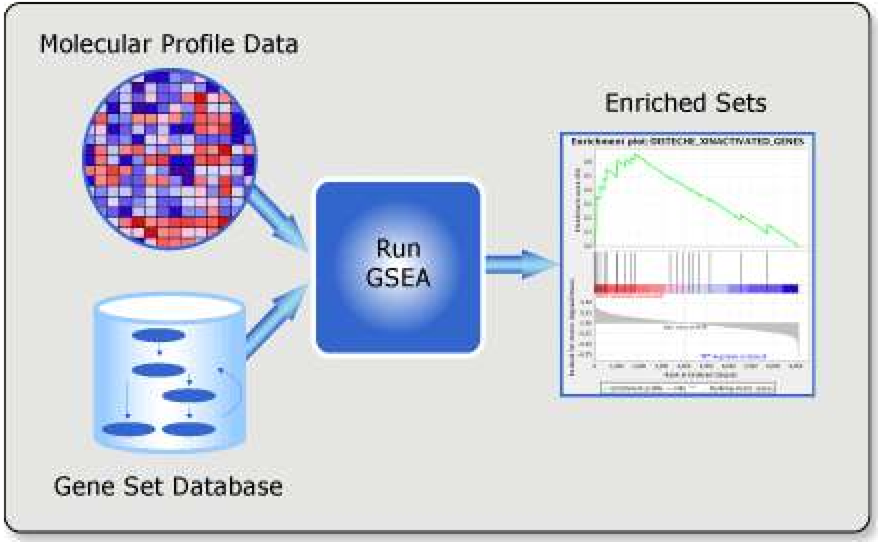


# Gene Set Analysis:

## почему интерпретировать глобальные генетические изменения труднее, чем кажется

Александр Предеус

Институт Биоинформатики

# Outline

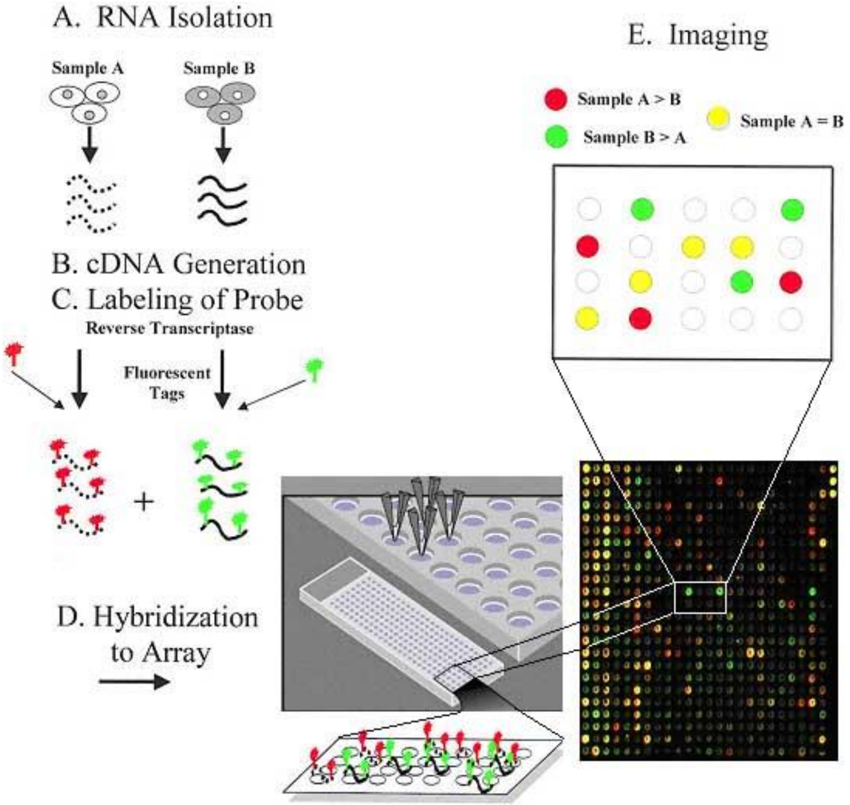
- Formulating the problem
- What are the references?
- Overrepresentation methods
- Gene set enrichment analysis
- Gene set analysis generalization

# Outline

- **Formulating the problem**
- What are the references?
- Overrepresentation methods
- Gene set enrichment analysis
- Gene set analysis generalization

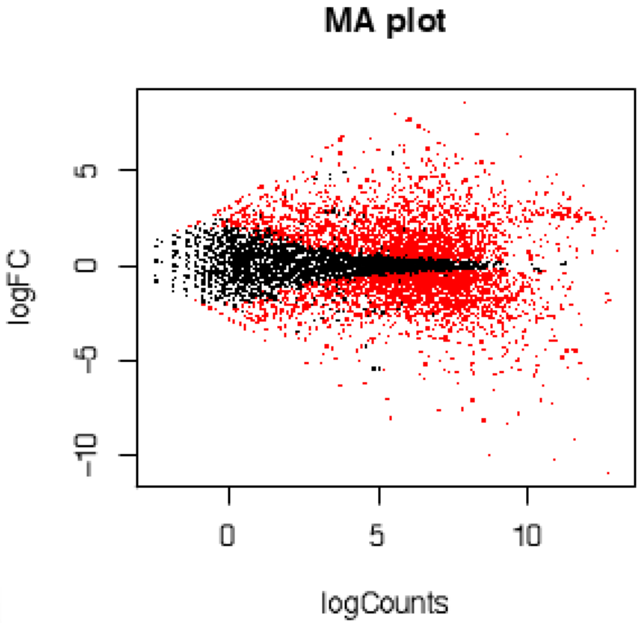
# Дифференциальная экспрессия

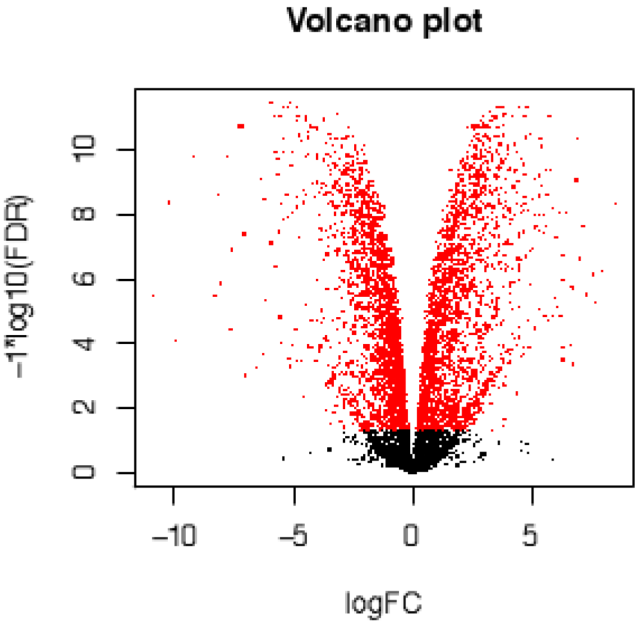
- Several experimental samples
- Several controls
- Statistical analysis gives sets of up- and down-regulated genes



# Volcano & MA plots

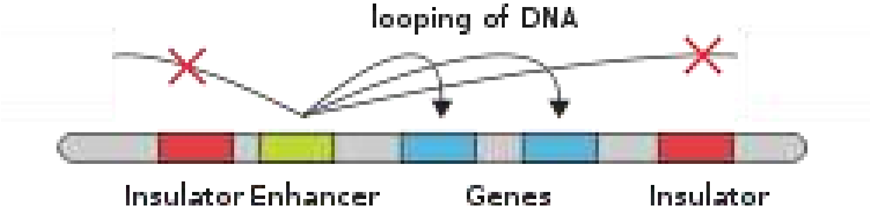
- logFC is actually log2





# ChIP-seq too

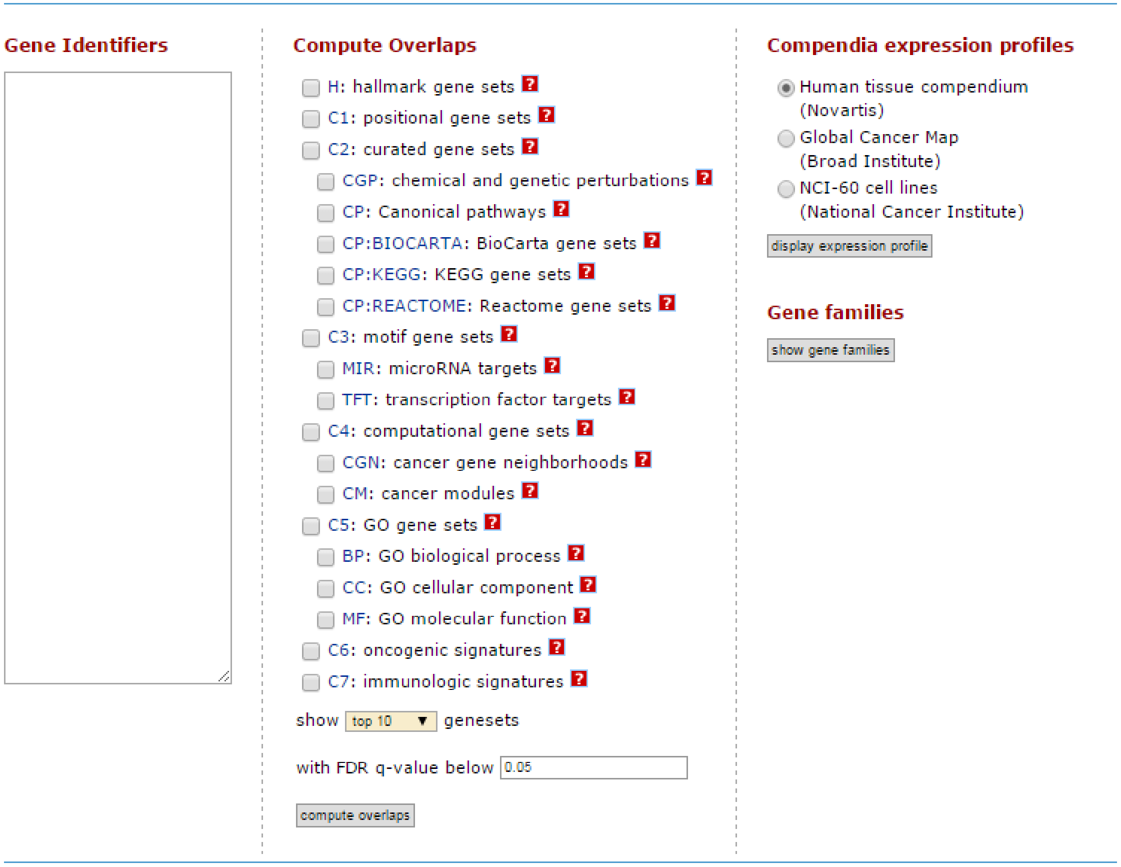
- Analysis of ChIP-seq gives a set of (regulated) genes as well!
- Hypergeometric methods
- GREAT



# Outline

- Formulating the problem
- **What are the references?**
- Overrepresentation methods
- Gene set enrichment analysis
- Gene set analysis generalization

# A wealth of choices



**Gene Identifiers**

**Compute Overlaps**

- □ H: hallmark gene sets
- □ C1: positional gene sets
- □ C2: curated gene sets
  - □ CGP: chemical and genetic perturbations
  - □ CP: Canonical pathways
  - □ CP:BIOCARTA: BioCarta gene sets
  - □ CP:KEGG: KEGG gene sets
  - □ CP:REACTOME: Reactome gene sets
- □ C3: motif gene sets
  - □ MIR: microRNA targets
  - □ TFT: transcription factor targets
- □ C4: computational gene sets
  - □ CGN: cancer gene neighborhoods
  - □ CM: cancer modules
- □ C5: GO gene sets
  - □ BP: GO biological process
  - □ CC: GO cellular component
  - □ MF: GO molecular function
- □ C6: oncogenic signatures
- □ C7: immunologic signatures

show top 10 ▼ genesets

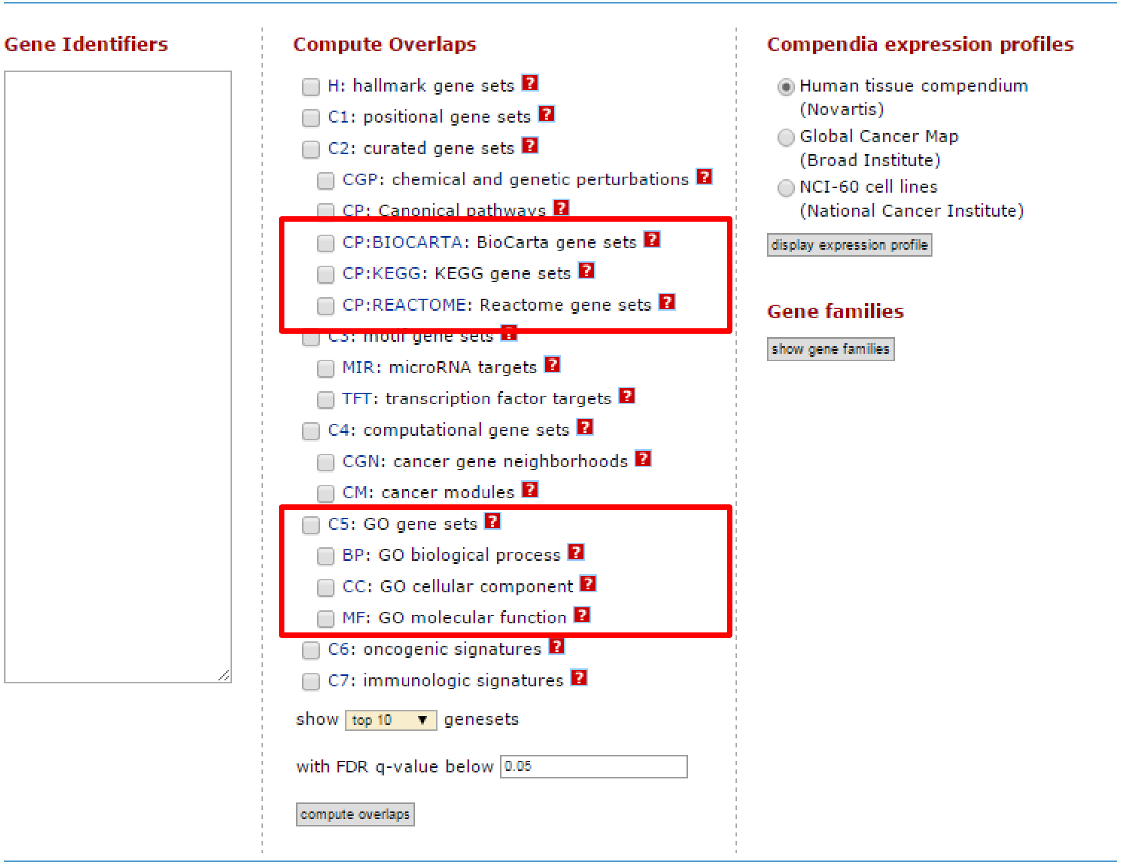
with FDR q-value below 0.05

**Compendia expression profiles**

- ● Human tissue compendium (Novartis)
- ○ Global Cancer Map (Broad Institute)
- ○ NCI-60 cell lines (National Cancer Institute)

