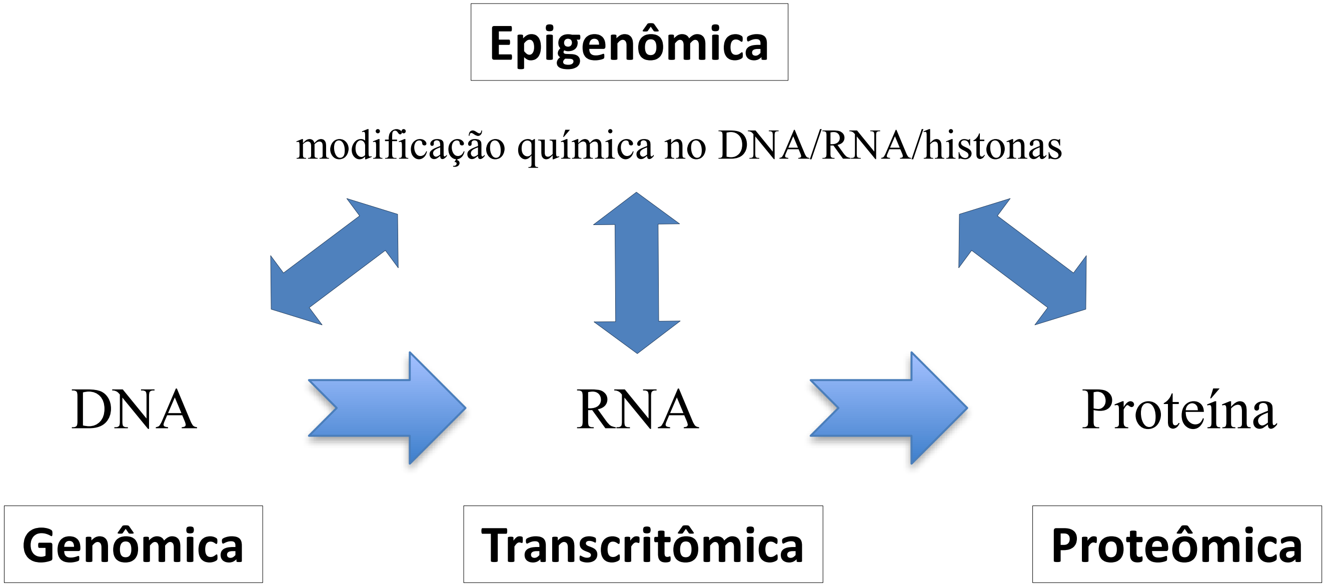
Biologia Molecular Computacional
IBI5035/QBQ2507 - 2023

Anotação funcional de genes

Eduardo Moraes Rego Reis
Instituto de Química - USP

Estudos ômicos em larga-escala Identificam listas com dezenas a centenas de genes com padrão de expressão alterada



Epigenômica

modificação química no DNA/RNA/histonas

DNA

RNA

Proteína

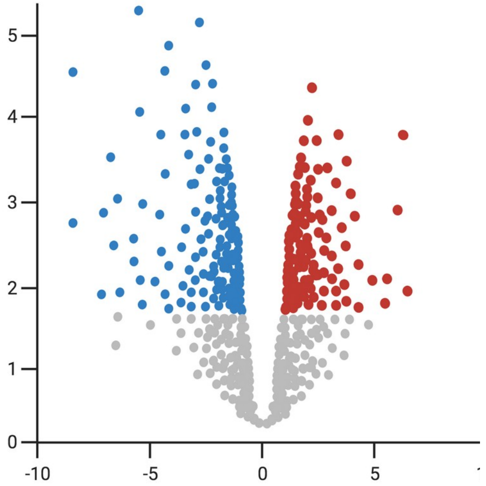
Genômica

Transcritômica

Proteômica

- Situações patológicas
- Estágio do desenvolvimento
- Tratamento com droga
- outros

A **anotação funcional** auxilia na interpretação dos resultados e na identificação das alterações mais relevantes para explicar o fenômeno biológico de interesse.



Anotação funcional e análise de enriquecimento gênico

Objetivo:

Atribuir significado biológico a um ou mais grupos de genes identificados durante experimentos.

Estratégia:

Identificar o enriquecimento de algum tipo de padrão entre os genes selecionados, acima do esperado ao acaso.

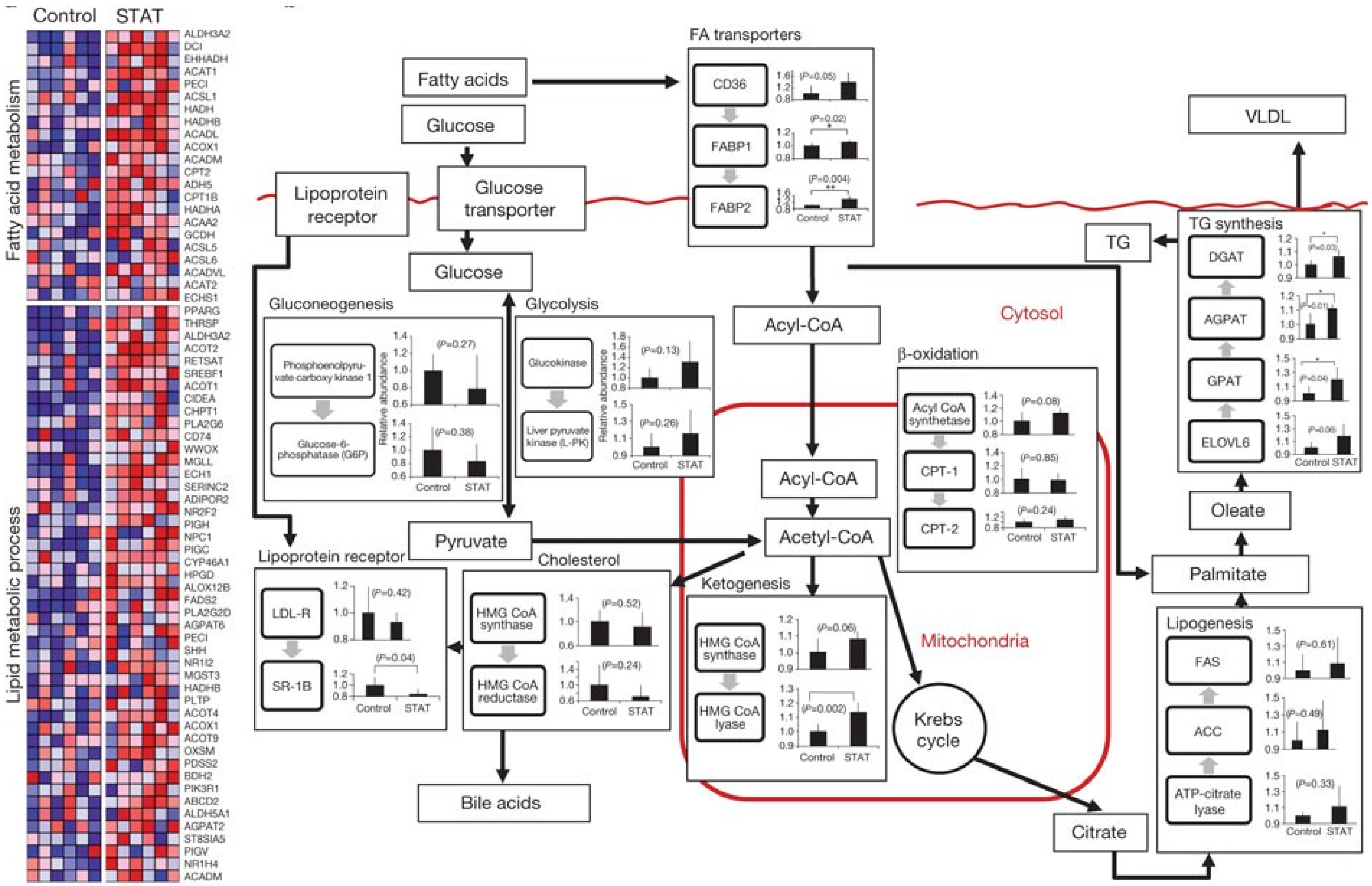
Exs.:

- Possuam as **mesmas funções moleculares** ou participem nos mesmo processos biológicos
- Codifiquem proteínas que se localizam nos **mesmos componentes celulares** (ex. Núcleo, mitocôndria, membrana plasmática)
- Participem das **mesmas vias metabólicas**
- Sejam ativados pelos **mesmos fatores de transcrição**
- Estejam envolvidos em uma mesma **doença**

Permite gerar hipóteses para experimentação adicional

Exemplo de aplicação de análise de enriquecimento gênico

Genes diferencialmente expressos no fígado de camundongos em um modelo de obesidade estão enriquecidos em genes relacionados ao metabolismo lipídico



- ➢ 45,000 genes interrogados
- ➢ 397 genes diferencialmente expressos (GDEs)
- ➢ 68 GDEs em vias relacionadas a lipogênese e síntese de triglicerídeos (na figura)

Cho et al., 2012 Nature 488, 621–626

Análises de enriquecimento dependem de informações (anotações) estruturadas de genes e suas funções

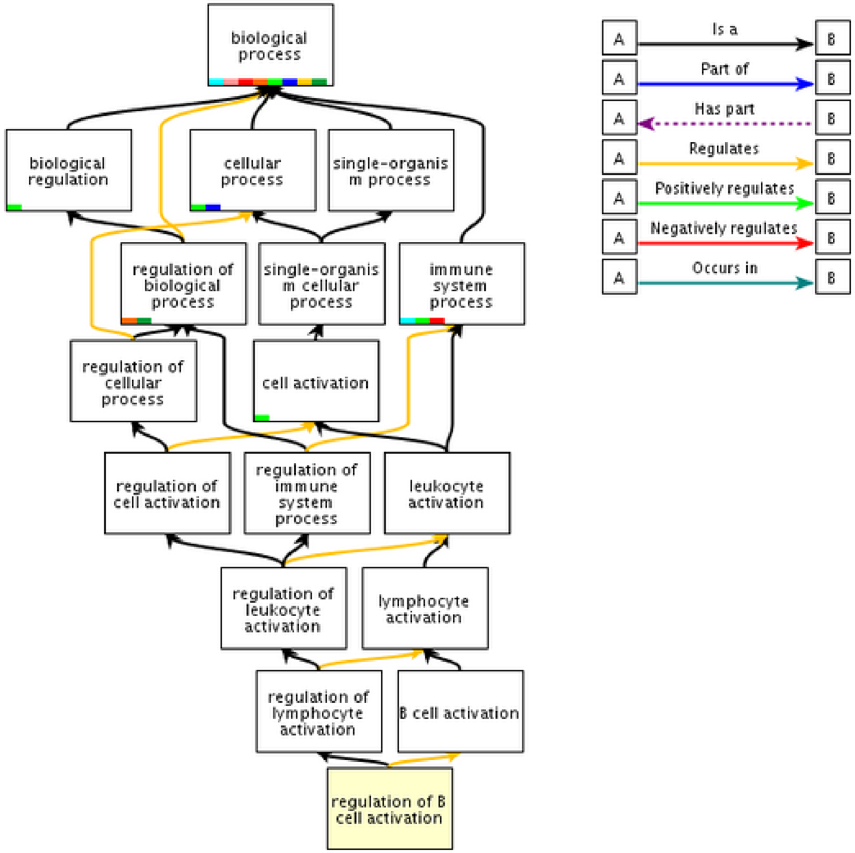
- Gene Ontology
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- Reactome

mais alguns bancos de dados biológicos

Gene Ontology

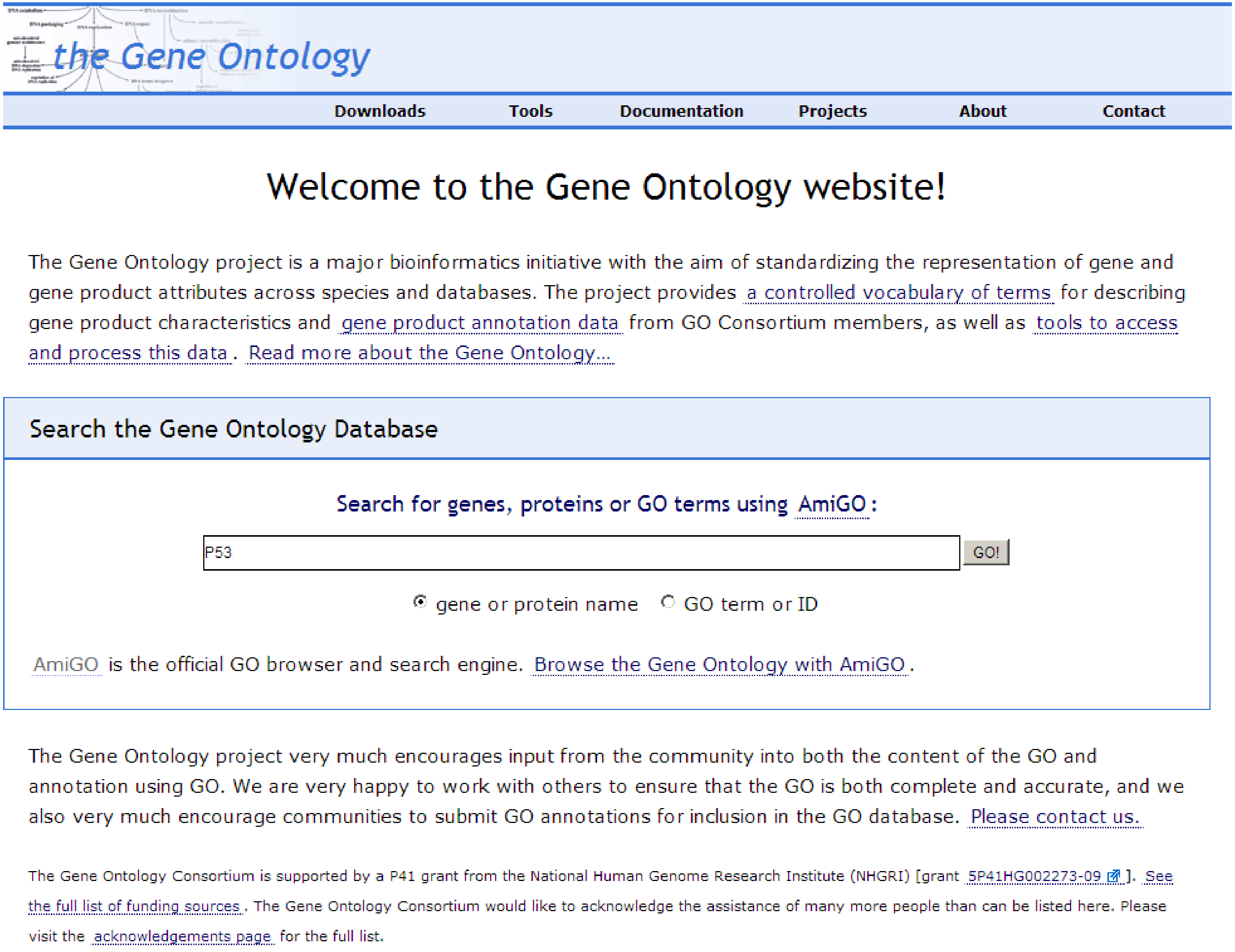
- Vocabulário estruturado e controlado que descreve produtos gênicos em termos de **processos biológicos, funções moleculares** e **componentes celulares**

Ex.: Term Neighborhood for regulation of B cell activation (GO:0050864)



QuickGO - http://www.ebi.ac.uk/QuickGO

http://www.geneontology.org/



Welcome to the Gene Ontology website!

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. Read more about the Gene Ontology...

