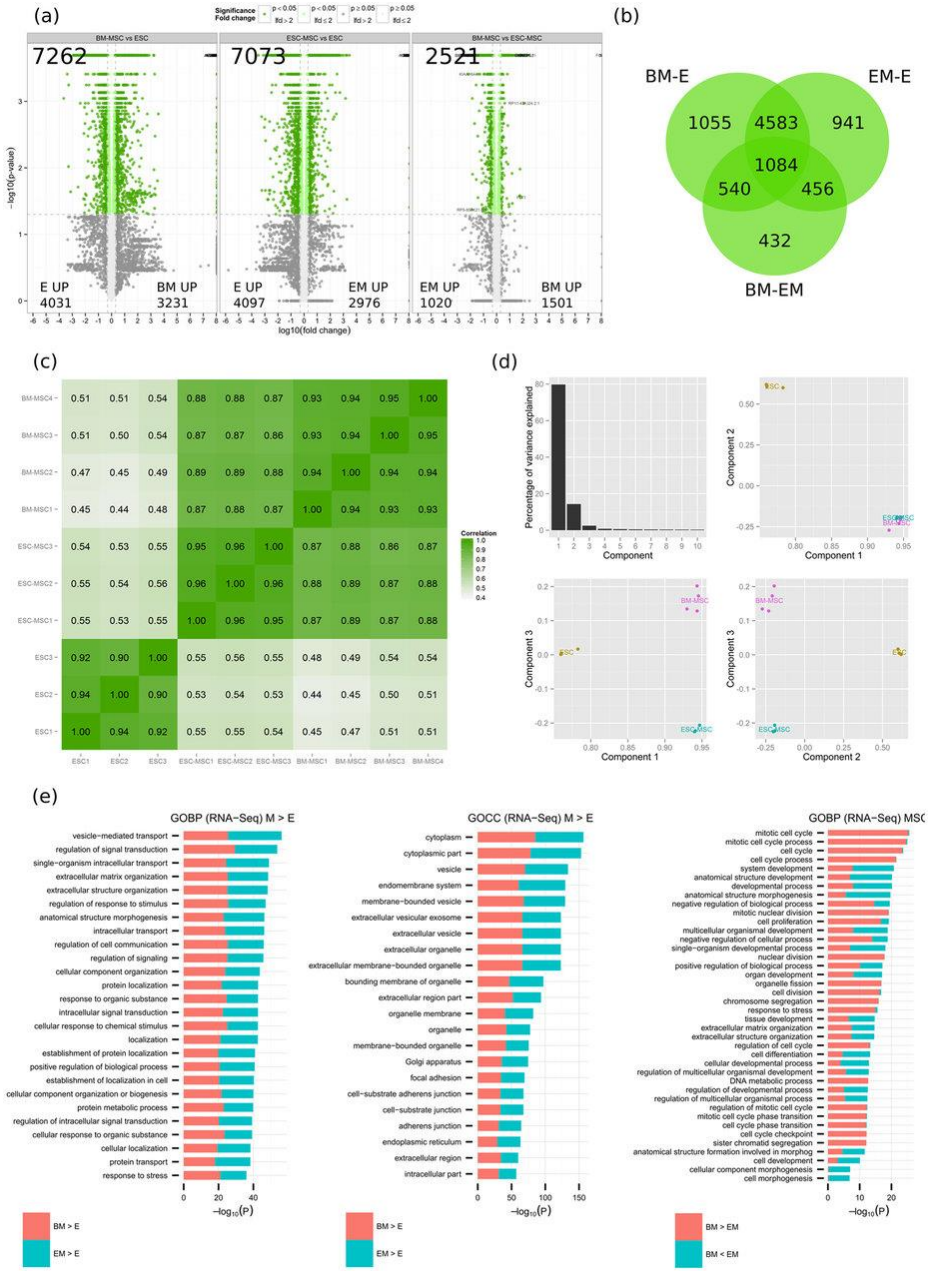
- ✓ **Intersection analysis** : leads to many list of genes
  - Venn diagram