**Gene families**

# Богатство выбора



**Gene Identifiers**

**Compute Overlaps**

- ☐ H: hallmark gene sets
- ☐ C1: positional gene sets
- ☐ C2: curated gene sets
  - ☐ CGP: chemical and genetic perturbations
  - ☐ CP: Canonical pathways
  - ☐ CP:BIOCARTA: BioCarta gene sets
  - ☐ CP:KEGG: KEGG gene sets
  - ☐ CP:REACTOME: Reactome gene sets
- ☐ C3: motif gene sets
  - ☐ MIR: microRNA targets
  - ☐ TFT: transcription factor targets
- ☐ C4: computational gene sets
  - ☐ CGN: cancer gene neighborhoods
  - ☐ CM: cancer modules
- ☐ C5: GO gene sets
  - ☐ BP: GO biological process
  - ☐ CC: GO cellular component
  - ☐ MF: GO molecular function
- ☐ C6: oncogenic signatures
- ☐ C7: immunologic signatures

show top 10 ▼ genesets

with FDR q-value below 0.05

compute overlaps

**Compendia expression profiles**

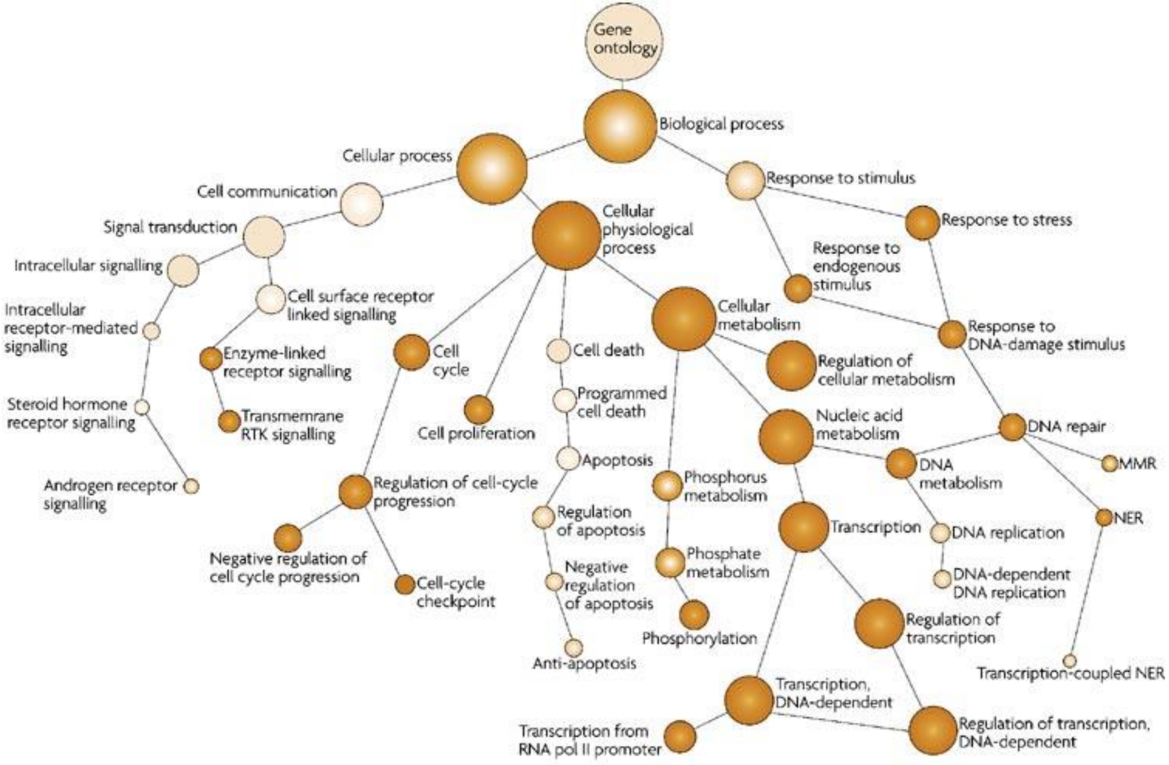
- ● Human tissue compendium (Novartis)
- ○ Global Cancer Map (Broad Institute)
- ○ NCI-60 cell lines (National Cancer Institute)

display expression profile

**Gene families**

show gene families

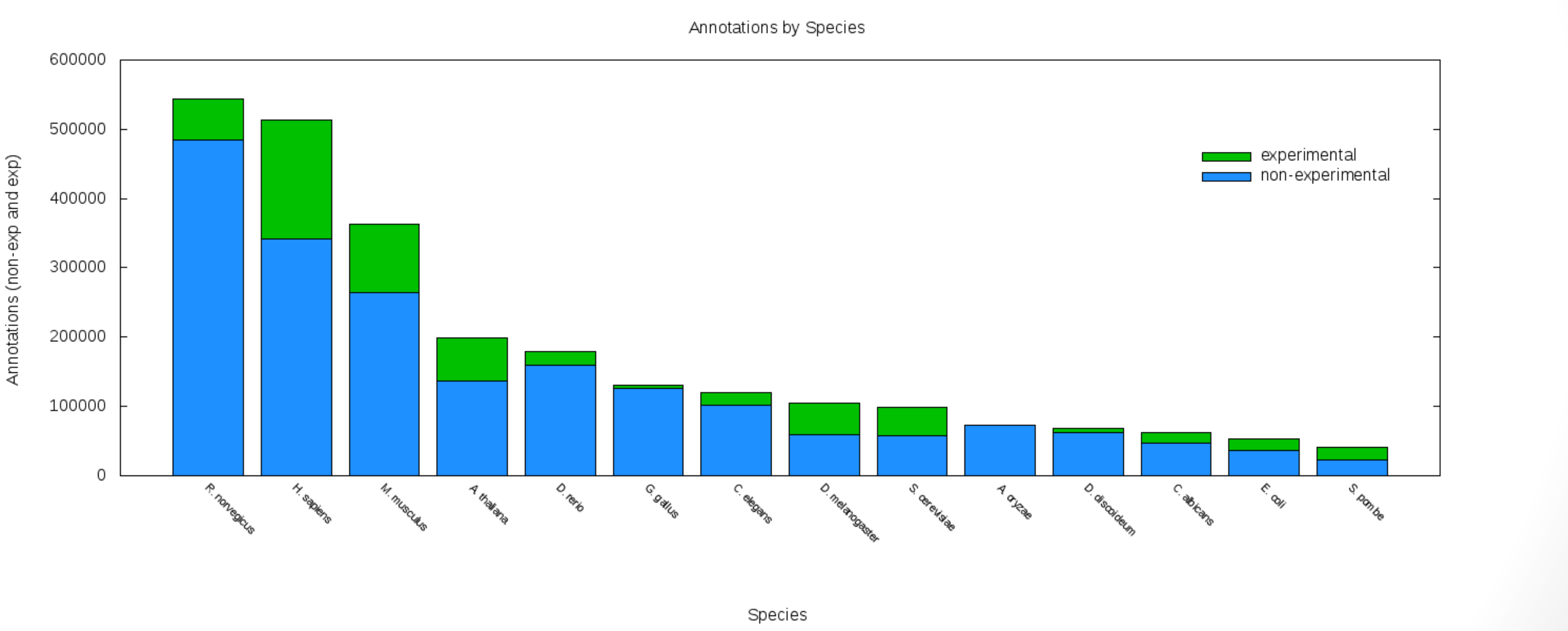
# GO = Gene ontology


Gene ontology
Biological process
Cellular process
Response to stimulus
Cell communication
Signal transduction
Cellular physiological process
Response to stress
Response to endogenous stimulus
Intracellular signalling
Cell surface receptor linked signalling
Intracellular receptor-mediated signalling
Cellular metabolism
Response to DNA-damage stimulus
Enzyme-linked receptor signalling
Cell cycle
Cell death
Regulation of cellular metabolism
Programmed cell death
Steroid hormone receptor signalling
Transmemrane RTK signalling
Cell proliferation
Nucleic acid metabolism
DNA repair
Apoptosis
DNA metabolism
MMR
Androgen receptor signalling
Regulation of cell-cycle progression
Phosphorus metabolism
Regulation of apoptosis
Transcription
NER
DNA replication
Negative regulation of cell cycle progression
Phosphate metabolism
Negative regulation of apoptosis
DNA-dependent DNA replication
Cell-cycle checkpoint
Regulation of transcription
Phosphorylation
Anti-apoptosis
Transcription-coupled NER
Transcription, DNA-dependent
Transcription from RNA pol II promoter
Regulation of transcription, DNA-dependent

Nature Reviews | Cancer

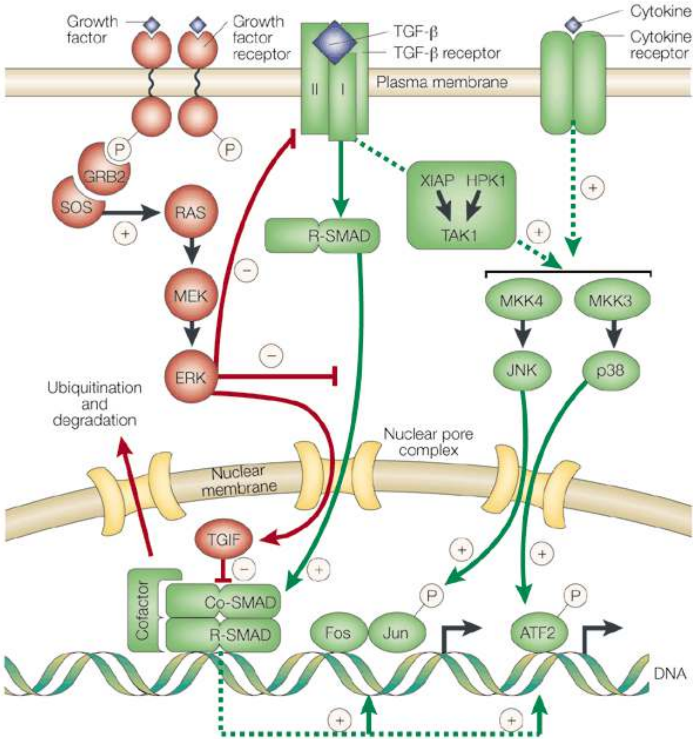
# GO = Gene ontology

- Mostly from UniProt



# Pathway annotation

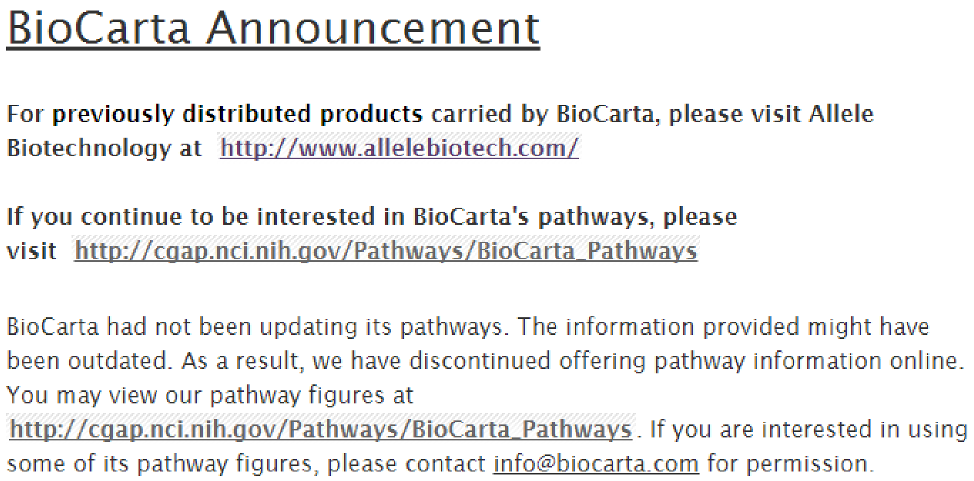
- Organism-specific
- Thoroughly curated (well...)
- Much more informative
- Much less overlapping



Nature Reviews | Molecular Cell Biology

# Biocarta

- http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
- Outdated/retired



BioCarta Announcement

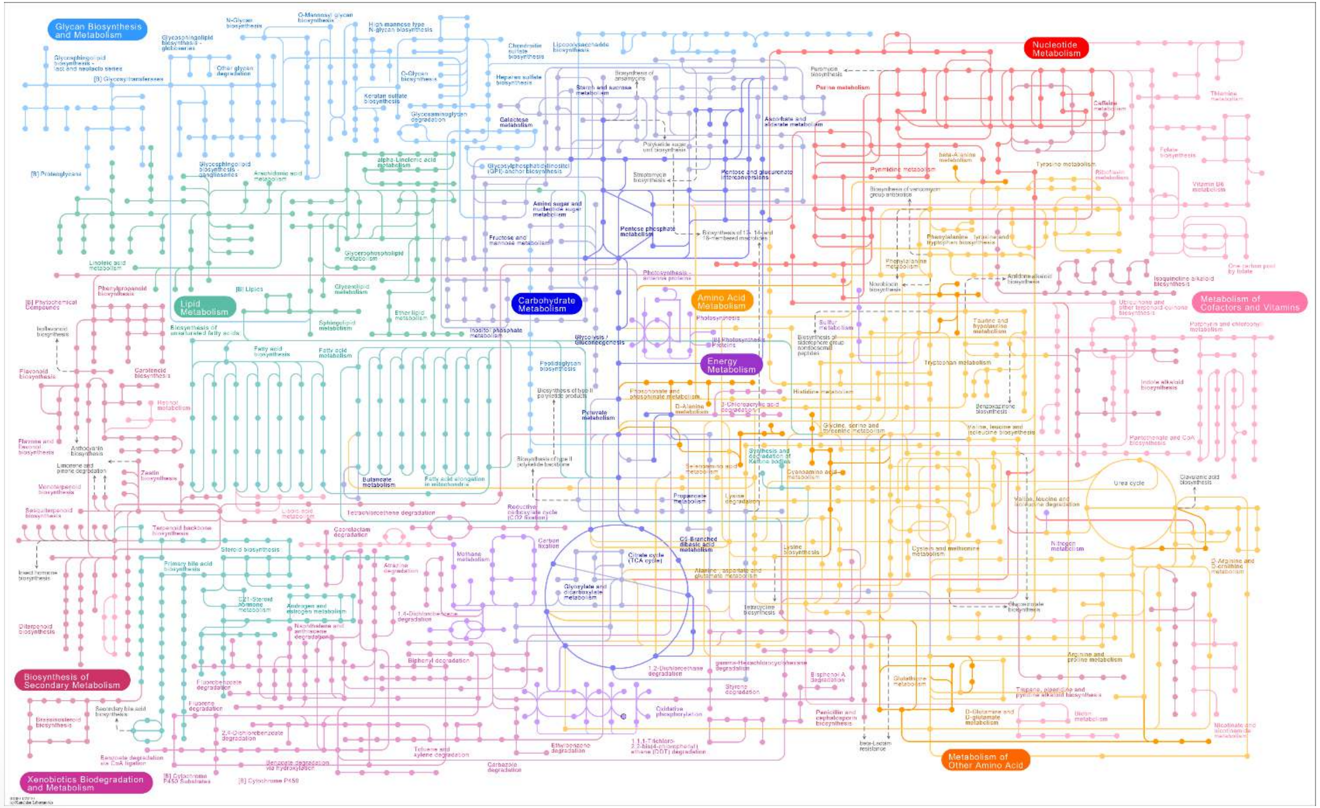
For **previously distributed products** carried by BioCarta, please visit Allele Biotechnology at http://www.allelebiotech.com/

**If you continue to be interested in BioCarta's pathways, please visit** http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways

BioCarta had not been updating its pathways. The information provided might have been outdated. As a result, we have discontinued offering pathway information online. You may view our pathway figures at http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways. If you are interested in using some of its pathway figures, please contact info@biocarta.com for permission.