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

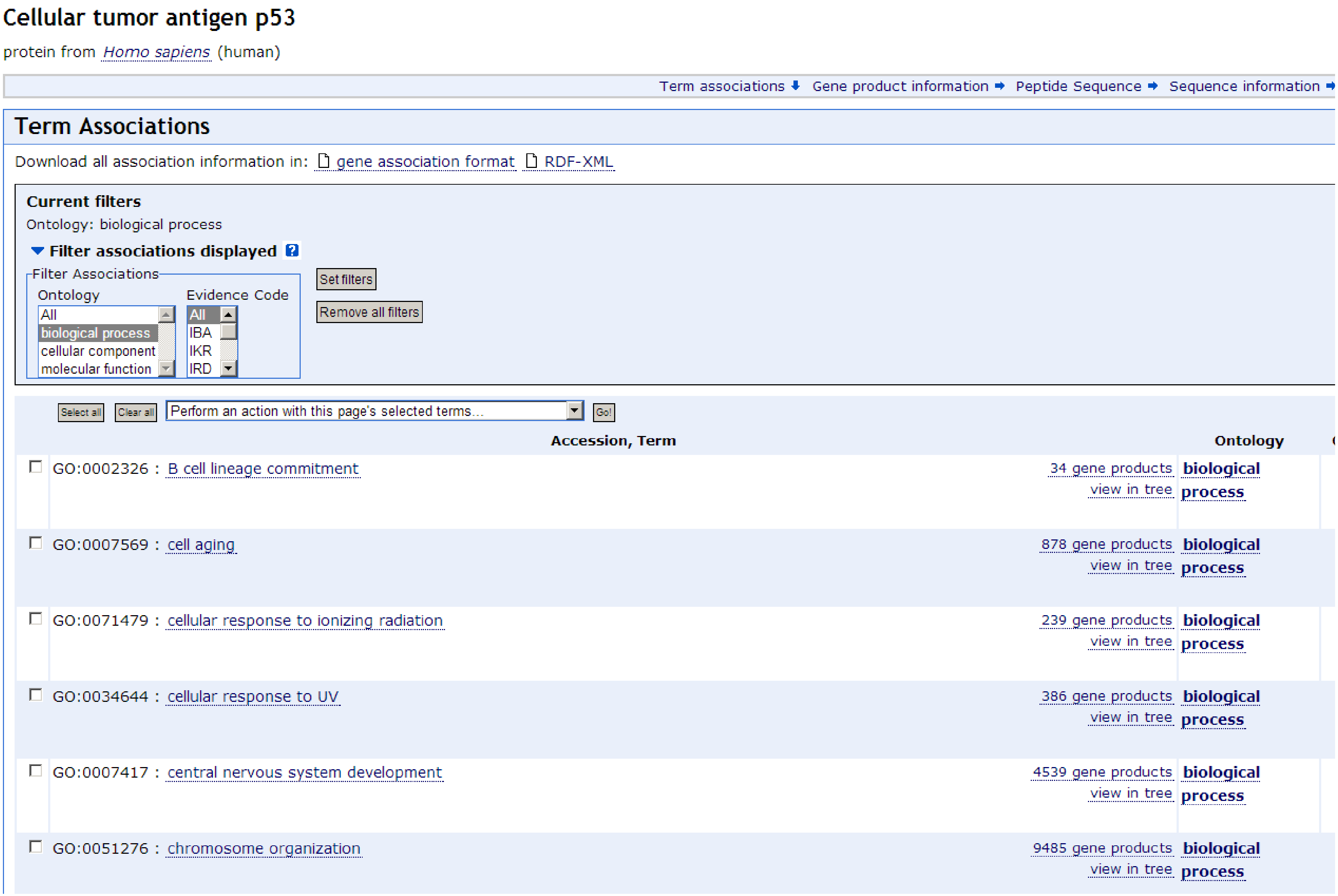
P53

● gene or protein name ○ GO term or ID

AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.

The Gene Ontology project very much encourages input from the community into both the content of the GO and annotation using GO. We are very happy to work with others to ensure that the GO is both complete and accurate, and we also very much encourage communities to submit GO annotations for inclusion in the GO database. Please contact us.

The Gene Ontology Consortium is supported by a P41 grant from the National Human Genome Research Institute (NHGRI) [grant 5P41HG002273-09]. See the full list of funding sources. The Gene Ontology Consortium would like to acknowledge the assistance of many more people than can be listed here. Please visit the acknowledgements page for the full list.



Cellular tumor antigen p53

protein from *Homo sapiens* (human)

Term associations ⬇ Gene product information ➡ Peptide Sequence ➡ Sequence information ➡

Term Associations

Download all association information in: gene association format RDF-XML

Current filters

Ontology: biological process

▼ Filter associations displayed ?

Filter Associations

Ontology	Evidence Code
All biological process cellular component molecular function	All IBA IKR IRD

Set filters

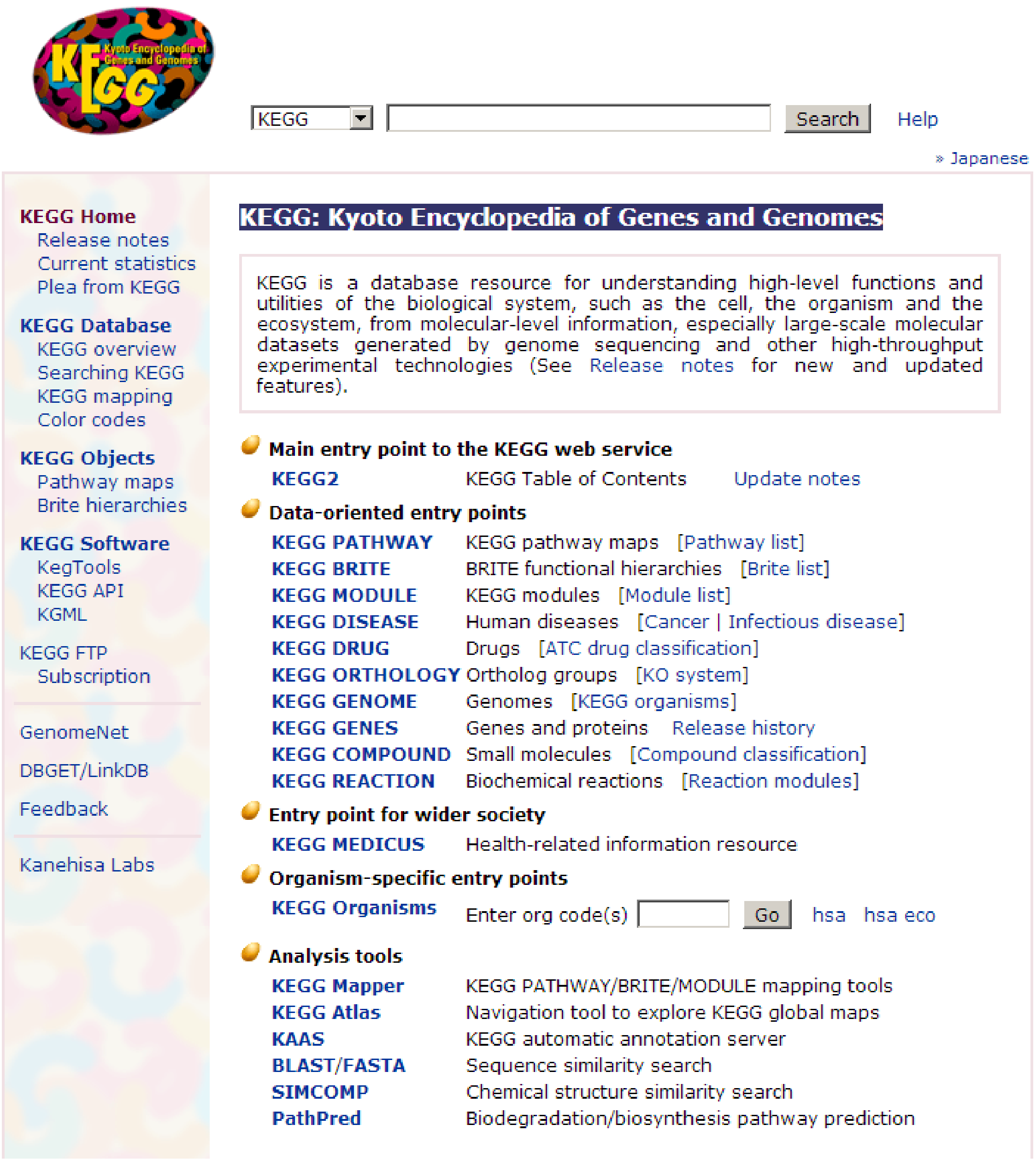
Remove all filters

Select all | Clear all | Perform an action with this page's selected terms... | Go!

	Accession, Term		Ontology
□	GO:0002326 : B cell lineage commitment	34 gene products view in tree	biological process
□	GO:0007569 : cell aging	878 gene products view in tree	biological process
□	GO:0071479 : cellular response to ionizing radiation	239 gene products view in tree	biological process
□	GO:0034644 : cellular response to UV	386 gene products view in tree	biological process
□	GO:0007417 : central nervous system development	4539 gene products view in tree	biological process
□	GO:0051276 : chromosome organization	9485 gene products view in tree	biological process

KEGG: Kyoto Encyclopedia of Genes and Genomes

http://www.genome.jp/kegg/



KEGG Home
- Release notes
- Current statistics
- Plea from KEGG

KEGG Database
- KEGG overview
- Searching KEGG
- KEGG mapping
- Color codes

KEGG Objects
- Pathway maps
- Brite hierarchies

KEGG Software
- KegTools
- KEGG API
- KGML

KEGG FTP
- Subscription

GenomeNet

DBGET/LinkDB

Feedback

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (See Release notes for new and updated features).

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents	Update notes

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps [Pathway list]
KEGG BRITE	BRITE functional hierarchies [Brite list]
KEGG MODULE	KEGG modules [Module list]
KEGG DISEASE	Human diseases [Cancer | Infectious disease]
KEGG DRUG	Drugs [ATC drug classification]
KEGG ORTHOLOGY	Ortholog groups [KO system]
KEGG GENOME	Genomes [KEGG organisms]
KEGG GENES	Genes and proteins Release history
KEGG COMPOUND	Small molecules [Compound classification]
KEGG REACTION	Biochemical reactions [Reaction modules]

Entry point for wider society

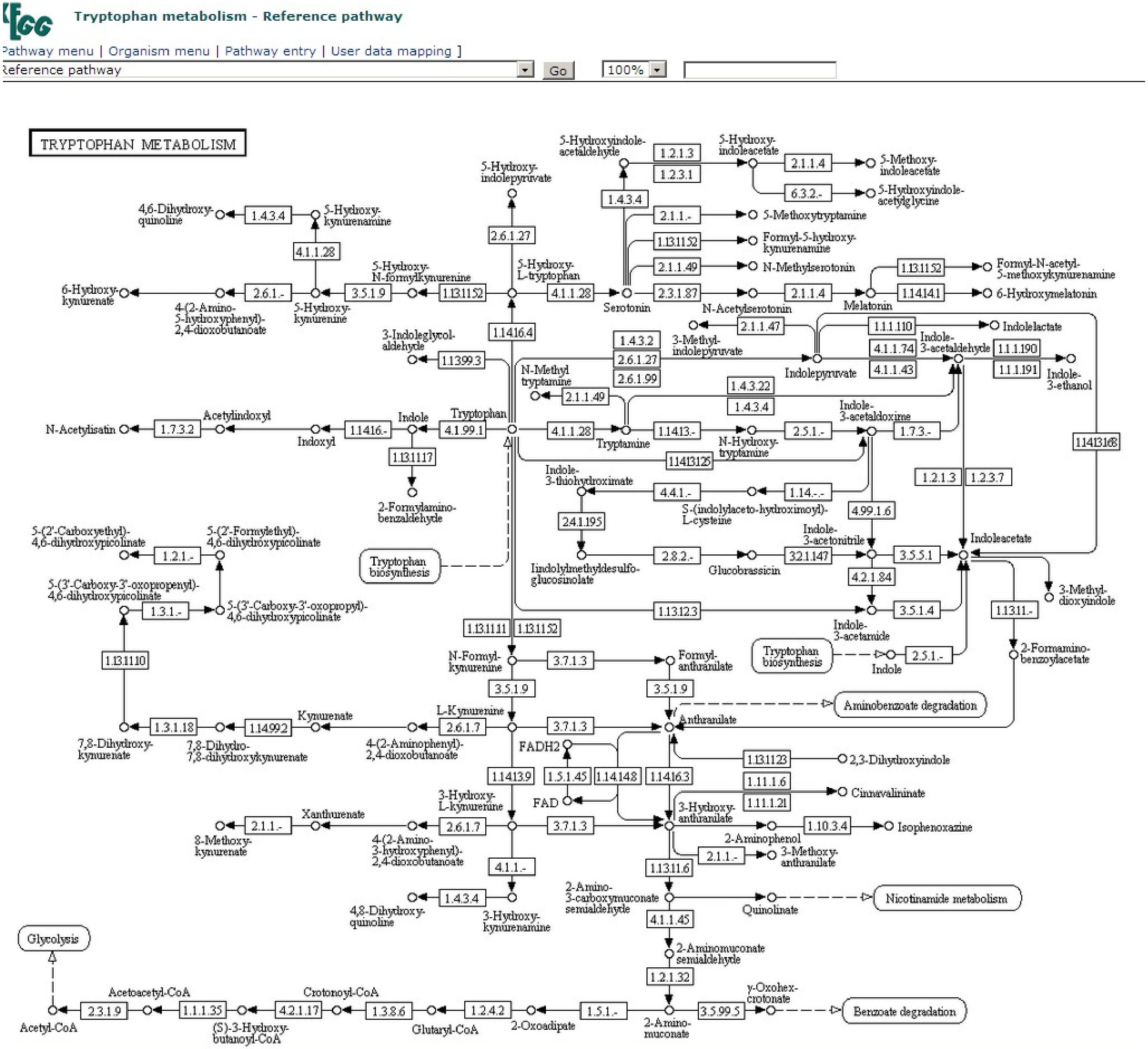
KEGG MEDICUS	Health-related information resource

Organism-specific entry points

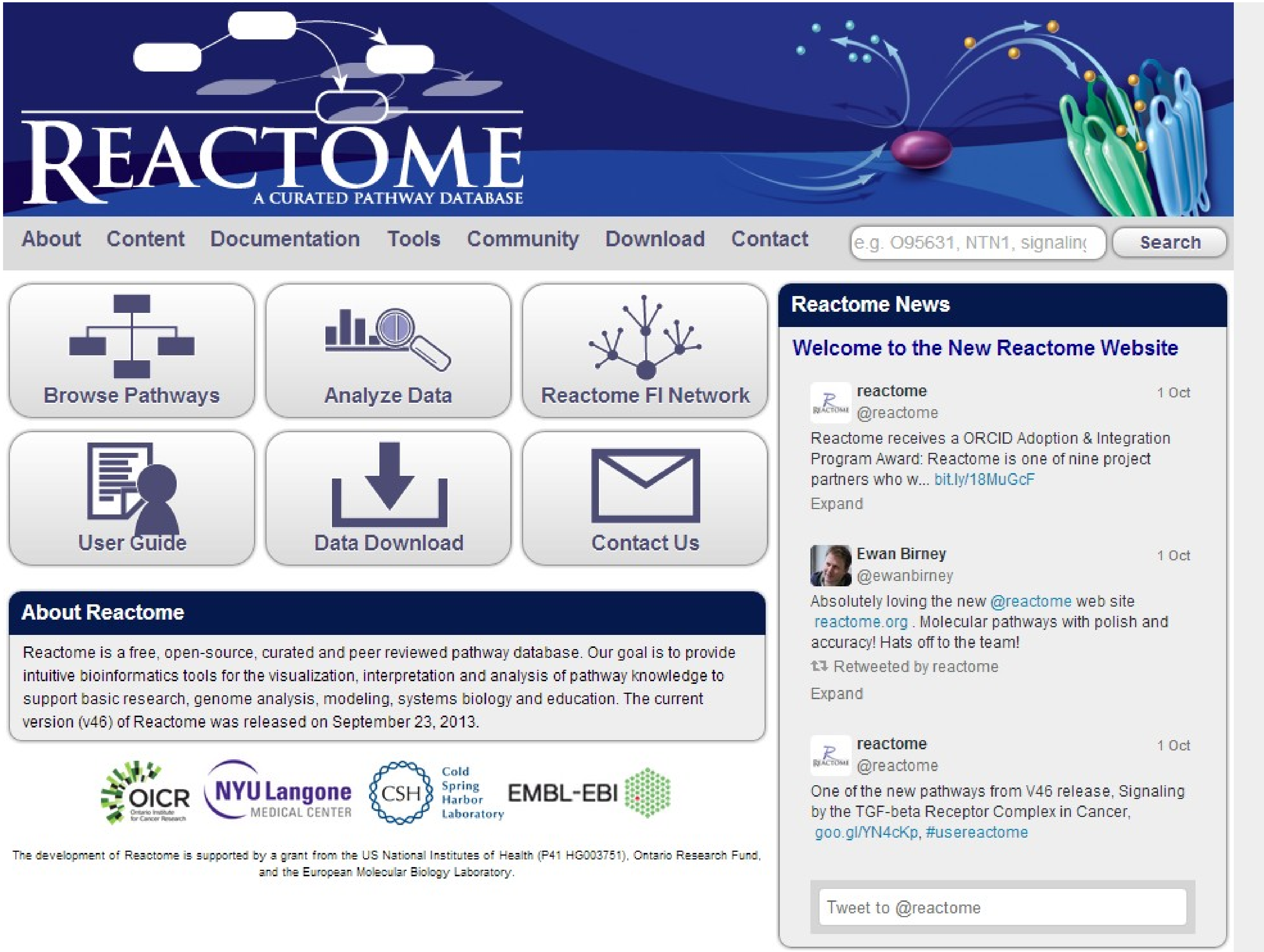
KEGG Organisms Enter org code(s) ▭ Go hsa hsa eco

Analysis tools

KEGG Mapper	KEGG PATHWAY/BRITE/MODULE mapping tools
KEGG Atlas	Navigation tool to explore KEGG global maps
KAAS	KEGG automatic annotation server
BLAST/FASTA	Sequence similarity search
SIMCOMP	Chemical structure similarity search
PathPred	Biodegradation/biosynthesis pathway prediction


