



# Know Your Transcriptome

Integrative Bioinformatic Approaches

Anil Jegga  
Biomedical Informatics

Contact Information:

Anil Jegga

Biomedical Informatics

Room # 232, S Building 10th Floor

CCHMC

Homepage: <http://anil.cchmc.org>

Tel: 513-636-0261

E-mail: [anil.jegga@cchmc.org](mailto:anil.jegga@cchmc.org)

**Slides and Example data sets available for download at:**

**<http://anil.cchmc.org/dhc.html>**

**Workshop Evaluation:** Please provide your valuable feedback on the evaluation sheet provided along with the hand-outs

This workshop is about the analysis of transcriptome and **does not** cover microarray data analysis

Contact Huan Xu ([huan.xu@cchmc.org](mailto:huan.xu@cchmc.org) for GeneSpring related questions or microarray data analysis

All the applications/servers/databases used in this workshop are **free** for academic-use. Applications that are not free for use (e.g. Ingenuity Pathway Analysis, MatInspector, etc.) are not covered here. However, we have licensed access to both of these and please contact us if you are interested in using them.

## I have a list of co-expressed mRNAs (Transcriptome)....

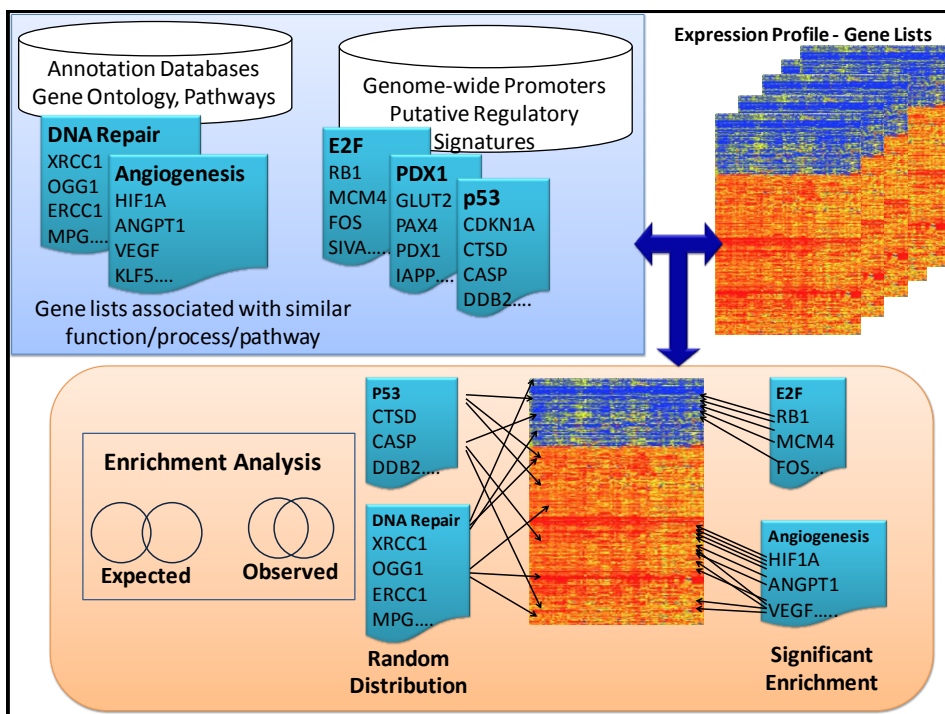
### Now what?

#### 1. Identify putative shared regulatory elements

- Known transcription factor binding sites (TFBS)
  - Conserved
  - Non-conserved
- Unknown TFBS or Novel motifs
  - Conserved
  - Non-conserved
- MicroRNAs

#### 2. Identify the underlying biological theme

- Gene Ontology
- Pathways
- Phenotype/Disease Association
- Protein Domains
- Protein Interactions
- Expression in other tissues/experiments
- Drug targets
- Literature co-citation...



**I have a list of co-expressed mRNAs (Transcriptome)....  
I want to find the shared cis-elements – Known and Novel**

Known transcription factor binding sites (TFBS)

- ❖ Conserved
  - oPOSSUM
  - DiRE
- ❖ Non-conserved
  - Pscan
  - **MatInspector** (\*Licensed)

Unknown TFBS or Novel motifs

- ❖ Conserved
  - oPOSSUM
  - **Weeder-H**
- ❖ Non-conserved
  - **MEME**
  - **Weeder**

1. Each of these applications support different forms of input. Very few support probeset IDs.
2. **Red Font**: Input sequence required; Do not support gene symbols, gene IDs, or accession numbers. The advantage is you can use them for scanning sequences from any species.
3. \*Licensed software: We have access to the licensed version.

**I have a list of co-expressed mRNAs (Transcriptome)....  
I want to find the shared cis-elements – Known and Novel**

Known transcription factor binding sites (TFBS)

- ❖ Conserved
  - oPOSSUM
  - DiRE
- ❖ Non-conserved
  - Pscan
  - **MatInspector** (\*Licensed)

Unknown TFBS or Novel motifs

- ❖ Conserved
  - oPOSSUM
  - **Weeder-H**
- ❖ Non-conserved
  - **MEME**
  - **Weeder**

## oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)

**oPOSSUM** (<http://www.cisreg.ca/oPOSSUM>)

Welcome to oPOSSUM

oPOSSUM is a web-based system for the detection of over-represented transcription factor binding sites in the promoters of sets of genes.

**Human SSA** Enter >>

Human Single Site Analysis (SSA) is designed to detect over-represented conserved single sites in human and mouse genes.

Reference: Ho Sui, et al. (2005). oPOSSUM: Identification of over-represented transcription factor binding sites in co-expressed genes. *NAR*, 33(10):3154-64. PMID: [15933209](https://pubmed.ncbi.nlm.nih.gov/15933209/)

**Human CSA (Module analysis)** Enter >>

Human Combination Site Analysis (CSA) identifies over-represented combinations of conserved transcription factor binding sites in sets of human and mouse genes.

Reference: Huang, S., Fulton, D., et al. (2006). Identification of over-represented combinations of transcription factor binding sites in sets of co-expressed genes. *In Advances in Bioinformatics and Computational Biology*, Vol. 3. Imperial College Press, London, UK. 247-56. [PDF](#).

**Worm SSA** Enter >>

Worm Single Site Analysis (SSA) identifies over-represented conserved transcription factor binding sites in sets of *C. elegans* and *C. briggsae* genes.

**Yeast SSA** Enter >>

Yeast Single Site Analysis (SSA) identifies over-represented transcription factor binding sites in sets of *S. cerevisiae* genes. Phylogenetic footprinting has not been used for yeast.

Supports human and mouse

## oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)

Select Analysis Parameters

STEP 1: Enter a list of co-expressed genes

**Species:**

human  mouse

**Gene ID type:**

Ensembl  HUGO/MGI Symbol/Alias  RefSeq  Entrez Gene

Paste gene IDs:

Use sample genes

259  
5265  
350  
335  
335  
1558

OR upload a file containing a list of gene identifiers:

Disadvantage: Supports either human or mouse only

## oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)

STEP 2: Select transcription factor binding site matrices

### JASPAR CORE Profiles

All profiles with a minimum specificity of  bits (min. 8 bits)

**OR select by taxonomic supergroup:**

plant  vertebrate  insect

**OR select specific profiles:**

ABI4  
Agamous  
AGL3  
Ar  
Arnt  
Arnt-Ahr  
ARR10  
Athb-1

The JASPAR PHYLOFACTS database consists of 174 profiles that were extracted from phylogenetically conserved gene upstream elements. They are a mix of known and as of yet undefined motifs.

#### When should it be used?

They are useful when one expects that other factors might determine promoter characteristics and/or tissue specificity.

### JASPAR PhyloFACTS Profiles

All profiles with a minimum specificity of  bits (min. 8 bits)

## oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)

STEP 3: Select parameters

Level of conservation:

Matrix match threshold:

%

Amount of upstream / downstream sequence:

Number of results to display:

Top  results

**OR** only results with **Z-score**  $\geq$   and **Fisher score**  $\leq$

Sort results by:

Z-score  Fisher score

The Fisher statistic reflects the proportion of genes that contain the TFBS compared to background.

The Z-score statistic reflects the occurrence of the TFBS in the promoters of the co-expressed set compared to background.

Press the **Submit** button to perform the analysis or **Reset** to reset the analysis seconds to a minute or more to perform. Please be patient.

# oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)

**Analysis Results**

Selected Parameters

Conservation level: Top 10% of conserved sites

Matrix match score: 80%

Upstream sequence length: 2000

Downstream sequence length: 2000

Number of genes submitted: 21

Number of genes included: 15

Number of genes excluded: 6

Target Genes

Analyzed: 1356 350 2158 259 383 335 3273 462 1571 3105 229 325 2168 2244 5055

Excluded: 5265 1558 125 3240 3827 5004

Download as a tab delimited text file (results will be kept on the server for 3 days after analysis)

**Genes Containing Conserved HNF1A Binding Sites:**

Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1356	ENSG0000024252	3	-1	150422269	150420270	150424269	GTTAATGTTTAA	150421319	951	150421332	938	1	15.334
350	ENSG0000019358	17	-1	61659974	61659975	61659974	GTTAATGTTTAA	616564022	-98	616564045	-71	-1	13.479
3273	ENSG00000113905	3	1	187864487	187864487	187864486	TGTAATGTTTAA	187864244	-143	187864357	-130	-1	9.708
1571	ENSG00000130648	10	1	135190857	135188857	135192856	GTTTATTATTAG	135190745	-112	135190758	-109	-1	14.409
5185	ENSG00000124253	20	1	55569543	55567543	55571542	AGATAATCATTGAA	55569396	-147	55569409	-134	-1	9.903
325	ENSG00000132783	1	1	157824239	15782239	157825284	AGTTATTATTGAA	157824079	-160	157824092	-147	-1	12.759
2148	ENSG00000163586	2	-1	88206893	88206894	88210493	AGTTATTATTGAA	88208792	-99	88208805	-112	-1	12.830
2244	ENSG00000121564	4	1	155703596	155701596	155705595	AGTTAATTTTAA	155703524	-72	155703537	-59	-1	14.863

Download as a tab delimited text file

**Genes Containing Conserved SRY Binding Sites:**

Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1356	ENSG0000024252	3	-1	150422269	150420270	150424269	TAAACATT	150421323	947	150421331	939	-1	6.961
350	ENSG0000019358	17	-1	61659974	61659975	61659974	TAAACATT	61654250	-92	61654259	-100	-1	7.793
3273	ENSG00000113905	3	1	187864487	187864487	18786486	TGACACAA	187864357	-99	187864450	-94	-1	9.474
1571	ENSG00000130648	10	1	135190857	135188857	135192856	TGACACAA	135190758	-112	135190758	-109	-1	14.409
5185	ENSG00000124253	20	1	55569543	55567543	55571542	TTGACAAA	55569396	-147	55569409	-134	-1	9.903
325	ENSG00000132783	1	1	157824239	15782239	157825284	TTGACAAA	157824079	-160	157824092	-147	-1	12.759
2148	ENSG00000163586	2	-1	88206893	88206894	88210493	TGACACAA	88208792	-99	88208805	-112	-1	12.830
2244	ENSG00000121564	4	1	155703596	155701596	155705595	TAAACATT	155703524	-72	155703537	-59	-1	14.863

Download as a tab delimited text file

**Genes Containing Conserved HNF1A Binding Sites:**

Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1356	ENSG0000024252	3	-1	150422269	150420270	150424269	GTTAATGTTTAA	150421319	951	150421332	938	1	15.334
350	ENSG0000019358	17	-1	61659974	61659975	61659974	GTTAATGTTTAA	616564022	-98	616564045	-71	-1	13.479
3273	ENSG00000113905	3	1	187864487	187864487	187864486	TGTAATGTTTAA	187864244	-143	187864357	-130	-1	9.708
1571	ENSG00000130648	10	1	135190857	135188857	135192856	GTTTATTATTAG	135190745	-112	135190758	-109	-1	14.409
5185	ENSG00000124253	20	1	55569543	55567543	55571542	AGATAATCATTGAA	55569396	-147	55569409	-134	-1	9.903
325	ENSG00000132783	1	1	157824239	15782239	157825284	AGTTATTATTGAA	157824079	-160	157824092	-147	-1	12.759
2148	ENSG00000163586	2	-1	88206893	88206894	88210493	AGTTATTATTGAA	88208792	-99	88208805	-112	-1	12.830
2244	ENSG00000121564	4	1	155703596	155701596	155705595	AGTTAATTTTAA	155703524	-72	155703537	-59	-1	14.863

Download as a tab delimited text file

**Genes Containing Conserved SRY Binding Sites:**

Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1356	ENSG0000024252	3	-1	150422269	150420270	150424269	TAAACATT	150421323	947	150421331	939	-1	6.961
350	ENSG0000019358	17	-1	61659974	61659975	61659974	TAAACATT	61654250	-92	61654259	-100	-1	7.793
3273	ENSG00000113905	3	1	187864487	187864487	18786486	TGACACAA	187864357	-99	187864450	-94	-1	9.474
1571	ENSG00000130648	10	1	135190857	135188857	135192856	TGACACAA	135190758	-112	135190758	-109	-1	14.409
5185	ENSG00000124253	20	1	55569543	55567543	55571542	TTGACAAA	55569396	-147	55569409	-134	-1	9.903
325	ENSG00000132783	1	1	157824239	15782239	157825284	TTGACAAA	157824079	-160	157824092	-147	-1	12.759
2148	ENSG00000163586	2	-1	88206893	88206894	88210493	TGACACAA	88208792	-99	88208805	-112	-1	12.830
2244	ENSG00000121564	4	1	155703596	155701596	155705595	TAAACATT	155703524	-72	155703537	-59	-1	14.863

Download as a tab delimited text file

# oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)

## Select Custom Analysis Parameters

STEP 1a: Enter a list of co-expressed genes

Species:

human  mouse

Gene ID type:

Ensembl  HUGO/MGI Symbol/Alias  RefSeq  Entrez Gene

Paste gene IDs (max. 1000 genes):

259

5265

350

335

335

1558

OR upload a file containing a list of gene identifiers:

STEP 1b: Enter a background list of genes

Use background set of 1000 random genes

OR Paste background gene IDs (max. 1000 genes):