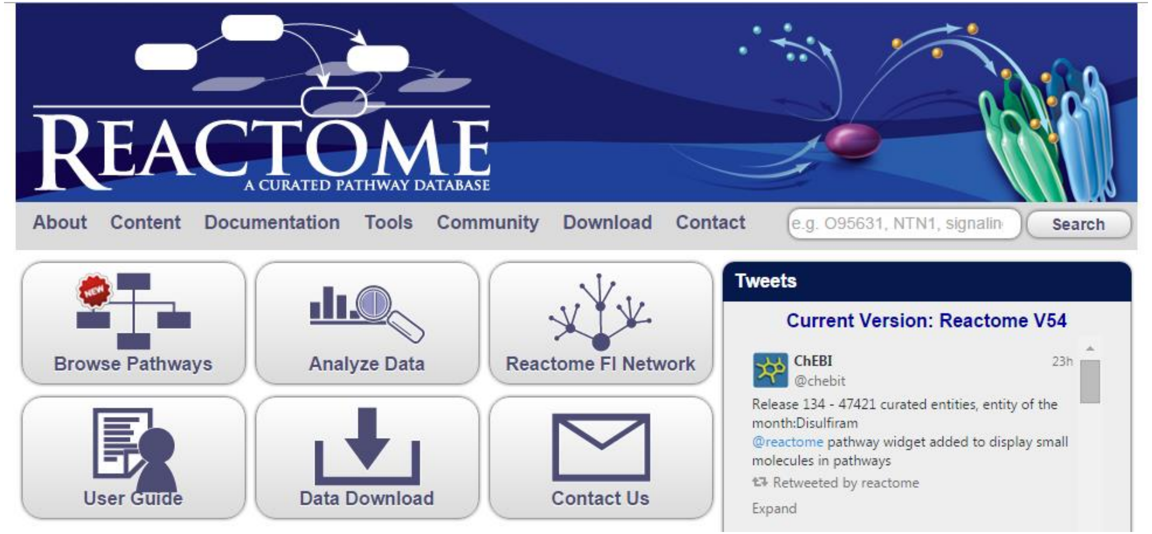
# KEGG

- Heavy on metabolism; commercial since 2008


Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Secondary Metabolism
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid

# Reactome

- Curated by EMBL
- System of pathway peer review
- Many apps



Current Version: Reactome V54

ChEBI @chebit 23h

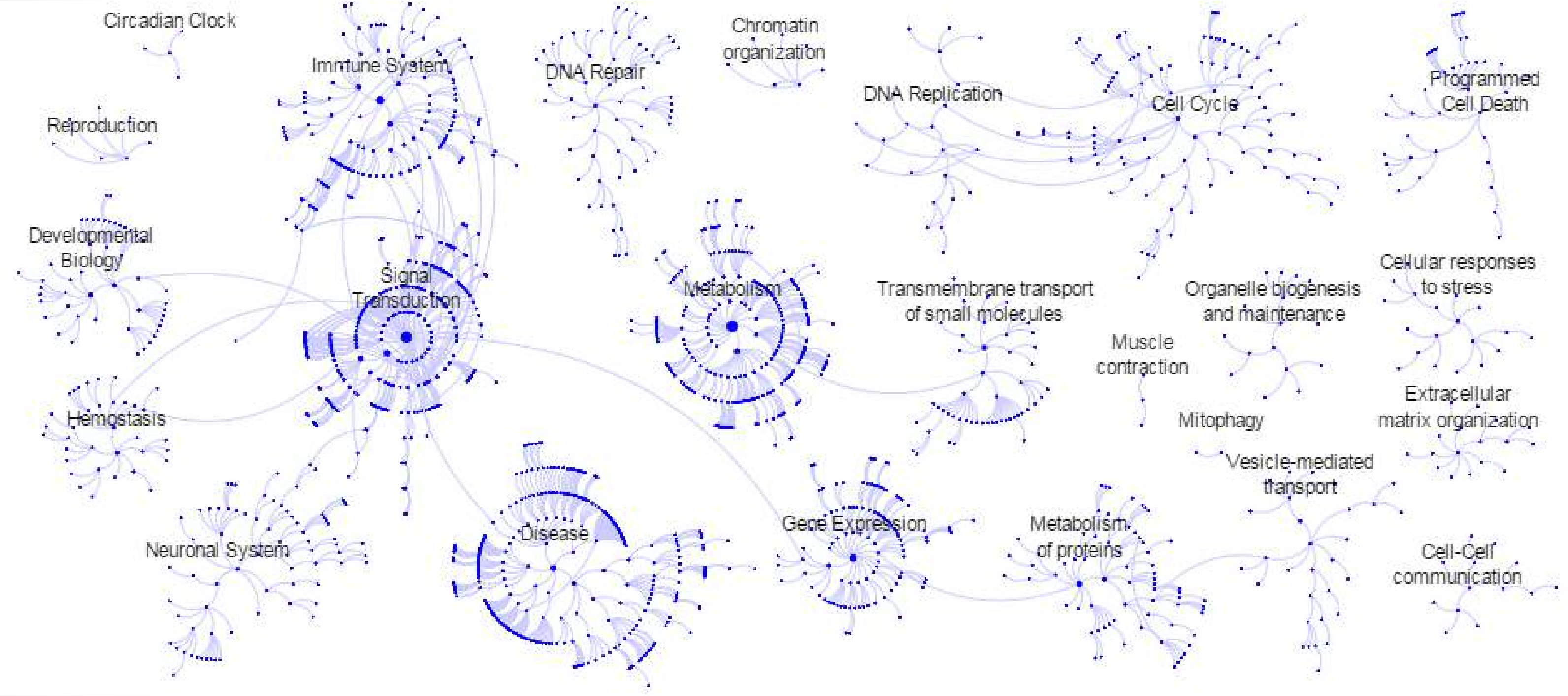
Release 134 - 47421 curated entities, entity of the month:Disulfiram
@reactome pathway widget added to display small molecules in pathways

Retweeted by reactome

Expand

http://www.ncbi.nlm.nih.gov/pubmed/24243840

# Pathway Browser



# Outline

- Formulating the problem
- What are the references?
- **Overrepresentation methods**
- Gene set enrichment analysis
- Gene set analysis generalization

http://www.ncbi.nlm.nih.gov/pubmed/24243840

# Method classification

- Review from 2009 counted 68 “enrichment” tools
- Algorithms split in three groups:
  - Singular enrichment analysis (SEA)
  - Gene set enrichment analysis (GSEA)
  - Modular enrichment analysis (MEA)

http://www.ncbi.nlm.nih.gov/pubmed/19033363

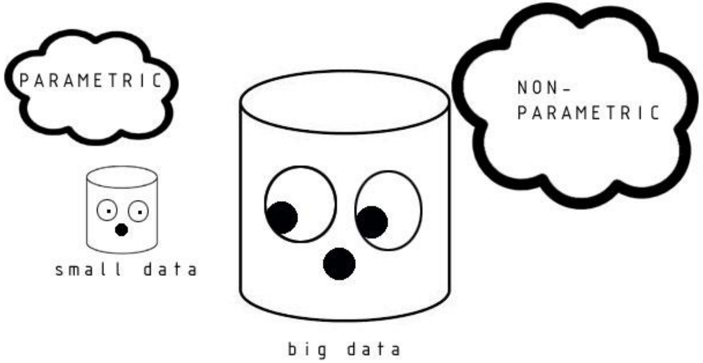
# Method classification

- Review from 2009 counted 68 “enrichment” tools
- Algorithms split in three groups:
  - Singular enrichment analysis (SEA)
  - Gene set enrichment analysis (GSEA)
  - Modular enrichment analysis (MEA)
- Major features:
  - Statistical algorithm
  - Uses all genes or only selected portion?
  - Uses weights or only presence/absence based?

http://www.ncbi.nlm.nih.gov/pubmed/19033363

# Underlying statistics

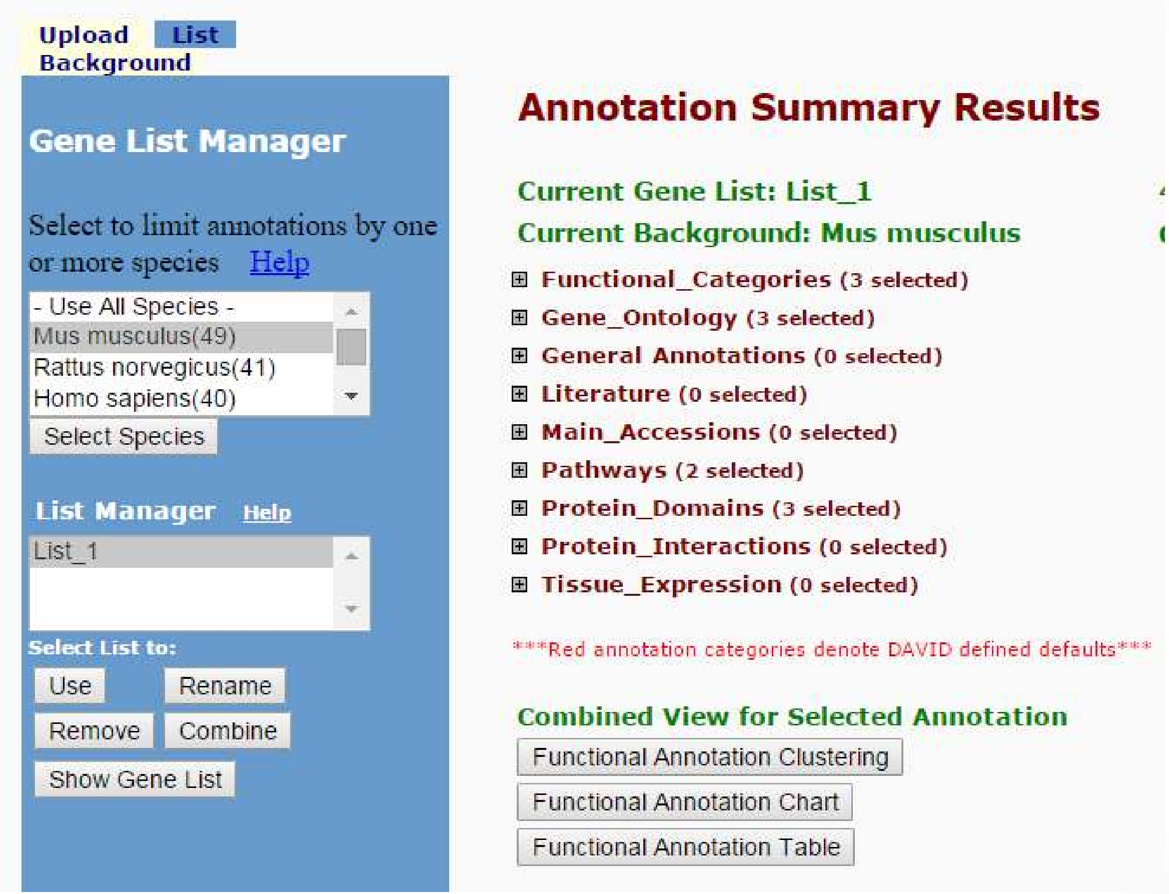
- Used distributions:
  - Hypergeometric distribution (Fisher's exact test)
  - Binomial distribution
  - Non-parametric (i.e. no distribution)



http://www.ncbi.nlm.nih.gov/pubmed/19033363

# DAVID

- Dramatically overloaded with, eh, things. Many things.



Gene List Manager

Select to limit annotations by one or more species Help

List Manager Help

Annotation Summary Results

Current Gene List: List_1

Current Background: Mus musculus

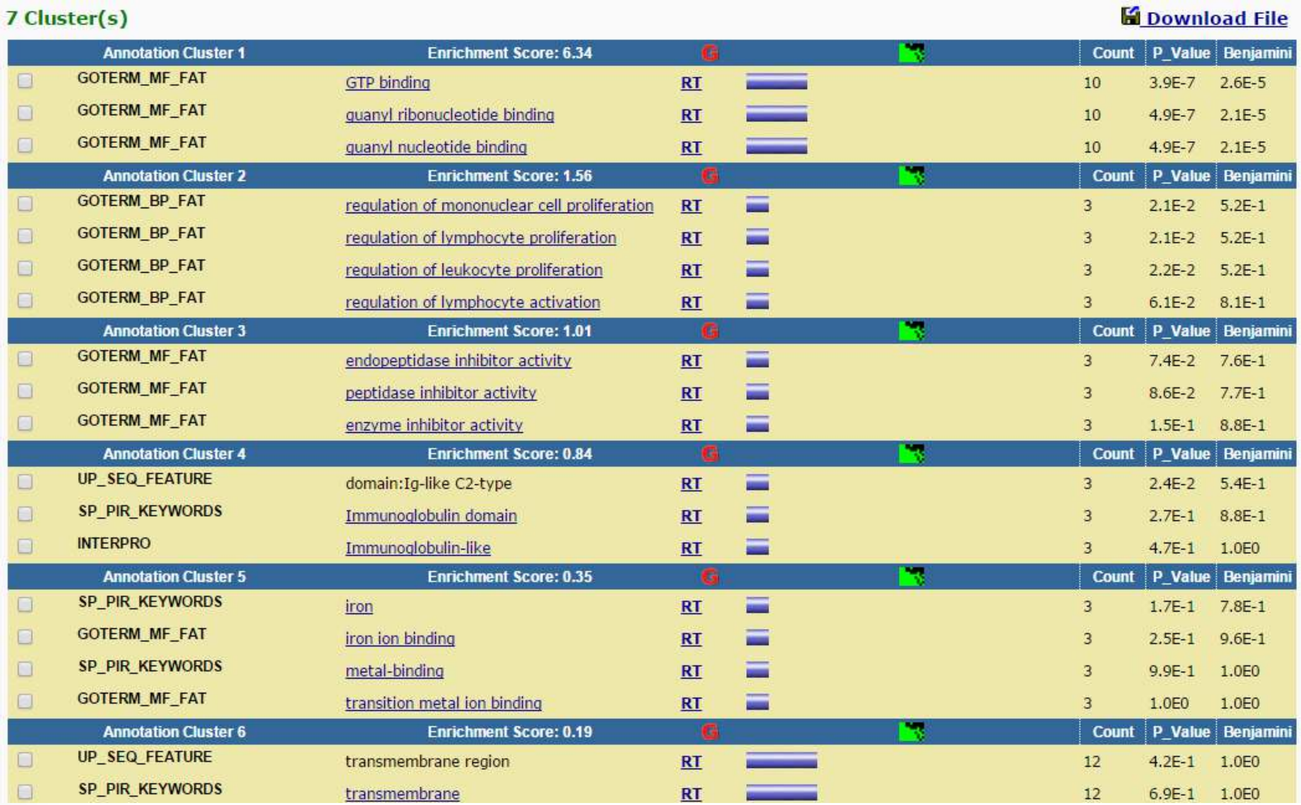
- ⊞ Functional_Categories (3 selected)
- ⊞ Gene_Ontology (3 selected)
- ⊞ General Annotations (0 selected)
- ⊞ Literature (0 selected)
- ⊞ Main_Accessions (0 selected)
- ⊞ Pathways (2 selected)
- ⊞ Protein_Domains (3 selected)
- ⊞ Protein_Interactions (0 selected)
- ⊞ Tissue_Expression (0 selected)

***Red annotation categories denote DAVID defined defaults***

Combined View for Selected Annotation

Functional Annotation Clustering
Functional Annotation Chart
Functional Annotation Table