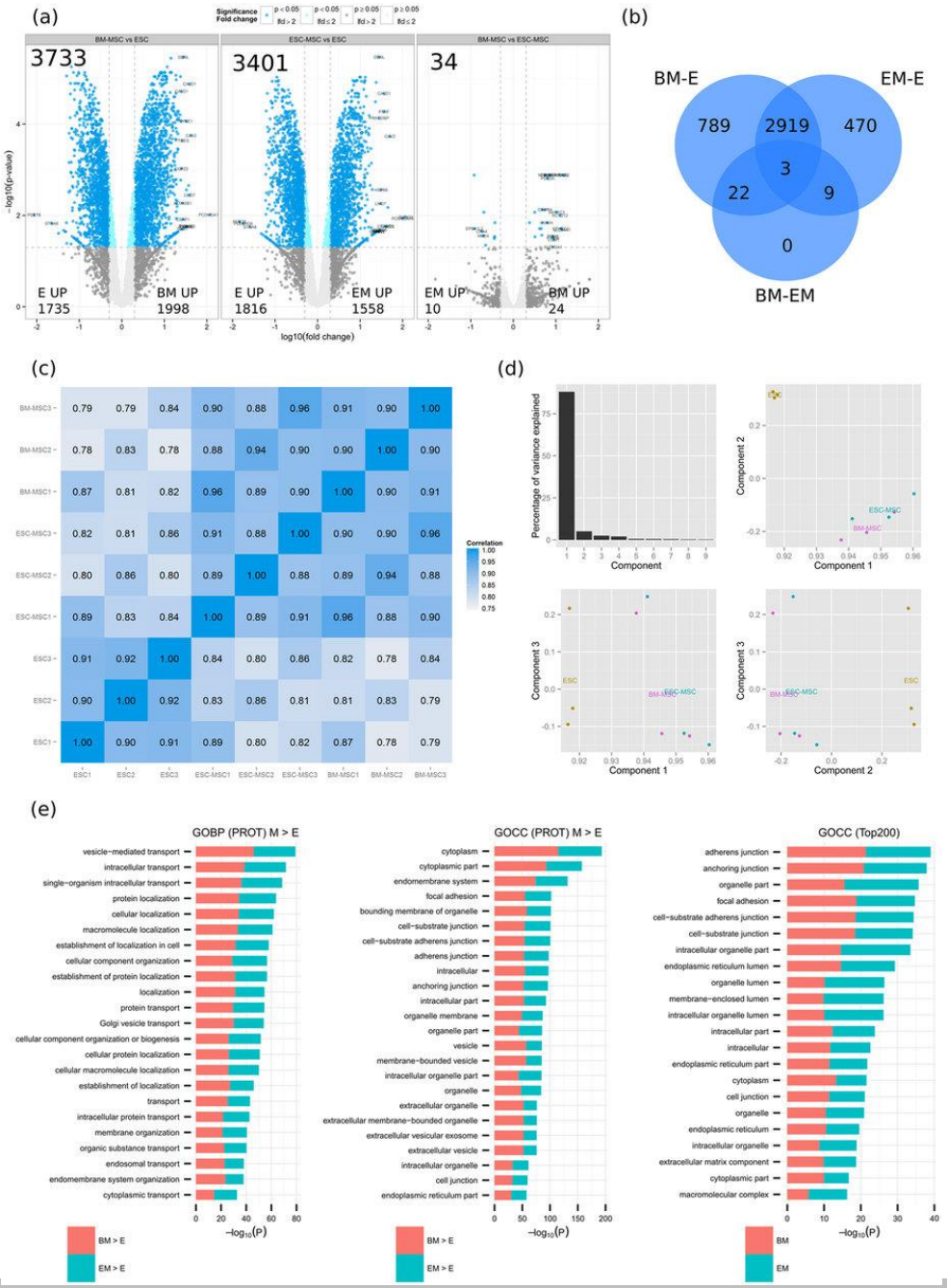


Transcriptome analysis (RNA) of ESC-MSC (EM), BM-MSC (BM) and ESC (E) by RNA-seq.



(**a**) Volcano plots for BM-MSC vs ESC (left panel), ESC-MSC vs ESC (middle panel) and BM-MSC vs ESC-MSC (right panel). Numbers are given for the total of differentially expressed genes as well as for the up-regulated per cell type. (**b**) Venn diagram of SDEs per comparisons (BM-MSC vs ESC, ESC-MSC vs ESC and BM-MSC vs ESC-MSC). (**c**) Pearson correlation matrix for all replicates. (**d**) PCA analysis based on FPKM values; principal components 1 to 3 were plotted against each other. (**e**) GO-term enrichment analysis for biological process (GOBP) and cellular component (GOCC). Bar charts represent the most significant top 20 terms of each category for each cell type sorted by mean $-\log_{10}$ p values.

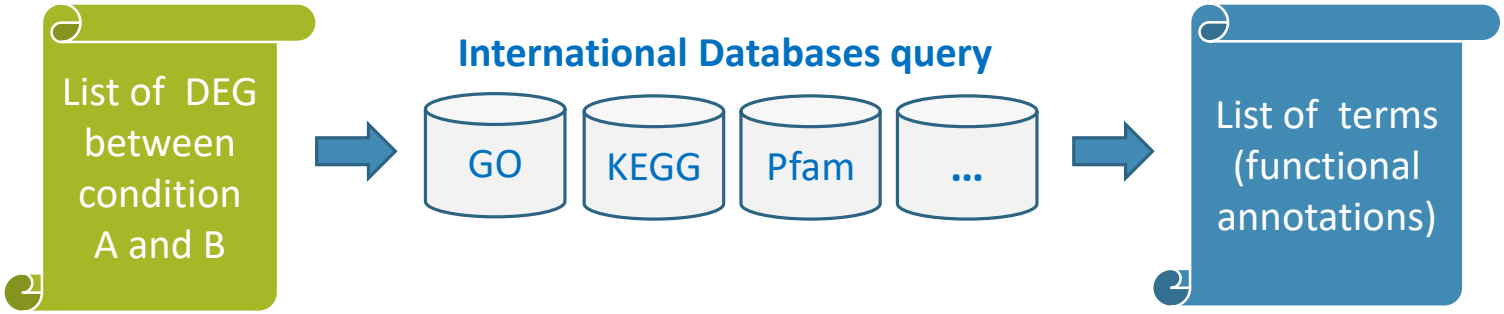
From: Comprehensive transcriptomic and proteomic characterization of human mesenchymal stem cells reveals source specific cellular markers



Proteome analysis (PROT) of ESC-MSC (EM), BM-MSC (BM) and ESC (E) by nano LC-MS/MS.

- **Enrichment analysis** : a set of **statistical approaches** to identify significantly enriched or depleted groups of genes .. supposing to be able to provide valuable insight into the collective biological function underlying a list of genes.

  - **First step : functional annotation by mapping** genes and proteins to their associated biological annotations : gene ontology [GO] terms or pathway membership etc.



- *Reminder your project last year : « scripting pour l'agrégation automatique d'annotations »*

- **Enrichment analysis** : a set of **statistical approaches** to identify significantly enriched or depleted groups of genes .. supposing to be able to provide valuable insight into the collective biological function underlying a list of genes.

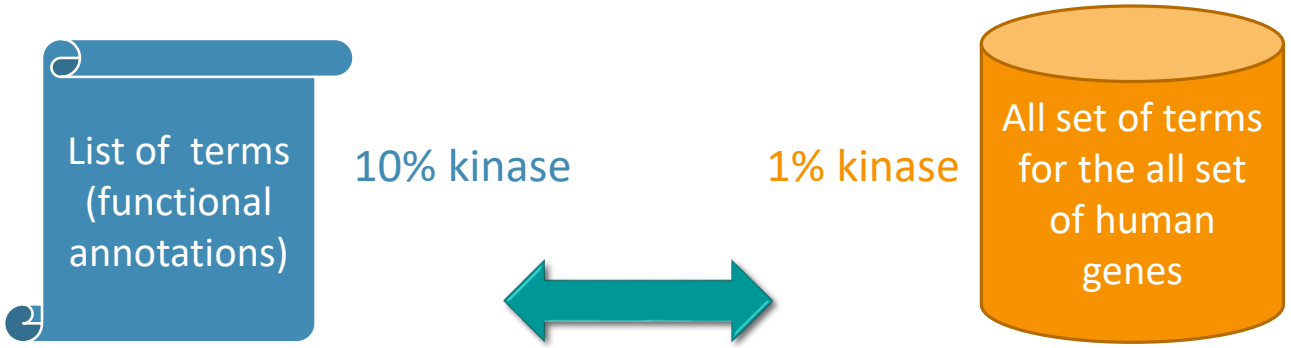
  - **First step : functional annotation by mapping** genes and proteins to their associated biological annotations : gene ontology [GO] terms or pathway membership etc.

  - **Second step : comparing the distribution of the terms** within a gene set of interest with the background distribution of these terms
    - enrichment analysis can identify terms which are statistically over-or under-represented within the list of interest, significantly function over or under represented.
    - It is inferred that such enriched terms describe some important underlying biological process or behaviour.
    - Evidence of biomarkers for the biological system studied

For example : if 10 % of the genes on the 'interesting' list are kinases, compared with 1 % of the genes in the human genome (the population background), by using **statistical methods**, it is possible to determine that kinases are enriched in the gene list and therefore have important functions in the biological study undertaken
(from Hum Genomics. 2010 Feb;4(3):202-6.)