1281

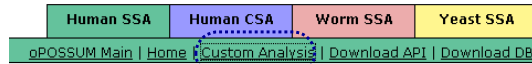
1281

1805

125

10551

OR upload a file containing a list of gene identifiers:



## oPOSSUM Analysis

TF	TF Class	TF Supergroup	IC	Background gene hits	Background gene num-hits	Target gene hits	Target gene num-hits	Background TFBS hits	Background TFBS rate	Target TFBS hits	Target TFBS rate	Z-score	Fisher score
MEF1A	HOMEO	vertebrate	15.548	1	10	8	7	1	0.0025	8	0.0116	20.32	2.452e-02
ME	NSP	vertebrate	11.847	3	10	2	4	1	0.0025	2	0.0116	21.88	1.863e-02
MOXD-1	HOMEO	vertebrate	11.127	3	8	8	4	3	0.0037	11	0.0110	18.95	1.042e-01
MEIS1	HOMEO	vertebrate	8.542	4	7	10	5	5	0.0079	20	0.0218	14.26	1.286e-01
MEIS2	HOMEO	vertebrate	10.941	3	8	8	7	5	0.0079	11	0.0120	6.171	1.775e-01
SOX1	HOMEO	vertebrate	9.680	7	4	13	2	28	0.0295	10	0.0362	2.531	1.825e-01
SOX2	HOMEO	vertebrate	9.183	7	4	13	2	28	0.0443	20	0.0361	-2.203	1.826e-01
MEIS3	HOMEO	vertebrate	8.270	7	4	13	2	28	0.0344	20	0.0442	4.845	1.826e-01
SOX11	POU/WRW	vertebrate	12.183	4	7	8	6	3	0.0148	11	0.0149	5.055	1.242e-01
SOX11	POU/WRW	vertebrate	13.190	1	10	4	11	1	0.0018	5	0.0061	9.231	2.739e-01

## Exercise 1: Use oPOSSUM to find shared conserved cis-elements in a group of co-expressed genes

1. Download the example dataset (file "Example-Set-1.xls" – rt. click and "save as" from <http://anil.cchmc.org/dhc.html>)
2. Copy 20 or 25 gene IDs from the downloaded file and use them for oPOSSUM analysis

### oPOSSUM Summary:

1. For **conserved** common cis-elements in a group of genes
2. Supports human or mouse only
3. Uses JASPAR matrices only which are not exhaustive
4. Options to select the regions (max. 10 kb flanking region)
5. Results indicate the TFBSs' positions relative to the TSS and the coordinates are from the current genome assembly
6. Supports selection of background set
7. Does not support upload of your sequences; Input should be standard gene symbols or IDs or accession numbers

**oPOSSUM Summary:**

4. Options to select the regions (max. 10 kb flanking region)

## DiRE (<http://dire.dcode.org/>)

**Transcription Factors Occurrence Importance**

TF	Occurrence	Importance
1 IRF7	14.2%	0.3176
2 VMAF	10.00%	0.34781
3 HSF2	2.88%	0.29571
4 YY1	11.14%	0.27550
5 ALS4	7.14%	0.27679
6 IRF	5.71%	0.19643
7 RUSHHA	5.71%	0.19643
8 ISRE	10.00%	0.19375
9 R	5.71%	0.17681
10 HMG1Y	11.43%	0.15571
11 CDP	8.57%	0.15286
12 LXR	5.71%	0.16161
13 POU6F1	5.71%	0.15732
14 PPARA	15.71%	0.15714
15 CEBP	21.43%	0.15000
16 STAT	11.43%	0.14857

**68 Potential Regulatory Elements**

**19 top TFs**

**Candidate Transcription Factors**

#	Regulatory element	Type	Score	Locus	Gene	Candidate transcription factor binding sites (relative positions)
1	chr1:157799292-157799916	Intergenic	11.860	chr1:157772437-157948651	APCS	15 = HES1(71) MEF2(105) NIKQ2(160) ATATA(184) AFP1(166) LHX3(166) TEL2(216) HES1(240) E2F1DP1(259) MYOD(265) HEN1(301) HMG1Y(344) CREBP1CJUN(448) YY1(611) CDP(611)
2	chr1:157822774-157823321	Promoter	1.327	chr1:157772437-157948651	APCS	4 = AREB6(80) TBX5(82) CHY1(365) TBX5(520)
3	chr1:157822347-157823820	Promoter	0.843	chr1:157772437-157948651	APCS	3 = ARF1(12) LEF1(20) CMAF(53)
4	chr1:157823967-157824176	Promoter	6.780	chr1:157772437-157948651	APCS	8 = STAT3(62) RORA2(76) PPARA(78) LXR(82) ER(83) PPARA(83) LXR_DR4(83) HNF1(117) 39 = PAV2(56) STAT3(65) OSF2(108) PEBP(107) AML1(107) RFX1(340) NFMUE1(396) TAL1(674) E2F1(259) MYOD(265) HEN1(301) HMG1Y(344) CREBP1CJUN(448) YY1(611) CDP(611)

## DiRE (<http://dire.dcode.org/>)

#	Regulatory element	Type	Score	Locus	Gene	Candidate transcription factor binding sites (relative positions)
1	chr1:157799292-157799916	Intergenic	11.860	chr1:157772437-157948651	APCS	15 = HES1(71) MEF2(105) NIKQ2(160) ATATA(184) AFP1(166) LHX3(166) TEL2(216) HES1(240) E2F1DP1(259) MYOD(265) HEN1(301) HMG1Y(344) CREBP1CJUN(448) YY1(611) CDP(611)
2	chr1:157822774-157823321	Promoter	1.327	chr1:157772437-157948651	APCS	4 = AREB6(80) TBX5(82) CHY1(365) TBX5(520)
3	chr1:157822347-157823820	Promoter	0.843	chr1:157772437-157948651	APCS	3 = ARF1(12) LEF1(20) CMAF(53)
4	chr1:157823967-157824176	Promoter	6.780	chr1:157772437-157948651	APCS	8 = STAT3(62) RORA2(76) PPARA(78) LXR(82) ER(83) PPARA(83) LXR_DR4(83) HNF1(117) 39 = PAV2(56) STAT3(65) OSF2(108) PEBP(107) AML1(107) RFX1(340) NFMUE1(396) TAL1(674) E2F1(259) MYOD(265) HEN1(301) HMG1Y(344) CREBP1CJUN(448) YY1(611) CDP(611)

### ECR-Browser (<http://ecrbrowser.dcode.org/>)

**ECR Browser on Human (hg18)**

Parameters: Graph length 100, ECR similarity 70, Layer height 55, Coordinate system relative

825 bps gene or position (chrN from-to) chr1:157799292-157799916

**DIRE regulatory element prediction (score: 11.860)**

**RelSeq Genes**

**Conservation**

100% 50% 0% Human GalGal3 XenTro3 MonDom4 PanPan2 Mm9 HmNec2

ENSEMBL Genes, UCSC Known Genes, RefSeq Genes, GNF Expression Atlas 2, SNPs

Ins 10508734




## Exercise 2: Use DiRE to find shared conserved cis-elements in a group of co-expressed genes

Use the same example dataset (downloaded file "Example-Set-1.xls") and identify putative distant regulatory regions using DiRE

### DiRE Summary:

1. DiRE's unique feature is the detection of **conserved** REs outside of proximal promoter regions, as it takes advantage of the full gene locus to conduct the search.
2. Supports human, mouse, and rat
3. Uses TRANSFAC matrices which are more exhaustive than JASPAR matrices
4. Limited options to select the regions for scanning
5. Results indicate the context (promoter, intronic, or UTR, etc.) and the coordinates are from the current genome assembly
6. Supports selection of background set
7. Does not support upload of your sequences; Input should be standard gene symbols or IDs or accession numbers
8. Connects to genome browser

## Pscan (<http://159.149.109.9/pscan>)

<p>Insert Gene/Sequence ID list: (<a href="#">help</a>) <b>PSCAN</b></p> <p>Select Organism: Homo sapiens</p> <p>Select Region: -450 +50</p> <p>Select Descriptors:          Jaspar <input checked="" type="radio"/>  Jaspar_Fam <input type="radio"/>  Transfac <input type="radio"/>  User Defined <input type="radio"/> </p> <p>Run! Undo changes Reset</p> <p>Messages:</p>	 <h3>Pscan Web Interface</h3> <p>Use the input form on the left to set up your query. The results will be displayed in this window.</p> <p><a href="#">If you need HELP please click here.</a></p> <p><b>Source:</b> <a href="#">Download Pscan source code</a></p> <p><b>Reference:</b> F.Zambelli, G.Pesole, G.Pavesi <a href="#">Pscan: Finding Over-represented Transcription Factor Binding Site Motifs in Sequences from Co-Regulated or Co-Expressed Genes.</a> <i>Nucleic Acids Research</i> 2009 37(Web Server issue):W247-W252.</p> <p><b>Contacts:</b> <a href="mailto:giulio.pavesi@unimi.it">giulio.pavesi@unimi.it</a> <a href="mailto:federico.zambelli@unimi.it">federico.zambelli@unimi.it</a></p>	<h3>Sample data</h3> <p>List of MYC target genes. MYCxx indicates that xx percent of the genes in the list are MYC targets, while the others are random genes added to the set to assess the performance of the algorithm.</p> <p><a href="#">MYC100</a> <a href="#">MYC90</a> <a href="#">MYC80</a> <a href="#">MYC75</a> <a href="#">MYC65</a> <a href="#">MYC55</a></p> <p>List of NFkB target genes, collected from literature. NFkBxx should be read as in the MYC dataset.</p> <p><a href="#">NFkB100</a> <a href="#">NFkB90</a> <a href="#">NFkB80</a> <a href="#">NFkB70</a> <a href="#">NFkB60</a> <a href="#">NFkB50</a> <a href="#">NFkB40</a></p> <p>List of NRF1 target genes. NRF1xx should be read as in the MYC dataset. Use the NRF1 matrix with the link provided below to test these datasets (save the matrix as a text file).</p> <p><a href="#">NRF1_100</a> <a href="#">NRF1_90</a> <a href="#">NRF1_80</a> <a href="#">NRF1_70</a> <a href="#">NRF1_60</a> <a href="#">NRF1_50</a> <a href="#">NRF1_40</a></p> <p><a href="#">NRF1 Matrix</a></p>
---	--	---

### Pscan (http://159.149.109.9/pscan)

Insert Gene/Sequence ID list: ([help](#))

NM\_006408  
 NM\_006418  
 NM\_006439  
 NM\_006475  
 NM\_001285  
 NM\_000668  
 NM\_000667  
 NM\_000669  
 NM\_000668

Select Organism: Homo sapiens

Select Region: -450 +50

Select Descriptors:

Jaspar  
 Jaspar\_Fam  
 Transfac  
 User Defined

Jaspar

Jaspar\_Fam

Transfac

User Defined

Messages:

6 (out of 84) gene ID(s) not found:  
 NM\_138298  
 NM\_138299  
 NM\_024416  
 XM\_936565  
 XM\_941953  
 XM\_930062  
 Working on 78 gene promoter(s).

Select Organism: Human and Mouse

Select Region: Homo sapiens  
Mus musculus  
**Human and Mouse**  
Drosophila melanogaster  
Arabidopsis thaliana  
Saccharomyces cerevisiae

Select Descriptors:

### Pscan (http://159.149.109.9/pscan)

[View Text Results](#)

97 TF profiles used

Matrix Name	P-value
IBP	1.59074e-08
Foxa2	0.000274079
FOXL1	0.000657034
MZF2A	0.000657227
Hand1-Tcf2a	0.000697277
Nobox	0.000790445
FOX11	0.000804377
PBX1	0.00124224
SRE	0.00124647
Evi1	0.00128699
TEAD1	0.00212536
Lhx3	0.00303459
Foxq1	0.00355502
Prx2	0.00486451
Lhx3	0.00527407
Nkx3-1	0.00590862
NFIL3	0.00642618
REL	0.00685234
Pax6	0.00765503
Foxd3	0.00776631
HNF1A	0.00783389
Cebpa	0.00920516
Nkx2-5	0.00940038

Matrix Info

ID	MA0047
Name	Foxa2
Class	FORKHEAD
Species	Rattus norvegicus
Inf. Content	12.43
SuperGroup	vertebrate
Protein Acc.	P32182
Type	COMPILED
PMID	8139574
Report Occurrences	<input type="button" value="Go!"/>

MA0047

	1	2	3	4	5	6	7	8	9	10	11	12
A	6	11	8	0	11	0	0	0	9	0	1	0
C	7	3	3	1	0	1	1	0	0	9	1	5
G	3	1	1	0	6	0	0	6	8	0	1	0
T	1	2	5	16	0	16	16	11	0	8	14	12

Sample Mean Score

0.975 0.95 0.925 0.9 0.875 0.85 0.825 0.8 0.775 0.75 0.725 0.7 0.675 0.65 0.625 0.6 0.575 0.55 0.525 0.5 0.475 0.45 0.425 0.4 0.375 0.35 0.325 0.3 0.275 0.25 0.225 0.2 0.175 0.15 0.125 0.1 0.075 0.05 0.025 0 0

0.836

Background Score Distribution

Sample Statistics

p-value	0.000274079
Bonferroni p-value	0.026585663
Mean	0.861172
Std Dev	0.0563177
Size	60

Compare with... (using Welch's t-test) [help](#)

Mean	<input type="text"/>	<input type="button" value="Go!"/>
Std Dev	<input type="text"/>	
Size	<input type="text"/>	

10

## Pscan (http://159.149.109.9/pscan)

**View Text Results**  
97 TF profiles used

Matrix Name	P-value
TBP	1.000746e-08
Foxa2	0.000274079
FOXD1	0.000670394
MEF2a	0.000697271
Hand1-Tcf2a	0.000697271
Nobox	0.000790445
FOXL1	0.000834377
PBX1	0.00134224
SRF	0.00124647
Evi1	0.00128699
TEAD1	0.00212538
Lhx3	0.00303459
Foxq1	0.00355502
HDZ2	0.00496451
Lhx3	0.00521407
NKX3-1	0.00590862
NFYA3	0.00642618
DEL	0.00685234
Pax5	0.00755503
Foxp2	0.00776631
HNF1A	0.00783389
Cespa	0.00920516
Nkx2-4	0.00940093