# M1 macrophages vs DAVID

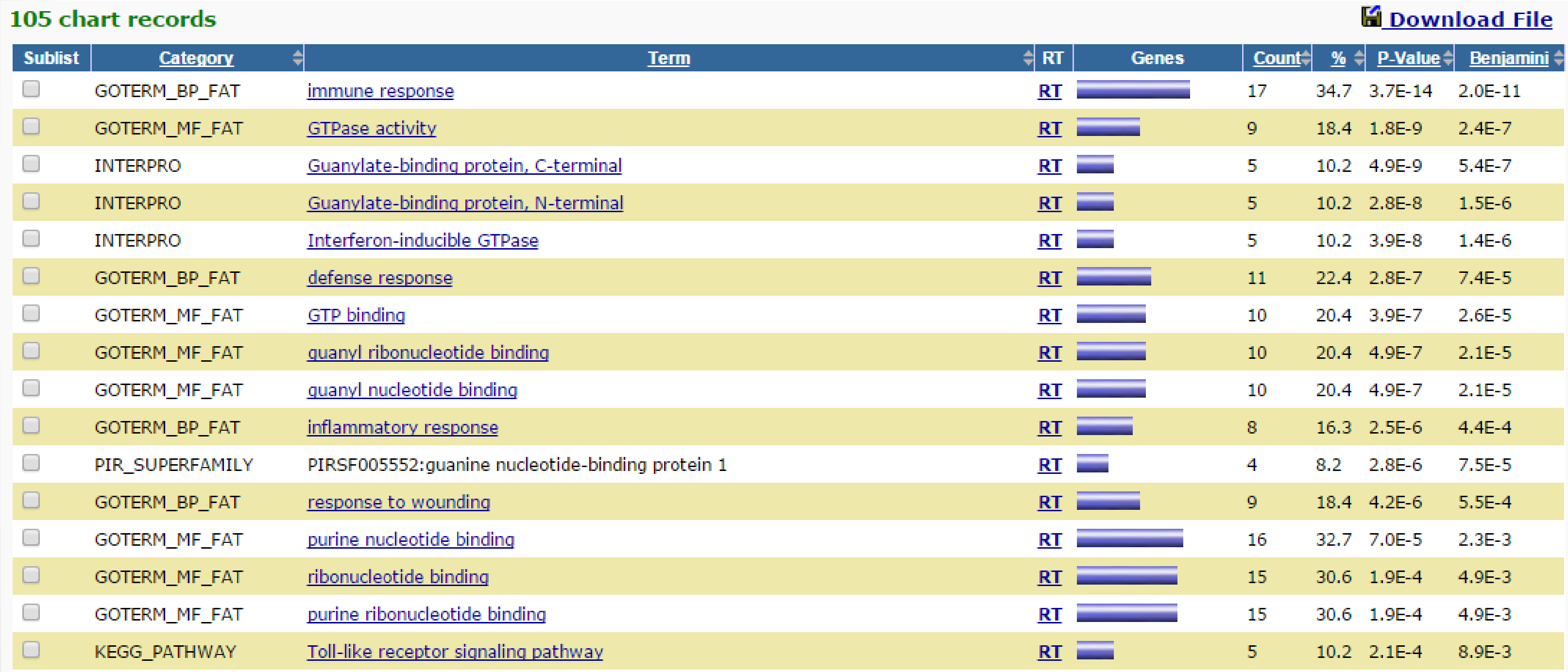


7 Cluster(s)

Download File

| | Annotation Cluster 1 | Enrichment Score: 6.34 | | Count | P_Value | Benjamini |
|---|---|---|---|---|---|---|
| □ | GOTERM_MF_FAT | GTP binding | RT | 10 | 3.9E-7 | 2.6E-5 |
| □ | GOTERM_MF_FAT | guanyl ribonucleotide binding | RT | 10 | 4.9E-7 | 2.1E-5 |
| □ | GOTERM_MF_FAT | guanyl nucleotide binding | RT | 10 | 4.9E-7 | 2.1E-5 |
| | Annotation Cluster 2 | Enrichment Score: 1.56 | | Count | P_Value | Benjamini |
| □ | GOTERM_BP_FAT | regulation of mononuclear cell proliferation | RT | 3 | 2.1E-2 | 5.2E-1 |
| □ | GOTERM_BP_FAT | regulation of lymphocyte proliferation | RT | 3 | 2.1E-2 | 5.2E-1 |
| □ | GOTERM_BP_FAT | regulation of leukocyte proliferation | RT | 3 | 2.2E-2 | 5.2E-1 |
| □ | GOTERM_BP_FAT | regulation of lymphocyte activation | RT | 3 | 6.1E-2 | 8.1E-1 |
| | Annotation Cluster 3 | Enrichment Score: 1.01 | | Count | P_Value | Benjamini |
| □ | GOTERM_MF_FAT | endopeptidase inhibitor activity | RT | 3 | 7.4E-2 | 7.6E-1 |
| □ | GOTERM_MF_FAT | peptidase inhibitor activity | RT | 3 | 8.6E-2 | 7.7E-1 |
| □ | GOTERM_MF_FAT | enzyme inhibitor activity | RT | 3 | 1.5E-1 | 8.8E-1 |
| | Annotation Cluster 4 | Enrichment Score: 0.84 | | Count | P_Value | Benjamini |
| □ | UP_SEQ_FEATURE | domain:Ig-like C2-type | RT | 3 | 2.4E-2 | 5.4E-1 |
| □ | SP_PIR_KEYWORDS | Immunoglobulin domain | RT | 3 | 2.7E-1 | 8.8E-1 |
| □ | INTERPRO | Immunoglobulin-like | RT | 3 | 4.7E-1 | 1.0E0 |
| | Annotation Cluster 5 | Enrichment Score: 0.35 | | Count | P_Value | Benjamini |
| □ | SP_PIR_KEYWORDS | iron | RT | 3 | 1.7E-1 | 7.8E-1 |
| □ | GOTERM_MF_FAT | iron ion binding | RT | 3 | 2.5E-1 | 9.6E-1 |
| □ | SP_PIR_KEYWORDS | metal-binding | RT | 3 | 9.9E-1 | 1.0E0 |
| □ | GOTERM_MF_FAT | transition metal ion binding | RT | 3 | 1.0E0 | 1.0E0 |
| | Annotation Cluster 6 | Enrichment Score: 0.19 | | Count | P_Value | Benjamini |
| □ | UP_SEQ_FEATURE | transmembrane region | RT | 12 | 4.2E-1 | 1.0E0 |
| □ | SP_PIR_KEYWORDS | transmembrane | RT | 12 | 6.9E-1 | 1.0E0 |

# M1 macrophages vs DAVID



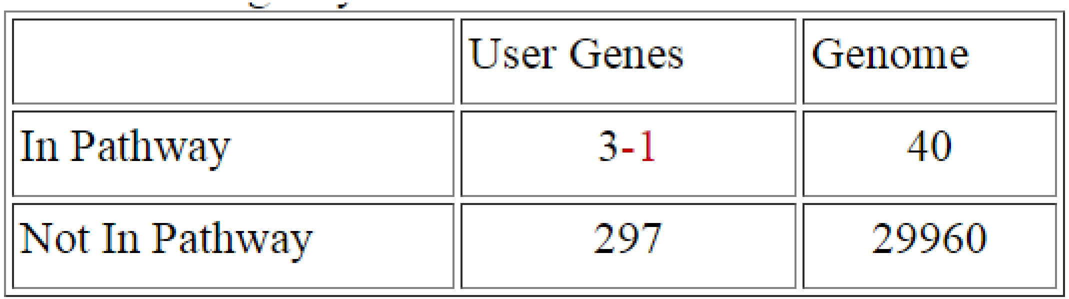
**105 chart records**

**Download File**

| Sublist | Category | Term | RT | Genes | Count | % | P-Value | Benjamini |
|---|---|---|---|---|---|---|---|---|
| □ | GOTERM_BP_FAT | immune response | RT | ▬ | 17 | 34.7 | 3.7E-14 | 2.0E-11 |
| □ | GOTERM_MF_FAT | GTPase activity | RT | ▬ | 9 | 18.4 | 1.8E-9 | 2.4E-7 |
| □ | INTERPRO | Guanylate-binding protein, C-terminal | RT | ▬ | 5 | 10.2 | 4.9E-9 | 5.4E-7 |
| □ | INTERPRO | Guanylate-binding protein, N-terminal | RT | ▬ | 5 | 10.2 | 2.8E-8 | 1.5E-6 |
| □ | INTERPRO | Interferon-inducible GTPase | RT | ▬ | 5 | 10.2 | 3.9E-8 | 1.4E-6 |
| □ | GOTERM_BP_FAT | defense response | RT | ▬ | 11 | 22.4 | 2.8E-7 | 7.4E-5 |
| □ | GOTERM_MF_FAT | GTP binding | RT | ▬ | 10 | 20.4 | 3.9E-7 | 2.6E-5 |
| □ | GOTERM_MF_FAT | guanyl ribonucleotide binding | RT | ▬ | 10 | 20.4 | 4.9E-7 | 2.1E-5 |
| □ | GOTERM_MF_FAT | guanyl nucleotide binding | RT | ▬ | 10 | 20.4 | 4.9E-7 | 2.1E-5 |
| □ | GOTERM_BP_FAT | inflammatory response | RT | ▬ | 8 | 16.3 | 2.5E-6 | 4.4E-4 |
| □ | PIR_SUPERFAMILY | PIRSF005552:guanine nucleotide-binding protein 1 | RT | ▬ | 4 | 8.2 | 2.8E-6 | 7.5E-5 |
| □ | GOTERM_BP_FAT | response to wounding | RT | ▬ | 9 | 18.4 | 4.2E-6 | 5.5E-4 |
| □ | GOTERM_MF_FAT | purine nucleotide binding | RT | ▬ | 16 | 32.7 | 7.0E-5 | 2.3E-3 |
| □ | GOTERM_MF_FAT | ribonucleotide binding | RT | ▬ | 15 | 30.6 | 1.9E-4 | 4.9E-3 |
| □ | GOTERM_MF_FAT | purine ribonucleotide binding | RT | ▬ | 15 | 30.6 | 1.9E-4 | 4.9E-3 |
| □ | KEGG_PATHWAY | Toll-like receptor signaling pathway | RT | ▬ | 5 | 10.2 | 2.1E-4 | 8.9E-3 |

# EASE score

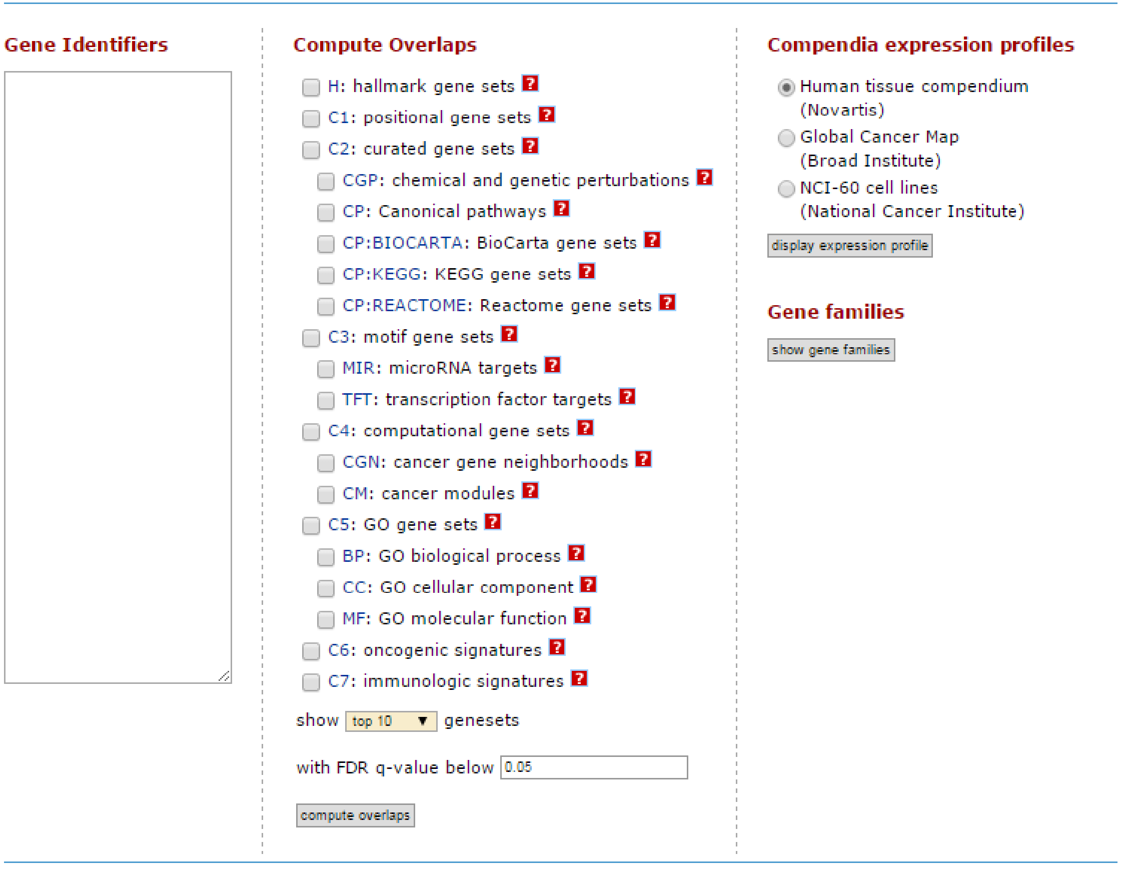
- Fisher with “jackknifing” correction



| | User Genes | Genome |
|---|---|---|
| In Pathway | 3-1 | 40 |
| Not In Pathway | 297 | 29960 |

# MsigDB

- Go-to overrepresentation tool



**Gene Identifiers**

**Compute Overlaps**

- □ H: hallmark gene sets ?
- □ C1: positional gene sets ?
- □ C2: curated gene sets ?
  - □ CGP: chemical and genetic perturbations ?
  - □ CP: Canonical pathways ?
  - □ CP:BIOCARTA: BioCarta gene sets ?
  - □ CP:KEGG: KEGG gene sets ?
  - □ CP:REACTOME: Reactome gene sets ?
- □ C3: motif gene sets ?
  - □ MIR: microRNA targets ?
  - □ TFT: transcription factor targets ?
- □ C4: computational gene sets ?
  - □ CGN: cancer gene neighborhoods ?
  - □ CM: cancer modules ?
- □ C5: GO gene sets ?
  - □ BP: GO biological process ?
  - □ CC: GO cellular component ?
  - □ MF: GO molecular function ?
- □ C6: oncogenic signatures ?
- □ C7: immunologic signatures ?

show top 10 ▼ genesets

with FDR q-value below 0.05

compute overlaps

**Compendia expression profiles**

● Human tissue compendium (Novartis)

○ Global Cancer Map (Broad Institute)

○ NCI-60 cell lines (National Cancer Institute)

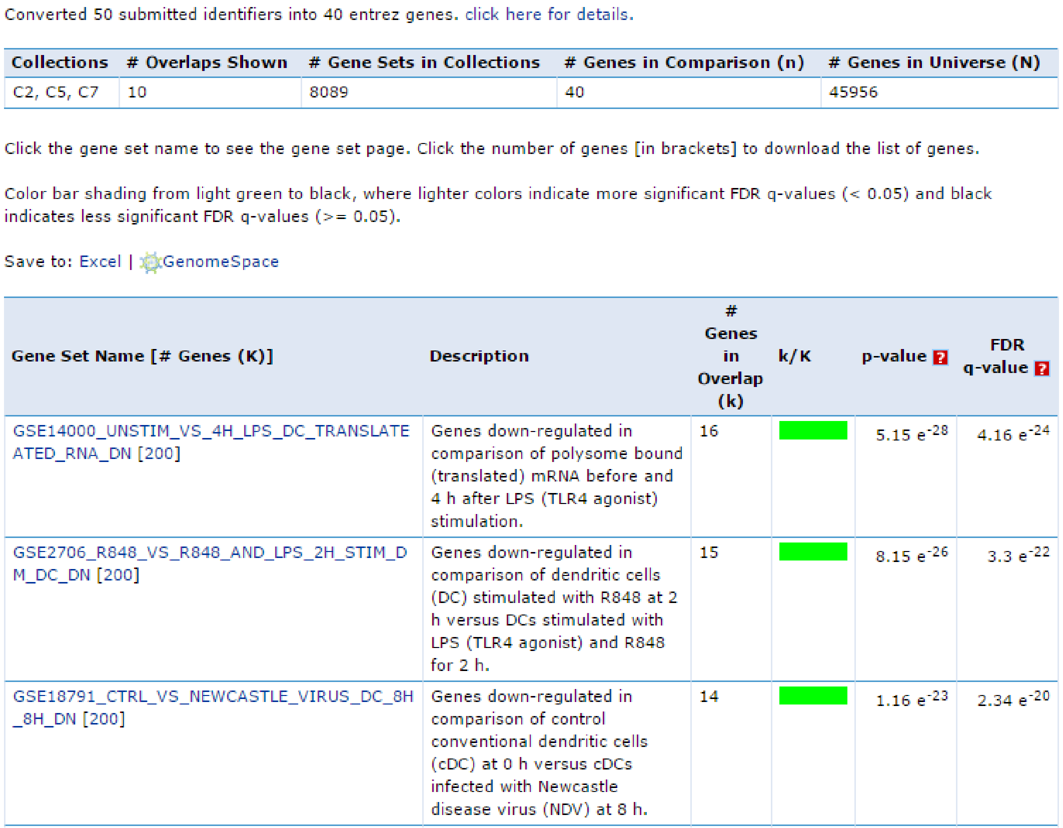
display expression profile

**Gene families**

show gene families

# M1 macrophages vs MsigDB

- Fisher’s exact test
- FDR correction



Converted 50 submitted identifiers into 40 entrez genes. click here for details.

| Collections | # Overlaps Shown | # Gene Sets in Collections | # Genes in Comparison (n) | # Genes in Universe (N) |
|---|---|---|---|---|
| C2, C5, C7 | 10 | 8089 | 40 | 45956 |

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates less significant FDR q-values (>= 0.05).