Result of functional annotation of a list of human gene IDs

Genome background



Enrichment analysis by statistical methods

❑ Statistical Methods : tests

Enrichment p-value (and adj p-value) calculated by comparing the observed frequency of annotation term with the frequency expected by chance (from the reference background). Null hypothesis is : genes of the list are picked at random from the total gene population (background)

- ✓ χ2-test
- ✓ Fisher's exact test
- ✓ binomial probability
- ✓ hypergeometric test

❑ Three classes of enrichment algorithms

- ✓ Singular enrichment analysis (SEA) : preselected interesting genes
- ✓ Gene set enrichment analysis (GSEA) : all genes
- ✓ Modular enrichment analysis (MEA) : consider relationship between annotation terms

Exemple for GO terms

Category	Term	Genes	Count	%	P-Value
GOTERM_BP_ALL	response to chemical stimulus		14	8.2%	6.1E-5
GOTERM_BP_ALL	response to abiotic stimulus		15	8.8%	6.5E-5
GOTERM_MF_ALL	protein binding		55	32.2%	8.8E-5
GOTERM_BP_ALL	response to bacteria		7	4.1%	1.7E-4
GOTERM_MF_ALL	iron ion binding		10	5.8%	2.6E-4
GOTERM_BP_ALL	cell-cell signaling		15	8.8%	4.0E-4
GOTERM_BP_ALL	defense response to bacteria		6	3.5%	5.4E-4
GOTERM_BP_ALL	regulation of hydrolase activity		6	3.5%	6.1E-4
GOTERM_BP_ALL	regulation of GTPase activity		5	2.9%	9.8E-4
GOTERM_BP_ALL	response to stress		22	12.9%	9.9E-4
GOTERM_BP_ALL	response to other organism		15	8.8%	1.2E-3
GOTERM_MF_ALL	heme binding		6	3.5%	1.3E-3
GOTERM_MF_ALL	tetrapyrrole binding		6	3.5%	1.3E-3
GOTERM_BP_ALL	response to stimulus		40	23.4%	1.4E-3
GOTERM_MF_ALL	receptor binding		14	8.2%	1.6E-3
GOTERM_BP_ALL	response to pest, pathogen or parasite		14	8.2%	2.0E-3
GOTERM_BP_ALL	behavior		8	4.7%	2.2E-3
GOTERM_BP_ALL	defense response		23	13.5%	2.2E-3
GOTERM_MF_ALL	oxygen binding		4	2.3%	2.7E-3
GOTERM_BP_ALL	inflammatory response		8	4.7%	3.3E-3
GOTERM_MF_ALL	sodium ion binding		5	2.9%	3.6E-3
GOTERM_BP_ALL	response to biotic stimulus		23	13.5%	3.8E-3
GOTERM_MF_ALL	carbohydrate binding		8	4.7%	3.9E-3
GOTERM_BP_ALL	sodium ion transport		6	3.5%	4.0E-3

From DAVID

❑ Visualization

- ✓ Bar plots
- ✓ Pie charts
- ✓ Bloc charts
- ✓ Semantic spaces (dotplot)
- ✓ GO graphs
- ✓ Pathways maps with fold-changes on regulated genes