**View Text Results**

Name	Score	Position	Sequence	Strand
hg18_refGene_NM_002345	0.98343	-197	CAATATTGATT	-
hg18_refGene_NM_000668	0.982619	-145	AAATATTGACTT	-
hg18_refGene_NM_006408	0.951013	-258	CTTTATTTACTT	-
hg18_refGene_NM_000667	0.944804	-34	ATTTATTTATTT	-
hg18_refGene_NM_000609	0.939791	-428	ACTTGTTCCTT	+
hg18_refGene_NM_00103388	0.939791	-428	ACTTGTTCCTT	+
hg18_refGene_NM_199168	0.939791	-428	ACTTGTTCCTT	+
hg18_refGene_NM_021010	0.935606	-261	GAGTATTTACTT	-
hg18_refGene_NM_133477	0.932417	-356	AAACATTTATTT	+
hg18_refGene_NM_194435	0.931745	-216	CTTTGTTCCTT	+
hg18_refGene_NM_003381	0.931745	-216	CTTTGTTCCTT	+
hg18_refGene_NM_005603	0.922637	-143	GAAATATTCAT	+
hg18_refGene_NM_004616	0.9222	-56	ATCTGTTTACTT	+
hg18_refGene_NM_004705	0.91989	220	ATTTATTTACTT	-

**Matrix Info**

ID	MA0047	MA0047
Name	Foxa2	
Class	FORHEAD	
Species	Rattus norvegicus	
Inf. Content	12.43	
SuperGroup	vertebrate	
Protein Acc.	P32122	
Type	COMPLED	
PMID	8139574	
Occurrences	Gal	

**Sample Mean Score**: 0.851

**Background Score Distribution**

**Sample Statistics**

p-value	0.000274079
Bonferroni p-value	0.026585663
Mean	0.861172
Std Dev	0.0563177
Size	60

**Compare with... using Welch's t-test!**

**Sequence Logo**

**Occurrences Position Distribution (score >=0.836)**

**Occurrences Score Distribution**

## Pscan (http://159.149.109.9/pscan)

**View Text Results**  
97 TF profiles used

Matrix Name	P-value
TBP	1.000746e-08
Foxa2	0.000274079
FOXD1	0.000670394
MEF2a	0.000697271
Hand1-Tcf2a	0.000697271
Nobox	0.000790445
FOXL1	0.000834377
PBX1	0.00134224
SRF	0.00124647
Evi1	0.00128699
TEAD1	0.00212538
Lhx3	0.00303459
Foxq1	0.00355502
HDZ2	0.00496451
Lhx3	0.00521407
NKX3-1	0.00590862
NFYA3	0.00642618
DEL	0.00685234
Pax5	0.00755503
Foxp2	0.00776631
HNF1A	0.00783389
Cespa	0.00920516
Nkx2-4	0.00940093

**View Text Results**

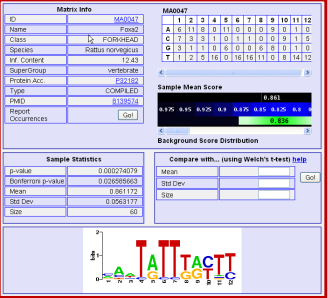
TF	Score	Position	Sequence	Strand
TBP	1.000746	-100	CAATATTGATT	-
Foxa2	0.98343	-197	CAATATTGATT	-
FOXD1	0.951013	-258	CTTTATTTACTT	-
MEF2a	0.944804	-34	ATTTATTTATTT	-
Hand1-Tcf2a	0.939791	-428	ACTTGTTCCTT	+
Nobox	0.939791	-428	ACTTGTTCCTT	+
FOXL1	0.935606	-261	GAGTATTTACTT	-
PBX1	0.932417	-356	AAACATTTATTT	+
SRF	0.931745	-216	CTTTGTTCCTT	+
Evi1	0.931745	-216	CTTTGTTCCTT	+
TEAD1	0.922637	-143	GAAATATTCAT	+
Lhx3	0.9222	-56	ATCTGTTTACTT	+
Foxq1	0.91989	220	ATTTATTTACTT	-

**TF NAME MATRIX\_ID Z\_SCORE P\_VALUE SAMPLE\_AVERAGE BACKGROUND\_AVERAGE SAMPLE\_DEVSTD SAMPLE\_SIZE**

TBP	MA0108	5.52625	1.59074e-08	0.859494	0.816811	0.0493175	60
Foxa2	MA0047	3.45141	0.000274079	0.861172	0.935541	0.0563177	60
FOXD1	MA0033	3.23019	0.000657034	0.918922	0.891892	0.0523769	60
MEF2a	MA0052	3.20941	0.000657227	0.810217	0.777574	0.0710495	60
Hand1-Tcf2a	MA0092	3.19081	0.000697277	0.895392	0.879315	0.0428231	60
Nobox	MA0125	3.15645	0.000790445	0.880691	0.854103	0.0553916	60
FOXL1	MA0042	3.15094	0.000804377	0.85506	0.825768	0.0653533	60
PBX1	MA0070	3.02199	0.00124224	0.802727	0.781876	0.048527	60
SRF	MA0083	3.02072	0.00124647	0.759912	0.740766	0.0462187	60
Evi1	MA0029	3.01195	0.00128699	0.769219	0.746618	0.0548341	60
TEAD1	MA0090	2.85352	0.00212538	0.827634	0.810446	0.0565813	60
Lhx3	MA0134	2.7417	0.00303459	0.834887	0.804852	0.0683236	60
Foxq1	MA0040	2.68955	0.00355502	0.825721	0.802203	0.0522346	60
HDZ2	MA0075	2.58254	0.00496451	0.82424	0.809708	0.0697038	60
Lhx3	MA0135	2.55439	0.00521407	0.798641	0.773954	0.0704547	60
NKX3-1	MA0124	2.51437	0.00590862	0.853125	0.829214	0.0728452	60
NF1L3	MA0025	2.49463	0.00642618	0.80459	0.783263	0.0663844	60
RE1	MA0101	2.46099	0.00685234	0.8843	0.866993	0.0500624	60
Pax6	MA0069	2.42112	0.00755503	0.778947	0.766632	0.0436411	60
Foxd3	MA0041	2.4171	0.00776631	0.857889	0.837233	0.0604468	60
HNF1A	MA0046	2.41362	0.00783389	0.790719	0.770942	0.0615964	60

**Heatmap**

## Pscan (<http://159.149.109.9/pscan>)



**Matrix Info**

ID: MANDL2  
Name: Foa2  
Class: F0904640  
Species: *Saltia roveogus*  
VT Content: 12.43  
Sequence: vntest08  
Protein Acc: P21102  
Type: COMPLET  
MFO: 1310514  
Report Occurrences: [Go!]

**Sample Statistics**

p-value: 0.00274079  
Enrichment p-value: 0.02026663  
Mean: 0.81172  
Std Dev: 0.0583177  
Size: 68

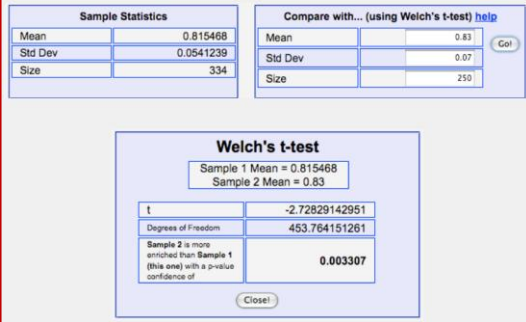
**Background Score Distribution**

Compare with... (using Welch's t-test) help

Mean: [ ] [Go!]  
Std Dev: [ ]  
Size: [ ]

**Comparing different input gene sets:**

- In the detailed output for a given matrix, you can compare the results obtained with the matrix on the gene set just submitted with the results the matrix had produced on another gene set. The latter could be a "negative" gene set (or vice versa).
- To perform the comparison, you have to fill in the "Compare with..." box fields with mean, standard deviation and sample size values of the other analysis - for the current one you can find them in the "Sample Data Statistics" box or in the overall text output that can be downloaded from the main output page.
- Warning:** Make sure that the values you input are correct, and especially that they were obtained by using the same matrix. Once you have clicked the "Go!" button, an output window will pop up and report if either of the two means is significantly higher than the other, together with a confidence p-value computed with a Welch t-test.



**Sample Statistics**

Mean: 0.815468  
Std Dev: 0.0541239  
Size: 334

**Compare with... (using Welch's t-test) help**

Mean: 0.83 [Go!]  
Std Dev: 0.07  
Size: 250

**Welch's t-test**

Sample 1 Mean = 0.815468  
Sample 2 Mean = 0.83

t	-2.72829142951
Degrees of Freedom	453.764151261
Sample 2 is more enriched than Sample 1 (this one) with a p-value confidence of	0.003307

[Close!]

## Exercise 3: Use Pscan to find shared cis-elements (Transfac) in a group of co-expressed genes

- Use the same example data set (downloaded file "Example-Set-1.xls") and find the enriched JASPAR and Transfac TFBS. How do the outputs differ?

### PScan Summary:

- Pscan supports a variety of TFBS matrices (e.g. JASPAR, Transfac) including user input matrix.
- Supports human, mouse, drosophila, and yeast
- Limited options to select the regions for scanning
- Cannot select the background set although comparisons can be computed
- Does not support upload of your sequences; Input options are very limited
- Variety of user-friendly output formats including heat map view

**I have a list of co-expressed mRNAs (Transcriptome)....  
I want to find the shared cis-elements – Known and Novel**

Known transcription factor binding sites (TFBS)

- ❖ Conserved
  - oPOSSUM
  - DiRE
- ❖ Non-conserved
  - Pscan
  - **MatInspector** (\*Licensed)

Unknown TFBS or Novel motifs

- ❖ Conserved
  - oPOSSUM
  - **Weeder-H**
- ❖ Non-conserved
  - **MEME**
  - **Weeder**

**oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)**

Select Analysis Parameters

STEP 1: Enter a list of co-expressed genes

**Species:**  
 human  mouse

**Gene ID type:**  
 Ensembl  HUGO/MGI Symbol/Alias  RefSeq  Entrez Gene

Paste gene IDs:

259  
5265  
350  
335  
335  
1558

OR upload a file containing a list of gene identifiers:

## oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)

### STEP 2: Select transcription factor binding site matrices

#### JASPAR CORE Profiles

All profiles with a minimum specificity of  bits (min. 8 bits)

OR select by taxonomic supergroup:

plant  vertebrate  insect

OR select specific profiles:

ABI4  
Agamous  
AGL3  
Ar  
Arnt  
Arnt-Ahr  
ARR10  
Athb-1

The JASPAR PHYLOFACTS database consists of 174 profiles that were extracted from phylogenetically conserved gene upstream elements. They are a mix of known and as of yet undefined motifs.

#### When should it be used?

They are useful when one expects that other factors might determine promoter characteristics and/or tissue specificity.

#### JASPAR PhyloFACTS Profiles

All profiles with a minimum specificity of  bits (min. 8 bits)

## oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)

### STEP 3: Select parameters

Level of conservation:

Matrix match threshold:

%

Amount of upstream / downstream sequence:

Number of results to display:

Top  results

OR only results with Z-score  $\geq$   and Fisher score  $\leq$

Sort results by:

Z-score  Fisher score

Press the **Submit** button to perform the analysis or **Reset** to reset the analysis seconds to a minute or more to perform. Please be patient.

## oPOSSUM (<http://www.cisreg.ca/oPOSSUM>)

### oPOSSUM Analysis

TF	TF Class	TF Supergroup	IC	Background gene hits	Background gene non-hits	Target gene hits	Target gene non-hits	Background TFBS hits	Background TFBS rate	Target TFBS hits	Target TFBS rate	Z-score	Fisher score
<b>RYTAAWNNNTGAY</b>	Unknown	mammals	16.655	2636	12514	<b>9</b>	6	3909	0.0041	<b>15</b>	0.0237	27.75	2.676e-04
<b>TAATTA</b>	Unknown	mammals	12.000	7400	7750	<b>13</b>	2	27227	0.0132	<b>31</b>	0.0226	7.458	2.949e-03
<b>YATAAB</b>	Unknown	mammals	12.000	9219	5931	<b>14</b>	1	47951	0.0232	<b>47</b>	0.0342	6.638	6.209e-03
<b>BGITAMNNATT</b>	Unknown	mammals	17.072	2277	12873	<b>6</b>	9	3189	0.0028	<b>8</b>	0.0107	13.33	1.705e-02
<b>RTAAACA</b>	Unknown	mammals	13.000	7918	7232	<b>12</b>	3	25209	0.0142	<b>29</b>	0.0246	7.953	2.670e-02
<b>YATTNATC</b>	Unknown	mammals	13.061	6858	8292	<b>11</b>	4	18528	0.0119	<b>19</b>	0.0185	5.394	2.682e-02
<b>CTTTGA</b>	Unknown	mammals	12.000	10591	4559	<b>14</b>	1	54148	0.0262	<b>47</b>	0.0342	4.553	3.478e-02
<b>YCATTAA</b>	Unknown	mammals	13.004	7484	7666	<b>11</b>	4	22958	0.0129	<b>22</b>	0.0187	4.57	5.404e-02
<b>AACWVCAANK</b>	Unknown	mammals	15.858	3060	12090	<b>6</b>	9	4631	0.0037	<b>8</b>	0.0097	8.817	6.381e-02
<b>TGGAAA</b>	Unknown	mammals	12.000	11182	3968	<b>14</b>	1	67892	0.0328	<b>43</b>	0.0313	-0.7882	6.656e-02