Save to: Excel | GenomeSpace

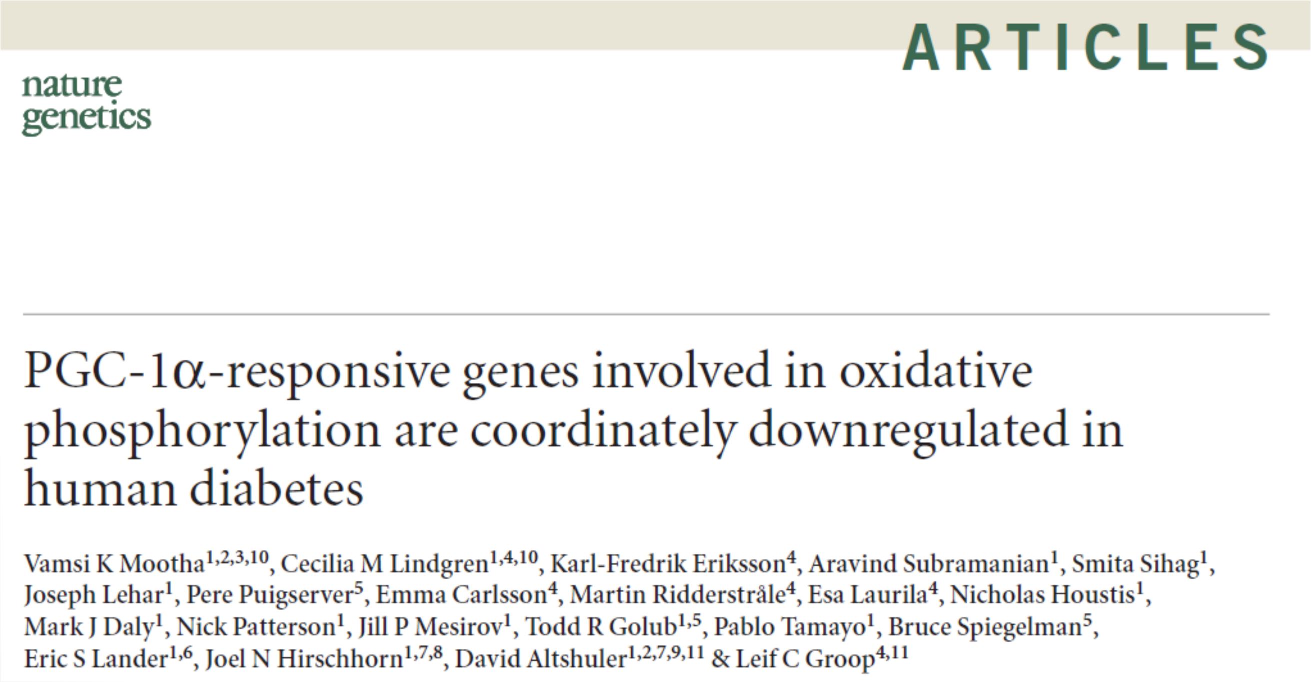
| Gene Set Name [# Genes (K)] | Description | # Genes in Overlap (k) | k/K | p-value ? | FDR q-value ? |
|---|---|---|---|---|---|
| GSE14000_UNSTIM_VS_4H_LPS_DC_TRANSLATED_RNA_DN [200] | Genes down-regulated in comparison of polysome bound (translated) mRNA before and 4 h after LPS (TLR4 agonist) stimulation. | 16 | 🟩 | 5.15 e$^{-28}$ | 4.16 e$^{-24}$ |
| GSE2706_R848_VS_R848_AND_LPS_2H_STIM_DM_DC_DN [200] | Genes down-regulated in comparison of dendritic cells (DC) stimulated with R848 at 2 h versus DCs stimulated with LPS (TLR4 agonist) and R848 for 2 h. | 15 | 🟩 | 8.15 e$^{-26}$ | 3.3 e$^{-22}$ |
| GSE18791_CTRL_VS_NEWCASTLE_VIRUS_DC_8H_8H_DN [200] | Genes down-regulated in comparison of control conventional dendritic cells (cDC) at 0 h versus cDCs infected with Newcastle disease virus (NDV) at 8 h. | 14 | 🟩 | 1.16 e$^{-23}$ | 2.34 e$^{-20}$ |

# Outline

- Formulating the problem
- What are the references?
- Overrepresentation methods
- **Gene set enrichment analysis**
- Gene set analysis generalization

# GSEA

- Published in 2003 as a side-method in Nature Genetics



ARTICLES

nature genetics

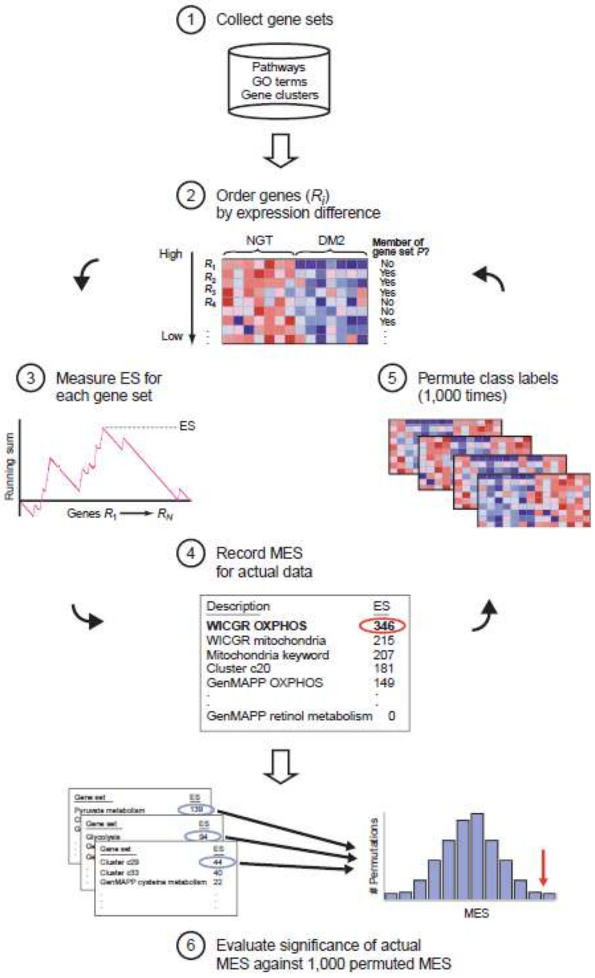
PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes

Vamsi K Mootha[1,2,3,10], Cecilia M Lindgren[1,4,10], Karl-Fredrik Eriksson[4], Aravind Subramanian[1], Smita Sihag[1], Joseph Lehar[1], Pere Puigserver[5], Emma Carlsson[4], Martin Ridderstråle[4], Esa Laurila[4], Nicholas Houstis[1], Mark J Daly[1], Nick Patterson[1], Jill P Mesirov[1], Todd R Golub[1,5], Pablo Tamayo[1], Bruce Spiegelman[5], Eric S Lander[1,6], Joel N Hirschhorn[1,7,8], David Altshuler[1,2,7,9,11] & Leif C Groop[4,11]

http://www.ncbi.nlm.nih.gov/pubmed/12808457

# Original GSEA

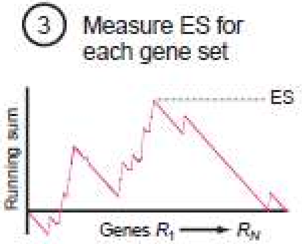
- Used Kolmogorov-Smirnov test
- Nonparametric in nature – uses rank
- Uses all genes (not just selected set)

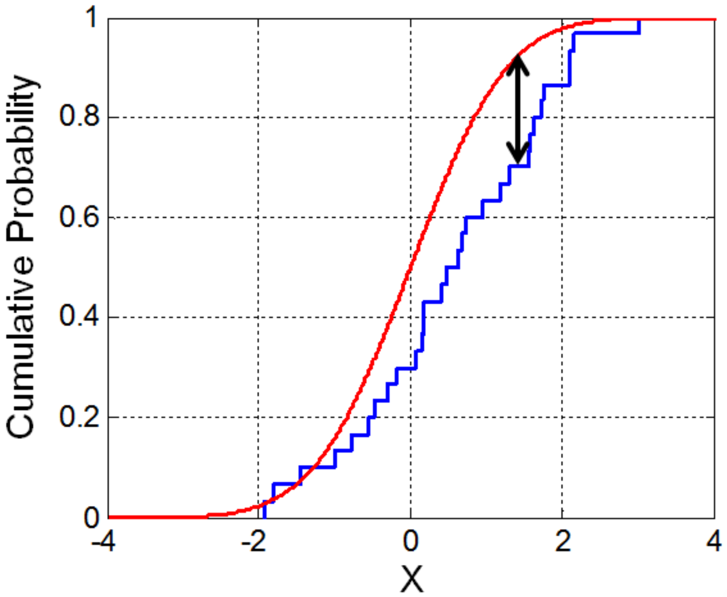


| Description | ES |
|---|---|
| **WICGR OXPHOS** | **346** |
| WICGR mitochondria | 215 |
| Mitochondria keyword | 207 |
| Cluster c20 | 181 |
| GenMAPP OXPHOS | 149 |
| ⋮ | ⋮ |
| GenMAPP retinol metabolism | 0 |

# Kolmogorov-Smirnov test

- Quantifies the distance between
  - Empirical distribution
  - Reference CDF
- ES = enrichment score
- Defined as highest running sum





# P-value

- P-value is calculated via permutations
- Labels (exp, control) are shuffled randomly 1000 times
- Number of times larger ES is obtained recorded (n)
- Nominal pval = n/1000

# Criticism

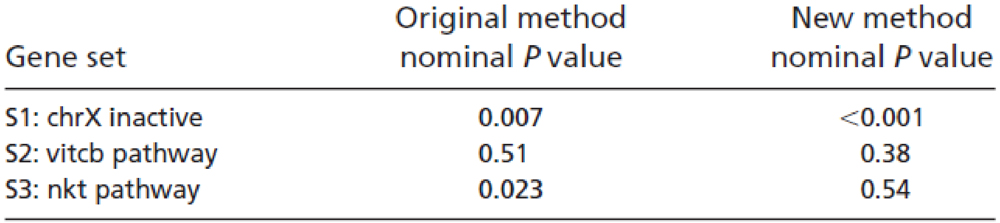
- Concern that few dramatic changes are lost in large pool of insignificantly changing genes
- Too dependent on pre-determined gene sets

# Reply to criticism

- Significance should be dependent on size: more measurements = less variance
- Dependence on a priori defined gene sets is declared and expected

# New re-vamped GSEA

- Correlation-weighted KS statistic (more power to more differential genes)
- ES normalization (NES)
- Compute FDR-like adjusted significance measure instead of FWER

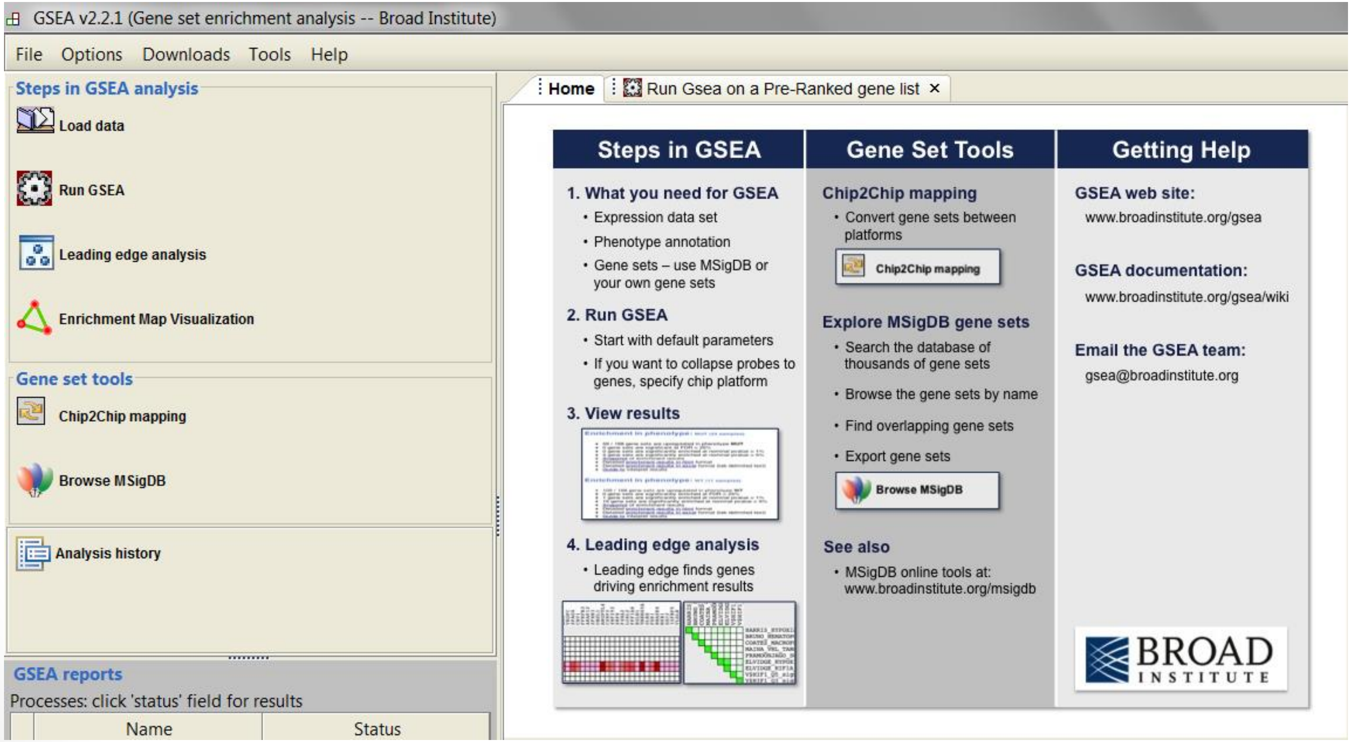


| Gene set | Original method nominal *P* value | New method nominal *P* value |
|---|---|---|
| S1: chrX inactive | 0.007 | <0.001 |
| S2: vitcb pathway | 0.51 | 0.38 |
| S3: nkt pathway | 0.023 | 0.54 |

http://www.ncbi.nlm.nih.gov/pubmed/16199517

# GSEA application

- Optimized for microarrays



GSEA v2.2.1 (Gene set enrichment analysis -- Broad Institute)
File
Options
Downloads
Tools
Help
Home
Run Gsea on a Pre-Ranked gene list ×
Chip2Chip mapping
Browse MSigDB
Name
Status