Tryptophan metabolism - Reference pathway
Pathway menu | Organism menu | Pathway entry | User data mapping]
Reference pathway
Go
100%
TRYPTOPHAN METABOLISM
5-Hydroxyindole-acetaldehyde
1.2.1.3
1.2.3.1
5-Hydroxy-indoleacetate
2.1.1.4
5-Methoxy-indoleacetate
6.3.2.-
5-Hydroxyindole-acetylglycine
5-Hydroxy-indolepyruvate
1.4.3.4
2.1.1.-
5-Methoxytryptamine
2.6.1.27
1.13.11.52
Formyl-5-hydroxy-kynurenamine
4,6-Dihydroxy-quinoline
1.4.3.4
5-Hydroxy-kynurenamine
4.1.1.28
2.1.1.49
N-Methylserotonin
1.13.11.52
Formyl-N-acetyl-5-methoxykynurenamine
5-Hydroxy-N-formylkynurenine
5-Hydroxy-L-tryptophan
6-Hydroxy-kynurenate
2.6.1.-
3.5.1.9
1.13.11.52
4.1.1.28
2.3.1.87
2.1.1.4
1.14.14.1
6-Hydroxymelatonin
4-(2-Amino-5-hydroxyphenyl)-2,4-dioxobutanoate
5-Hydroxy-kynurenine
Serotonin
N-Acetylserotonin
Melatonin
1.14.16.4
2.1.1.47
1.1.1.110
Indolelactate
3-Indoleglycol-aldehyde
1.4.3.2
3-Methyl-indolepyruvate
4.1.1.74
Indole-3-acetaldehyde
1.1.1.190
1.13.99.3
2.6.1.27
1.1.1.191
Indole-3-ethanol
N-Methyl tryptamine
2.6.1.99
Indolepyruvate
4.1.1.43
2.1.1.49
1.4.3.22
1.4.3.4
Indole-3-acetaldoxime
Acetylindoxyl
Indole
Tryptophan
N-Acetylisatin
1.7.3.2
1.14.16.-
4.1.99.1
4.1.1.28
1.14.13.-
2.5.1.-
1.7.3.-
1.14.13.168
Indoxyl
Tryptamine
N-Hydroxy-tryptamine
1.13.11.17
1.14.13.125
Indole-3-thiohydroximate
4.4.1.-
1.14.-.-
1.2.1.3
1.2.3.7
2-Formylamino-benzaldehyde
S-(indolylaceto-hydroximoyl)-L-cysteine
4.99.1.6
2.4.1.195
Indole-3-acetonitrile
Indoleacetate
5-(2'-Carboxyethyl)-4,6-dihydroxypicolinate
5-(2'-Formylethyl)-4,6-dihydroxypicolinate
1.2.1.-
Tryptophan biosynthesis
2.8.2.-
3.2.1.147
3.5.5.1
Indolylmethyldesulfo-glucosinolate
Glucobrassicin
4.2.1.84
3-Methyl-dioxyindole
5-(3'-Carboxy-3'-oxopropenyl)-4,6-dihydroxypicolinate
1.3.1.-
5-(3'-Carboxy-3'-oxopropyl)-4,6-dihydroxypicolinate
1.13.12.3
3.5.1.4
1.13.11.-
Indole-3-acetamide
1.13.11.11
1.13.11.52
Tryptophan biosynthesis
2.5.1.-
1.13.11.10
N-Formyl-kynurenine
3.7.1.3
Formyl-anthranilate
Indole
2-Formamino-benzoylacetate
3.5.1.9
3.5.1.9
Aminobenzoate degradation
Kynurenate
L-Kynurenine
1.3.1.18
1.14.99.2
2.6.1.7
3.7.1.3
Anthranilate
7,8-Dihydroxy-kynurenate
7,8-Dihydro-7,8-dihydroxykynurenate
4-(2-Aminophenyl)-2,4-dioxobutanoate
FADH2
1.13.11.23
2,3-Dihydroxyindole
1.14.13.9
1.5.1.45
1.14.14.8
1.14.16.3
1.11.1.6
3-Hydroxy-L-kynurenine
FAD
1.11.1.21
Cinnavalininate
Xanthurenate
3-Hydroxy-anthranilate
8-Methoxy-kynurenate
2.1.1.-
2.6.1.7
3.7.1.3
2-Aminophenol
1.10.3.4
Isophenoxazine
4-(2-Amino-3-hydroxyphenyl)-2,4-dioxobutanoate
4.1.1.-
2.1.1.-
3-Methoxy-anthranilate
1.13.11.6
2-Amino-3-carboxymuconate semialdehyde
Quinolinate
Nicotinamide metabolism
4,8-Dihydroxy-quinoline
1.4.3.4
3-Hydroxy-kynurenamine
4.1.1.45
Glycolysis
2-Aminomuconate semialdehyde
1.2.1.32
γ-Oxohex-crotonate
Acetoacetyl-CoA
Crotonoyl-CoA
2.3.1.9
1.1.1.35
4.2.1.17
1.3.8.6
1.2.4.2
1.5.1.-
3.5.99.5
Benzoate degradation
Acetyl-CoA
(S)-3-Hydroxy-butanoyl-CoA
Glutaryl-CoA
2-Oxoadipate
2-Amino-muconate

http://www.reactome.org/



About Reactome

Reactome is a free, open-source, curated and peer reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education. The current version (v46) of Reactome was released on September 23, 2013.

The development of Reactome is supported by a grant from the US National Institutes of Health (P41 HG003751), Ontario Research Fund, and the European Molecular Biology Laboratory.

Reactome News

Welcome to the New Reactome Website

reactome @reactome 1 Oct

Reactome receives a ORCID Adoption & Integration Program Award: Reactome is one of nine project partners who w... bit.ly/18MuGcF

Expand

Ewan Birney @ewanbirney 1 Oct

Absolutely loving the new @reactome web site reactome.org . Molecular pathways with polish and accuracy! Hats off to the team!

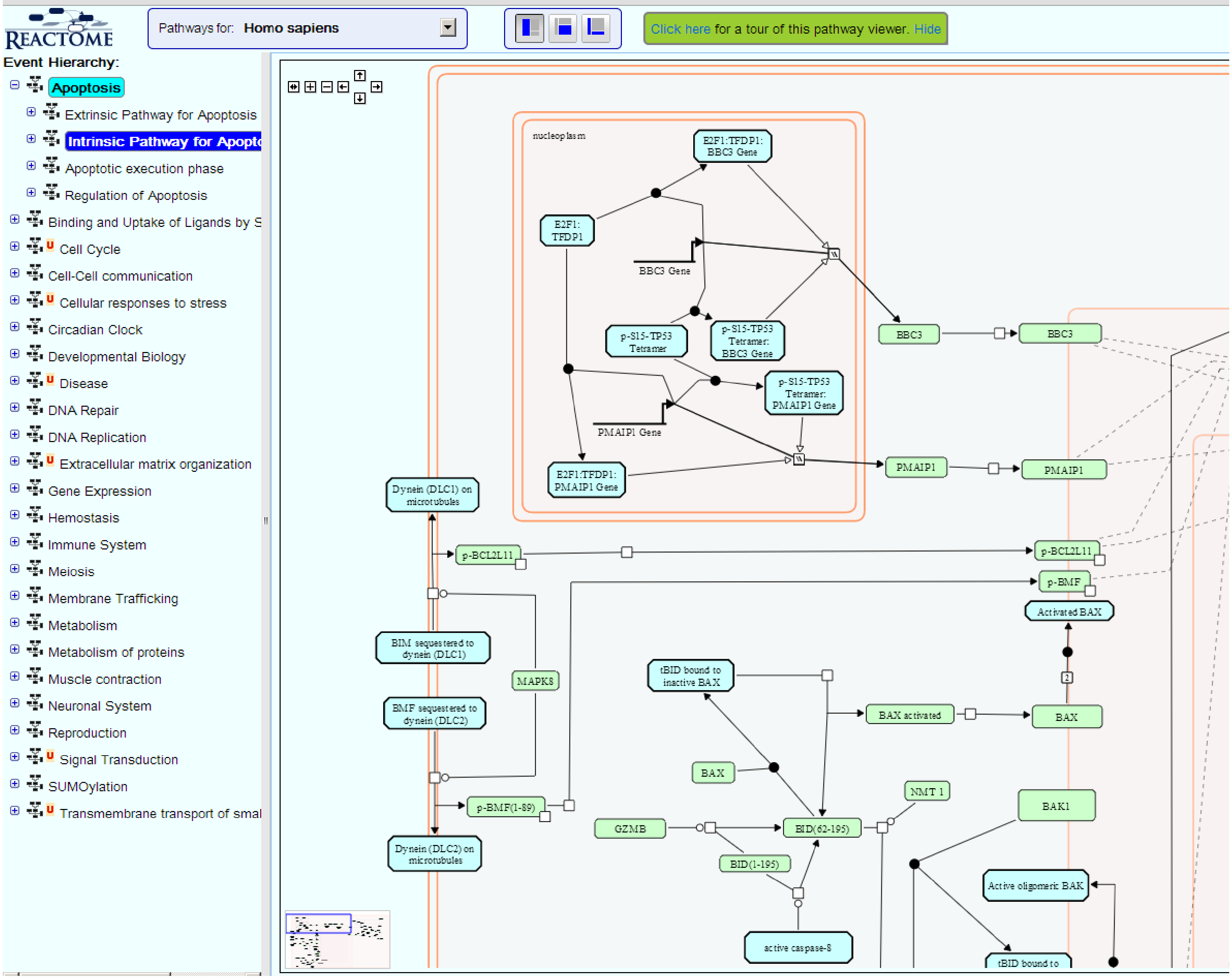
Retweeted by reactome

Expand

reactome @reactome 1 Oct

One of the new pathways from V46 release, Signaling by the TGF-beta Receptor Complex in Cancer, goo.gl/YN4cKp, #usereactome

Tweet to @reactome



Pathways for: Homo sapiens

Click here for a tour of this pathway viewer. Hide

Event Hierarchy:

- Apoptosis
  - Extrinsic Pathway for Apoptosis
  - Intrinsic Pathway for Apopto
  - Apoptotic execution phase
  - Regulation of Apoptosis
- Binding and Uptake of Ligands by S
- Cell Cycle
- Cell-Cell communication
- Cellular responses to stress
- Circadian Clock
- Developmental Biology
- Disease
- DNA Repair
- DNA Replication
- Extracellular matrix organization
- Gene Expression
- Hemostasis
- Immune System
- Meiosis
- Membrane Trafficking
- Metabolism
- Metabolism of proteins
- Muscle contraction
- Neuronal System
- Reproduction
- Signal Transduction
- SUMOylation
- Transmembrane transport of smal

Métodos para análise de enriquecimento de categorias gênicas

Programas para análises de enriquecimento de categorias funcionais

- DAVID (http://david.abcc.ncifcrf.gov/)
- G:Profiler (http://biit.cs.ut.ee/gprofiler/)
- GSEA (Gene Set Enrichment Analysis - www.broadinstitute.org/gsea/)
- Ingenuity Pathway Analysis (Comercial)

Identificação de categorias enriquecidas entre genes de interesse

- parte de uma lista de genes selecionada com algum critério (expressão diferencial, abundância, outros)
- utiliza conhecimento *a priori* (ex. GO, vias moleculares, anotações funcionais, outras…)
- Testa a probabibilidade de uma determinada categoria estar sobre-representada na lista de genes selecionada em relação ao universo de genes.
- assume uma distribuição hipergeométrica (= teste exato de Fisher (chi-quadrado) mono-caudal)

	Genes selecionados	Total de genes
Pertencem a categoria X	10 (k)	70 (K)
Não pertencem a categoria X	90 (n - k)	930 (N – K)
Total	100 (n)	1000 (N)

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

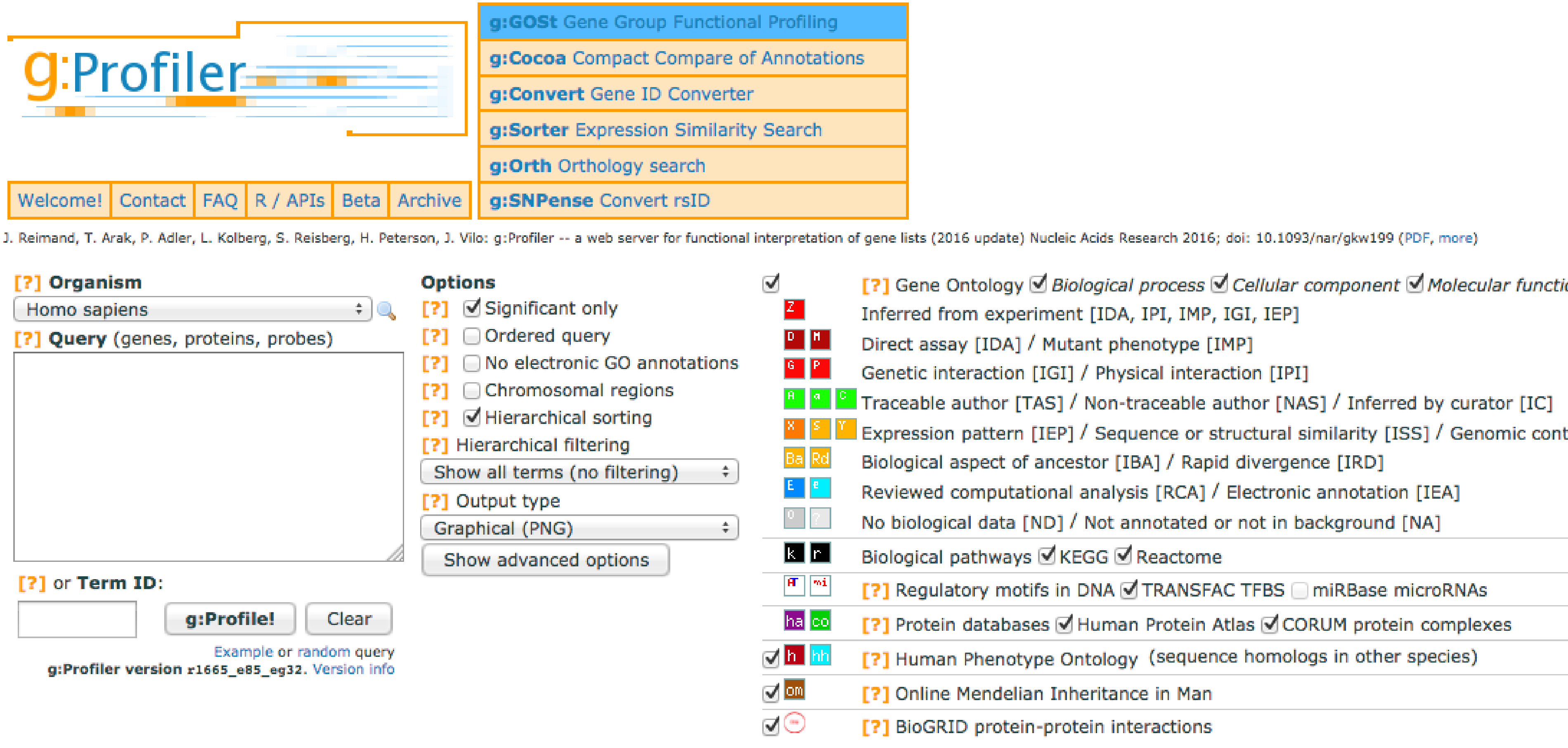
Testar se a frequência de genes da categoria x na lista selecionada (10/100; 10%) é maior que o esperado ao acaso (70/1000; 7%)

Passos em uma análise de enriquecimento de categoria gênica

- Definir a lista de genes de interesse e o universo de genes avaliados ("background”). Nota: No caso do RNAseq ou outra análise global, o total de genes anotados pode ser usado como referência
- Selecionar os genes de interesse para verificar o enriquecimento de termos/categorias (ex. DEGs)
- Executar teste de enriquecimento com correção para testes múltiplos (ex. Bonferroni, Benjamini-Hochberg) para controlar o número de falsos -positivos.

g:Profiler
a web server for functional interpretation of gene lists

http://biit.cs.ut.ee/gprofiler/



g:GOSt Gene Group Functional Profiling

g:Cocoa Compact Compare of Annotations

g:Convert Gene ID Converter

g:Sorter Expression Similarity Search

g:Orth Orthology search

g:SNPense Convert rsID

Welcome! | Contact | FAQ | R / APIs | Beta | Archive

J. Reimand, T. Arak, P. Adler, L. Kolberg, S. Reisberg, H. Peterson, J. Vilo: g:Profiler -- a web server for functional interpretation of gene lists (2016 update) Nucleic Acids Research 2016; doi: 10.1093/nar/gkw199 (PDF, more)

[?] **Organism**

Homo sapiens

[?] **Query** (genes, proteins, probes)

[?] or **Term ID**:

g:Profile! | Clear

Example or random query

g:Profiler version r1665_e85_eg32. Version info

Options

[?] ☑ Significant only

[?] ☐ Ordered query

[?] ☐ No electronic GO annotations

[?] ☐ Chromosomal regions

[?] ☑ Hierarchical sorting

[?] Hierarchical filtering

Show all terms (no filtering)