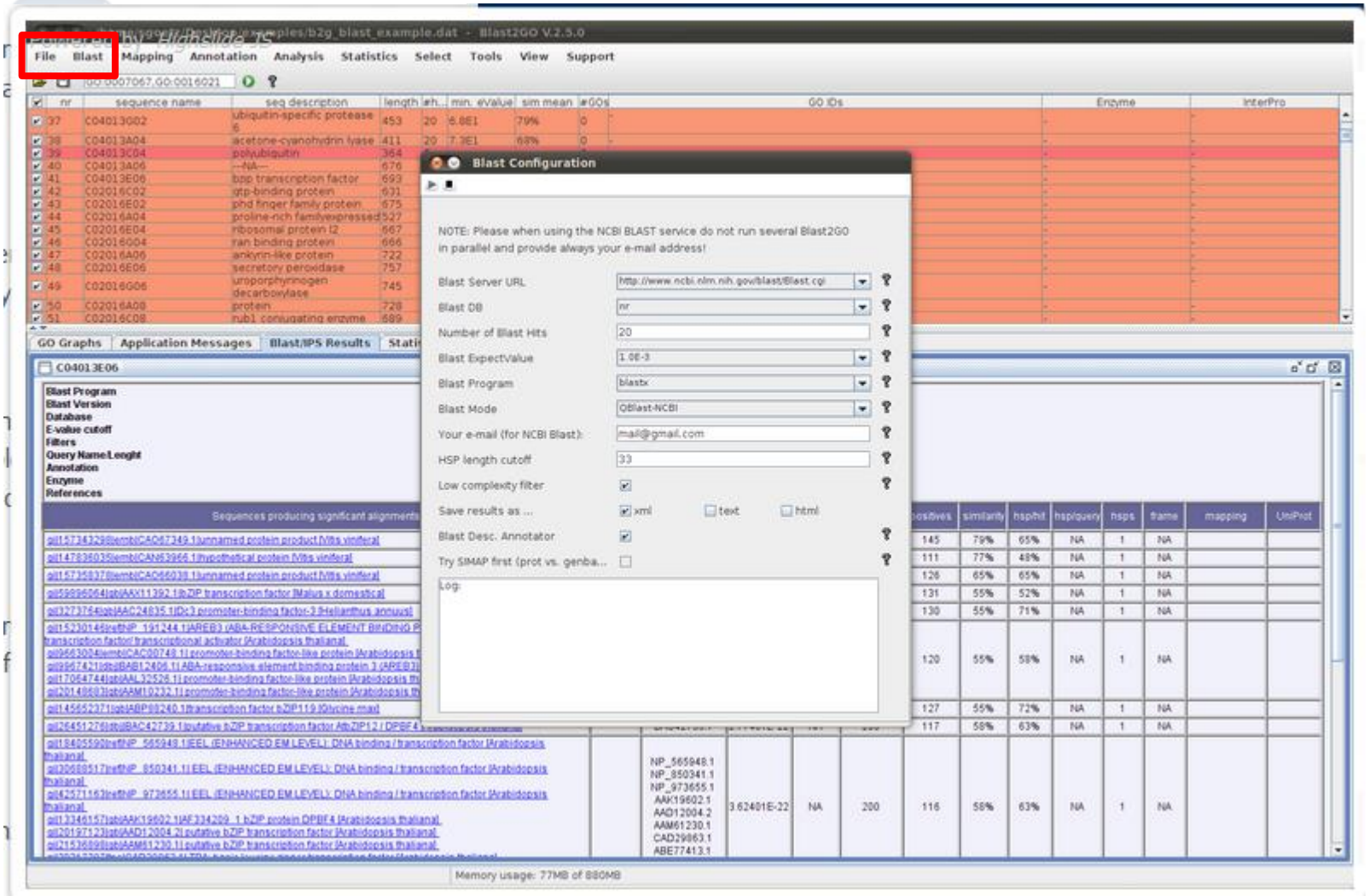
❑ Annotation enrichment Tools

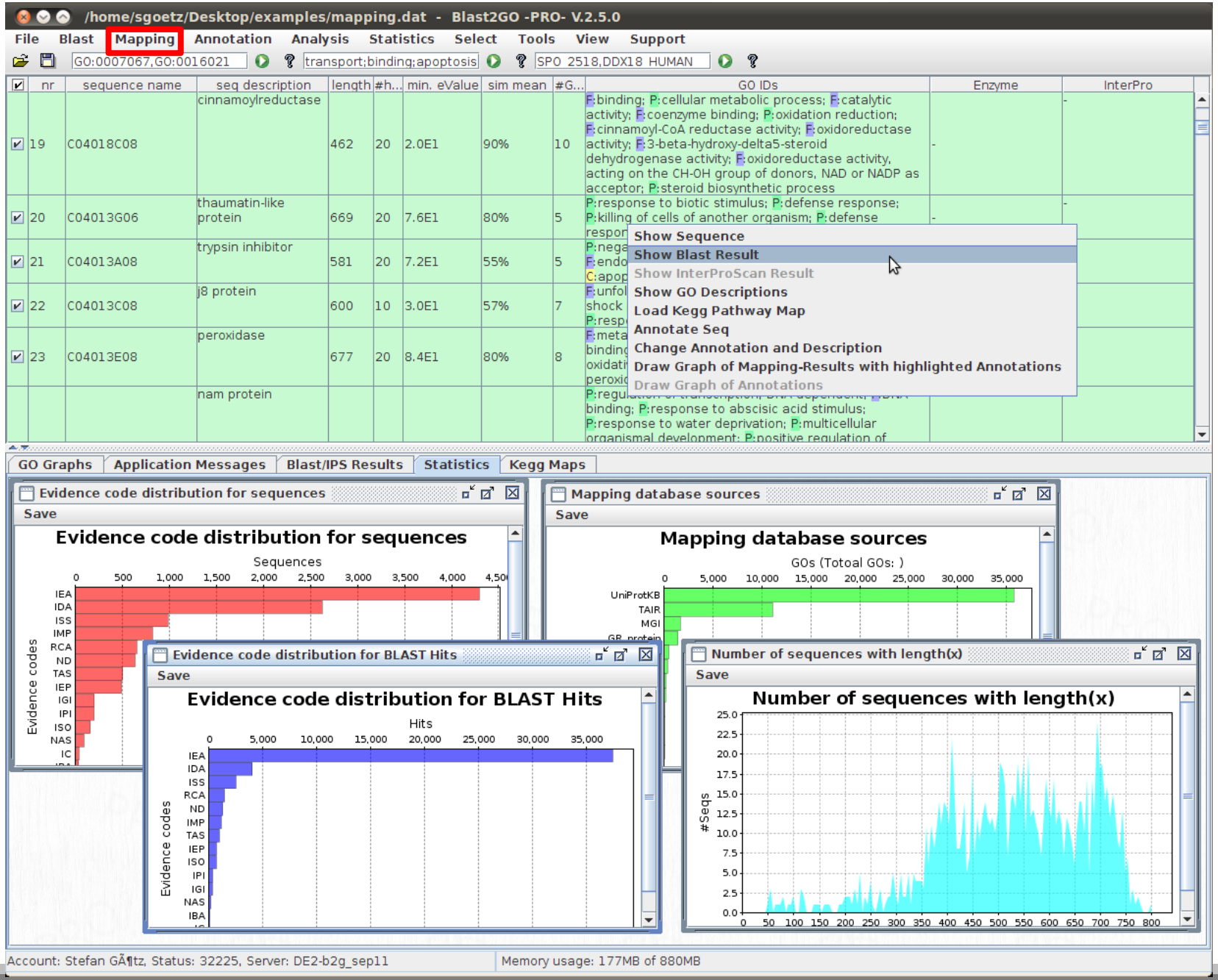
- ✓ R Packages : **clusterProfiler, Pathview, biomaRt** (for ensembl BioMart.)
- ✓ Dedicated softwares :
  - ✓ most GO terms : DAVID, GSEA, GO consortium, GOStat, FatiGO, GOrilla, gProfiler, PANTHER, ReviGO…
  - ✓ few for pathways : DAVID, GSEA, KEGG (KO), Reactome, PANTHER
  - ✓ Domains : Blast2GO, other ?
- ✓ The special place of blast2GO : for not well annotated genomes (background problem)