**Steps in GSEA analysis**

- Load data
- Run GSEA
- Leading edge analysis
- Enrichment Map Visualization

**Gene set tools**

- Chip2Chip mapping
- Browse MSigDB

Analysis history

**GSEA reports**

Processes: click 'status' field for results

**Steps in GSEA**

1. What you need for GSEA
   - Expression data set
   - Phenotype annotation
   - Gene sets – use MSigDB or your own gene sets
2. Run GSEA
   - Start with default parameters
   - If you want to collapse probes to genes, specify chip platform
3. View results
4. Leading edge analysis
   - Leading edge finds genes driving enrichment results

**Gene Set Tools**

Chip2Chip mapping

- Convert gene sets between platforms

Explore MSigDB gene sets

- Search the database of thousands of gene sets
- Browse the gene sets by name
- Find overlapping gene sets
- Export gene sets

See also

- MSigDB online tools at: www.broadinstitute.org/msigdb

**Getting Help**

GSEA web site:
www.broadinstitute.org/gsea

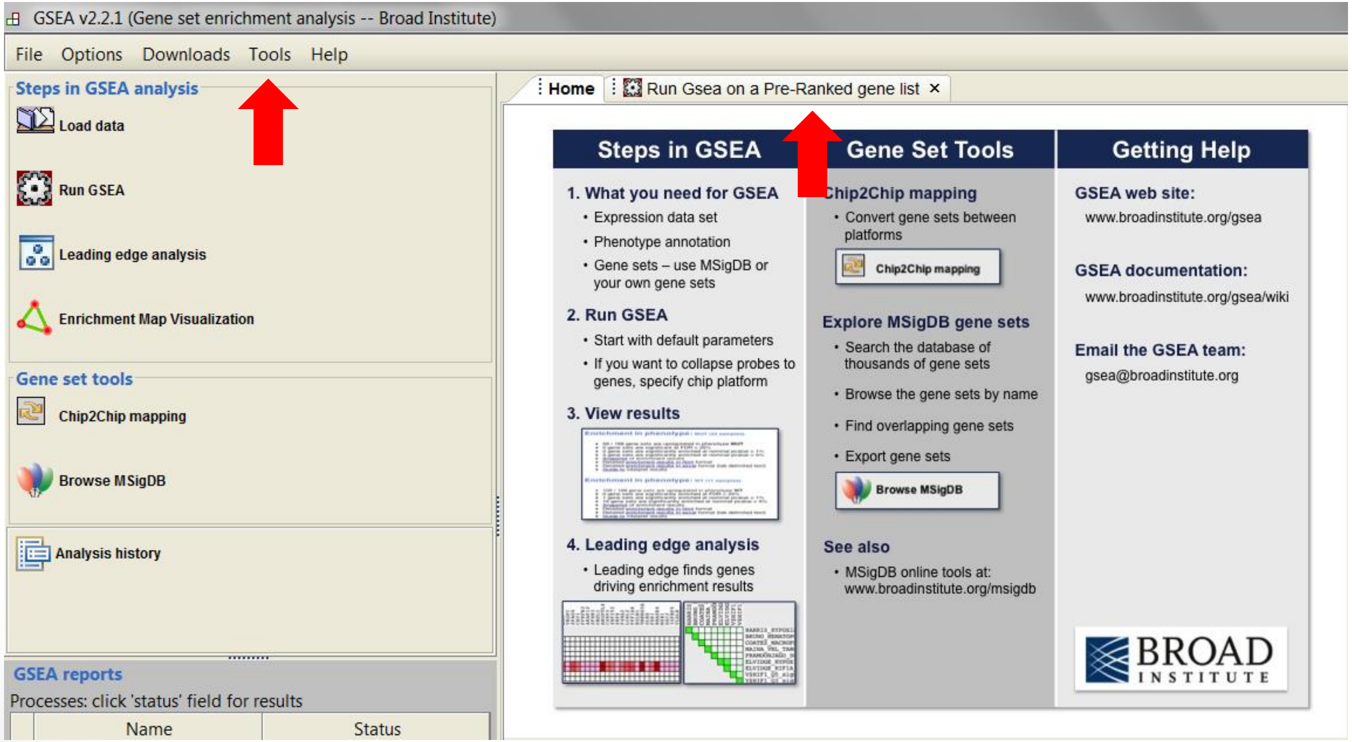
GSEA documentation:
www.broadinstitute.org/gsea/wiki

Email the GSEA team:
gsea@broadinstitute.org

BROAD INSTITUTE

# GSEA application

- Use Gsea Pre-ranked tool for RNA-seq!



**Steps in GSEA**

1. What you need for GSEA
   - Expression data set
   - Phenotype annotation
   - Gene sets – use MSigDB or your own gene sets
2. Run GSEA
   - Start with default parameters
   - If you want to collapse probes to genes, specify chip platform
3. View results
4. Leading edge analysis
   - Leading edge finds genes driving enrichment results

**Gene Set Tools**

Chip2Chip mapping

- Convert gene sets between platforms

Explore MSigDB gene sets

- Search the database of thousands of gene sets
- Browse the gene sets by name
- Find overlapping gene sets
- Export gene sets

See also

- MSigDB online tools at: www.broadinstitute.org/msigdb

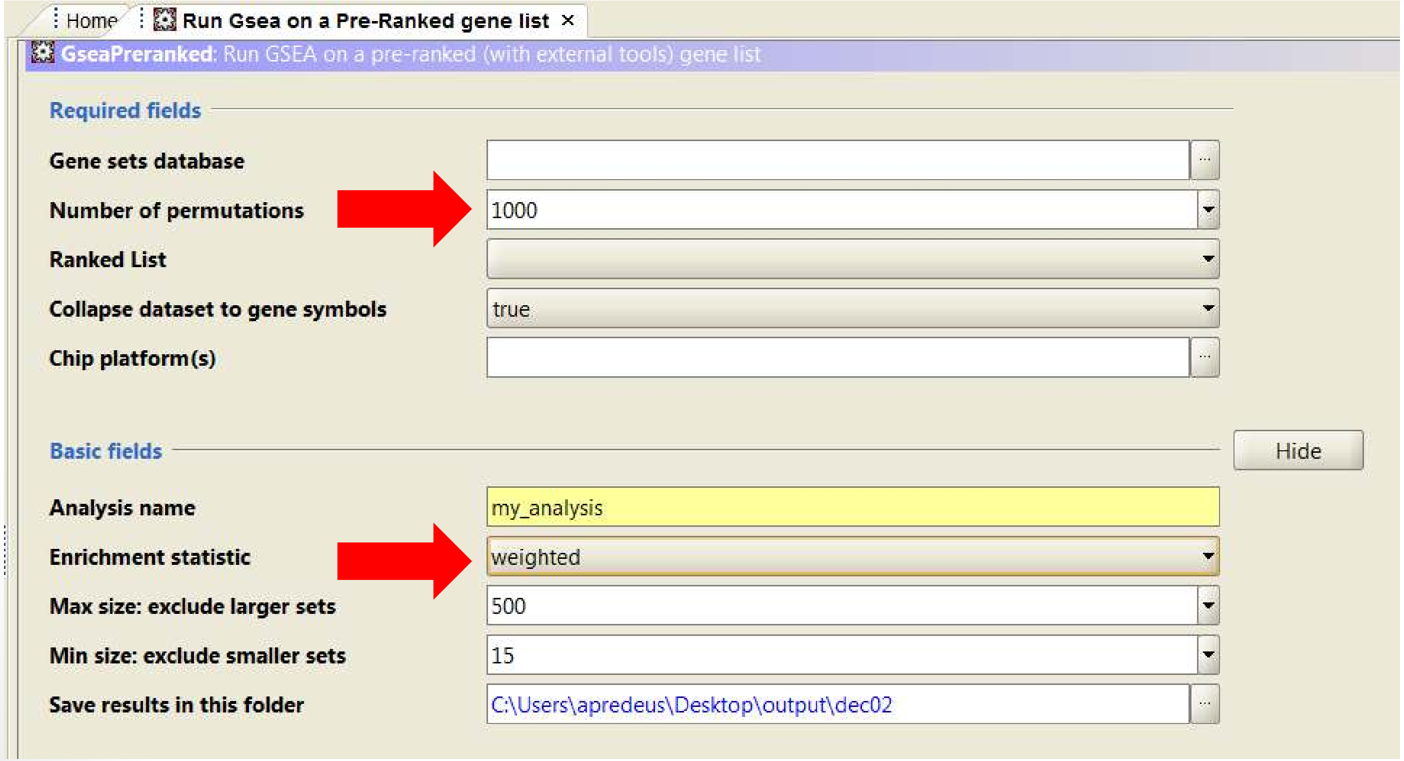
**Getting Help**

GSEA web site:
www.broadinstitute.org/gsea

GSEA documentation:
www.broadinstitute.org/gsea/wiki

Email the GSEA team:
gsea@broadinstitute.org

# Permutations & statistic are crucial



**GseaPreranked**: Run GSEA on a pre-ranked (with external tools) gene list

**Required fields**

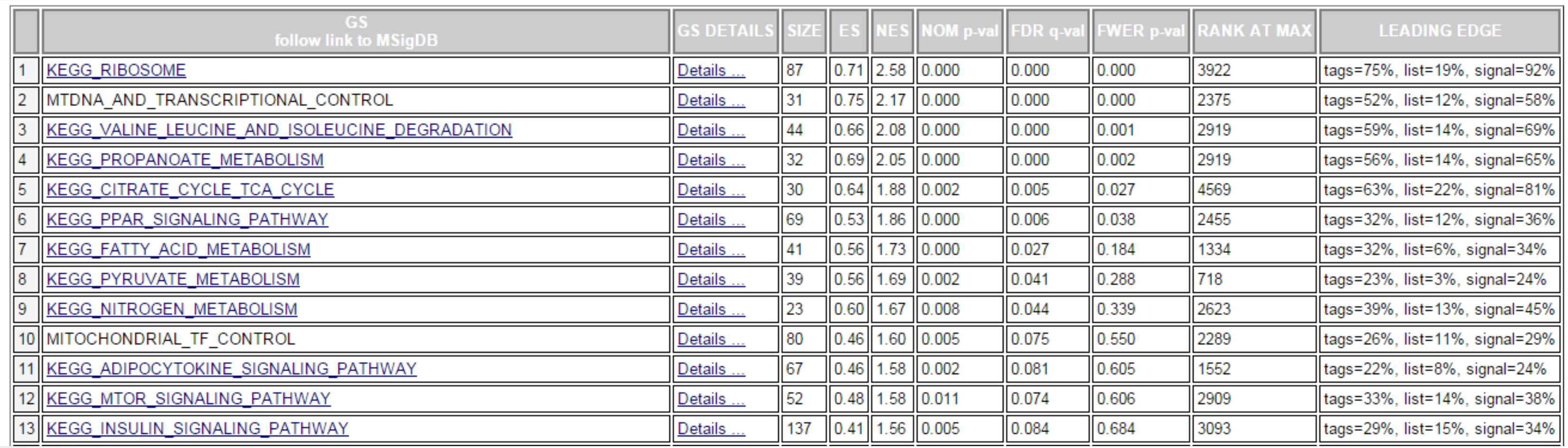
| | |
|---|---|
| Gene sets database | |
| Number of permutations | 1000 |
| Ranked List | |
| Collapse dataset to gene symbols | true |
| Chip platform(s) | |

**Basic fields**

| | |
|---|---|
| Analysis name | my_analysis |
| Enrichment statistic | weighted |
| Max size: exclude larger sets | 500 |
| Min size: exclude smaller sets | 15 |
| Save results in this folder | C:\Users\apredeus\Desktop\output\dec02 |

# Output

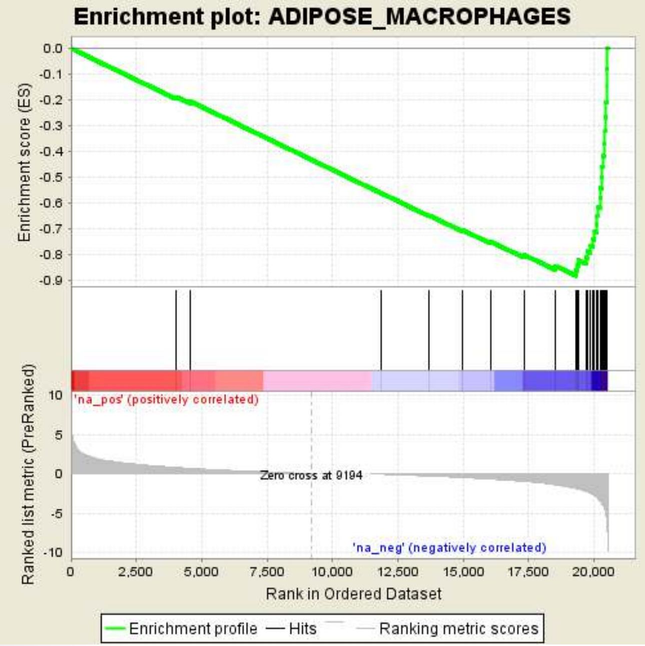
- Folder with results
- Separate .html files for up- and down-regulated

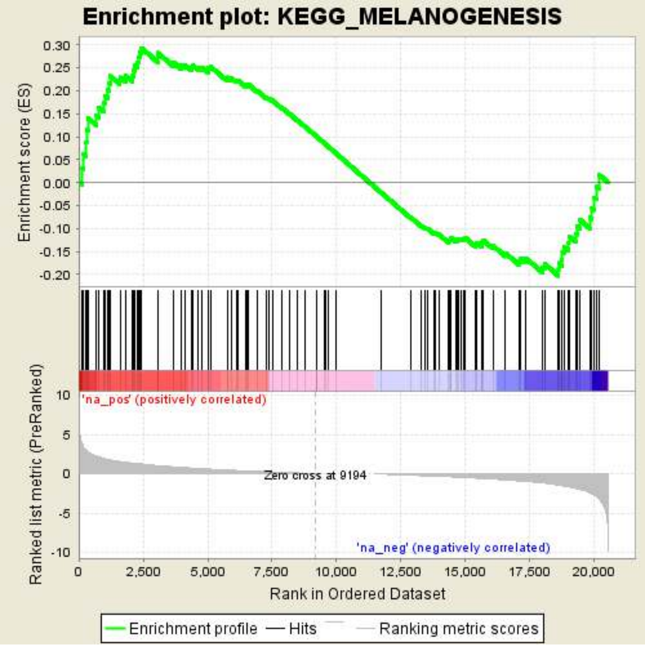


| | GS<br>follow link to MSigDB | GS DETAILS | SIZE | ES | NES | NOM p-val | FDR q-val | FWER p-val | RANK AT MAX | LEADING EDGE |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | KEGG_RIBOSOME | Details ... | 87 | 0.71 | 2.58 | 0.000 | 0.000 | 0.000 | 3922 | tags=75%, list=19%, signal=92% |
| 2 | MTDNA_AND_TRANSCRIPTIONAL_CONTROL | Details ... | 31 | 0.75 | 2.17 | 0.000 | 0.000 | 0.000 | 2375 | tags=52%, list=12%, signal=58% |
| 3 | KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION | Details ... | 44 | 0.66 | 2.08 | 0.000 | 0.000 | 0.001 | 2919 | tags=59%, list=14%, signal=69% |
| 4 | KEGG_PROPANOATE_METABOLISM | Details ... | 32 | 0.69 | 2.05 | 0.000 | 0.000 | 0.002 | 2919 | tags=56%, list=14%, signal=65% |
| 5 | KEGG_CITRATE_CYCLE_TCA_CYCLE | Details ... | 30 | 0.64 | 1.88 | 0.002 | 0.005 | 0.027 | 4569 | tags=63%, list=22%, signal=81% |
| 6 | KEGG_PPAR_SIGNALING_PATHWAY | Details ... | 69 | 0.53 | 1.86 | 0.000 | 0.006 | 0.038 | 2455 | tags=32%, list=12%, signal=36% |
| 7 | KEGG_FATTY_ACID_METABOLISM | Details ... | 41 | 0.56 | 1.73 | 0.000 | 0.027 | 0.184 | 1334 | tags=32%, list=6%, signal=34% |
| 8 | KEGG_PYRUVATE_METABOLISM | Details ... | 39 | 0.56 | 1.69 | 0.002 | 0.041 | 0.288 | 718 | tags=23%, list=3%, signal=24% |
| 9 | KEGG_NITROGEN_METABOLISM | Details ... | 23 | 0.60 | 1.67 | 0.008 | 0.044 | 0.339 | 2623 | tags=39%, list=13%, signal=45% |
| 10 | MITOCHONDRIAL_TF_CONTROL | Details ... | 80 | 0.46 | 1.60 | 0.005 | 0.075 | 0.550 | 2289 | tags=26%, list=11%, signal=29% |
| 11 | KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY | Details ... | 67 | 0.46 | 1.58 | 0.002 | 0.081 | 0.605 | 1552 | tags=22%, list=8%, signal=24% |
| 12 | KEGG_MTOR_SIGNALING_PATHWAY | Details ... | 52 | 0.48 | 1.58 | 0.011 | 0.074 | 0.606 | 2909 | tags=33%, list=14%, signal=38% |
| 13 | KEGG_INSULIN_SIGNALING_PATHWAY | Details ... | 137 | 0.41 | 1.56 | 0.005 | 0.084 | 0.684 | 3093 | tags=29%, list=15%, signal=34% |