[?] Output type

Graphical (PNG)

Show advanced options

☑ [?] Gene Ontology ☑ *Biological process* ☑ *Cellular component* ☑ *Molecular functi*

Z Inferred from experiment [IDA, IPI, IMP, IGI, IEP]

D M Direct assay [IDA] / Mutant phenotype [IMP]

G P Genetic interaction [IGI] / Physical interaction [IPI]

A a C Traceable author [TAS] / Non-traceable author [NAS] / Inferred by curator [IC]

X S Y Expression pattern [IEP] / Sequence or structural similarity [ISS] / Genomic cont

Ba Rd Biological aspect of ancestor [IBA] / Rapid divergence [IRD]

E e Reviewed computational analysis [RCA] / Electronic annotation [IEA]

0 ? No biological data [ND] / Not annotated or not in background [NA]

k r Biological pathways ☑ KEGG ☑ Reactome

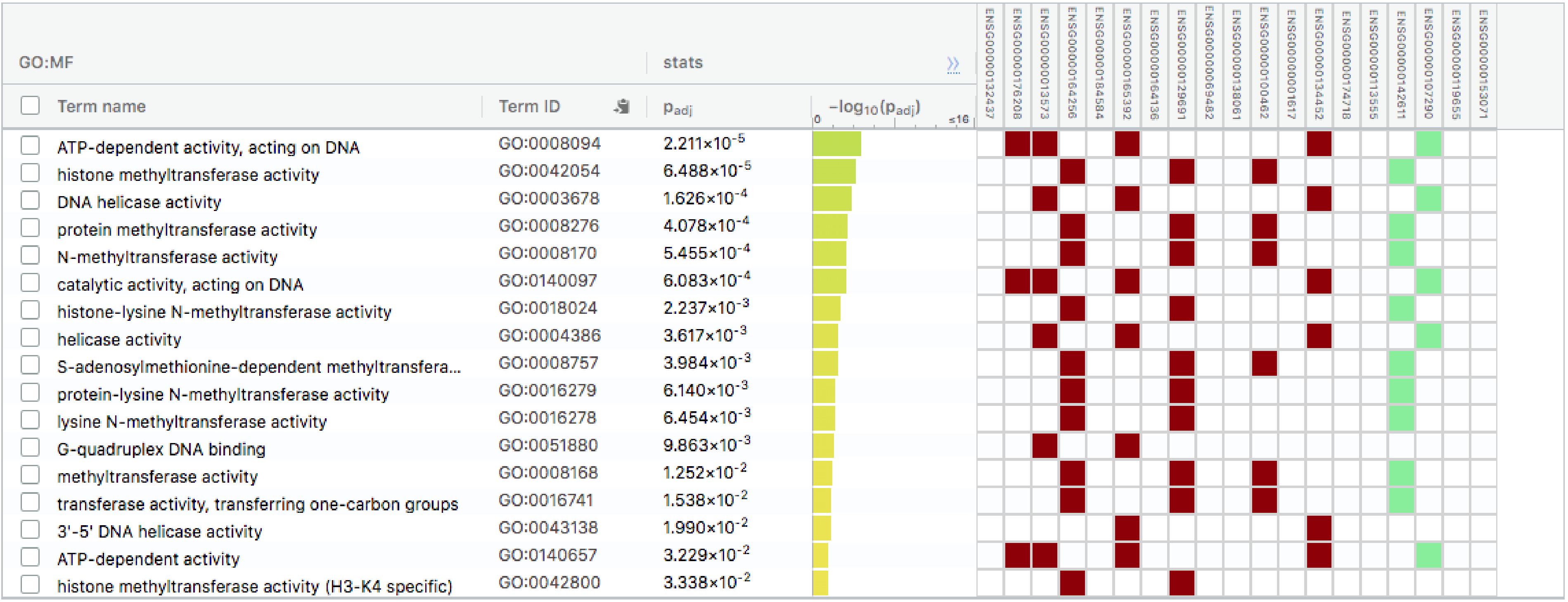
ᴬT ᵐi [?] Regulatory motifs in DNA ☑ TRANSFAC TFBS ☐ miRBase microRNAs

ha co [?] Protein databases ☑ Human Protein Atlas ☑ CORUM protein complexes

☑ h hh [?] Human Phenotype Ontology (sequence homologs in other species)

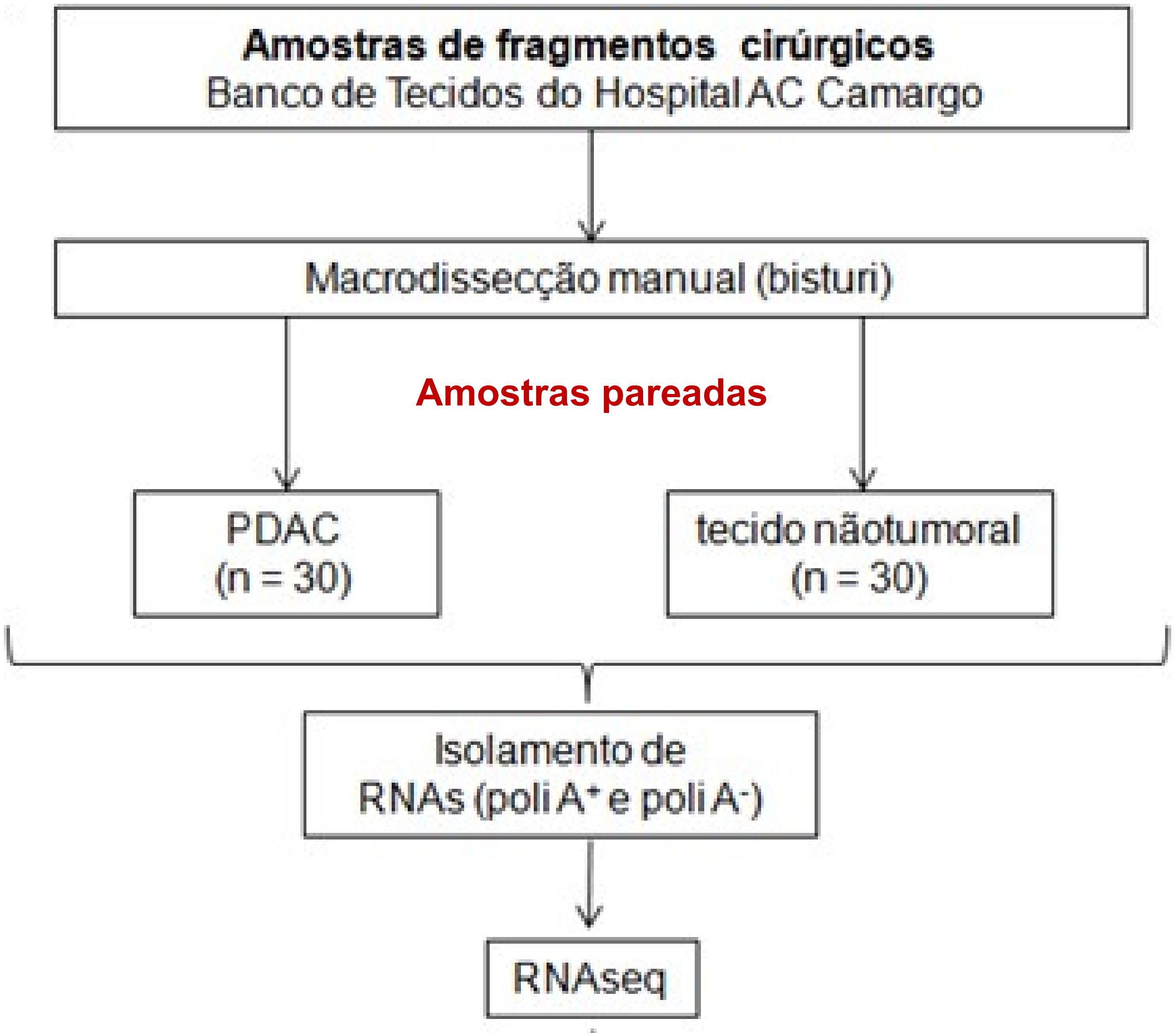
☑ om [?] Online Mendelian Inheritance in Man

☑ ○ [?] BioGRID protein-protein interactions



GO:MF		stats																						
☐	Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16	ENSG00000132437	ENSG00000176208	ENSG00000013573	ENSG00000164256	ENSG00000184584	ENSG00000165392	ENSG00000164136	ENSG00000129691	ENSG00000069482	ENSG00000138061	ENSG00000100462	ENSG00000001617	ENSG00000134452	ENSG00000174718	ENSG00000113555	ENSG00000142611	ENSG00000107290	ENSG00000119655	ENSG00000153071
☐	ATP-dependent activity, acting on DNA	GO:0008094	2.211×10^{-5}			🟥	🟥			🟥							🟥				🟩		
☐	histone methyltransferase activity	GO:0042054	6.488×10^{-5}					🟥				🟥			🟥					🟩			
☐	DNA helicase activity	GO:0003678	1.626×10^{-4}				🟥			🟥							🟥				🟩		
☐	protein methyltransferase activity	GO:0008276	4.078×10^{-4}					🟥				🟥			🟥					🟩			
☐	N-methyltransferase activity	GO:0008170	5.455×10^{-4}					🟥				🟥			🟥					🟩			
☐	catalytic activity, acting on DNA	GO:0140097	6.083×10^{-4}			🟥	🟥			🟥							🟥				🟩		
☐	histone-lysine N-methyltransferase activity	GO:0018024	2.237×10^{-3}					🟥				🟥								🟩			
☐	helicase activity	GO:0004386	3.617×10^{-3}				🟥			🟥							🟥				🟩		
☐	S-adenosylmethionine-dependent methyltransfera...	GO:0008757	3.984×10^{-3}					🟥				🟥			🟥					🟩			
☐	protein-lysine N-methyltransferase activity	GO:0016279	6.140×10^{-3}					🟥				🟥								🟩			
☐	lysine N-methyltransferase activity	GO:0016278	6.454×10^{-3}					🟥				🟥								🟩			
☐	G-quadruplex DNA binding	GO:0051880	9.863×10^{-3}				🟥			🟥													
☐	methyltransferase activity	GO:0008168	1.252×10^{-2}					🟥				🟥			🟥					🟩			
☐	transferase activity, transferring one-carbon groups	GO:0016741	1.538×10^{-2}					🟥				🟥			🟥					🟩			
☐	3'-5' DNA helicase activity	GO:0043138	1.990×10^{-2}							🟥							🟥						
☐	ATP-dependent activity	GO:0140657	3.229×10^{-2}			🟥	🟥			🟥							🟥				🟩		
☐	histone methyltransferase activity (H3-K4 specific)	GO:0042800	3.338×10^{-2}					🟥				🟥											
]

Comparação do transcritoma de tumores de pâncreas com tecido não tumoral por RNAseq



Amostras de fragmentos cirúrgicos
Banco de Tecidos do Hospital AC Camargo

Macrodissecção manual (bisturi)

Amostras pareadas

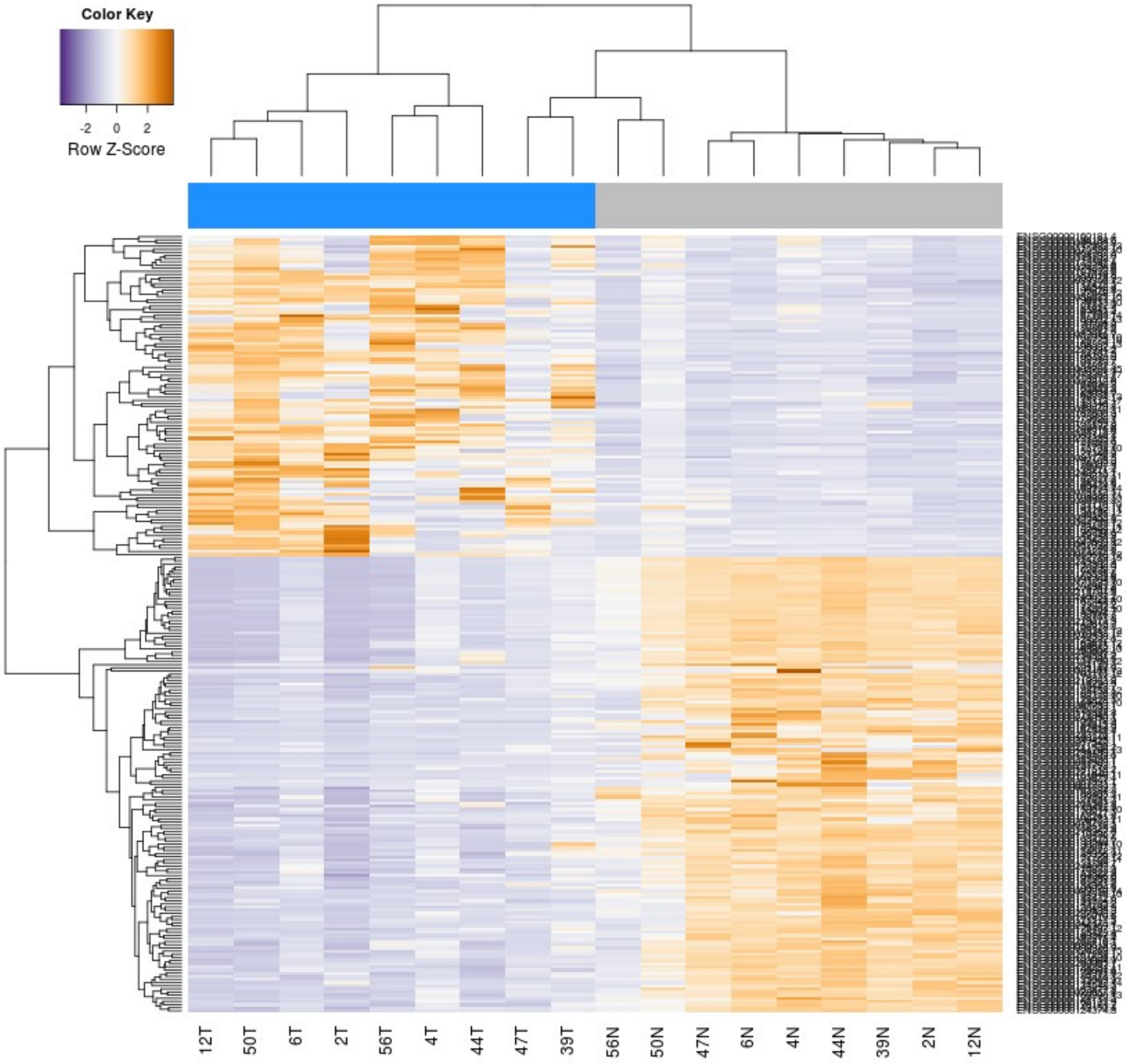
PDAC
(n = 30)

tecido nãotumoral
(n = 30)

Isolamento de
RNAs (poli A^+ e poli A^-)

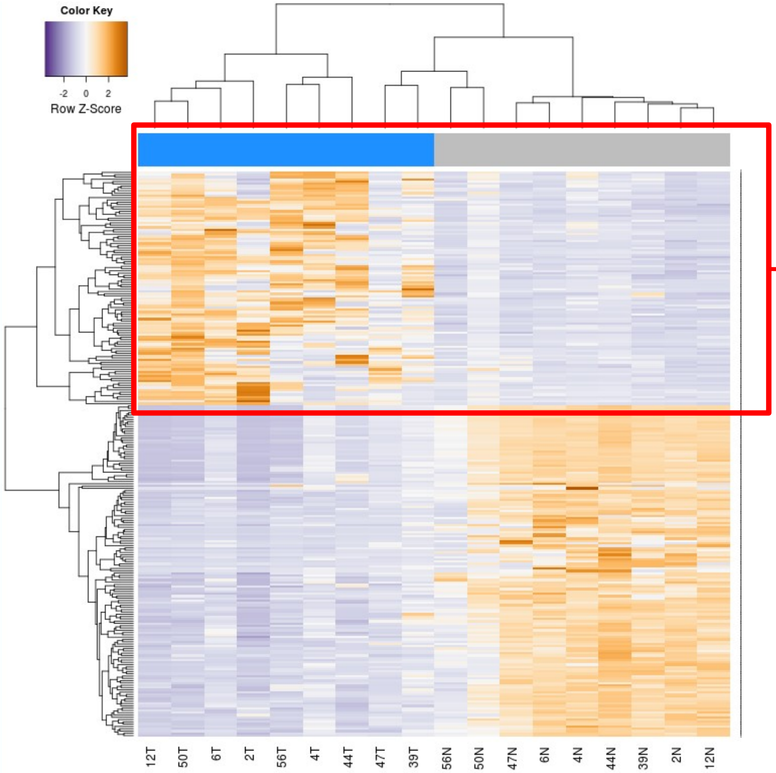
RNAseq

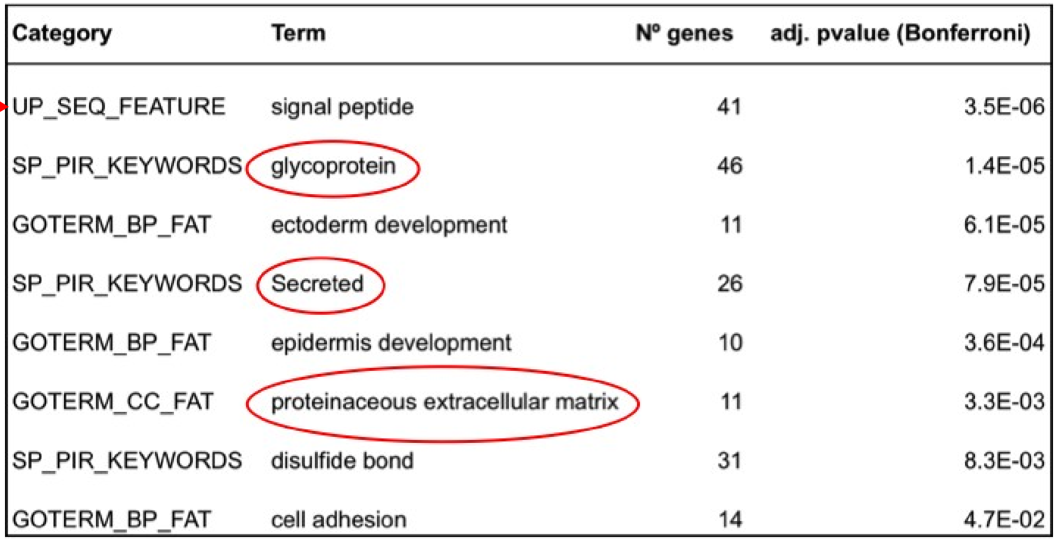
398 genes codificadores de proteína (GENCODE v.22) diferencialmente expressos no PDAC (padj < 0.001, FC > |10|)