### Genes Containing Conserved RYTAAWNNNTGAY Binding Sites:

Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1385	<b>ENSG00000124052</b>	3	-1	150422269	150420270	150424269	ACTAAATTGTGTC	150422362	-98	150422375	-104	-1	8.802
183	<b>ENSG00000118526</b>	6	1	131936059	131934059	131938058	GTACAACTTTGAC	131937510	1440	131937531	1473	1	6.292
3273	<b>ENSG00000113985</b>	3	1	187866487	187864487	187868486	ACTAATCATTAC	187866344	-143	187866357	-130	1	8.969
462	<b>ENSG00000112881</b>	1	-1	172146654	172144655	172148654	CTTAATCTGTCT	172144991	1664	172145004	1651	1	6.144
				172153139	172151140	172155139	GTCAAAGGCTGAT	172153165	-26	172153178	-39	1	11.059
1571	<b>ENSG00000138849</b>	10	1	135190897	135188897	135192896	TTCAAAGGCTGAT	135190716	-141	135190729	-128	-1	6.038
5185	<b>ENSG00000124253</b>	20	1	55569543	55567943	55571542	ACTAAACTTTGAC	55569306	-237	55569319	-224	-1	13.888
				55569543	55567543	55571542	GTTAATGAATGCT	55569374	-169	55569387	-156	-1	8.174
				55569543	55567543	55571542	GATAATCATTGAA	55569396	-147	55569409	-134	-1	6.601
325	<b>ENSG00000132783</b>	1	1	157824239	157822239	157825284	ATTAATACAGAC	157822921	-1318	157822934	-1305	-1	10.324
2188	<b>ENSG00000163588</b>	2	-1	88208693	88206893	88210693	GTTAATGTTTGA	88208792	-99	88208805	-112	-1	12.880
				88208693	88206694	88210693	CTTATCATTGAC	88208819	-126	88208832	-139	-1	6.864
				88208693	88206694	88210693	ATTAATGTTTCT	88208867	-174	88208880	-187	-1	11.146
2244	<b>ENSG00000171564</b>	4	1	155703596	155701596	155705595	GTTAATTTAAT	155703524	-72	155703537	-59	-1	11.267
				155703596	155701596	155705595	GCTAATGAAGAT	155703971	376	155703984	389	1	7.064

## I have a list of co-expressed mRNAs (Transcriptome).... I want to find the shared cis-elements – Known and Novel

### Known transcription factor binding sites (TFBS)

- ❖ Conserved
  - oPOSSUM
  - DiRE
- ❖ Non-conserved
  - Pscan
  - **MatInspector** (\*Licensed)

### Unknown TFBS or Novel motifs

- ❖ Conserved
  - oPOSSUM
  - **Weeder-H**
- ❖ Non-conserved
  - **MEME**
  - **Weeder**

1. Each of these applications support different forms of input. Very few support probeset IDs.
2. **Red Font:** Input sequence required; Do not support gene symbols, gene IDs, or accession numbers. The advantage is you can use them for scanning sequences from any species.
3. \*Licensed software: We have access to the licensed version.

How to fetch promoter/upstream sequence – single/multiple?

# Genome Browser (<http://genome.ucsc.edu>)

**UCSC Genome Bioinformatics**

**Genomes** | [Blat](#) | [Tables](#) | [Gene Sorter](#) | [PCR](#) | [VisiGene](#) | [Proteome](#) | [Session](#) | [FAQ](#) | [Help](#)

**Genome Browser**

**About the UCSC Genome Bioinformatics Site**

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides a portal to the ENCODE project.

We encourage you to explore these sequences with our tools. The [Genome Browser](#) zooms and scrolls over chromosomes, showing the work of annotators worldwide. The [Gene Sorter](#) shows expression, homology and other information on groups of genes that can be related in many ways. [Blat](#) quickly maps your sequence to the genome. The [Table Browser](#) provides convenient access to the underlying database. [VisiGene](#) lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. [Genome Graphs](#) allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our [public mailing list](#).

**News** News Archives ►

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the [genome-announce](#) mailing list.

**9 September 2009 - Changes to the bigBed/bigWig data formats**

If you have been taking advantage of the new bigBed format (for very large data sets), you'll be happy to hear that we have considerably slimmed down the memory footprint of the program that converts BED files into bigBed files: bedToBigBed. Because it now uses a multi-pass approach, it now takes only 1/4 the amount of RAM as the size of the uncompressed BED input file (instead of the 5x RAM it needed previously!). Read more [here](#). Pick up the new bedToBigBed executable [here](#).

In conjunction with this change, there is also a change to the way you must specify your bigBed or bigWig Custom Track. When you specify the location of your local bigBed/bigWig file (on your web-accessible http, https, or ftp server), use this designation: bigDataUri (instead of the old designation: dataUri).

e.g. track type=bigBed name="My Big Bed" description="Some Data from My Lab" bigDataUri=http://myorg.edu/mylab/myBigBed.bb

Additionally, we would like to announce a companion program to the previously-announced wigToBigWig program: bedGraphToBigWig. This program converts bedGraph files into bigWig files. The bedGraph format allows display of sparse or varying-size data. Read more [here](#). You can download the new bedGraphToBigWig utility [here](#).

The main advantage of the bigBed and bigWig formats is that only the portions of the files needed to display a particular region are transferred to UCSC, so for large data sets, displaying bigBed/bigWig data is considerably faster than regular BED/wig data. The bigBed/bigWig file remains on your web accessible server (http, https, or ftp), not on the UCSC server. Consequently, creating your Custom Track is very fast. Only the portion that is needed for the chromosomal position you are currently viewing is locally cached at UCSC as a "sparse file".

# Genome Browser (<http://genome.ucsc.edu>)

**Human (*Homo sapiens*) Genome Browser Gateway**

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz. Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term	image width
Vertebrate	Human	Mar. 2006	<input type="text" value="put symbol, keyword, ID here"/>	620 <input type="button" value="submit"/>
Vertebrate	Human			
Deuterostome	Chimp			
Insect	Rhesus			
Nematode	Mouse			
Other	Rat			
	Cat			
	Dog			
	Cow			
	Opossum			
	Chicken			
	X. tropicalis			
	Zebrafish			
	Tetraodon			
	Fugu			

Let the browser user interface settings to their defaults.

**Configure Image**

image width: 620 text size: small

Display chromosome ideograms small  main graphic.

Show light blue vertical guide medium

Display labels to the left of tracks. large

Display track description above each track. huge

**Genome Browser Gateway choices:**

- Select Clade
- Select genome/species: You can search only one species at a time
- Assembly: the official backbone DNA sequence
- Position: location in the genome to examine or search term (gene symbol, accession number, etc.)
- Image width: how many pixels in display window; 5000 max
- Configure: make fonts bigger + other options



## Genome Browser (http://genome.ucsc.edu)

**UCSC Genome Bioinformatics**

**Genomes** Blat Tables Gene Sorter PCR VisiGene Proteome Session FAQ Help

Genome Browser  
ENCODE  
Blat  
Table Browser

**About the UCSC Genome Bioinformatics Site**

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working d

We encourage you to explore these sequences with our tools. The [Genome Browser](#) zooms and scrolls over views. [Blat](#) quickly maps your sequence to the genome. The [Table Browser](#) provides convenient access to upload and display genome-wide data sets.

clade genome assembly position or search term image width

Vertebrate Human Mar. 2006 chrX:151,073,054-151,383,976 620 submit

[Click here to reset](#) the browser user interface settings to their defaults.

add custom tracks configure tracks and display clear position

clade genome assembly position or search term image width

Vertebrate Human Mar. 2006 PDX1 620 submit

[Click here to reset](#) the browser user interface settings to their defaults.

add custom tracks configure tracks and display clear position

## Genome Browser (http://genome.ucsc.edu)

**UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)**

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr9:103,218,857-103,241,888 jump clear size 22,832 bp. configure

Scale chr9: 10 kb 100 kb 1 Mb 10 Mb 100 Mb 1 Gb 10 Gb 100 Gb 1 Tg

RepeatMasker

move start < 2.0 > move end < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click gray/blue bars on left for track options and descriptions.

collapse all expand all

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

- Mapping and Sequencing Tracks refresh
- Phenotype and Disease Associations refresh
- Genes and Gene Prediction Tracks refresh
- mRNA and EST Tracks refresh
- Expression refresh
- Regulation refresh
- Comparative Genomics refresh
- Variation and Repeats refresh
- Pilot ENCODE Regions and Genes refresh
- Pilot ENCODE Transcription refresh
- Pilot ENCODE Chromatin Immunoprecipitation refresh
- Pilot ENCODE Chromatin Structure refresh
- Pilot ENCODE Comparative Genomics and Variation refresh

refresh

Explore the tracks





**1. Select "refFlat" under "table"**

**2. Ensure that "region" is "genome"**

**3. Click on "paste list"**

**4. Describe your track**

**Enter number of bp you want to analyze/download**

**3. Select the output format as "custom track"**

**1. Paste the gene symbols**

**2. Remember it is case-sensitive:**

- Human: all upper case (e.g. XRCC1)
- Mouse: lower case (first letter upper case. E.g. Xrcc1)

**5. Select "Variation and Repeats" under "Group"**

**6. Click on "create" under "intersection"**

**Change the "group" to "Custom Tracks" and select the appropriate "track" and "table"**

**7. Try GTF output too**

**8. Try GTF output too**

9

10

UCSC Genome Browser on Human Mar. 2006 Assembly

position search: chr1:157,820,957-157,825,966

size 5,010 bp

chr1 (chr1) 157822991 157823881 157824771

UCSC Gene Predictions Based on RefSeq, UniProt, Ensembl, and other sources

RefSeq Genes: *UCSC Gene Predictions Based on RefSeq, UniProt, Ensembl, and other sources*

Human mRNAs: *Human mRNAs from GenBank*

Human ESTs that Have Been Spliced: *Human ESTs that Have Been Spliced*

Vertebrate MULTIZ Alignment & PhyloP Conservation (20 Species)

SNPs: *SNPs*

RepeatMasker: *Repeating Elements by RepeatMasker*

Genome Browser view that lists all the SNPs lying within the upstream 1 kb (the region we queried) region of one of the genes analyzed.

One drawback with this output is it doesn't tell you which SNPs are in the upstream region of which gene. However, since the positions of SNPs are included, you can compare them with the gene coordinates and figure it out.

- Exercise 4: Download upstream 500 bp sequence for a list of genes (use the same list as before).**
- Exercise 5: Download all SNPs overlapping with these genes.**
- Exercise 6: Download the orthologous promoter sequences (human, mouse, and rat) for the gene SLC7A1.**
- Exercise 7: Are there any putative microRNA regulators for SLC7A1? If yes, download all of them using table browser.**

## I have a list of co-expressed mRNAs (Transcriptome).... I want to find the shared cis-elements – Known and Novel

### □ Known transcription factor binding sites (TFBS)

- ❖ Conserved
  - oPOSSUM
  - DiRE
- ❖ Non-conserved
  - Pscan
  - **MatInspector** (\*Licensed)

1. Each of these applications support different forms of input. Very few support probeset IDs.
2. **Red Font:** Input sequence required; Do not support gene symbols, gene IDs, or accession numbers. The advantage is you can use them for scanning sequences from any species.
3. \*Licensed software: We have access to the licensed version.

### □ Unknown TFBS or Novel motifs

- ❖ Conserved
  - oPOSSUM
  - **Weeder-H**
- ❖ Non-conserved
  - **MEME**
  - **Weeder**

Use the fetched promoter/upstream sequences for the following analyses

## WeederH (<http://159.149.109.9/pscan>)

**WeederH**  
Motif discovery in sequences from homologous genes  
Version beta running

[Click here to switch to Weeder](#)

Please, avoid submitting a large number of jobs (> 5) simultaneously. For large-scale analyses, you're welcome to download the standalone version.

**NEW** If you are looking for over-represented motifs in promoter sequences, perhaps you can also find our brand new tool, [Pscan](#) useful.

Enter your e.mail address:

Input exactly one sequence in each box

Reference sequence (FASTA)	<input type="text"/>	from Homo sapiens
Homologous sequence n. 1 (FASTA)	<input type="text"/>	from Homo sapiens
Homologous sequence n. 2 (FASTA)	<input type="text"/>	from Homo sapiens
Homologous sequence n. 3 (FASTA)	<input type="text"/>	from Homo sapiens Mus musculus Rattus norvegicus Canis familiaris Yeast (any) Drosophila (any) Caenorhabditis (any) Anopheles gambiae Arabidopsis thaliana Ciona intestinalis Danio rerio Pisca subspes Gallus gallus Xenopus tropicalis P. falciparum Magnaporthe oryzae

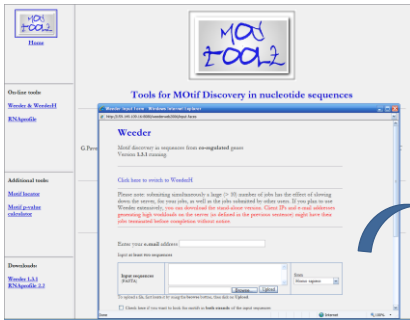
For technical reasons, it's better that you directly paste sequence in the text boxes, rather than uploading a file.

Your sequences are:

Name of this job:

1. Supports large number of species.
2. Does not support multiple sequences (multifasta) input. You have to enter each sequence separately.
3. Good for small number of sequences where you expect a potential novel (or not included in the TFBS libraries) conserved motif.

# Weeder (http://159.149.109.9:8080/weederweb2006)



## Weeder

Motif discovery in sequences from co-regulated genes  
Version 1.3.1 running.

[Click here to switch to WeederH](#)

Please note: submitting simultaneously a large (> 10) number of jobs will slow down the server, for your jobs, as well as the jobs submitted by other users. Weeder extensively, you can download the stand-alone Weeder generating high workloads on the server (as defined in the job) terminated before completion without notice.

Enter your e-mail address:

Input at least two sequences

Input sequences (FASTA)

```
ggtgtatgactacttggctgacagagcagagcagagagga
tctgtctgacagcagagagcagcgtgtgtctatcctcctga
```

To upload a file, first locate it by using the browser button, then click on Upload.

Check here if you want to look for motifs in both strands of the input sequences

Check here if you want motifs to appear in all the sequences (default is in some)  
Hint: don't try this option even if you're pretty much sure that all your sequences share a motif.

Check here if you think that the motif might appear more than once in a single sequence (without, you expect zero or one occurrence per sequence)

And, finally, you'd like:

a quick scan (short motifs, no longer than 8 nt) of your sequences  a normal scan of your sequences  a complete and thorough scan

Important input length: more than 20K will be limited to quick analysis. For large jobs, you can download the source code by following the link in the home page.  
Quick scan results will be ready in a few minutes Normal scan: results will be ready in one-two hours Thorough scan: results will be ready in a few hours However, try the normal scan first. If nothing interesting comes out, try the thorough one.

Name of this job:

Click submit once to start the computation. Click reset to clear all the fields.