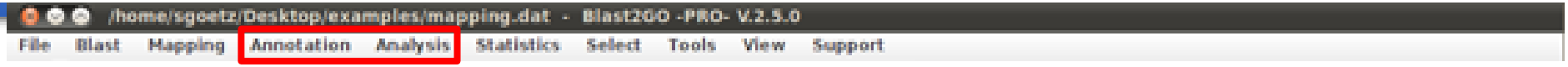
blast2go

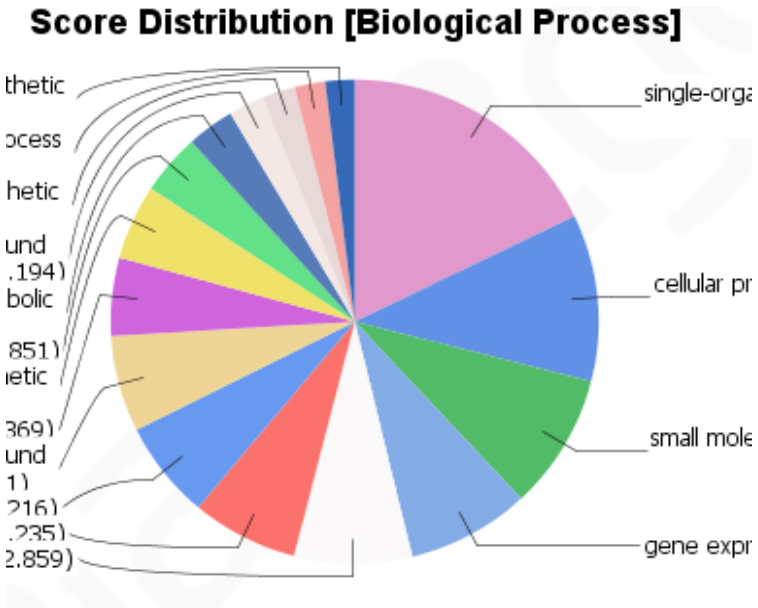
Pour les génomes mal ou peu annotés

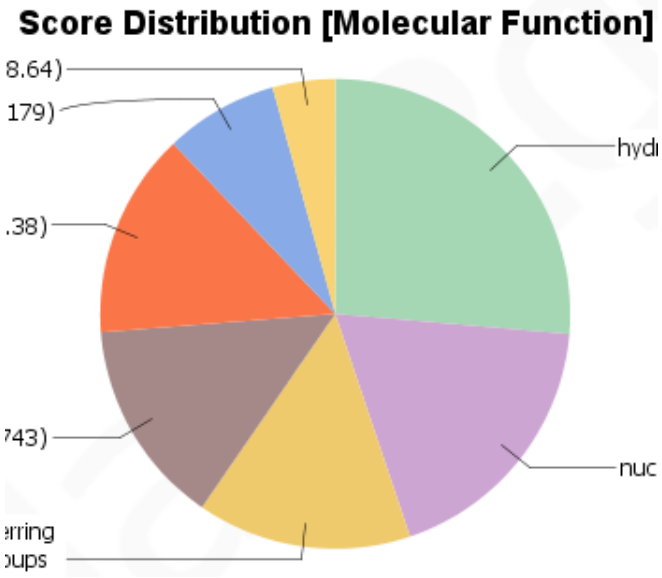
Gene Ontology, KEGG maps, InterPro and Enzyme Codes