Genes com expressão aumentada no câncer de pâncreas estão enriquecidos em proteínas com potencial para biomarcador de diagnóstico

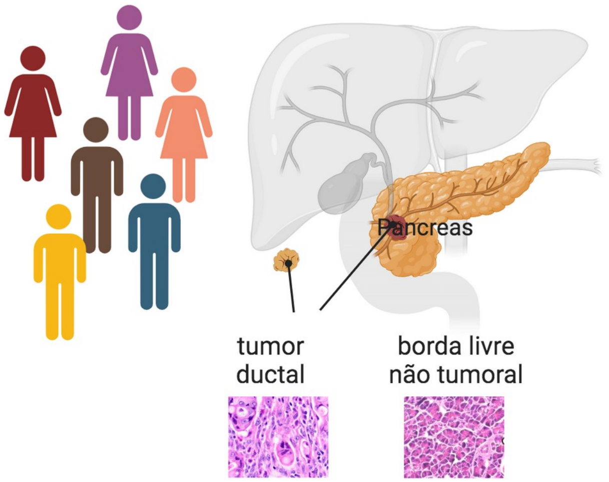
398 genes mRNAs codificadores de proteína
(padj < 0.001, FC > |10|)



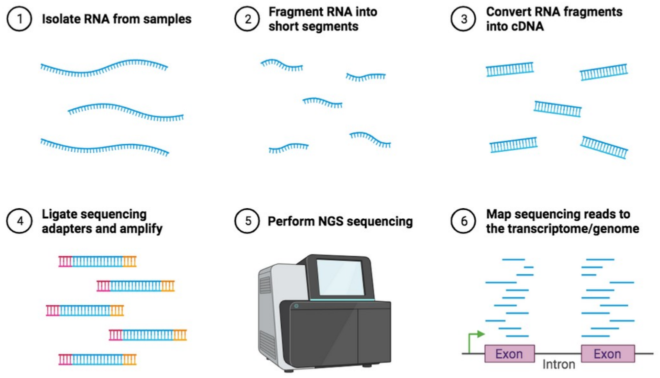


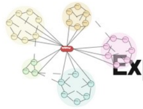
Category	Term	Nº genes	adj. pvalue (Bonferroni)
UP_SEQ_FEATURE	signal peptide	41	3.5E-06
SP_PIR_KEYWORDS	glycoprotein	46	1.4E-05
GOTERM_BP_FAT	ectoderm development	11	6.1E-05
SP_PIR_KEYWORDS	Secreted	26	7.9E-05
GOTERM_BP_FAT	epidermis development	10	3.6E-04
GOTERM_CC_FAT	proteinaceous extracellular matrix	11	3.3E-03
SP_PIR_KEYWORDS	disulfide bond	31	8.3E-03
GOTERM_BP_FAT	cell adhesion	14	4.7E-02

Tutorial enriquecimento de categorias gênicas - gProfiler



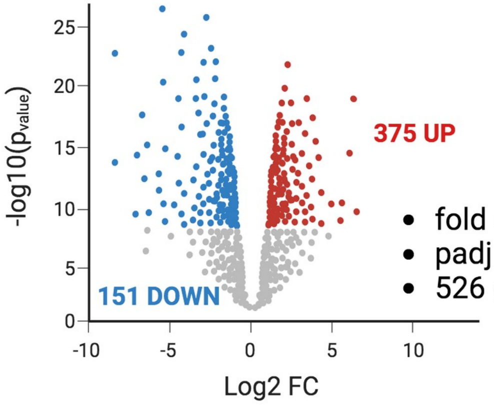
RNA-Seq


1 Isolate RNA from samples
2 Fragment RNA into short segments
3 Convert RNA fragments into cDNA
4 Ligate sequencing adapters and amplify
5 Perform NGS sequencing
6 Map sequencing reads to the transcriptome/genome
Exon
Intron
Exon



Expressão gênica diferencial

Volcano plot



- fold change > 2x
- padj < 10 -5
- 526 genes

Analise de enriquecimento de categorias gênicas

Vias moleculares, processos biológicos ativadas ou reprimidas em tumores

Identificação de vulnerabilidades e alvos para terapia

Tutorial - gProfiler

Analisar lista de genes diferencialmente expressos em tumores de pâncreas Identificados através de RNAseq (Paixão et al., Cellular Oncology 2022).
Critérios de seleção: razão exoressão Tumor / nãotumor > |2x| (>= |1| log2|) e padj < 0,00001

Lista disponível na pagina da disciplina: DEGs – tumor de pâncreas. Abrir a planilha (Excel, csv).
Na coluna 1, selecionar genes com expressão aumentada ou diminuida nos tumores.
No gProfiler, analise separadamente os genes aumentados (razão > 1 log2) e os genes diminuídos (razão < -1 log2) nos tumores.

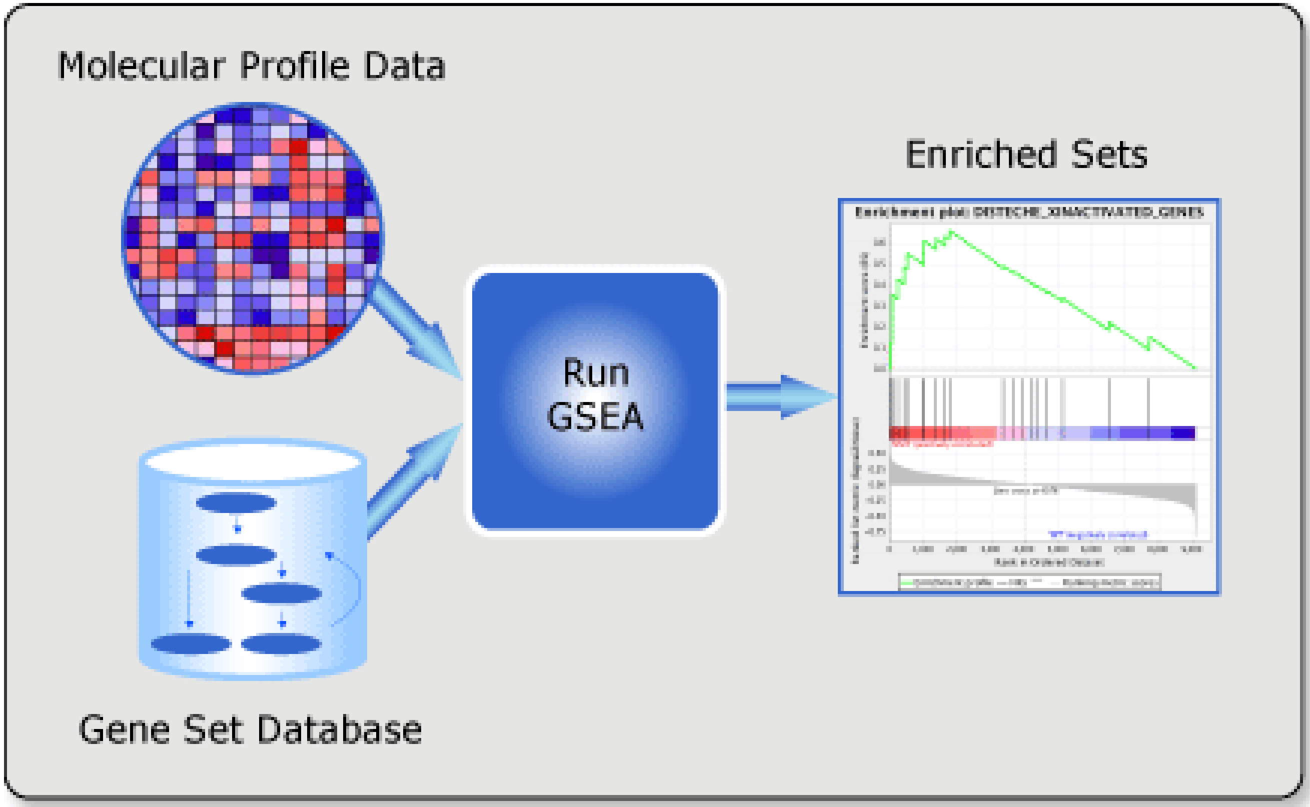
Investigue se existem termos enriquecidos (padj < 0.05) entre genes com expressão aumentada ou diminuída. Utilize diferentes ontologias:

- GO Processos biológicos, Funções moleculares, Componente celular
- KEGG
- BioCarta
- Outros

- Reporte no relatório uma tabela com o nome e estatísticas das 5 categorias mais significativas (ou que você considere mais relevante no contexto do câncer). Pode ser uma categoria de cada ontologia.

“Gene Set Enrichment Analysis”

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



- Estratégia alternativa que parte de uma lista genes ranqueada em função do fenótipo de interesse (expressão gênica, outros).
- Evita a utilização de um critério arbitrário na seleção dos genes de interesse. Ex. genes diferencialmente expressos X vezes

O que são “gene sets” ?

Conjuntos de genes definidos a partir de conhecimento biológico prévio

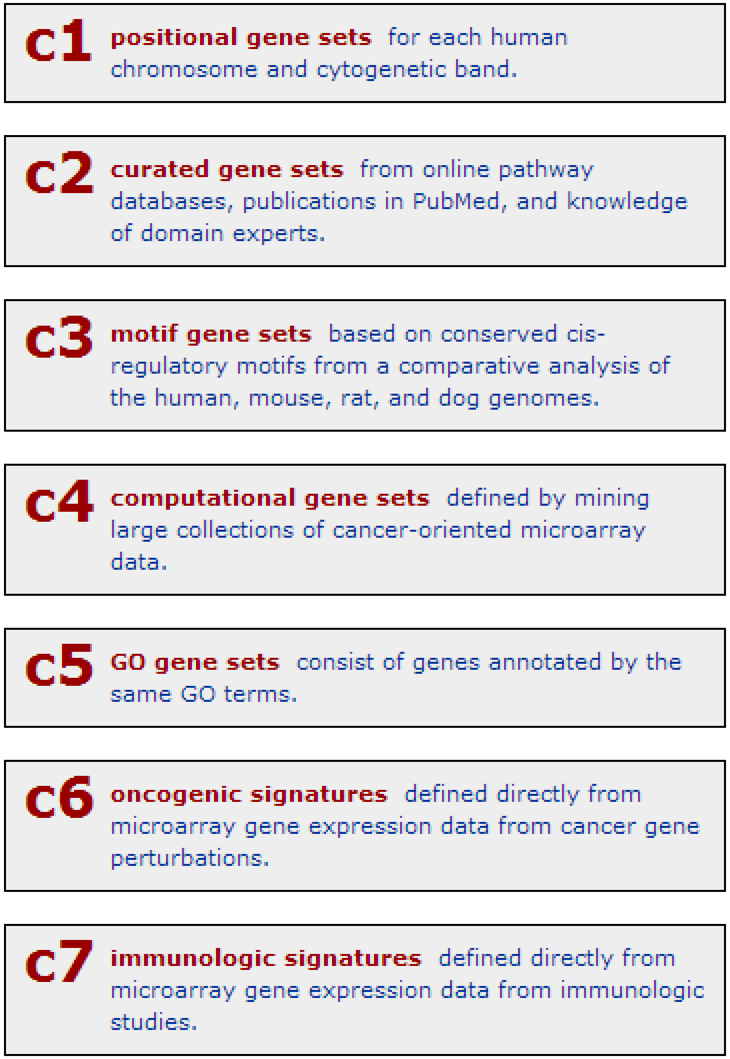
Ex.:

- Publicações científicas sobre vias bioquímicas
- Padrões de co-expressão observados em experimentos prévios

O programa GSEA pode usar conjuntos curados de “gene sets” disponíveis publicamente, ou fornecidos pelo usuário

“Molecular Signatures Database” - Conjunto curado de gene sets

http://www.broadinstitute.org/gsea/msigdb/index.jsp



c1 **positional gene sets** for each human chromosome and cytogenetic band.

c2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

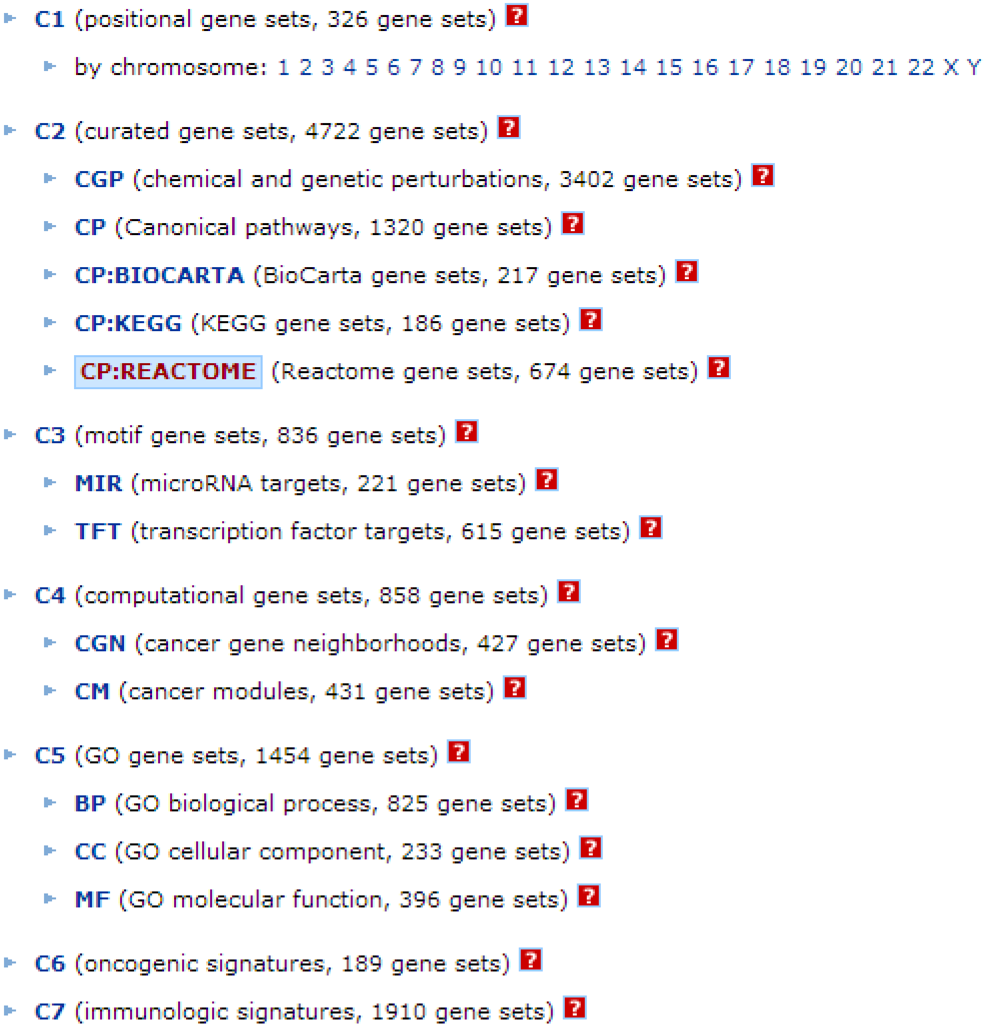
c3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

c4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

c5 **GO gene sets** consist of genes annotated by the same GO terms.

c6 **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.