# Output

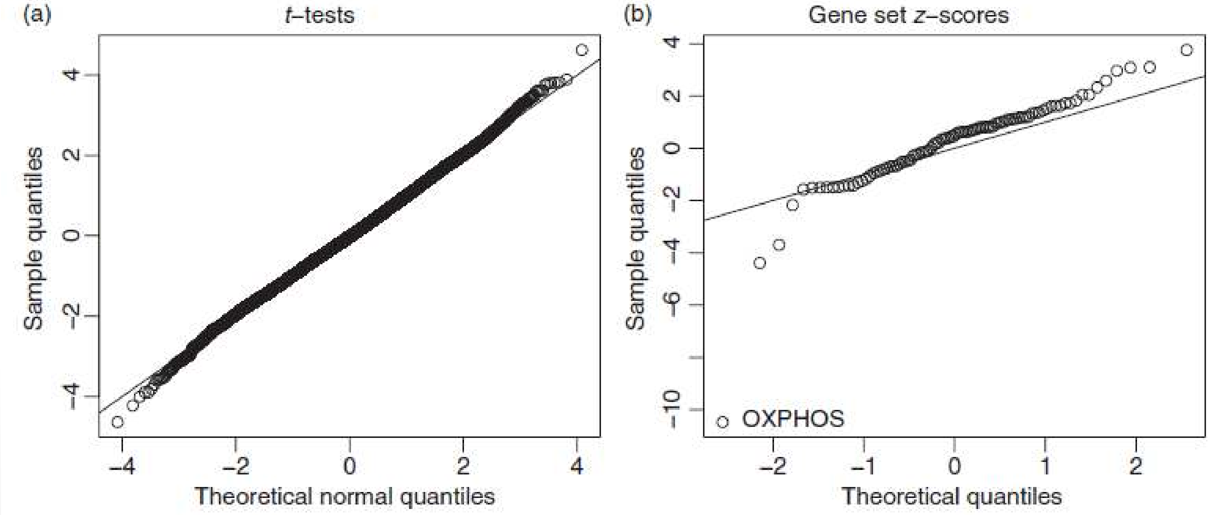
- ES as the main illustration of significance





# Simple GSEA

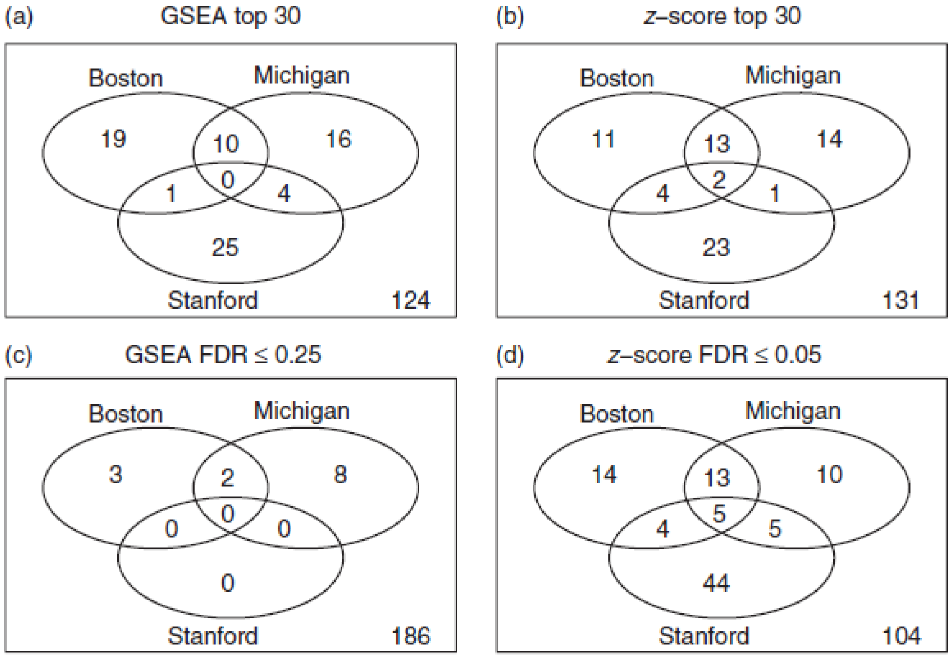
- Irizarry et al
- Assume gene independence
- Use "one sample t-test" to estimate enrichment



http://www.ncbi.nlm.nih.gov/pubmed/20048385

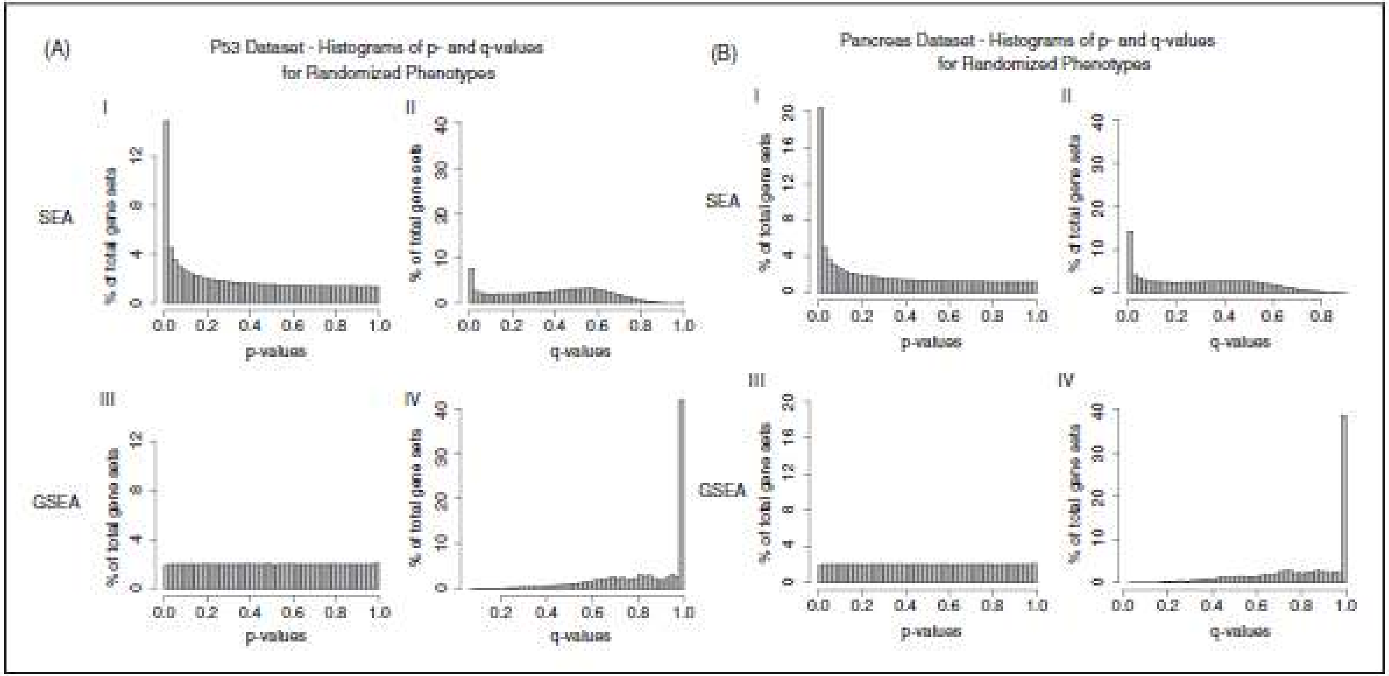
# Simple GSEA

- Cancer dataset – better agreement?



# Not-so-simple GSEA

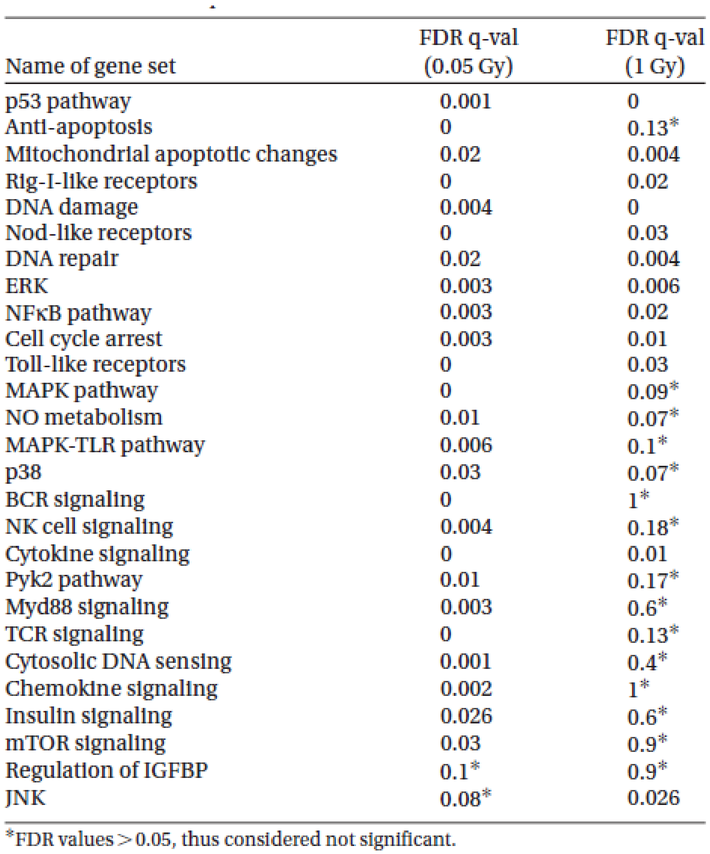
- Refuted by Mesirov in 2012



http://www.ncbi.nlm.nih.gov/pubmed/23070592

# Example 1: Mutant blood!!!

- Compared 0, 0.05, and 1 Gy X-rays treated blood
- Microarray PMBC
- More inflammation at low dose
- More p53 and DNA repair at high dose

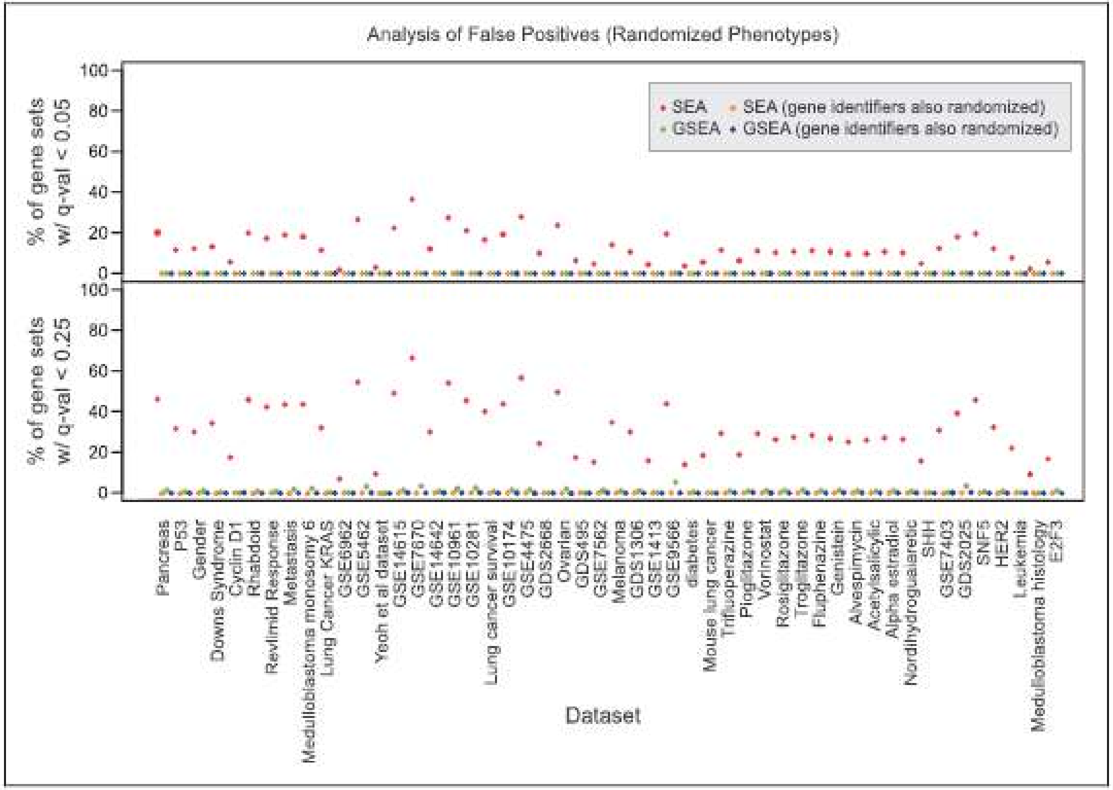


| Name of gene set | FDR q-val (0.05 Gy) | FDR q-val (1 Gy) |
|---|---|---|
| p53 pathway | 0.001 | 0 |
| Anti-apoptosis | 0 | 0.13* |
| Mitochondrial apoptotic changes | 0.02 | 0.004 |
| Rig-I-like receptors | 0 | 0.02 |
| DNA damage | 0.004 | 0 |
| Nod-like receptors | 0 | 0.03 |
| DNA repair | 0.02 | 0.004 |
| ERK | 0.003 | 0.006 |
| NFκB pathway | 0.003 | 0.02 |
| Cell cycle arrest | 0.003 | 0.01 |
| Toll-like receptors | 0 | 0.03 |
| MAPK pathway | 0 | 0.09* |
| NO metabolism | 0.01 | 0.07* |
| MAPK-TLR pathway | 0.006 | 0.1* |
| p38 | 0.03 | 0.07* |
| BCR signaling | 0 | 1* |
| NK cell signaling | 0.004 | 0.18* |
| Cytokine signaling | 0 | 0.01 |
| Pyk2 pathway | 0.01 | 0.17* |
| Myd88 signaling | 0.003 | 0.6* |
| TCR signaling | 0 | 0.13* |
| Cytosolic DNA sensing | 0.001 | 0.4* |
| Chemokine signaling | 0.002 | 1* |
| Insulin signaling | 0.026 | 0.6* |
| mTOR signaling | 0.03 | 0.9* |
| Regulation of IGFBP | 0.1* | 0.9* |
| JNK | 0.08* | 0.026 |

*FDR values > 0.05, thus considered not significant.