Do not use Groupwise mail when submitting large number of sequences because the results are sent "in the mail" and not as an attachment. And Groupwise mail truncates messages if they are very long. Use Gmail instead. A link to the results page used to be sent earlier.

## Weeder

Thank you!  
You submitted 33 sequences from Homo sapiens

You asked to process both strands of the input sequences  
You asked for a normal scan

A confirmation e-mail and the final results will be sent to the following e-mail address:  
[aui.jegga@gmail.com](mailto:aui.jegga@gmail.com)

# Weeder (http://159.149.109.9:8080/weederweb2006)

## \*\*\* Your Weeder Web Results \*\*\*

The name of this job was `Fetal_Liver_33_27`

Input sequences from `H. sapiens`

You asked to include both strands of the input sequences  
You asked for a normal scan of your sequences

Confused about this output? [Click here](#)

Searching for motifs of length 6 with 1 mutations....

- 1) CAATTA 0.81
- 2) TAAACG 0.70
- 3) AITGAT 0.67
- 4) TATGAT 0.63
- 5) GAITTA 0.61
- 6) ATGATA 0.60
- 7) TCAITG 0.59
- 8) TGGTAT 0.59
- 9) TGGTAA 0.59
- 10) TGTAT 0.58

Searching for motifs of length 8 with 2 mutations....

- 1) CGTITAGA 0.93
- 2) ACTAAGC 0.88
- 3) GATAAACT 0.87
- 4) TAIGGTAT 0.87
- 5) CTAAAGCT 0.87
- 6) AGTATTC 0.84
- 7) ACATTTGAT 0.82
- 8) GTAATACT 0.80
- 9) CTAGCAAT 0.79
- 10) ATAGTTCG 0.78

## \*\*\* Interesting motifs (highest-ranking) seem to be :

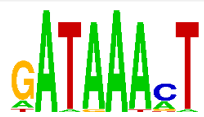
```
GATAAACT
AGTITATC
0 redundant motifs found:
```

Best occurrences (match percentage):

Seq	5' oligo	pos	match
1 +	..AAAAACT	205	(92.85)
1 +	[AAATAAAT]	676	(85.28)
1 +	[ATATAACT]	922	(88.60)
1 -	.ATATAACT	786	(92.79)
1 -	.AATAAACT	697	(92.36)
1 -	[ATATAAT]	169	(85.17)
2 +	[TAAAAACT]	508	(85.63)
2 +	[TATAAAT]	944	(85.73)
2 -	[AAAAAT]	956	(85.28)
2 -	[AAATAAT]	776	(85.28)
2 -	.ATATAACT	652	(90.33)
4 +	[AAATAAAT]	546	(87.13)
4 +	[CAAAAAAT]	788	(85.24)
8 +	[AATAAAT]	393	(85.29)
8 -	[AATAAAT]	260	(85.24)
6 +	[TATAAAT]	733	(87.56)
7 -	.ATATAAAT	430	(94.77)
8 +	[AAATAAAT]	307	(87.13)
8 +	[AATAAAT]	791	(87.13)
8 -	[AAAAACT]	808	(85.19)
8 -	[AATAAAT]	884	(85.28)
8 -	[TATAAAT]	285	(85.24)
8 -	[TATAAAT]	13	(87.56)
9 -	.ATATAAAT	603	(100.00)
9 +	[AATAAAT]	615	(85.24)
9 -	.ATATAAAT	438	(94.77)
10 +	.ATATAAAT	603	(100.00)
10 +	[AATAAAT]	615	(85.24)
10 -	.ATATAAAT	438	(94.77)
11 +	[ATATAAAT]	148	(85.10)
11 +	[AAAAAAT]	205	(85.77)
12 +	.ATATAAAT	143	(92.93)
12 +	.TATAAAT	271	(92.79)
12 +	[AATAAAT]	286	(87.13)
12 +	.ATATAAAT	523	(90.60)
12 +	.ATATAAAT	896	(94.77)
12 -	[AATAAAT]	347	(85.29)
13 +	.AATAAAT	549	(92.36)
13 -	[ATAA-CT]	532	(85.34)
13 -	[ATAA-CT]	577	(88.34)
14 +	.AATAAAT	549	(92.36)
14 -	[ATAA-CT]	532	(85.34)
14 -	[ATAA-CT]	577	(88.34)
16 +	.ATATAAAT	163	(94.77)
16 +	[AGATAACT]	316	(89.17)
16 +	[AATAAAT]	814	(85.29)
16 -	.AATAAAT	947	(92.36)
16 -	.ATATAAAT	943	(90.40)
18 -	[TATAAAT]	637	(89.17)
18 -	.ATATAAAT	555	(94.77)

### Frequency Matrix

	All Occs				Best Occs			
	A	C	G	T	A	C	G	T
1	28	16	167	31	4	0	20	2
2	201	8	17	16	26	0	0	0
3	33	14	19	176	1	0	0	25
4	201	6	21	14	25	0	1	0
5	208	6	9	19	26	0	0	0
6	198	10	13	21	25	0	1	1
7	43	146	25	28	7	16	2	1
8	22	17	5	198	1	0	0	25



# MEME (http://meme.sdsc.edu)

MEME takes as input a group of DNA or protein sequences and outputs as many motifs as requested. MEME uses statistical modeling techniques to automatically choose the best width, number of occurrences, and description for each motif.

Your MEME results consist of:

- your MEME results in HTML format
- your MEME results in XML format
- your MEME results in TEXT format
- and the MAST results of searching your input sequences for the motifs found by MEME using MAST.

Your job id is: **app1254080196482**  
 You can view your job results at: [http://meme.sdsc.edu/meme4\\_1\\_1/cgi-bin/querystatus.cgi?jobid=app1254080196482&service=MEME](http://meme.sdsc.edu/meme4_1_1/cgi-bin/querystatus.cgi?jobid=app1254080196482&service=MEME)  
 You can view server activity [here](#).

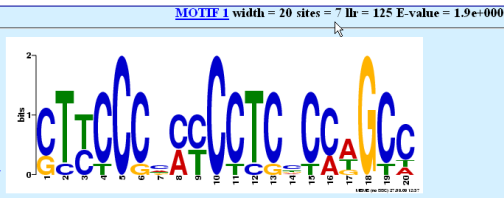
- Sequence file pasted, sequences
- Distribution of motif occurrences: Zero or one per sequence
- Number of different motifs: 20
- Minimum motif width: 5
- Maximum motif width: 20
- Statistics on your dataset

type of sequence	dna
number of sequences	20
shortest sequence (residues)	1000
longest sequence (residues)	1000
average sequence length (residues)	1000.0
total dataset size (residues)	20000

You will also receive a confirming message at your email address: [will.jaggi@chmc.org](mailto:will.jaggi@chmc.org)

# MEME (http://meme.sdsc.edu)

**SEQUENCE LOGO**  
[Information Content](#)  
 24.4 (bits)  
[Relative Entropy](#)  
 25.8 (bits)  
[Download LOGO](#)  
 Without SSC: [EFSIFNGI]  
 With SSC: [EFSIFNGI]



NAME	STRAND	START	P-VALUE	SITES
SERPINA1	+	949	8.38e-12	CTCTGCAAG CTCCCGCCCTCCCGAGCC TACTGCTCC
ADH1B	-	752	3.07e-10	CTTTCCCTCA CTCCCAACCCCGCCGCCC TCTGGAATTC
APOA1	+	650	8.10e-10	CCGACGCTCC CTCCCTCCCTCCTCTGCC AACACAATGG
AMBP	+	927	9.50e-09	AGGCCAGGT GCTCCCATCCTCGCATCC CTCTGTGGGG
SERPINC1	+	979	1.00e-08	TTTGACCTCA GTTCCCTCCCTGACCAAGT C
APOH	+	259	1.09e-08	GACAAACCCC CTTCGAACTCTCTCAAGCA ACAACATCAG
ALDOB	-	262	2.63e-08	AAATCATGT CTCTCCCATCTCTCCAGTC CTCCAAAACC

**Motif 1 block diagrams**

Name	Lowest p-value	Motifs
SERPINA1	8.38e-12	+1
ADH1B	3.07e-10	-1
APOA1	8.10e-10	+1
AMBP	9.50e-09	+1
SERPINC1	1.00e-08	+1
APOH	1.09e-08	+1
ALDOB	2.63e-08	-1

SCALE 1 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775 800 825 850 875 900 925 950 975

**MEME Job app1254080196482**

- [/home/fstall/har/app/meme\\_4.1.1/bin/norm\\_sequences -E-pas](#)

**Results**

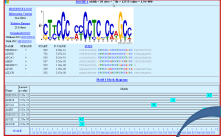
- MEME output as HTML
- MEME output as plain text
- MEME output as XML
- XSLT Stylesheet for converting MEME XML to HTML
- MAST output as HTML
- input sequences

**Messages**

- Processing Messages
- Error Messages



## MEME (<http://meme.sdsc.edu>)



[Motif 1 position-specific probability matrix](#)

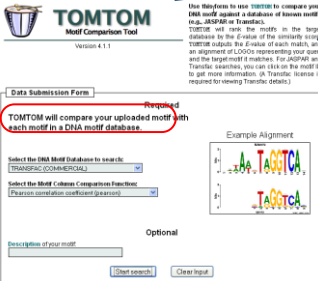
View PSPM 1

FIMO PSPM 1 Scan sequence databases for all matches with this motif using [FIMO](#).

TOMTOM PSPM 1 Compare to known motifs in motif databases using [Tomtom](#).

GOMO PSPM 1 Find Genome Ontology terms associated with upstream regions matching this motif using [GOMO](#).

TOMTOM can be used to find out if an overrepresented motif in your sequences matches or is similar to a known TFBS





**TOMTOM OUTPUT**

Query File: /opt/path/to/meme\_44592614\_99603686/query.fasta

Target File: /opt/path/to/meme\_44592614\_99603686/target.fasta

Distance Measure: jaccard

All Motif Matches with q-value at most: 0.5. The q-value is the estimated false discovery rate of the occurrence is accepted as significant. See Storey JD, Tibshirani R. Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA* (2003) 100:9440-9445

Motif ID	Target Motif	Target Description	Query Motif	Query Description	p-value	q-value	Overlap	Query Offset	Orientation	Figure
M00049	MAZ	query	query	query	0.00015	0.24	0	-8	-	
M00083	MZF1	query	query	query	0.00034	0.17	0	-1	-	

**Exercise 8: Use the downloaded SLC7A1 ortholog promoter sequences to find out common motifs using WeederH**

**Exercise 9: Use the downloaded promoter sequences (from Exercise 4) to find out common motifs using Weeder and MEME**

**Exercise 10: Does any of the motifs found by Meme match known TFBS?**

I have found a miRNA enriched in my gene list or I am interested in a specific gene and I want to identify putative regulatory regions for miRNA/gene

GenomeTrafac: <http://genometrafac.cchmc.org>

**GenomeTraFaC**

A comparative genomics-based resource for initial characterization of gene models and the identification of putative cis-regulatory regions of RefSeq Gene Orthologs

- [Cis-element clusters within BlastZ Alignments](#)  
Find conserved cis-element clusters within BlastZ-identified conserved sequence alignment blocks.
- [Cis-elements shared between any gene pair](#)  
Find shared cis-elements between user-selected gene segment pairs.
- [Conserved Cis-Element Scanner](#)  
Genome-wide ortholog conserved Cis-element module search

**Note:** If you publish results obtained using GenomeTrafac, please cite  
[Jegga et al., Nucleic Acids Res. 2006 Dec 18; \[Epub ahead of print\]](#)  
 OR  
[Jegga et al., Genome Research 12: 1408-1417, September 2002](#)

GenomeTrafac: <http://genometrafac.cchmc.org>

**Basic Search**

Description ▼

mir-122a

---

**Search by disease, gene ontology, pathway, gene family, or custom groups**

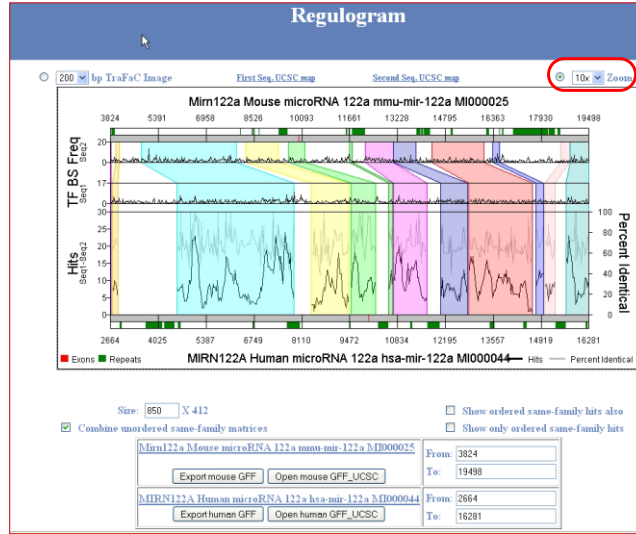
Select	Query
<input type="radio"/>	Disease <i>(Always use <a href="#">Disease Selector</a>)</i>
<input type="radio"/>	Pathway <i>(Always use <a href="#">Pathway Selector</a>)</i>
<input type="radio"/>	Gene ontology <i>(Always use <a href="#">Dntology Selector</a>)</i>
<input type="radio"/>	Mammalian phenotype <i>(Always use <a href="#">Phenotype Selector</a>)</i>
<input type="radio"/>	Select gene family from the list
<input type="radio"/>	Select custom group from the list

Query took 1.514 s  
(2 genes meet the search criteria)