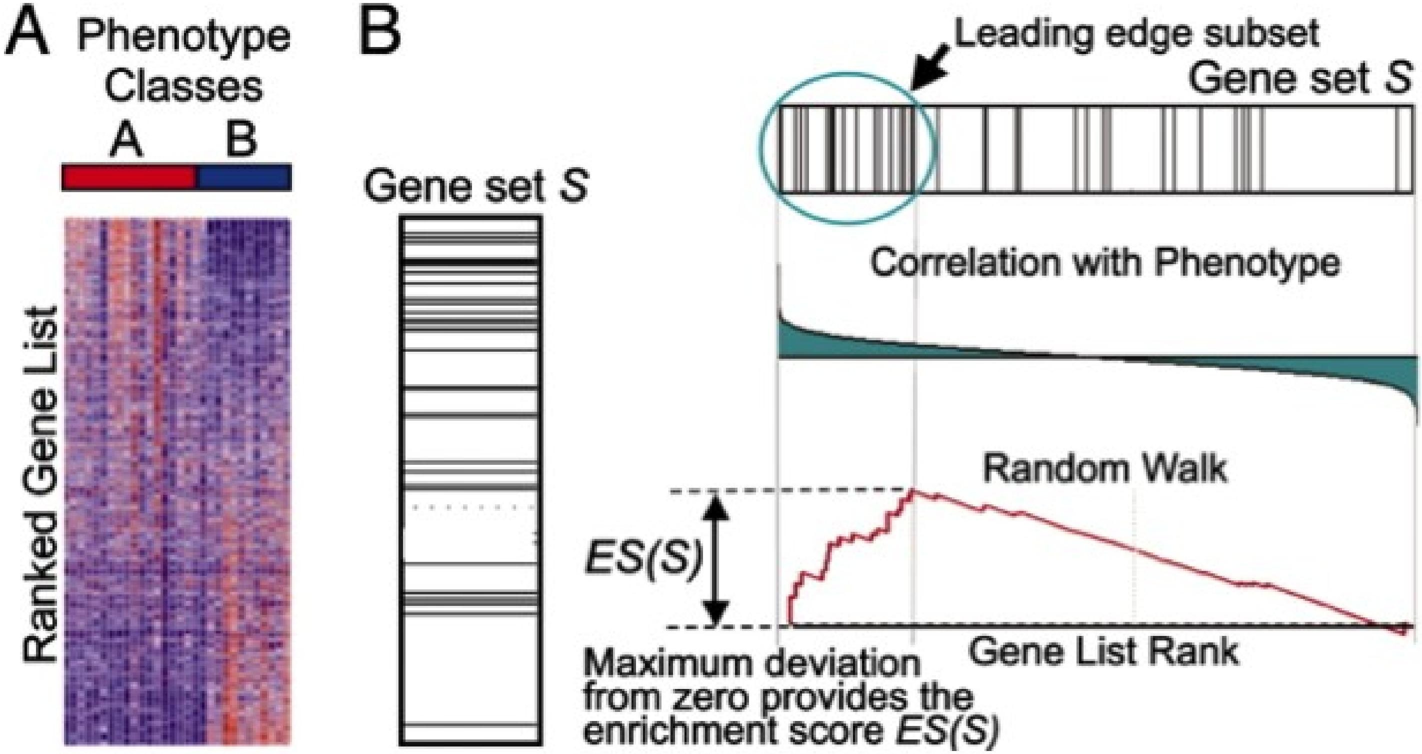
c7 **immunologic signatures** defined directly from microarray gene expression data from immunologic studies.



- **C1** (positional gene sets, 326 gene sets) ?
  - by chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
- **C2** (curated gene sets, 4722 gene sets) ?
  - **CGP** (chemical and genetic perturbations, 3402 gene sets) ?
  - **CP** (Canonical pathways, 1320 gene sets) ?
  - **CP:BIOCARTA** (BioCarta gene sets, 217 gene sets) ?
  - **CP:KEGG** (KEGG gene sets, 186 gene sets) ?
  - **CP:REACTOME** (Reactome gene sets, 674 gene sets) ?
- **C3** (motif gene sets, 836 gene sets) ?
  - **MIR** (microRNA targets, 221 gene sets) ?
  - **TFT** (transcription factor targets, 615 gene sets) ?
- **C4** (computational gene sets, 858 gene sets) ?
  - **CGN** (cancer gene neighborhoods, 427 gene sets) ?
  - **CM** (cancer modules, 431 gene sets) ?
- **C5** (GO gene sets, 1454 gene sets) ?
  - **BP** (GO biological process, 825 gene sets) ?
  - **CC** (GO cellular component, 233 gene sets) ?
  - **MF** (GO molecular function, 396 gene sets) ?
- **C6** (oncogenic signatures, 189 gene sets) ?
- **C7** (immunologic signatures, 1910 gene sets) ?

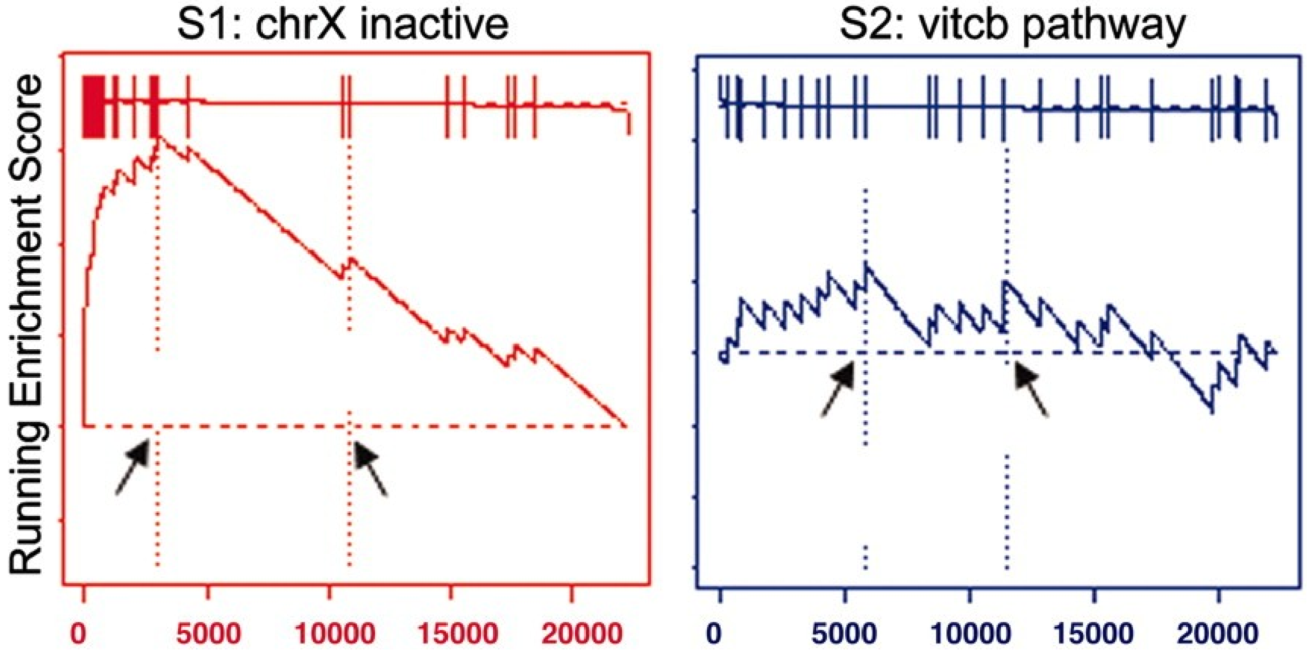
GSEA - Etapas na identificação de "gene sets" significativamente enriquecidos

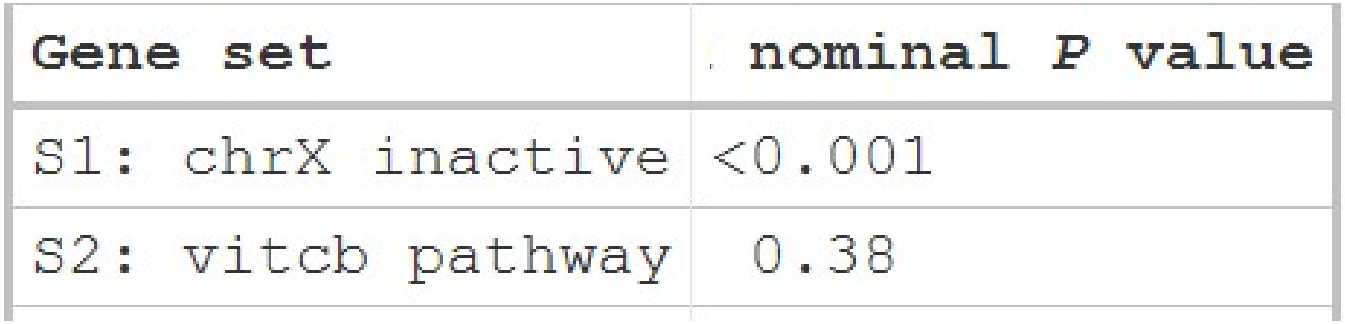
- Passo 1: Cálculo do valor de enriquecimento ("Enrichment Score" – ES)
- Passo 2: Estimativa da significância estatística de ES (comparação com distribuição ao acaso)
- Passo 3: Correção para testes múltiplos ("False Discovery Rate")



Subramanian A et al. PNAS 2005;102:15545-15550

Exemplo de uso do GSEA: Identificação de “gene sets” relacionados com inativação do cromossomo X em listas de genes expressos em linhagens celulares de machos e fêmeas





Gene set	nominal *P* value
S1: chrX inactive	<0.001
S2: vitcb pathway	0.38

Subramanian A et al. PNAS 2005;102:15545-15550

Baixa sobreposição entre os genes com expressão correlacionada à sobrevida do paciente identificados em 3 estudos de câncer de pulmão

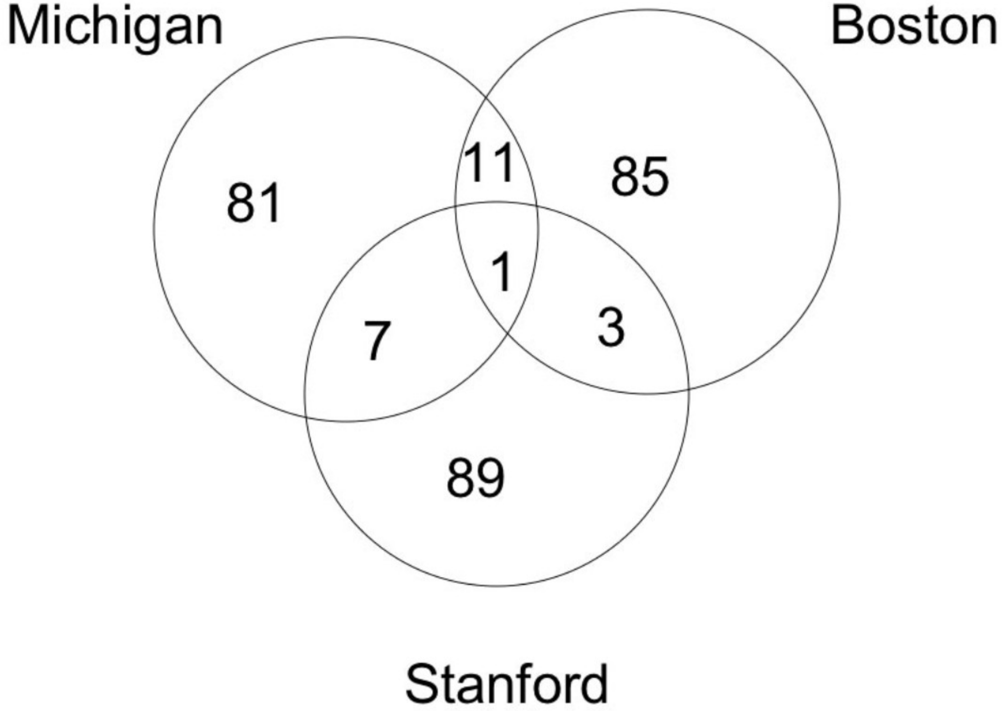
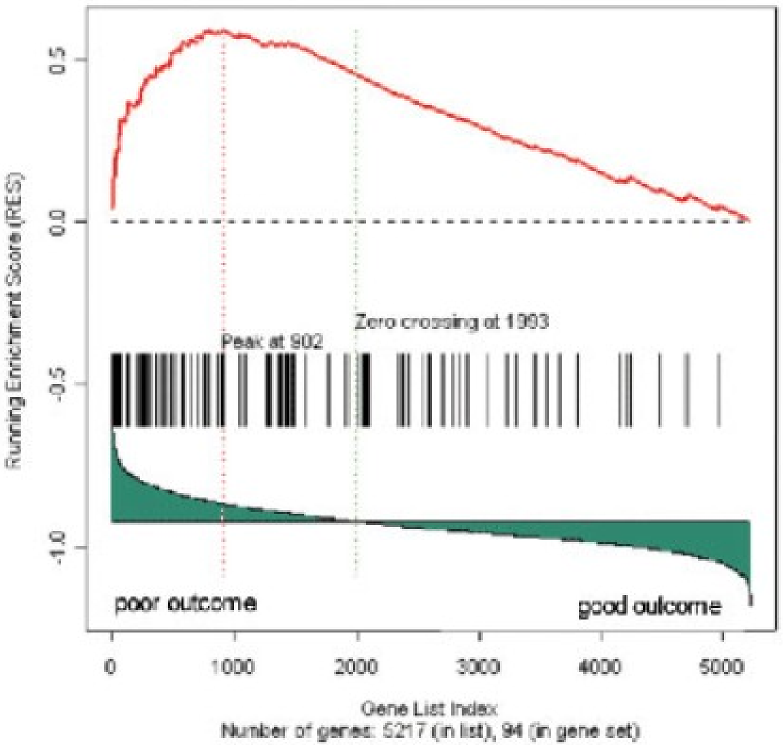


Fig. 5. Single gene overlaps in lung cancer studies. This Venn diagram shows the pairwise and three-way overlap between the top 100 genes correlated with poor outcome in the Michigan, Boston, and Stanford data sets. Pairwise overlap is determined by using genes that appear on the technology platforms of both studies. Three-way overlap is the overlap of the pairwise overlaps. Restricting to genes on all three platforms would reduce the gene space by 50% in the Michigan study and by 70% in the Boston and Stanford studies.