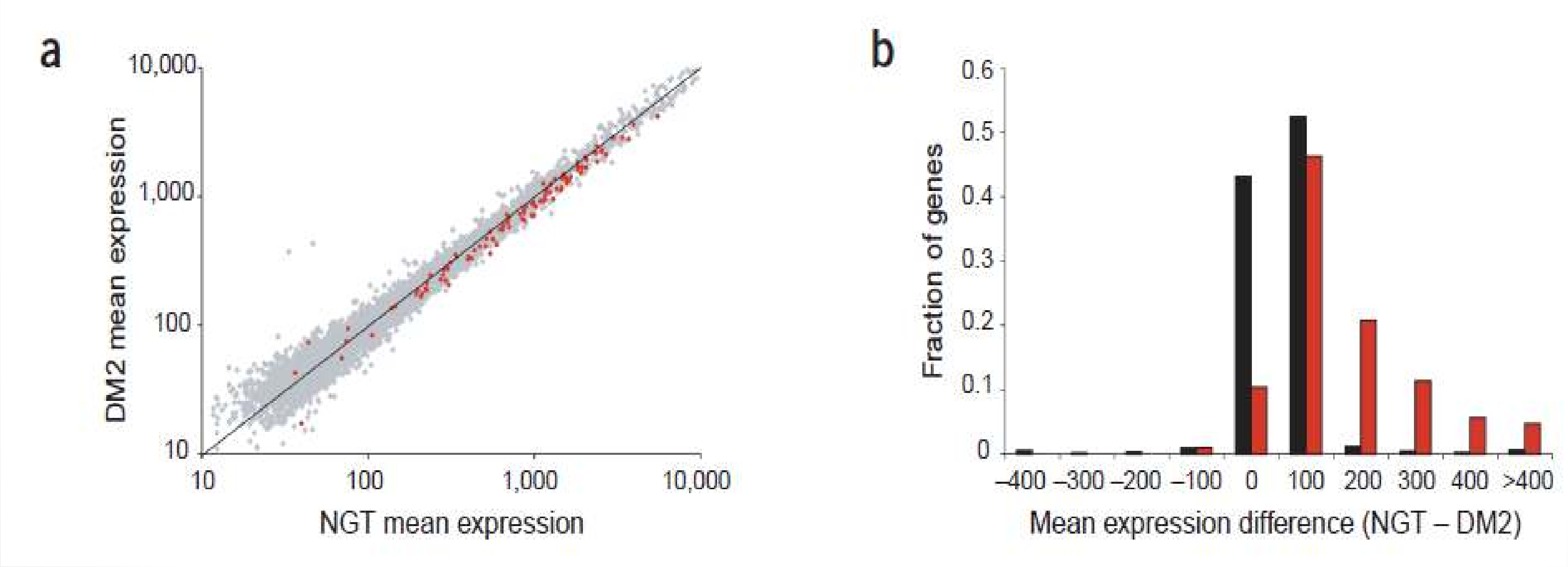
http://www.ncbi.nlm.nih.gov/pubmed/23484538

# Inflated false positives in SEA



# Example 2: diabetic PGC1a

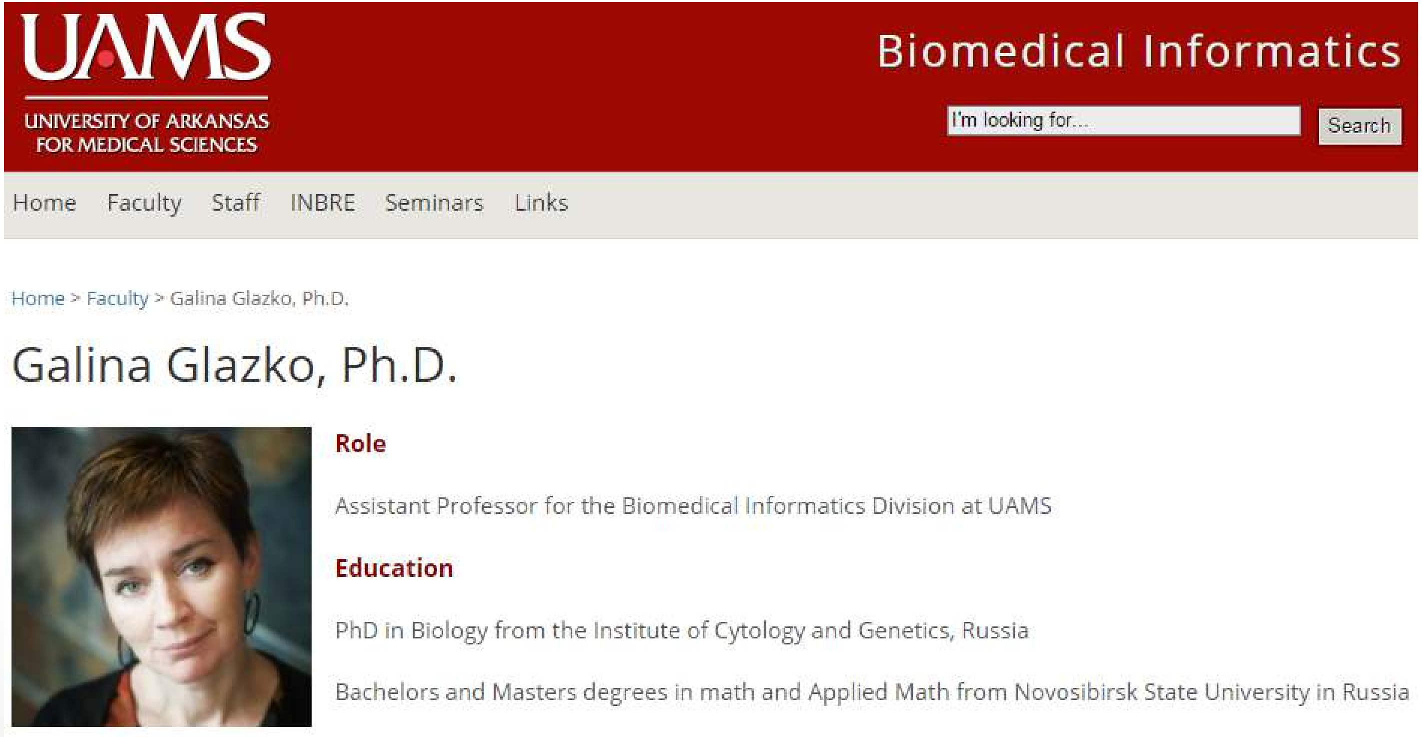
- Individual changes are small in metabolic adjustments
- Overall changes are significant



# Outline

- Formulating the problem
- What are the references?
- Overrepresentation methods
- Gene set enrichment analysis
- **Gene set analysis generalization**

# Gene Set Analysis



Home > Faculty > Galina Glazko, Ph.D.

## Galina Glazko, Ph.D.

### Role

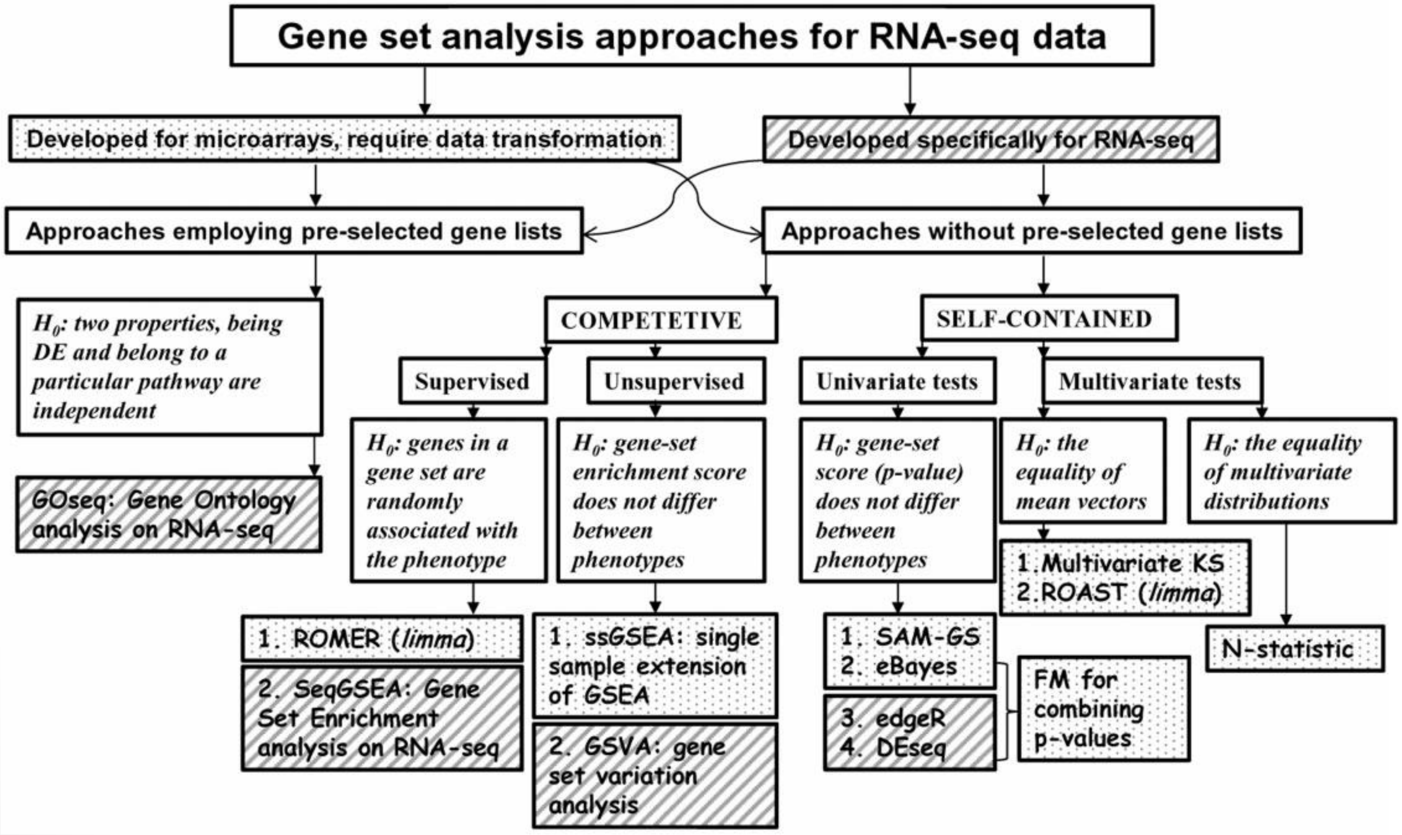
Assistant Professor for the Biomedical Informatics Division at UAMS

### Education

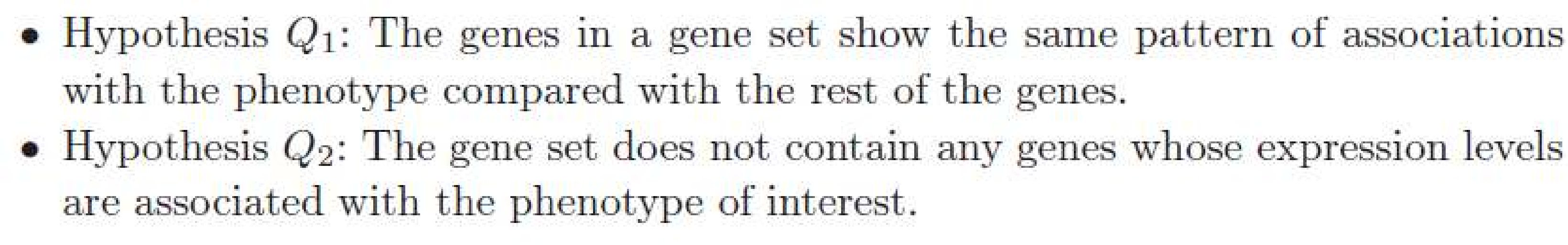
PhD in Biology from the Institute of Cytology and Genetics, Russia

Bachelors and Masters degrees in math and Applied Math from Novosibirsk State University in Russia

# GSA framework


Gene set analysis approaches for RNA-seq data
Developed for microarrays, require data transformation
Developed specifically for RNA-seq
Approaches employing pre-selected gene lists
Approaches without pre-selected gene lists
$H_0$: two properties, being DE and belong to a particular pathway are independent
GOseq: Gene Ontology analysis on RNA-seq
COMPETETIVE
SELF-CONTAINED
Supervised
Unsupervised
Univariate tests
Multivariate tests
$H_0$: genes in a gene set are randomly associated with the phenotype
$H_0$: gene-set enrichment score does not differ between phenotypes
$H_0$: gene-set score (p-value) does not differ between phenotypes
$H_0$: the equality of mean vectors
$H_0$: the equality of multivariate distributions
1. Multivariate KS
2. ROAST (limma)
1. ROMER (limma)
2. SeqGSEA: Gene Set Enrichment analysis on RNA-seq
1. ssGSEA: single sample extension of GSEA
2. GSVA: gene set variation analysis
1. SAM-GS
2. eBayes
3. edgeR
4. DEseq
FM for combining p-values
N-statistic

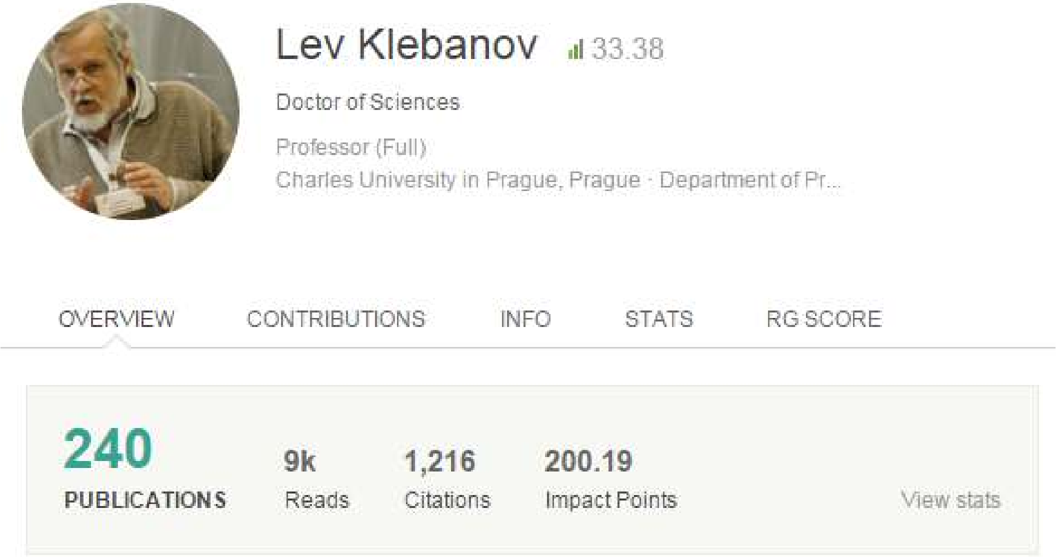
# Competitive vs self-contained null



- Hypothesis $Q_1$: The genes in a gene set show the same pattern of associations with the phenotype compared with the rest of the genes.
- Hypothesis $Q_2$: The gene set does not contain any genes whose expression levels are associated with the phenotype of interest.

# Multivariate GSEA

- Lev Klebanov
- Uses N-statistic
- More sensitive than the generic version



Lev Klebanov 33.38

Doctor of Sciences

Professor (Full)
Charles University in Prague, Prague · Department of Pr...

240 PUBLICATIONS

9k Reads

1,216 Citations

200.19 Impact Points

That's all Folks!