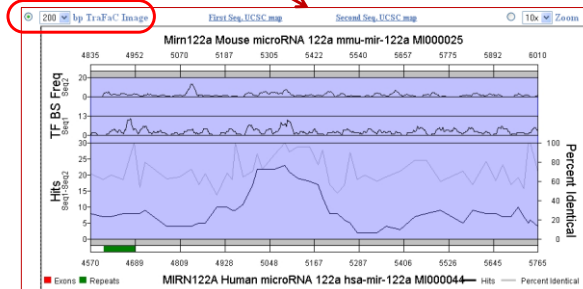
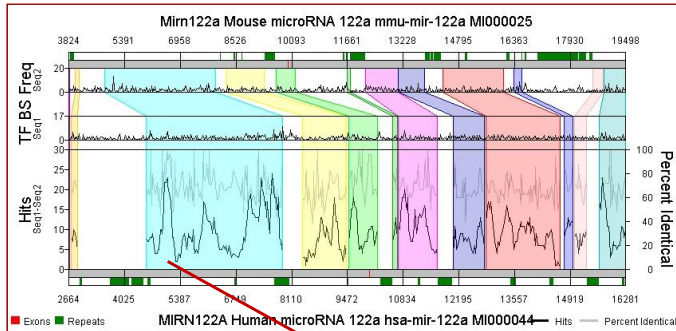
✓	Query Term	Accession Number	Name
<input checked="" type="checkbox"/>	MIR-122A	hgMIRN122A	MIRN122A Human microRNA 122a hsa-mir-122a MI000044
<input checked="" type="checkbox"/>	MIR-122A	mgMim122a	Mim122a Mouse microRNA 122a mma-mir-122a MI000025

# GenomeTrafac: <http://genometrafac.cchmc.org>

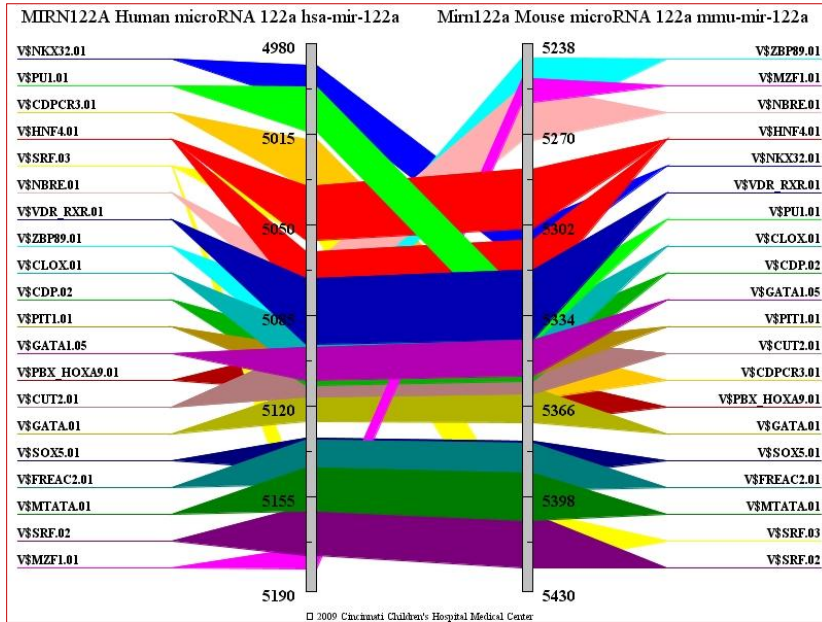
First Sequence	Second Sequence	Timestamp	Action
hgMIRN122A, MIRN122A Human microRNA 122a hsa-mir-122a MI000044	mmMir122a, Mirn122a Mouse microRNA 122a mmu-mir-122a MI000025	07/25/2006 12:00	<a href="#">View</a> <a href="#">Regulogram</a>



# GenomeTrafac: <http://genometrafac.cchmc.org>



## GenomeTrafac: <http://genometrafac.cchmc.org>



## GenomeTrafac: <http://genometrafac.cchmc.org>

### Shared *Cis*-elements

(Genomatix Matrix Family Library Version 5.0 (January 2005))  
 (For details and annotations of TFBS-PWMs, please register at [Genomatix](http://Genomatix))

Family/Matrix	Description	hgMIRN122A			mmuMir122a			
		Begin	End	Sequence	Begin	End	Sequence	
<a href="#">V\$NKXH/V\$NKX32.01</a>	Homeo-domain protein NKX3.2 (BAPX1, NKX3B, Bagpipe homolog)	4993	5007	CCCCACTCAGCAGA	-	5301 5315	CTGACTTAGTGGACT	+
<a href="#">V\$SETSF/V\$PUL1.01</a>	Fu 1 (Pu120) Ets-like transcription factor identified in lymphoid B-cells	5001	5017	CAGCAGAGGAATGGACT	+	5326 5342	CCTCTCTCCCCACAA	-
<a href="#">V\$CLOX/V\$CDPCR3.01</a>	Ctcf-like homeo-domain protein	5020	5038	CCAATCTTGCTGAGTGTGT	-	5343 5361	TCGATAATTTAATGTGACT	-
<a href="#">V\$HNF4/V\$HNF4.01</a>	Hepatic nuclear factor 4	5037	5057	GTTTGACCAAAGGTGGTCTG	+	5283 5303	GTTTGACCAAAGGTGACTCTG	+
<a href="#">V\$SRF/V\$SRF.03</a>	Serum responsive factor	5038	5056	TTTGACCAAAGGTGGTCT	-	5399 5417	GGATCCATAAAGGGAGAG	-
<a href="#">V\$HNF4/V\$HNF4.01</a>	Hepatic nuclear factor 4	5061	5081	TAGTGGCCTAAGTCTGTGCC	+	5307 5327	TAGTGGACTAAGTCTGTGCC	+
<a href="#">V\$RORA/V\$NBRE.01</a>	Monomers of the nuclear subfamily of nuclear receptors (nurr77, nurr1, nor-1)	5065	5083	GGCCTAAGTCTGTGCCCTC	+	5255 5273	GGGAGCTGGACCTTCGGTIT	-
<a href="#">V\$RXRF/V\$VDR_RXR.01</a>	VDR/RXR Vitamin D receptor RXR heterodimer site	5071	5095	AGGTGTGGCCTCCCTCCCCACTG	-	5317 5341	AGGTGATGCCCTCTCTCCCCACA	-
<a href="#">V\$ZBP/V\$ZBP89.01</a>	Zinc finger transcription factor ZBP-89	5077	5099	TGCCCTCCCTCCCCACTGAATC	+	5245 5267	GGGGCATGGGGGAGCTGGACCT	-
<a href="#">V\$CLOX/V\$CLOX.01</a>	Cloz	5089	5107	CCCACTGAATCGATAAATA	+	5334 5352	CCCCACAATCGATAAATT	+

## I have a list of co-expressed mRNAs (Transcriptome)....

### Now what?

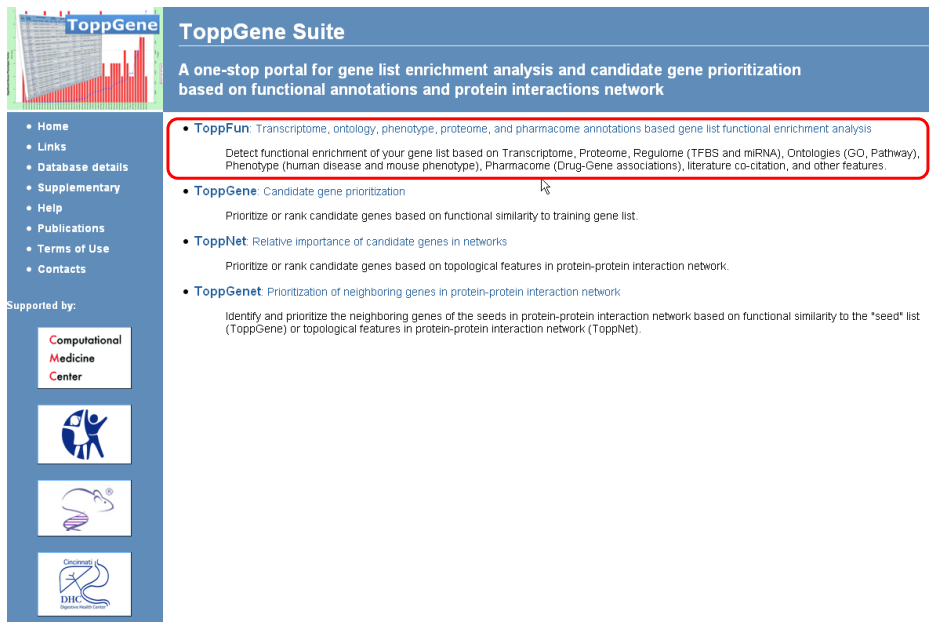
#### 1. Identify putative shared regulatory elements

- Known transcription factor binding sites (TFBS)
  - Conserved
  - Non-conserved
- Unknown TFBS or Novel motifs
  - Conserved
  - Non-conserved
- MicroRNAs

#### 2. Identify the underlying biological theme

- Gene Ontology
- Pathways
- Phenotype/Disease Association
- Protein Domains
- Protein Interactions
- Expression in other tissues/experiments
- Drug targets
- Literature co-citation...

## Toppgene Suite (<http://toppgene.cchmc.org>)



**Toppgene Suite**

A one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network

- **Toppgene**: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis  
Detect functional enrichment of your gene list based on Transcriptome, Proteome, Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacome (Drug-Gene associations), literature co-citation, and other features.
- **Toppgene**: Candidate gene prioritization  
Prioritize or rank candidate genes based on functional similarity to training gene list.
- **Toppgene**: Relative importance of candidate genes in networks  
Prioritize or rank candidate genes based on topological features in protein-protein interaction network.
- **Toppgene**: Prioritization of neighboring genes in protein-protein interaction network  
Identify and prioritize the neighboring genes of the seeds in protein-protein interaction network based on functional similarity to the "seed" list (Toppgene) or topological features in protein-protein interaction network (Toppgene).

Supported by:

Computational  
Medicine  
Center

Cincinnati Children's Hospital Medical Center

## TopGene Suite (<http://toppgene.cchmc.org>)

TopFun: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis

Select your gene identifier type, paste your sets below or select example set, then submit.

Entry Type:

Example gene sets:    
(click on "HGNC Symbol" or "Entrez ID" to use the example training and test set of genes)

Training Gene Set:

259  
5265  
350  
335  
335  
1558  
1571  
229  
462  
125  
3240  
5105  
5265  
3273  
2244  
2158  
5053  
125  
1356  
3827  
383

Clear

Submit Query

Input Gene List (81 / 97)

Entered	Human Symbol	Gene ID
259	AMBP	259
5265	SERPINA1	5265
350	APOH	350
335	APOA1	335
1558	CYP2C8	1558
1571	CYP2E1	1571
229	ALDOB	229
462	SERPINC1	462
125	ADH1B	125
3240	HP	3240
5105	PCK1	5105
3273	HRG	3273
2244	FGF	2244
2158	F9	2158
5053	PAH	5053
1356	CP	1356
3827	KNG1	3827
383	ARG1	383
5004	ORM1	5004
2168	FABP1	2168
325	APCS	325

Genes Not Found

Entered	Status
335	Duplicated
5265	Duplicated
125	Duplicated
1571	Duplicated
1571	Duplicated
1373	Duplicated
1356	Duplicated
462	Duplicated
2243	Duplicated
3827	Duplicated
125	Duplicated
229	Duplicated
8822	Duplicated
2328	Duplicated

1. Supports variety of inputs
2. Supports symbol correction
3. Eliminates any duplicates
4. Drawback: Supports human and mouse genes only

## TopGene Suite (<http://toppgene.cchmc.org>)

Calculations

Feature	Correction	p-Value cutoff	Gene Limits
<input checked="" type="checkbox"/> All	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> GO: Molecular Function	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> GO: Biological Process	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> GO: Cellular Component	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Human Phenotype	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Mouse Phenotype	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Domain	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Pathway	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Pubmed	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Interaction	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Cytoband	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> TFBS	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Gene Family	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Coexpression	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Computational	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> MicroRNA	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Drug	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Disease	Bonferroni	0.05	1 ≤ n ≤ 1500

Home

Modify Query

Submit

1. Gene list analyzed for as many as 17 features!
2. Single-stop enrichment analysis server for both regulatory elements (TFBSs and miRNA) and biological themes
3. Back-end has an exhaustive, normalized data resources compiled and integrated
4. Bonferroni correction is "too stringent"; FDR with 0.05 is preferable.
5. TFBS are based on conserved cis-elements and motifs within ±2kb region of TSS in human, mouse, rat, and dog.
6. miRNA-targets are based on TargetScan and PicTar

## TopGene Suite (http://toppgene.cchmc.org)

<b>GO Biological Process</b>	<b>Human Phenotype</b>	<b>Mouse Phenotype</b>																																																
Annotations: 16,372 Genes: 15,079 <small>Updated Aug 26, 2009</small>	Annotations: 9,551 Genes: 2,531 <small>Updated Sep 10, 2009</small>	Annotations: 6,203 Genes: 5,690 <small>Updated Aug 25, 2009</small>																																																
<b>GO Cellular Component</b>																																																		
Annotations: 2,326 Genes: 16,726 <small>Updated Aug 26, 2009</small>																																																		
<b>GO Molecular Function</b>																																																		
Annotations: 8,593 Genes: 15,948 <small>Updated Aug 26, 2009</small>																																																		
<b>Pathways</b>	<b>Domains</b>	<b>Pubmed</b>																																																
Annotations: 1,672 Genes: 164 <small>Aug 25, 2009</small>	Annotations: 10,223 Genes: 286 <small>Updated Sep 10, 2009</small>	Annotations: 221,282 Genes: 22,176 <small>Updated Aug 25, 2009</small>																																																
<table border="1"> <tr><td>Aug 25, 2009</td><td>BioCyc</td><td>314</td></tr> <tr><td></td><td>CGAP BioCarta</td><td>57</td></tr> <tr><td></td><td>GenMAP</td><td>202</td></tr> <tr><td>Apr 15, 2009</td><td>KEGG pathway</td><td>431</td></tr> <tr><td>May 10, 2009</td><td>MSigDB</td><td>150</td></tr> <tr><td>Aug 25, 2009</td><td>PantherDB</td><td>336</td></tr> <tr><td></td><td>Pathway Ontology</td><td>35</td></tr> <tr><td></td><td>Reactome</td><td>2</td></tr> <tr><td></td><td>SigAldrich</td><td>11</td></tr> <tr><td></td><td>Signaling Transduction KE</td><td>6,897</td></tr> </table>	Aug 25, 2009	BioCyc	314		CGAP BioCarta	57		GenMAP	202	Apr 15, 2009	KEGG pathway	431	May 10, 2009	MSigDB	150	Aug 25, 2009	PantherDB	336		Pathway Ontology	35		Reactome	2		SigAldrich	11		Signaling Transduction KE	6,897	<table border="1"> <tr><td></td><td>Gene3D</td><td>4,959</td></tr> <tr><td></td><td>InterPro</td><td>1,351</td></tr> <tr><td></td><td>PROSITE</td><td>2,774</td></tr> <tr><td></td><td>Plan</td><td>365</td></tr> <tr><td></td><td>ProDom</td><td>669</td></tr> <tr><td></td><td>SMART</td><td>12,430</td></tr> </table>		Gene3D	4,959		InterPro	1,351		PROSITE	2,774		Plan	365		ProDom	669		SMART	12,430	
Aug 25, 2009	BioCyc	314																																																
	CGAP BioCarta	57																																																
	GenMAP	202																																																
Apr 15, 2009	KEGG pathway	431																																																
May 10, 2009	MSigDB	150																																																
Aug 25, 2009	PantherDB	336																																																
	Pathway Ontology	35																																																
	Reactome	2																																																
	SigAldrich	11																																																
	Signaling Transduction KE	6,897																																																
	Gene3D	4,959																																																
	InterPro	1,351																																																
	PROSITE	2,774																																																
	Plan	365																																																
	ProDom	669																																																
	SMART	12,430																																																
<b>Interactions</b>	<b>Cytoband</b>	<b>TFBS</b>																																																
Annotations: 19,047 Genes: 4,370	Annotations: 362 Genes: 29,821	Annotations: 615 Genes: 9,770																																																
<b>miRNA</b>	<b>Gene Families</b>	<b>Coexpression</b>																																																
Annotations: 740 Genes: 313	Annotations: 151 Genes: 6,098	Annotations: 1,203 Genes: 23																																																
<b>Computational Gene Set</b>	<b>Drugs</b>	<b>Disease</b>																																																
Annotations: 427 Genes: 4,712	Annotations: 13,141 Genes: 4,977	Annotations: 3,789 Genes: 1,008																																																
<b>Master Gene Info File</b>																																																		
For All Annotations <small>Updated Aug 26, 2009</small>																																																		