Subramanian A et al. PNAS 2005;102:15545-15550

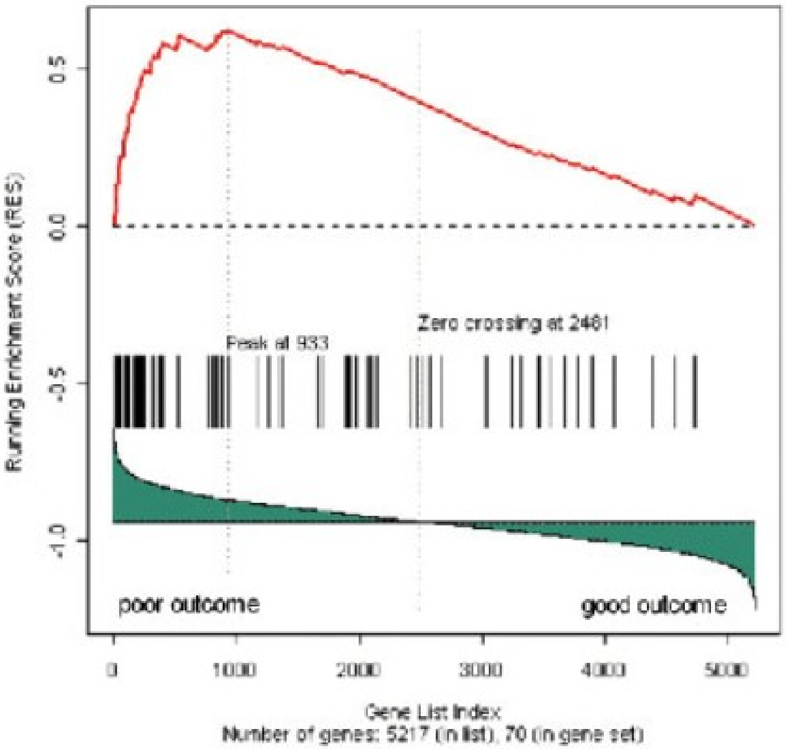
Boston Dataset

Gene Set: $S_{Michigan}$



Michigan Dataset

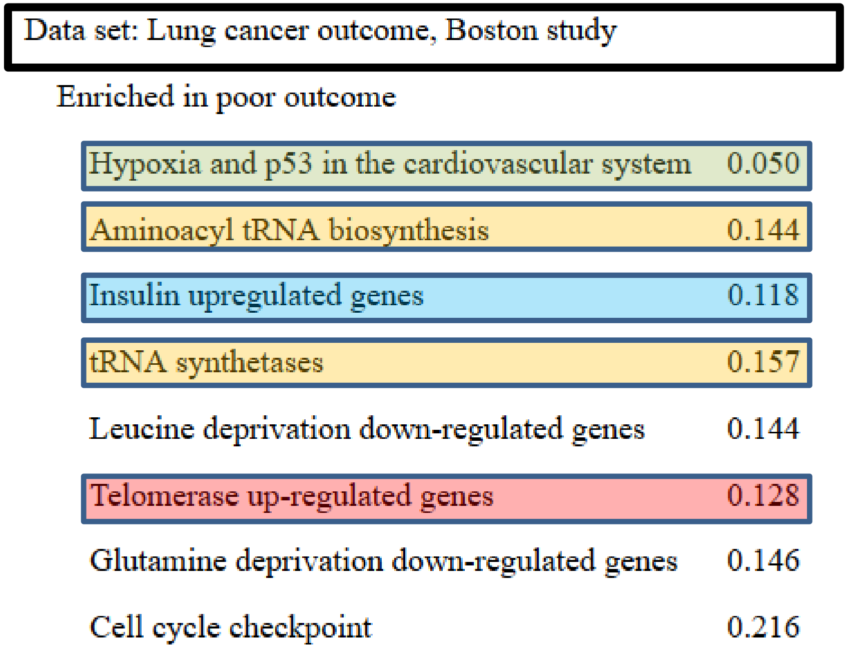
Gene Set: S_{Boston}



***P* < 0.001**

Subramanian A et al. PNAS 2005;102:15545-15550

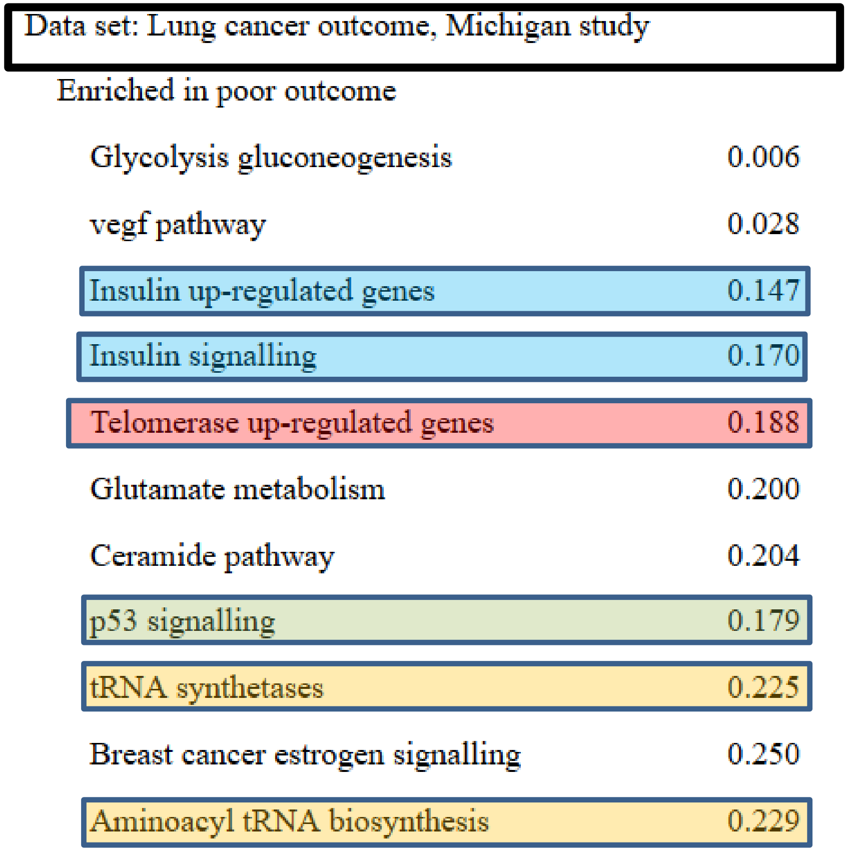
Alta sobreposição entre as vias correlacionadas à sobrevida do paciente nos diferentes estudos de câncer de pulmão



Data set: Lung cancer outcome, Boston study

Enriched in poor outcome

Hypoxia and p53 in the cardiovascular system	0.050
Aminoacyl tRNA biosynthesis	0.144
Insulin upregulated genes	0.118
tRNA synthetases	0.157
Leucine deprivation down-regulated genes	0.144
Telomerase up-regulated genes	0.128
Glutamine deprivation down-regulated genes	0.146
Cell cycle checkpoint	0.216



Data set: Lung cancer outcome, Michigan study

Enriched in poor outcome

Glycolysis gluconeogenesis	0.006
vegf pathway	0.028
Insulin up-regulated genes	0.147
Insulin signalling	0.170
Telomerase up-regulated genes	0.188
Glutamate metabolism	0.200
Ceramide pathway	0.204
p53 signalling	0.179
tRNA synthetases	0.225
Breast cancer estrogen signalling	0.250
Aminoacyl tRNA biosynthesis	0.229

FDR ≤ 0.25

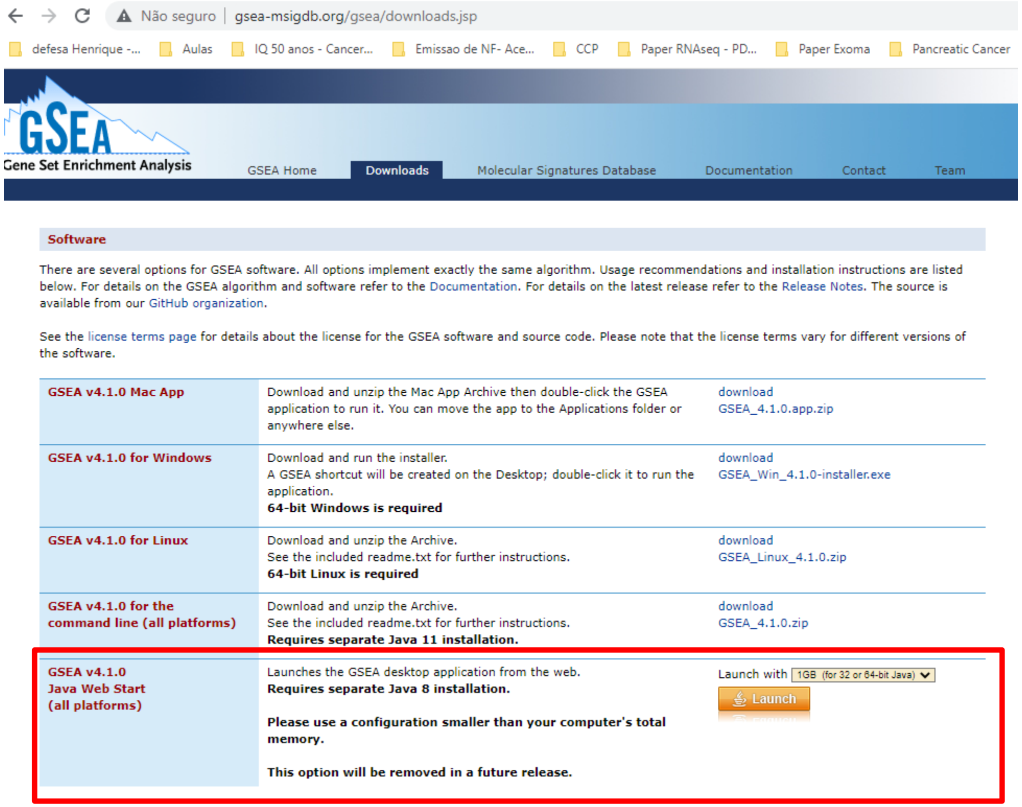
Tutorial - GSEA

- identificar “gene sets” com expressão aumentada (FDR < 25%) em pacientes com cancer de pulmão com pior prognóstico utilizando dados de expressão gênica gerados nos estudos de Boston e Michigan.

- verificar se existem “gene sets” em comum entre os dois estudos. Quais são eles?
- Escolher um “gene set” enriquecido nos dois estudos e verificar se existem genes diferencialmente expressos em comum. Rportar os resultados no relatório.

- o tutorial abaixo apresenta uma visão geral de como realizar análises utilizando o programa:
https://www.youtube.com/watch?v=KY6SS4vRchY

Baixar o programa no site http://www.gsea-msigdb.org/gsea/downloads.jsp


Não seguro | gsea-msigdb.org/gsea/downloads.jsp
defesa Henrique -... Aulas IQ 50 anos - Cancer... Emissao de NF- Ace... CCP Paper RNAseq - PD... Paper Exoma Pancreatic Cancer
GSEA
Gene Set Enrichment Analysis
GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

Software

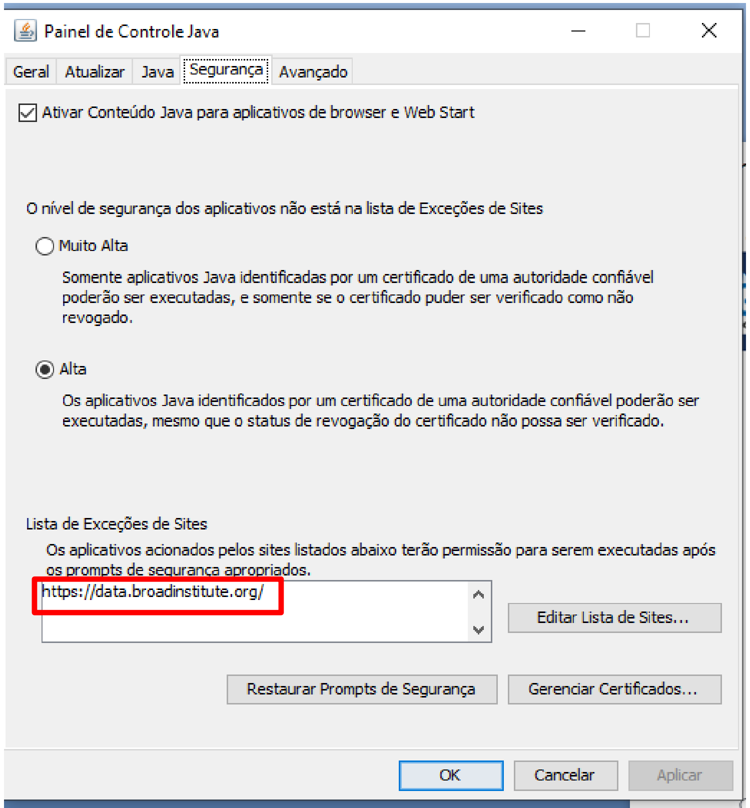
There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below. For details on the GSEA algorithm and software refer to the Documentation. For details on the latest release refer to the Release Notes. The source is available from our GitHub organization.

See the license terms page for details about the license for the GSEA software and source code. Please note that the license terms vary for different versions of the software.

GSEA v4.1.0 Mac App	Download and unzip the Mac App Archive then double-click the GSEA application to run it. You can move the app to the Applications folder or anywhere else.	download GSEA_4.1.0.app.zip
GSEA v4.1.0 for Windows	Download and run the installer. A GSEA shortcut will be created on the Desktop; double-click it to run the application. **64-bit Windows is required**	download GSEA_Win_4.1.0-installer.exe
GSEA v4.1.0 for Linux	Download and unzip the Archive. See the included readme.txt for further instructions. **64-bit Linux is required**	download GSEA_Linux_4.1.0.zip
GSEA v4.1.0 for the command line (all platforms)	Download and unzip the Archive. See the included readme.txt for further instructions. **Requires separate Java 11 installation.**	download GSEA_4.1.0.zip
GSEA v4.1.0 Java Web Start (all platforms)	Launches the GSEA desktop application from the web. **Requires separate Java 8 installation.** **Please use a configuration smaller than your computer's total memory.** **This option will be removed in a future release.**	Launch with 1GB (for 32 or 64-bit Java) Launch

A versão Java não necessita instalação no computador.

Dica: pode ser preciso adicionar o site do provedor do programa como exceção de segurança no Java



☑ Ativar Conteúdo Java para aplicativos de browser e Web Start

O nível de segurança dos aplicativos não está na lista de Exceções de Sites

○ Muito Alta

Somente aplicativos Java identificadas por um certificado de uma autoridade confiável poderão ser executadas, e somente se o certificado puder ser verificado como não revogado.

● Alta

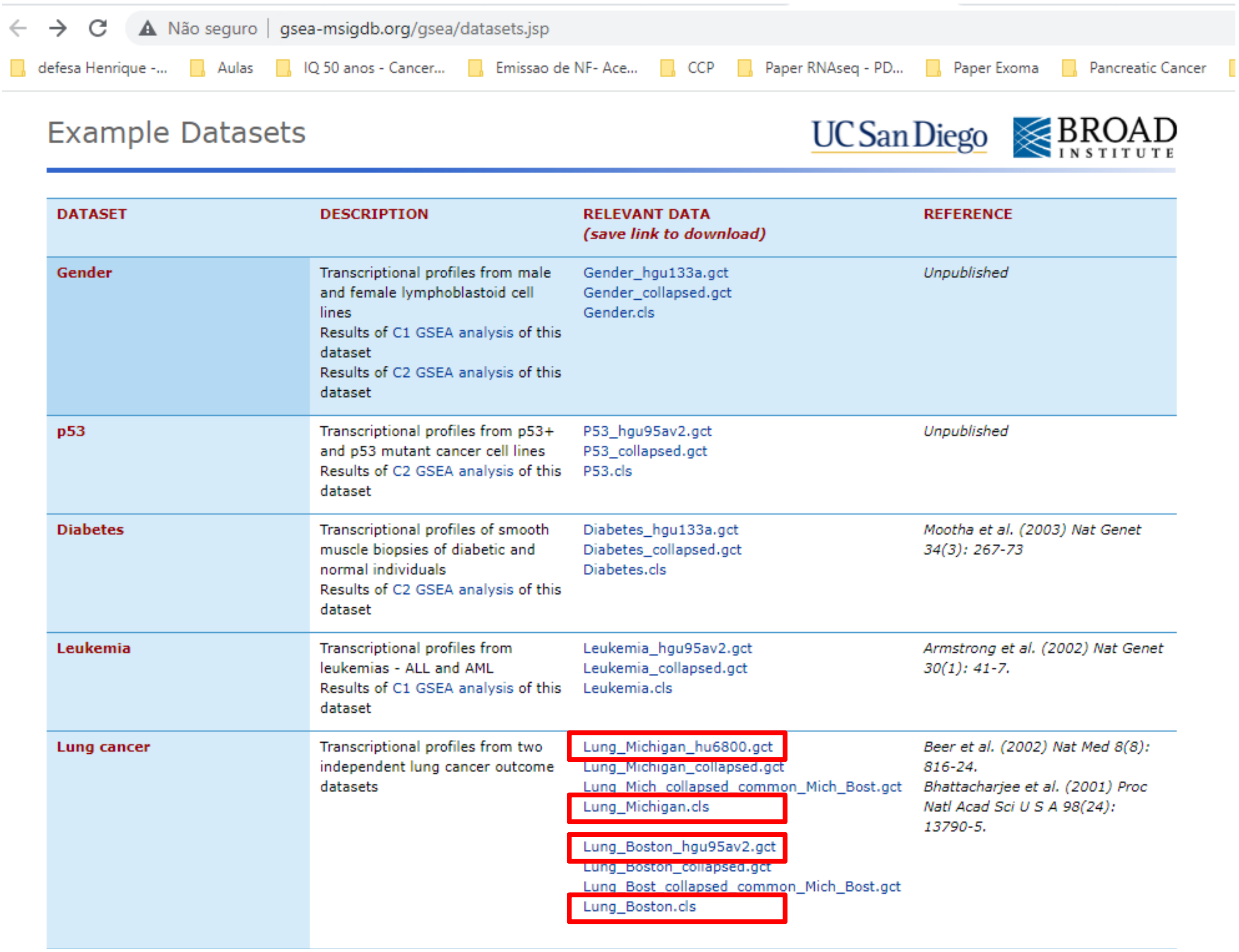
Os aplicativos Java identificados por um certificado de uma autoridade confiável poderão ser executadas, mesmo que o status de revogação do certificado não possa ser verificado.