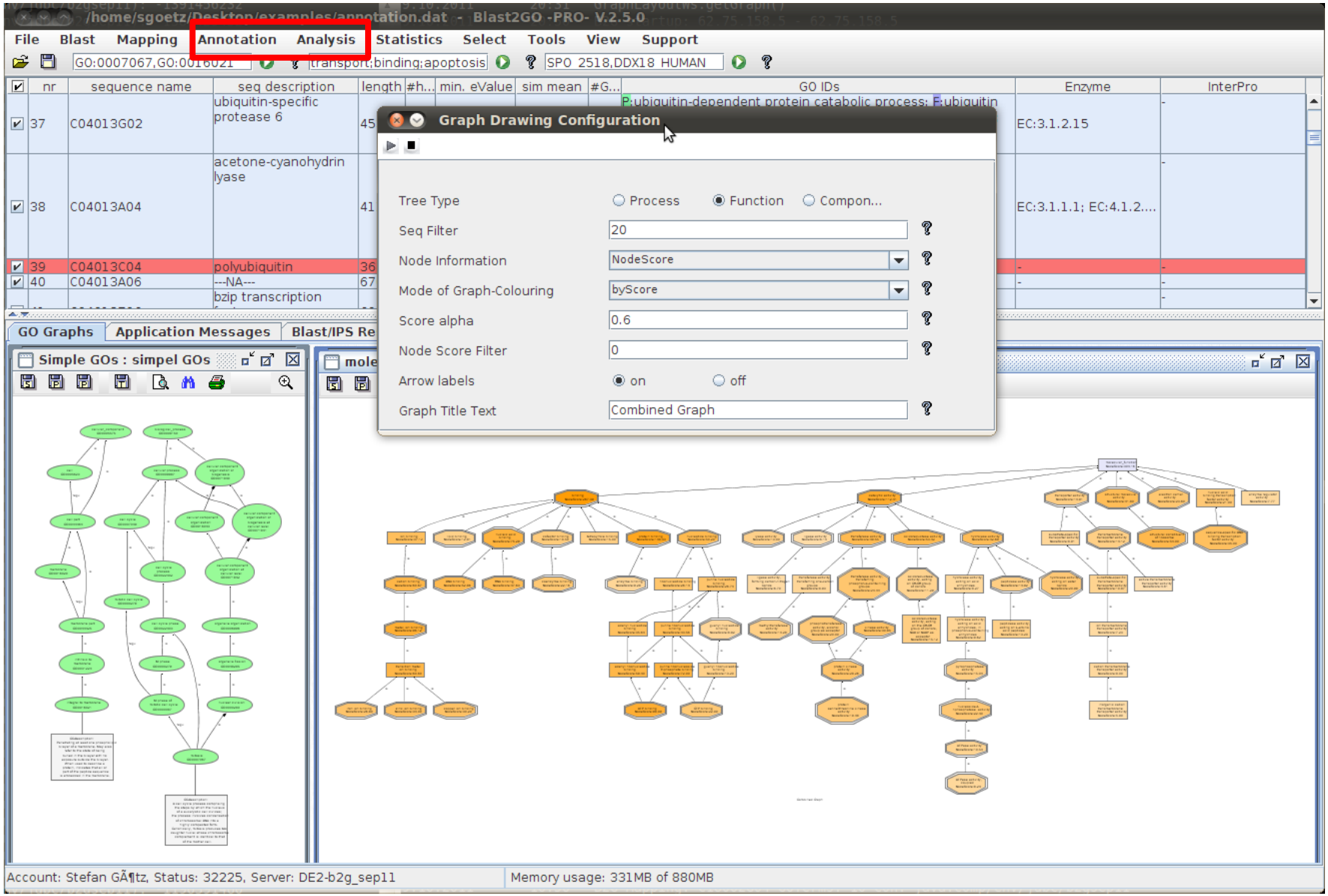


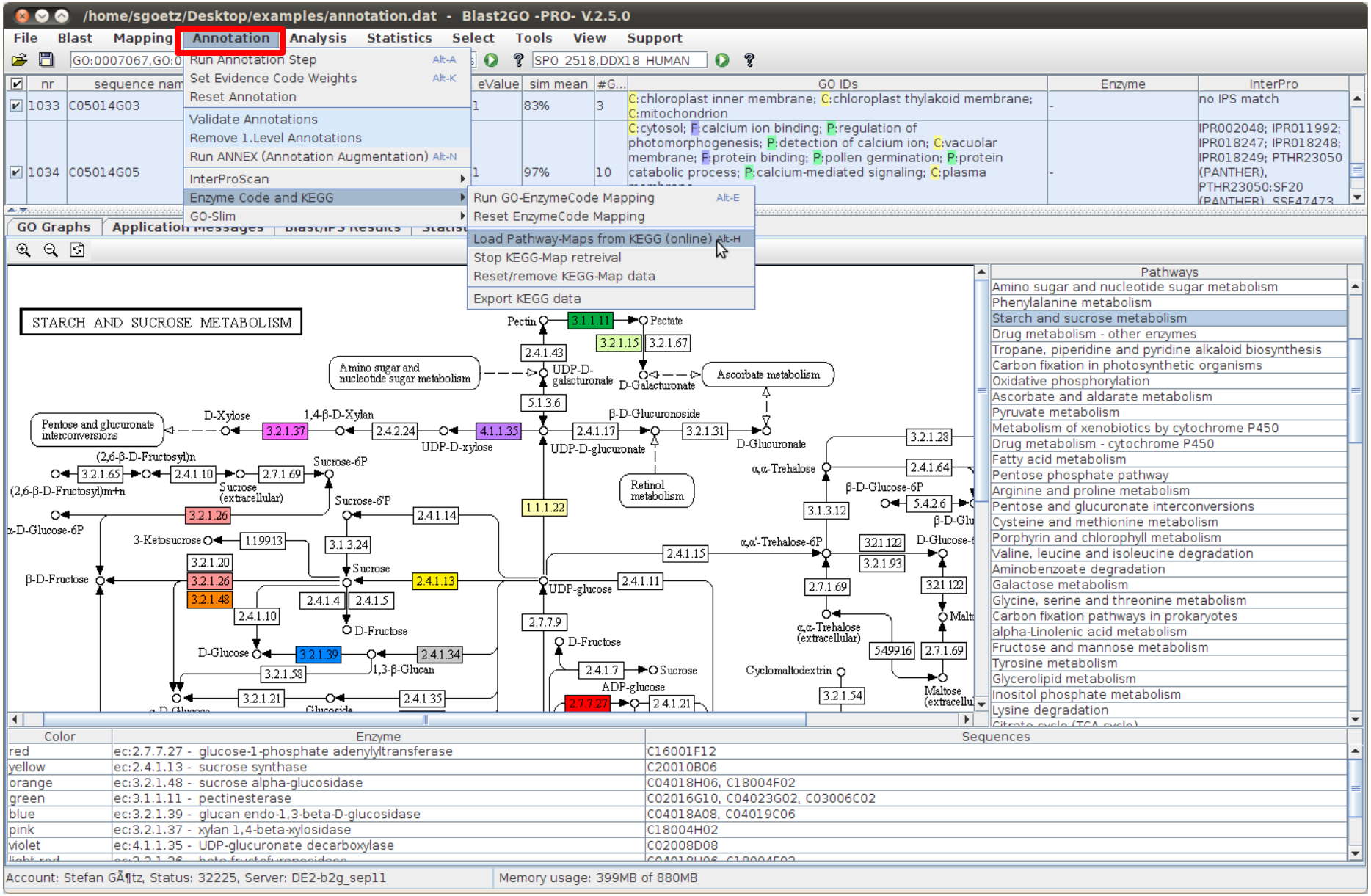

/home/sgoetz/Desktop/examples/mapping.dat - Blast2GO -PRO- V.2.5.0
File Blast Mapping Annotation Analysis Statistics Select Tools View Support
GO:0007067,GO:0016021
transport;binding;apoptosis
SPO 2518,DDX18 HUMAN
nr sequence name seq description length #h... min. eValue sim mean #G... GO IDs Enzyme InterPro
19 C04018C08 cinnamoylreductase 462 20 2.0E1 90% 10 F:binding; P:cellular metabolic process; F:catalytic activity; F:coenzyme binding; P:oxidation reduction; F:cinnamoyl-CoA reductase activity; F:oxidoreductase activity; F:3-beta-hydroxy-delta5-steroid dehydrogenase activity; F:oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor; P:steroid biosynthetic process
20 C04013G06 thaumatin-like protein 669 20 7.6E1 80% 5 P:response to biotic stimulus; P:defense response; P:killing of cells of another organism; P:defense
21 C04013A08 trypsin inhibitor 581 20 7.2E1 55% 5
22 C04013C08 j8 protein 600 10 3.0E1 57% 7
23 C04013E08 peroxidase 677 20 8.4E1 80% 8
nam protein
binding; P:response to abscisic acid stimulus; P:response to water deprivation; P:multicellular
Show Sequence
Show Blast Result
Show InterProScan Result
Show GO Descriptions
Load Kegg Pathway Map
Annotate Seq
Change Annotation and Description
Draw Graph of Mapping-Results with highlighted Annotations
Draw Graph of Annotations
GO Graphs Application Messages Blast/IPS Results Statistics Kegg Maps
Evidence code distribution for sequences
Save
Evidence code distribution for sequences
Sequences
0 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000
IEA IDA ISS IMP RCA ND TAS IEP IGI IPI ISO NAS IC
Evidence codes
Mapping database sources
Save
Mapping database sources
GOs (Totoal GOs:)
0 5,000 10,000 15,000 20,000 25,000 30,000 35,000
UniProtKB TAIR MGI
Evidence code distribution for BLAST Hits
Save
Evidence code distribution for BLAST Hits
Hits
0 5,000 10,000 15,000 20,000 25,000 30,000 35,000
IEA IDA ISS RCA ND IMP TAS IEP ISO IPI IGI NAS IBA
Evidence codes
Number of sequences with length(x)
Save
Number of sequences with length(x)
#Seqs
25.0 22.5 20.0 17.5 15.0 12.5 10.0 7.5 5.0 2.5 0.0
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800
Account: Stefan GÃ¶tz, Status: 32225, Server: DE2-b2g_sep11
Memory usage: 177MB of 880MB


/home/sgoetz/Desktop/examples/mapping.dat - Blast2GO -PRO- V.2.5.0
File
Blast
Mapping
Annotation
Analysis
Statistics
Select
Tools
View
Support


Score Distribution [Biological Process]
thetic
ocess
hetic
und
.194)
bolic
851)
ietic
369)
und
1)
216)
.235)
2.859)
single-orga
cellular pr
small mole
gene expr


Score Distribution [Molecular Function]
8.64)
179)
.38)
743)
erring
oups
hydr
nuc


/home/sgoetz/Desktop/examples/annotation.dat - Blast2GO -PRO- V.2.5.0
File Blast Mapping Annotation Analysis Statistics Select Tools View Support
GO:0007067,GO:0016021
transport;binding;apoptosis
SPO_2518,DDX18_HUMAN
nr sequence name seq description length #h... min. eValue sim mean #G... GO IDs Enzyme InterPro
37 C04013G02 ubiquitin-specific protease 6 45 EC:3.1.2.15
38 C04013A04 acetone-cyanohydrin lyase 41 EC:3.1.1.1; EC:4.1.2....
39 C04013C04 polyubiquitin 36
40 C04013A06 ---NA--- 67
bzip transcription
Graph Drawing Configuration
Tree Type Process Function Compon...
Seq Filter 20
Node Information NodeScore
Mode of Graph-Colouring byScore
Score alpha 0.6
Node Score Filter 0
Arrow labels on off
Graph Title Text Combined Graph
GO Graphs
Application Messages
Blast/IPS Re
Simple GOs : simpel GOs
Account: Stefan GÃ¶tz, Status: 32225, Server: DE2-b2g_sep11
Memory usage: 331MB of 880MB