1. Database updated regularly
2. Exhaustive collection of annotations

## TopGene Suite (http://toppgene.cchmc.org)

**Results**

Go To Start Page

**Input Parameters** [show Detail]

**Training Results** [show #] [Download #] [Sparse Matrix]

- 1: GO: Molecular Function [Display Chart] [Show Detail]
- 2: GO: Biological Process [Display Chart] [Show Detail]
- 3: GO: Cellular Component [Display Chart] [Show Detail]
- 4: Human Phenotype [Display Chart] [Show Detail]
- 5: Mouse Phenotype [Display Chart] [Show Detail]
- 6: Domain [Display Chart] [Show Detail]
- 7: Pathway [Display Chart] [Show Detail]
- 8: Pubmed [Display Chart] [Show Detail]
- 9: Interaction [Display Chart] [Show Detail]
- 10: Cytoband [Display Chart] [Show Detail]
- 11: TFBS [Display Chart] [Show Detail]
- 12: Gene Family [Display Chart] [Show Detail]
- 13: Coexpression [Display Chart] [Show Detail]
- 14: Computational [Display Chart] [Show Detail]
- 15: MicroRNA [Display Chart] [Show Detail]
- 16: Drug [Display Chart] [Show Detail]
- 17: Disease [Display Chart] [Show Detail]

**Input Parameters** [Hide Detail]

Number of genes in training set: 81

Number of genes in test set: 0

category	Correction	Cutoff	Min	Max
GO: Molecular Function	Bonferroni	0.05	1	1500
GO: Biological Process	Bonferroni	0.05	1	1500
GO: Cellular Component	Bonferroni	0.05	1	1500
Human Phenotype	Bonferroni	0.05	1	1500
Mouse Phenotype	Bonferroni	0.05	1	1500
Domain	Bonferroni	0.05	1	1500
Pathway	Bonferroni	0.05	1	1500
Pubmed	Bonferroni	0.05	1	1500
Interaction	Bonferroni	0.05	1	1500
Cytoband	Bonferroni	0.05	1	1500
TFBS	Bonferroni	0.05	1	1500
Gene Family	Bonferroni	0.05	1	1500
Coexpression	Bonferroni	0.05	1	1500
Computational	Bonferroni	0.05	1	1500
MicroRNA	Bonferroni	0.05	1	1500
Drug	Bonferroni	0.05	1	1500
Disease	Bonferroni	0.05	1	1500

Random sampling size in analysis: 0

Minimum feature count in test set: 2

Analysis tool: 2 seconds

Analysis finished at: Sun Sep 27 16:45:06 EDT 2009

**2: GO: Biological Process** [Display Chart] [Hide Detail]

ID	Name	Source	P-value	Term In Query	Term In Genome
1	GO:0009605 response to external stimulus		0	27	893
2	GO:0007518 blood coagulation		0	12	115
3	GO:0006229 lipid metabolic process		0	25	874
4	GO:0044255 cellular lipid metabolic process		0	23	720
5	GO:0050817 coagulation		0	12	119
6	GO:0007599 hemostasis		0	12	120
7	GO:0009611 response to wounding		0	20	542
8	GO:0042060 wound healing		0	13	185
9	GO:0050878 regulation of body fluid levels		0	12	151
10	GO:0055114 oxidation reduction		0	19	624
11	GO:0019752 carboxylic acid metabolic process		0	18	570

# ToppGene Suite (http://toppgene.cchmc.org)

2 GO: Biological Process [Display Chart] [Hide Detail]

ID	Name	Source	P-value	Term in Query	Term in Genome
1	GO:0008605	response to external stimulus	0	27	893
2	GO:0007544	blood coagulation	0	12	115
3	GO:0006629	lipid metabolic process	0	25	874
4	GO:0044255	cellular lipid metabolic process	0	23	720
5	GO:0050817	coagulation	0	12	119
6	GO:0007599	hemostasis	0	12	115
7	GO:0008611	response to wounding	0	20	542
8	GO:0042060	wound healing	0	13	185
9	GO:0050878	regulation of body fluid levels	0	12	151
10	GO:0055114	oxidation reduction	0	19	624
11	GO:0019752	carboxylic acid metabolic process	0	18	570

**Significant Terms For: Pathway**

**Entrez Gene ID Gene Symbol Gene Name Original Symbol**

Entrez Gene ID	Gene Symbol	Gene Name	Original Symbol
1128	ADH1C	alcohol dehydrogenase 1C (class I), gamma polypeptide	1128
335	APOA1	apolipoprotein A-I	335
380	APH1	apolipoprotein H (beta-2-glycoprotein I)	380
2158	F9	coagulation factor IX	2158
5950	REB4	retinol binding protein 4, plasma	5950
197	AHSG	alpha-2-HS-glycoprotein	197
2243	FGA	fibrinogen alpha chain	2243
213	ALB	albumin	213
2244	FBG	fibrinogen beta chain	2244
629	CFB	complement factor B	629
3158	HMOCS2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)	3158
5444	PCN1	paraoxonase 1	5444
1381	CFB2	carboxypeptidase B2 (plasma)	1381
3078	CFHR1	complement factor H-related 1	3078
5285	SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antitrypsin, antitrypsin), member 1	5285
3827	KMO1	kininogen 1	3827
325	APCS	amyloid P component, serum	325
2538	GSPC	glucose-6-phosphatase, catalytic subunit	2538
4153	MBL2	mannose-binding lectin (protein-C) 2, soluble (epsonic defect)	4153
735	CS9	complement component 9	735
462	SERPINC1	serpin peptidase inhibitor, clade C (antithrombin), member 1	462
3273	HRR	histidine-rich glycoprotein	3273
5340	PLG	plasminogen	5340
5004	ORM1	ornithinec1	5004
316	ALOX1	aldehyde oxidase 1	316
3053	SERPIND1	serpin peptidase inhibitor, clade D (heparin cofactor), member 1	3053
1356	CP	ceruloplasmin (ferroxidase)	1356

# ToppGene Suite (http://toppgene.cchmc.org)

**ToppGene Result Page**

Number of genes in training set: 351

Number of genes in test set: 0

category	Correction	Cutoff	Min	Max
GO: Molecular Function	Bonferroni	0.05	1	1500
GO: Biological Process	Bonferroni	0.05	1	1500
GO: Cellular Component	Bonferroni	0.05	1	1500
Human Phenotype	Bonferroni	0.05	1	1500
Mouse Phenotype	Bonferroni	0.05	1	1500
Domain	Bonferroni	0.05	1	1500
Pathway	Bonferroni	0.05	1	1500
Pubmed	Bonferroni	0.05	1	1500
Interaction	Bonferroni	0.05	1	1500
Cytoband	Bonferroni	0.05	1	1500
TFBS	Bonferroni	0.05	1	1500
Oeone Family	Bonferroni	0.05	1	1500
Congression	Bonferroni	0.05	1	1500
Computational	Bonferroni	0.05	1	1500
MicroRNA	Bonferroni	0.05	1	1500
Drug	Bonferroni	0.05	1	1500
Disease	Bonferroni	0.05	1	1500

Correction and Cutoff:

Random sampling size in analysis: 0

Minimum feature count in test set: 2

Analysis took: 2 seconds

Analysis finished at: Sun Sep 27 16:45:06 EDT 2009

**Training Results** [Show All] [Download] [Sparse Matrix]

- 1: GO: Molecular Function [Display Chart] [Show Detail]
- 2: GO: Biological Process [Display Chart] [Show Detail]
- 3: GO: Cellular Component [Display Chart] [Show Detail]
- 4: Human Phenotype [Display Chart] [Show Detail]
- 5: Mouse Phenotype [Display Chart] [Show Detail]
- 6: Domain [Display Chart] [Show Detail]
- 7: Pathway [Display Chart] [Show Detail]
- 8: Pubmed [Display Chart] [Show Detail]

Enter name of file to save to...

Save in: Desktop

File name: LiverGenes\_TopppUn.txt

Save as type: Text Document



## ToppGene Suite (<http://toppgene.cchmc.org>)

I have a list of 200 over-expressed genes and I want to prioritize them for experimental validation (apart from using the fold change as a parameter).....

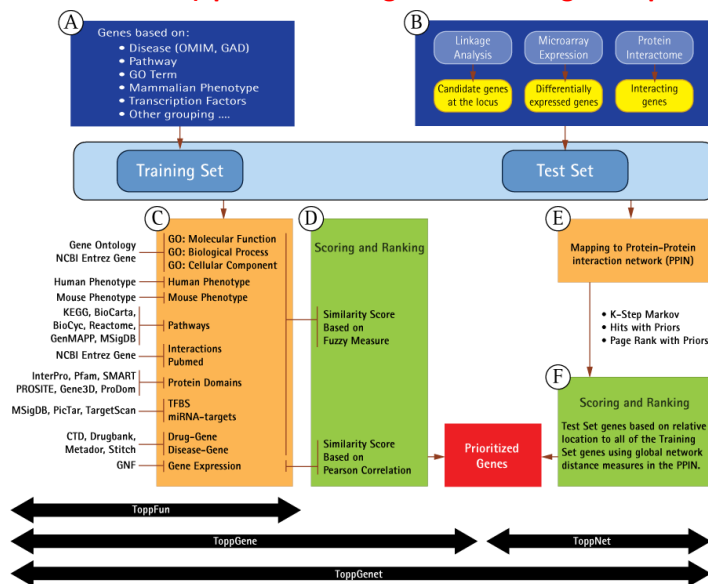
### ToppGene Suite

A one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network

- **ToppFun**: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis  
Detect functional enrichment of your gene list based on Transcriptome, Proteome, Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacome (Drug-Gene associations), literature co-citation, and other features.
- **ToppGene**: Candidate gene prioritization  
Prioritize or rank candidate genes based on functional similarity to training gene list.
- **ToppNet**: Relative importance of candidate genes in networks  
Prioritize or rank candidate genes based on topological features in protein-protein interaction network.
- **ToppGenet**: Prioritization of neighboring genes in protein-protein interaction network  
Identify and prioritize the neighboring genes of the seeds in protein-protein interaction network based on functional similarity to the "seed" list (ToppGene) or topological features in protein-protein interaction network (ToppNet).

## ToppGene Suite (<http://toppgene.cchmc.org>)

I have a list of 200 over-expressed genes and I want to prioritize them for experimental validation (apart from using the fold change as a parameter).....



# TopGene Suite (<http://toppgene.cchmc.org>)

## TopGene: Candidate gene prioritization

Select your gene identifier type, paste your training and test gene sets below or select example sets, then submit.