Lista de Exceções de Sites

Os aplicativos acionados pelos sites listados abaixo terão permissão para serem executadas após os prompts de segurança apropriados.

https://data.broadinstitute.org/

Baixar os arquivos com os dados de expressão gênica (*.gct) e identificação das amostras (*.cls)

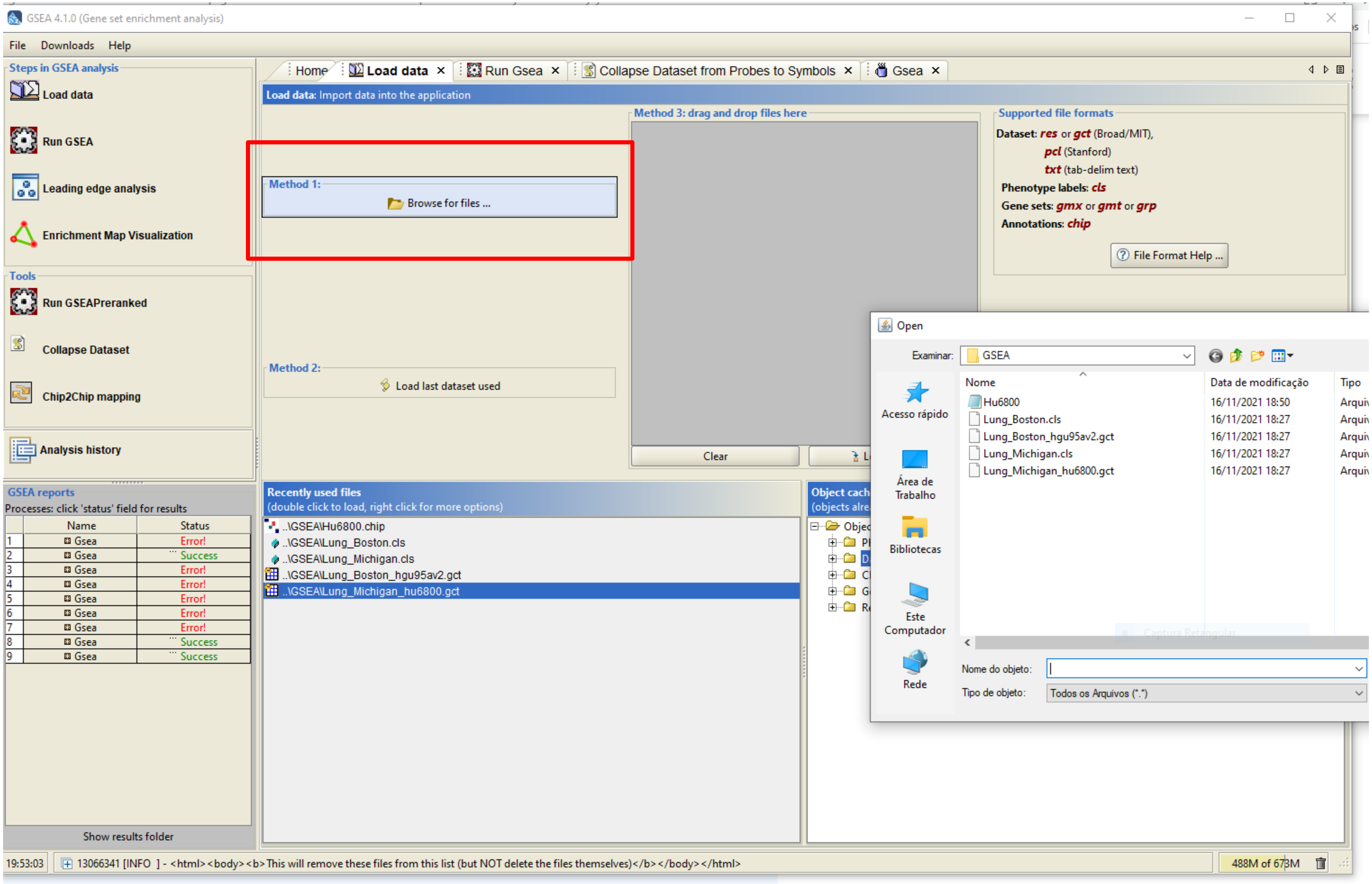


Example Datasets

DATASET	DESCRIPTION	RELEVANT DATA *(save link to download)*	REFERENCE
Gender	Transcriptional profiles from male and female lymphoblastoid cell lines Results of C1 GSEA analysis of this dataset Results of C2 GSEA analysis of this dataset	Gender_hgu133a.gct Gender_collapsed.gct Gender.cls	*Unpublished*
p53	Transcriptional profiles from p53+ and p53 mutant cancer cell lines Results of C2 GSEA analysis of this dataset	P53_hgu95av2.gct P53_collapsed.gct P53.cls	*Unpublished*
Diabetes	Transcriptional profiles of smooth muscle biopsies of diabetic and normal individuals Results of C2 GSEA analysis of this dataset	Diabetes_hgu133a.gct Diabetes_collapsed.gct Diabetes.cls	*Mootha et al. (2003) Nat Genet 34(3): 267-73*
Leukemia	Transcriptional profiles from leukemias - ALL and AML Results of C1 GSEA analysis of this dataset	Leukemia_hgu95av2.gct Leukemia_collapsed.gct Leukemia.cls	*Armstrong et al. (2002) Nat Genet 30(1): 41-7.*
Lung cancer	Transcriptional profiles from two independent lung cancer outcome datasets	Lung_Michigan_hu6800.gct Lung_Michigan_collapsed.gct Lung_Mich_collapsed_common_Mich_Bost.gct Lung_Michigan.cls Lung_Boston_hgu95av2.gct Lung_Boston_collapsed.gct Lung_Bost_collapsed_common_Mich_Bost.gct Lung_Boston.cls	*Beer et al. (2002) Nat Med 8(8): 816-24.* *Bhattacharjee et al. (2001) Proc Natl Acad Sci U S A 98(24): 13790-5.*

Dica: baixar também a anotação da plataforma de microarranjos de DNA Affy HU6800
https://data.broadinstitute.org/gsea-msigdb/msigdb/annotations_legacy/unconverted_chips/HU6800.chip

Carregar os arquivos do passo anterior no programa GSEA



Load data: Import data into the application

Dataset: *res* or *gct* (Broad/MIT),
pcl (Stanford)
txt (tab-delim text)
Phenotype labels: *cls*
Gene sets: *gmx* or *gmt* or *grp*
Annotations: *chip*

GSEA reports
Processes: click 'status' field for results

	Name	Status
1	Gsea	Error!
2	Gsea	Success
3	Gsea	Error!
4	Gsea	Error!
5	Gsea	Error!
6	Gsea	Error!
7	Gsea	Error!
8	Gsea	Success
9	Gsea	Success

Recently used files
(double click to load, right click for more options)

- ..\GSEA\Hu6800.chip
- ..\GSEA\Lung_Boston.cls
- ..\GSEA\Lung_Michigan.cls
- ..\GSEA\Lung_Boston_hgu95av2.gct
- ..\GSEA\Lung_Michigan_hu6800.gct

Nome	Data de modificação	Tipo
Hu6800	16/11/2021 18:50	Arqui
Lung_Boston.cls	16/11/2021 18:27	Arqui
Lung_Boston_hgu95av2.gct	16/11/2021 18:27	Arqui
Lung_Michigan.cls	16/11/2021 18:27	Arqui
Lung_Michigan_hu6800.gct	16/11/2021 18:27	Arqui

```
19:53:03   13066341 [INFO ] - <html><body><b>This will remove these files from this list (but NOT delete the files themselves)</b></body></html>
```

488M of 673M

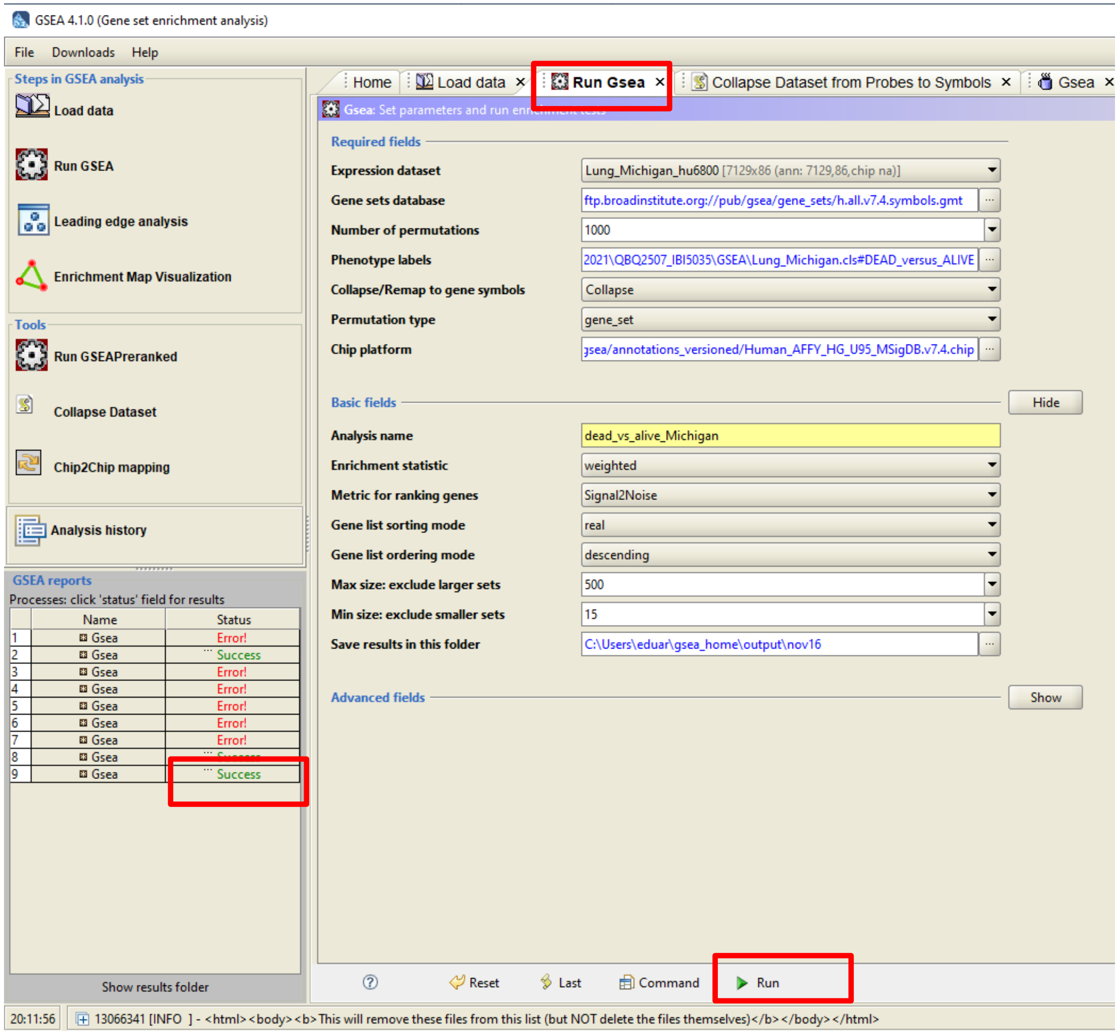
Na aba “Run GSEA”, selecionar:

- Item “Expression dataset”: selecionar o dado de expressão (Michigan ou Boston)
- Item Gene Set Database”: selecionar o gene set “Hallmarks”
 -Item “Phenotype”: selecionar DEAD vs ALIVE (Michigan ou Boston)
- Item Chip Platform”: Hu6800.chip (Michigan) ou Human_AFFY_HG_U95_MSigDB.v7.4.chip (Boston)
- Item “Analysis Name”: dead_vs_alive_Michigan ou dead_vs_alive_Boston

Clicar “Run” (rodar analises separadas para cada cada dataset).

O exemplo ao lado se refere a análise com os dados de Michigan

Para visualizar os resultados clicar no processo após finalizado



Required fields

Expression dataset	Lung_Michigan_hu6800 [7129x86 (ann: 7129,86,chip na)]
Gene sets database	ftp.broadinstitute.org://pub/gsea/gene_sets/h.all.v7.4.symbols.gmt
Number of permutations	1000
Phenotype labels	2021\QBQ2507_IBI5035\GSEA\Lung_Michigan.cls#DEAD_versus_ALIVE
Collapse/Remap to gene symbols	Collapse
Permutation type	gene_set
Chip platform	;sea/annotations_versioned/Human_AFFY_HG_U95_MSigDB.v7.4.chip

Basic fields

Analysis name	dead_vs_alive_Michigan
Enrichment statistic	weighted
Metric for ranking genes	Signal2Noise
Gene list sorting mode	real
Gene list ordering mode	descending
Max size: exclude larger sets	500
Min size: exclude smaller sets	15
Save results in this folder	C:\Users\eduar\gsea_home\output\nov16

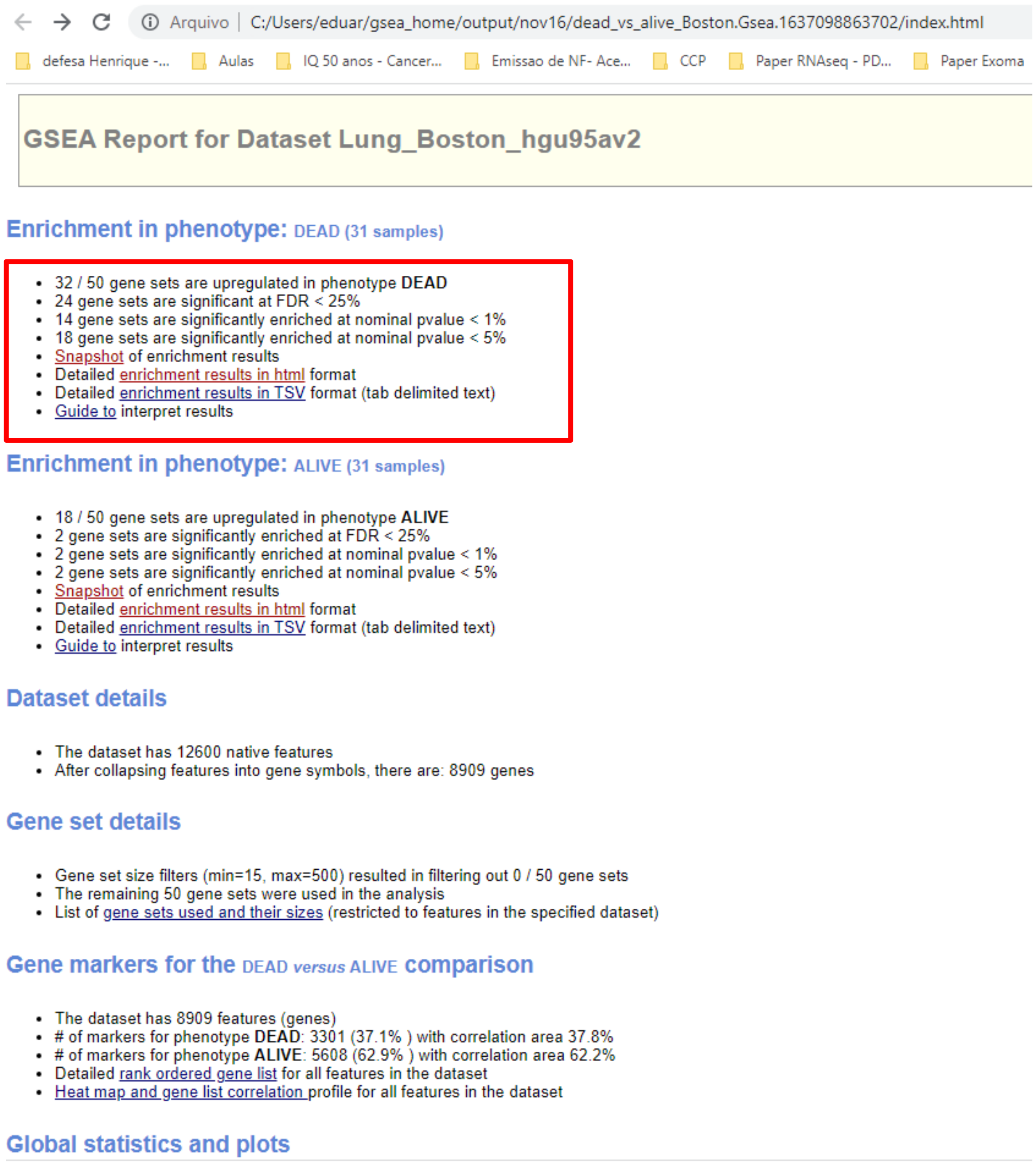
Advanced fields

GSEA reports

Processes: click 'status' field for results

	Name	Status
1	Gsea	Error!
2	Gsea	Success
3	Gsea	Error!
4	Gsea	Error!
5	Gsea	Error!
6	Gsea	Error!
7	Gsea	Error!
8	Gsea	Success
9	Gsea	Success

```
20:11:56 | 13066341 [INFO ] - <html><body><b>This will remove these files from this list (but NOT delete the files themselves)</b></body></html>
```



GSEA Report for Dataset Lung_Boston_hgu95av2

Enrichment in phenotype: DEAD (31 samples)

- 32 / 50 gene sets are upregulated in phenotype **DEAD**
- 24 gene sets are significant at FDR < 25%
- 14 gene sets are significantly enriched at nominal pvalue < 1%
- 18 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: ALIVE (31 samples)

- 18 / 50 gene sets are upregulated in phenotype **ALIVE**
- 2 gene sets are significantly enriched at FDR < 25%
- 2 gene sets are significantly enriched at nominal pvalue < 1%
- 2 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 12600 native features
- After collapsing features into gene symbols, there are: 8909 genes

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 0 / 50 gene sets
- The remaining 50 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the DEAD *versus* ALIVE comparison

- The dataset has 8909 features (genes)
- # of markers for phenotype **DEAD**: 3301 (37.1%) with correlation area 37.8%
- # of markers for phenotype **ALIVE**: 5608 (62.9%) with correlation area 62.2%
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset

Global statistics and plots