/home/sgoetz/Desktop/examples/annotation.dat - Blast2GO -PRO- V.2.5.0
File Blast Mapping Annotation Analysis Statistics Select Tools View Support
Run Annotation Step Alt-A
Set Evidence Code Weights Alt-K
Reset Annotation
Validate Annotations
Remove 1.Level Annotations
Run ANNEX (Annotation Augmentation) Alt-N
InterProScan
Enzyme Code and KEGG
GO-Slim
Run GO-EnzymeCode Mapping Alt-E
Reset EnzymeCode Mapping
Load Pathway-Maps from KEGG (online) Alt-H
Stop KEGG-Map retreival
Reset/remove KEGG-Map data
Export KEGG data
GO Graphs
STARCH AND SUCROSE METABOLISM
Pathways
Amino sugar and nucleotide sugar metabolism
Phenylalanine metabolism
Starch and sucrose metabolism
Drug metabolism - other enzymes
Tropane, piperidine and pyridine alkaloid biosynthesis
Carbon fixation in photosynthetic organisms
Oxidative phosphorylation
Ascorbate and aldarate metabolism
Pyruvate metabolism
Metabolism of xenobiotics by cytochrome P450
Drug metabolism - cytochrome P450
Fatty acid metabolism
Pentose phosphate pathway
Arginine and proline metabolism
Pentose and glucuronate interconversions
Cysteine and methionine metabolism
Porphyrin and chlorophyll metabolism
Valine, leucine and isoleucine degradation
Aminobenzoate degradation
Galactose metabolism
Glycine, serine and threonine metabolism
Carbon fixation pathways in prokaryotes
alpha-Linolenic acid metabolism
Fructose and mannose metabolism
Tyrosine metabolism
Glycerolipid metabolism
Inositol phosphate metabolism
Lysine degradation
Color
Enzyme
Sequences
red ec:2.7.7.27 - glucose-1-phosphate adenylyltransferase C16001F12
yellow ec:2.4.1.13 - sucrose synthase C20010B06
orange ec:3.2.1.48 - sucrose alpha-glucosidase C04018H06, C18004F02
green ec:3.1.1.11 - pectinesterase C02016G10, C04023G02, C03006C02
blue ec:3.2.1.39 - glucan endo-1,3-beta-D-glucosidase C04018A08, C04019C06
pink ec:3.2.1.37 - xylan 1,4-beta-xylosidase C18004H02
violet ec:4.1.1.35 - UDP-glucuronate decarboxylase C02008D08
Account: Stefan GÃ¶tz, Status: 32225, Server: DE2-b2g_sep11
Memory usage: 399MB of 880MB

Résumé du contexte scientifique

Les traitements bioinformatiques et biostatistiques des données de séquençages à haut débit en transcriptomique à visée DEG (Differential Expression of Genes) fournissent en sortie de longues listes de gènes *différentiellement (significativement)* exprimés entre plusieurs conditions d'études ou phénotypes. Pour permettre aux biologistes une interprétation biologique des processus cellulaires en cause dans le mécanisme étudié, ces listes de gènes doivent encore être annotées fonctionnellement. Cette phase consiste non seulement à assigner des annotations aux gènes (Cf projet scripting d'agrégation d'annotations en M1S2) mais surtout dans une seconde étape à établir si des **fonctions biologiques** sont **significativement sur- ou sous- représentées** parmi ces gènes. Cette étape est nommée « **Enrichment Analysis » (EA**), elle fait appel à des **méthodes statistiques pour effectuer des tests d'enrichissement [1-8]**. Les ressources en annotations fonctionnelles au cours de l'EA peuvent être au moins de trois types **: termes de la Gene Ontology, voies métaboliques et domaines protéiques.**

❖ Références

[1] Subramanian A, Tamayo P, Mootha VK, et al. **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. Proc Natl Acad Sci USA. 2005;102(43):15545–50. [PubMed]

[2] Rivals I, Personnaz L, Taing L, Potier MC. **Enrichment or depletion of a GO category within a class of genes: which test?** Bioinformatics. 2007 Feb 15;23(4):401-7. Epub 2006 Dec 20. PMID: 17182697.

[3] Huang da W, Sherman BT, Lempicki RA. **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. Nucleic Acids Res. 2009 Jan;37(1):1-13. PubMed PMID:19033363

[4] Tipney H, Hunter L. **An introduction to effective use of enrichment analysis software**. Hum Genomics. 2010 Feb;4(3):202-6. PubMed PMID: 20368141

[5] Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C. **Gene set enrichment analysis:performance evaluation and usage guidelines**. Brief Bioinform. 2012 May;13(3):281-91. Review. PubMed PMID: 21900207

[6] Khatri P, Sirota M, Butte AJ. **Ten years of pathway analysis: current approaches and outstanding challenges**. PLoS Comput Biol. 2012;8(2):e1002375. Epub 2012 Feb 23. Review. PubMed PMID: 22383865

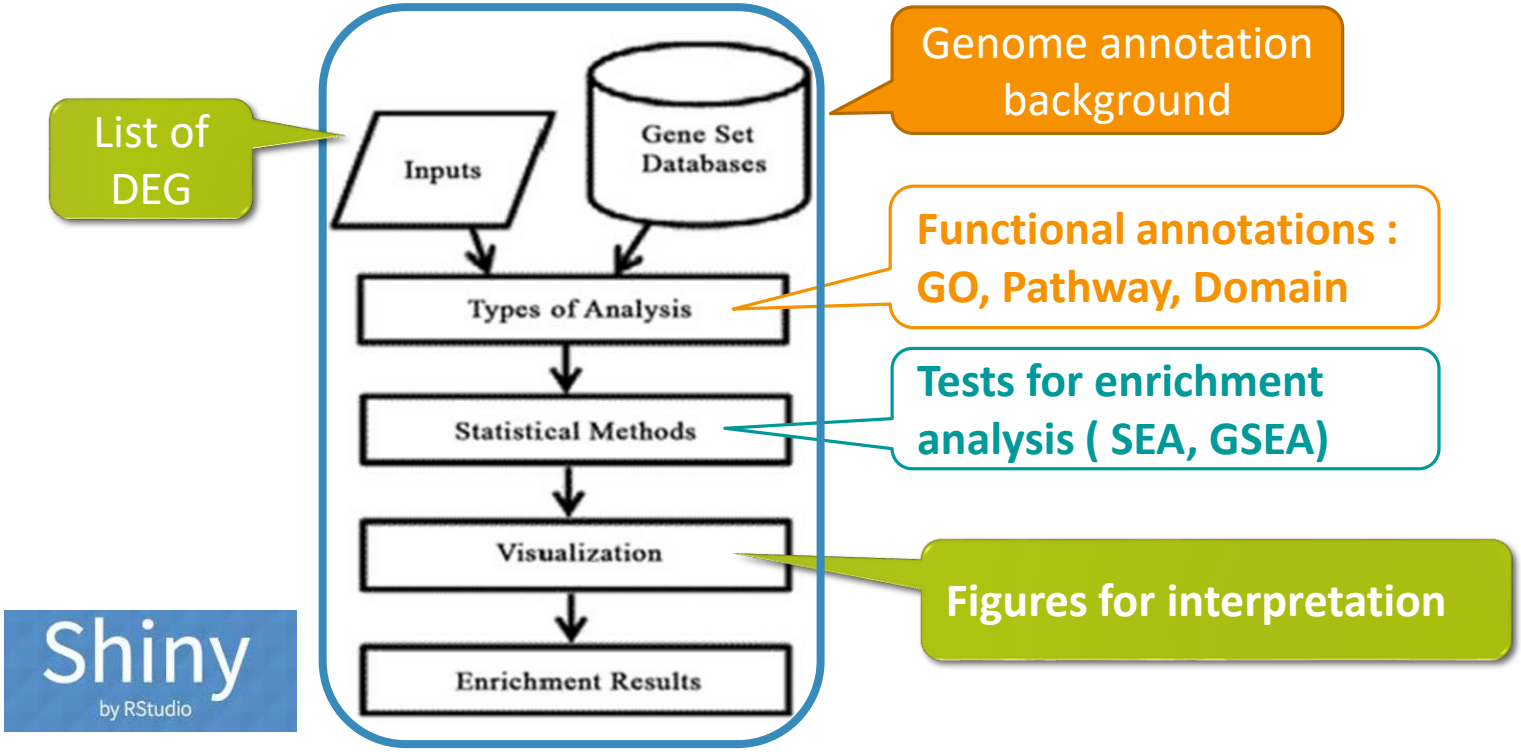
[7] Luo, Weijun, Brouwer and Cory (2013). "**Pathview: an R/Bioconductor package for pathway-based data integration and visualization**." *Bioinformatics*, **29**(14), pp. 1830-1831. doi: 10.1093/bioinformatics/btt285.

[8] Yu G, Wang L, Han Y and He Q (2012). "**clusterProfiler: an R package for comparing biological themes among gene clusters**." *OMICS: A Journal of Integrative Biology*, **16**(5), pp. 284-287. doi: 10.1089/omi.2011.0118.

A Shiny application for enrichment analysis

« Développement d'une chaîne de traitement et d'une application web en R pour l'analyse d'enrichissement fonctionnel en RNA-seq »

Le pipeline AEA que vous développerez devra à partir d'une liste de gènes issus d'une expérience RNA-seq (DEG) en entrée fournir au biologiste en sortie les résultats d'enrichissement et représentations graphiques associées. Vous créerez une interface web sous la forme d'une application Shiny afin de permettre l'utilisation aisée par un non programmeur.



Cahier des charges

- Input : un jeu de données dans un format imposé (CSV ou TSV) et entête imposées (GeneName, GeneID, baseMean, Log2FC, pval, padj)
- Whole data inspection : Vous mettrez en œuvre des fonctions R existantes pour générer une figure d'estimation de la dispersion (volcano plot, MA plot optionnel); avec représentation seuillée possible).
- Procédure EA :
  - Vous mettrez en œuvre les packages R suivants : biomaRt , ClusterProfiler et Pathview pour réaliser les deux types d'annotations (GO et voies métaboliques) pour les méthodes SEA et GSEA.
  - Vous créerez vous-mêmes un script permettant l'analyse d'enrichissement en domaines fonctionnels protéiques mettant en œuvre l'approche SEA uniquement.
- Output :
  - Tableaux de données
  - Représentations graphiques dédiées (pie, bar plot, graphes, voies...)

- Interfaçage de votre pipeline grâce à Shiny, un package R développé par RStudio.

Onglets du menu

- Whole data inspection
- GO term enrichment
- Pathway enrichemnt
- Protein enrichemnt

Choix des paramètres input

- origine gene IDs :
  - Gene NCBI
  - Ensembl
- Nom de l'organisme
  - biomaRt
  - autre

Choix des méthodes statistiques

- SEA
- GSEA

Choix de paramètres statistiques

- Seuil de p-Value
- Méthode et seuil de p-value adj

Travail préparatoire

1. Comprendre les modèles et tests statistiques utilisées pour la mise en évidence des enrichissements fonctionnels : SEA, GSEA, MEA

2. Tester les outils disponibles en ligne - identifier les méthodes utilisées

3. Trouver un jeu de données tests dans le bon format à partir d'un article qui permettra la comparaison de vos résultats

4. Vers la mise en oeuvre
   - ✓ Comprendre le fonctionnement de Shiny
   - ✓ Rechercher les sources des packages
   - ✓ Identifier les spécifications techniques et installations nécessaires

5. Modéliser votre solution (vue biologiste et vue développeur)

2. Project description

- **Modalités** ➢ Le travail se réalise en équipe : 2 quatuors et un trinome
- **Aide au cahier des charges** ➢ Voir le fichier .XLS
- **Durée- Calendrier 2018-2019**

Mois/semaines	Hrs Edt	Type	Travail à rendre – à présenter	
			Interface	Traitement
Oct S42/ S43	7	- Présentiel : 1h intro (HD) - Présentiel : 6h TD/TP Shiny (MS)		
	3, 5 2	- Autonomie URN en équipe - Présentiel collectif (bilan 1 HD)	- Diaporama outils existants - **Interface v1** : visuel graphique	- Jeu de données (par équipe)
Nov S47-S48	5,5 1,5	- Autonomie URN en équipe - Présentiel par équipe (bilan 2 HD) 25 min x 3 par équipe	- **Interface v2** : développée mais non fonctionnelle (non connectée aux traitements)	- **Traitement v1** : modèle + preuve de fonctionnement partiel - **Traitement v2** : modèle v2 si nécessaire
Jan S2/S3	~5 1,5	- Autonomie URN en équipe - Présentiel par équipe (bilan 3 HD + MS); 25 min x 3 par équipe	- **Interface v3** connectée au traitement v2	- **Traitement v2** quasi-complet - Traitement SEA domaine : proposition de codage R
Mars S10/11/12	~5	Autonomie URN en équipe	- **Interface v4** connectée au **traitement v3** complet - Documentation	
	1,5	Livraison finale par équipe (bilan 4 HD); 25 min x 3 par équipe	- **Solution complète** : Démonstration et présentation orale - **Livraison** sources + documentation + diaporama	
Total	32 Edt	- Présentiel collectif : 7h (1h HD/6h MS) - Autonomie en équipe URN : 18,5h - Bilan [1-4] : 3,5 h (etu); 6,5h (6,5h HD/2h MS)		