Example gene sets: [HGNC Symbol](#) [Entrez ID](#)  
(click on "HGNC Symbol" or "Entrez ID" to use the example training and test set of genes)

Symbol Types: HGNC Symbol HGNC Symbol

Training Gene Set: NKX2-5  
MEF2A  
GATA4  
HAND1  
HAND2  
TBX5  
SRF

Test gene set: ACVR1  
ACVR2B  
ADAM19  
ADM  
ADRA1A  
ADRA1B  
ADRBK1  
ALDH1A2  
ALPK3  
ATP6V0A1  
BMP10  
BMP2  
BMP4  
BMPR1A  
CALCRL  
CASP3  
CASP7  
CASP8  
CASQ2  
CENTA2  
CHD7

Clear Submit Query

# TopGene Suite (<http://toppgene.cchmc.org>)

Training set (7 / 7)			Test set (146 / 158)		
Entered	Human Symbol	Gene ID	Entered	Human Symbol	Gene ID
NKX2-5	NKX2-5	1482	ACVR1	ACVR1	90
MEF2A	MEF2A	4205	ACVR2B	ACVR2B	93
GATA4	GATA4	2626	ADAM19	ADAM19	8728
HAND1	HAND1	9421	ADM	ADM	133
HAND2	HAND2	9464	ADRA1A	ADRA1A	148
TBX5	TBX5	6910	ADRA1B	ADRA1B	147
SRF	SRF	6722	ADRBK1	ADRBK1	156
			ALDH1A2	ALDH1A2	8854
			ALPK3	ALPK3	57538
			ATP6V0A1	ATP6V0A1	535
			BMP10	BMP10	27302
			BMP2	BMP2	650
			BMP4	BMP4	652
			BMPR1A	BMPR1A	667
			CALCRL	CALCRL	10203
			CASP3	CASP3	836
			CASP7	CASP7	840
			CASP8	CASP8	841
			CASQ2	CASQ2	845
			CHD7	CHD7	55636
			CITED2	CITED2	10370

Entered	Suggestions
CENTA2	<input checked="" type="checkbox"/> ADAP2 - ArfGAP with dual PH domains 2 Human Synonym
CMYA1	<input checked="" type="checkbox"/> XIRP1 - xin actin-binding repeat containing 1 Human Synonym
GJA7	<input checked="" type="checkbox"/> GJC1 - gap junction protein, gamma 1, 45kDa Human Synonym
HOP	<input checked="" type="checkbox"/> HOPX - HOP homeobox Human Synonym
	<input type="checkbox"/> ST13 - suppression of tumorigenicity 13 (colon carcinoma) (Hsp70 interacting protein) Human Synonym
	<input type="checkbox"/> STIP1 - stress-induced-phosphoprotein 1 Human Synonym
PPARBP	<input checked="" type="checkbox"/> MED1 - mediator complex subunit 1 Human Synonym
RBPSUH	RBPU Duplicated

Update  Check All

Entered	Status
CENTA2	Not Found
CMYA1	Not Found
GATA4	In Training Set
GJA7	Not Found
HAND1	In Training Set
HAND2	In Training Set
HOP	Not Found
NKX2-5	In Training Set
PPARBP	Not Found
RBPSUH	Not Found
SRF	In Training Set
TBX5	In Training Set

[Find alternatives for missing symbols](#)

## ToppGene Suite (<http://toppgene.cchmc.org>)

Training parameters

Feature	Correction	p-Value cutoff	Gene Limits
<input checked="" type="checkbox"/> All	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> GO: Molecular Function	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> GO: Biological Process	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> GO: Cellular Component	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Human Phenotype	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Mouse Phenotype	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Domain	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Pathway	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Pubmed	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Interaction	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Cytoband	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> TFBS	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Gene Family	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Coexpression	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Computational	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> MicroRNA	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Drug	Bonferroni	0.05	1 ≤ n ≤ 1500
<input checked="" type="checkbox"/> Disease	Bonferroni	0.05	1 ≤ n ≤ 1500

Test parameter

Random sampling size: 1500 (6% of genome)  
 Min. feature count: 2

[Home](#)

[Modify Query](#)

[Start prioritization](#)

ToppGene is processing your query

Estimating p-Values  
 To see the training results before the test set is complete, click here.

## ToppGene Suite (<http://toppgene.cchmc.org>)

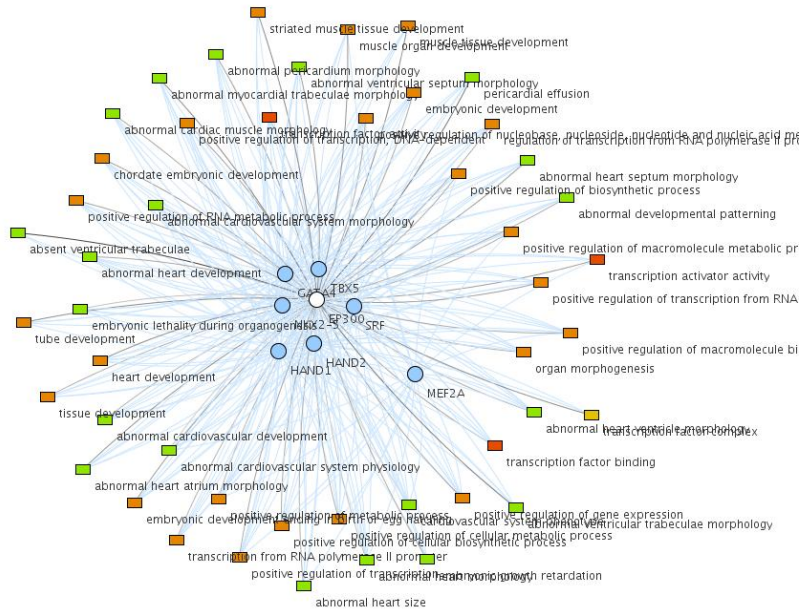
Rank	Gene Symbol	Gene ID	GO: Molecular Function		GO: Biological Process		GO: Cellular Component		Human Phenotype		Mouse Phenotype		Domain		Pathway		Pubmed		Interaction		Cytoband		
			Score	p-value	Score	p-value	Score	p-value	Score	p-value	Score	p-value	Score	p-value	Score	p-value	Score	p-value	Score	p-value	Score	p-value	
1	EP300	2033	0.713689	0.023541	0.999726	0.002945	0.4305914	0.001	0	0.5	0.9999891	0.005											
2	TEAD1	7003	0.5504123	0.009270	0.997771	0.0103093	0.4305914	0.001	0	0.5	0.9999337	0.035	0	0.5									
3	HIF1A	3091	0.3291513	0.0029455	1	0.001	0.4305914	0.001			0.9997067	0.02	0.8885697	0.001									
4	CTNNB1	1494	0.713689	0.023541	1	0.001	0.4305914	0.001	0	0.5	0.9529218	0.06			0.6371253	0.001							
5	TBX20	57957	0.5504123	0.009270	0.9999964	0.0014726	0	0.5022091			0.9999902	0.005		0.5									
6	ZFPM2	23414	0.6308709	0.0250368	0.9999978	0.001	0	0.5022091	0	0.5	1	0.001	0.5	0.5									
7	BMP4	662	0.5493373	0.001	0.999997	0.001	0	0.5022091	0	0.5	0.9999435	0.01	0.5	0.5	0.6478057	0.001	0.7993714	0.001			0	0.505	0
8	TBX1	699	0.9660807	0.001	0.999997	0.001	0	0.5022091	0	0.5	0.9996966	0.02	0.5	0.5									
9	TBX2	6906	0.5418852	0.0397644	0.9943991	0.0162003	0.4305914	0.001			0.9993508	0.035	0.5	0.5	0	0.5049834		0	0.4995093				0
10	TGFB2	7042	0.8603852	0.009891	1	0.001	0	0.5022091			0.9999998	0.005		0.5	0.3937178	0.0033223		0	0.4995093				0

Rank	Gene Symbol
1	EP300
2	TEAD1
3	HIF1A
4	CTNNB1
5	TBX20
6	ZFPM2
7	BMP4
8	TBX1
9	TBX2
10	TGFB2

Average score	Overall P-value
0.3417445	0.0000003
0.3015437	0.0000058
0.3041435	0.0000062
0.2489552	0.0000788
0.3207447	0.0000893
0.2749466	0.000112
0.229608	0.0001787
0.2395215	0.0002528
0.2566618	0.0002615
0.2503156	0.0002619
0.3307561	0.0002975

## TopGene Suite (<http://toppgene.cchmc.org>)

### Why is a test set gene ranked higher?



## TopGene Suite (<http://toppgene.cchmc.org>)

I have a list of 200 over-expressed genes and I want to prioritize them for experimental validation (apart from using the fold change as a parameter)....

### TopGene Suite

A one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network

- **TopFun**: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based gene list functional enrichment analysis  
Detect functional enrichment of your gene list based on Transcriptome, Proteome, Regulome (TFBS and miRNA), Ontologies (GO, Pathway), Phenotype (human disease and mouse phenotype), Pharmacome (Drug-Genes associations), literature co-citation, and other features.
- **TopGene**: Candidate gene prioritization  
Prioritize or rank candidate genes based on functional similarity to training gene list.
- **TopNet**: Relative importance of candidate genes in networks  
Prioritize or rank candidate genes based on topological features in protein-protein interaction network.
- **TopGenet**: Prioritization of neighboring genes in protein-protein interaction network  
Identify and prioritize the neighboring genes of the seeds in protein-protein interaction network based on functional similarity to the "seed" list (TopGene) or topological features in protein-protein interaction network (TopNet).

## ToppGene Suite (<http://toppgene.cchmc.org>)

Graph prioritization parameters

Prioritization method: k-Step Markov

Step Size (normally 4-8): 6

Training gene neighborhood subnetwork visualization parameters

Neighborhood distance: 1

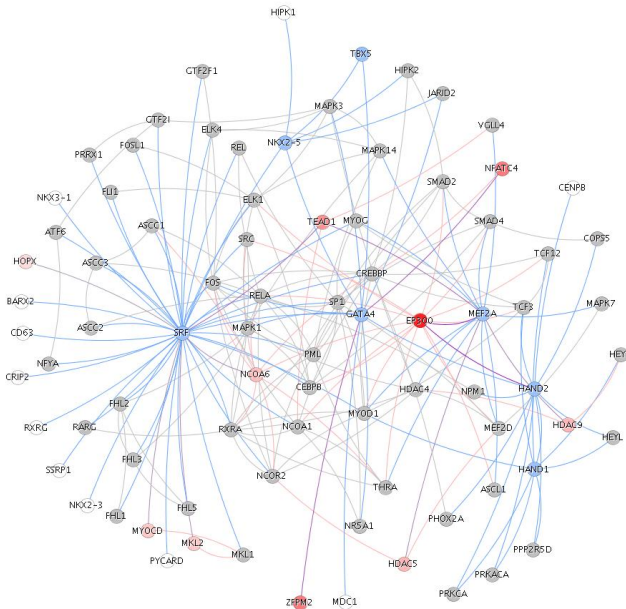
When training set is big, the training gene neighborhood subnetwork can be huge.

Home
Modify Query
Start prioritization

### Test Genes [Hide All]

Rank	ID	Name	Interactant count	Score
1	2033	EP300	129	0.008192
2	23414	ZFPM2	4	0.004724
3	4776	NFATC4	4	0.004615
4	7003	TEAD1	9	0.003739
5	9734	HDAC9	20	0.002319
6	10014	HDAC5	33	0.002317
7	23054	NCOA6	49	0.001991
8	93649	MYOCD	2	0.0016
9	57496	MKL2	2	0.0016
10	1499	CTNND1	138	0.001546

## ToppGene Suite (<http://toppgene.cchmc.org>)



### Exercise 11: Use the gene list from the downloaded file (“Example-Set-2”) and find out:

- How many of these genes are transcription factors?
- What are the enriched TFBSs and miRNAs?
- What gene families are enriched in this list?
- Are there any salivary gland development associated genes present in this list?
- How many and which genes from this list are associated with non-insulin dependent diabetes mellitus (NIDDM)?

### Exercise 12: Prioritize the 721 genes (“Example-Set-2”) using “stomach genes” from the “Example-Set-1”.

- What are the top 10 ranked genes using ToppGene and ToppNet?
- Why is TFF3 ranked among the top 5 in ToppGene prioritization? What is its rank in ToppNet?

### What if I want to compare several gene lists at a time?

#### ToppCluster (<http://toppcluster.cchmc.org>)

The screenshot shows the ToppCluster web application interface. On the left, there is a navigation sidebar with categories: Navigation (Main, Alternative Entry Methods, Cluster Dataset), Information (Disclaimer, ToppGene), and Support (Documentation). The main content area is titled "Paste Input list" and "Load Sample Data". It features a dropdown menu for "Symbols are" (currently set to "HGNC Symbol") and a text input field for "Cluster Name" containing "Cluster 1". A "remove" button is located below the input field. To the right of the input fields is a large, empty rectangular area labeled "Genes". At the bottom of the interface, there are two buttons: "Add Cluster" and "Next".

# TopCluster (http://toppcluster.cchmc.org)

**Paste input list** Load Sample Data

Symbols are Entrez ID  
Cluster Name Liver

remove

- 259
- 2565
- 350
- 335
- 335
- 1558
- 1571
- 229
- 462
- 125
- 3240
- 5105
- 5265
- 3273
- 2244
- 2158
- 5053
- 125
- 1356
- 3827
- 389
- 5004
- 2168
- 1571
- 325
- 5950
- 127
- 1373
- 213
- 735

Symbols are Entrez ID  
Cluster Name Salivary\_Glands

remove

- 64065
- 3855
- 100129410
- 64065
- 245973

Options

Feature	Correction	pValue cutoff	Gene Limits
All	Bonferoni	0.05	1 ≤ n ≤ 1500
GO: Molecular Function	Bonferoni	0.05	1 ≤ n ≤ 1500
GO: Biological Process	Bonferoni	0.05	1 ≤ n ≤ 1500
GO: Cellular Component	Bonferoni	0.05	1 ≤ n ≤ 1500
Human Phenotype	Bonferoni	0.05	1 ≤ n ≤ 1500
Mouse Phenotype	Bonferoni	0.05	1 ≤ n ≤ 1500
Domain	Bonferoni	0.05	1 ≤ n ≤ 1500
Pathway	Bonferoni	0.05	1 ≤ n ≤ 1500
Pubmed	Bonferoni	0.05	1 ≤ n ≤ 1500
Interaction	Bonferoni	0.05	1 ≤ n ≤ 1500
Cytoband	Bonferoni	0.05	1 ≤ n ≤ 1500
TFBS	Bonferoni	0.05	1 ≤ n ≤ 1500
Coexpression	Bonferoni	0.05	1 ≤ n ≤ 1500
Computational	Bonferoni	0.05	1 ≤ n ≤ 1500
MicroRNA	Bonferoni	0.05	1 ≤ n ≤ 1500
Drug	Bonferoni	0.05	1 ≤ n ≤ 1500
Disease	Bonferoni	0.05	1 ≤ n ≤ 1500

Annotations must have at least 2 gene(s)  
Chose Topcluster output format: Extended HTML Table (Cytoscape Link)

Gene Sets

Liver			Salivary_Glands		
Original	Human Symbol	Entrez ID	Original	Human Symbol	Entrez ID
1576	CYP3A4	1576	2591	GALNT3	2591
1577	CYP3A5	1577	9073	CLDN8	9073
7036	TP2	7036	466	FXYD2	466
229	ALDOB	229	54959	ODAM	54959
341	APOC1	341	5349	FXYD3	5349
126	ADH1C	126	155006	TMEM213	155006
125	ADH1B	125	100129410	LOC100129410	100129410
7448	VTN	7448	54097	FAM3B	54097
5053	PAH	5053	352999	C6orf58	352999
3840	HF	3840	57535	KIARA1324	57535
197	AHSG	197	26298	EHF	26298
3078	CFHR1	3078	999	CDH1	999
383	ARG1	383	360	AQP3	360

# TopCluster (http://toppcluster.cchmc.org)

Processing Salivary\_Glands

Jump To ...

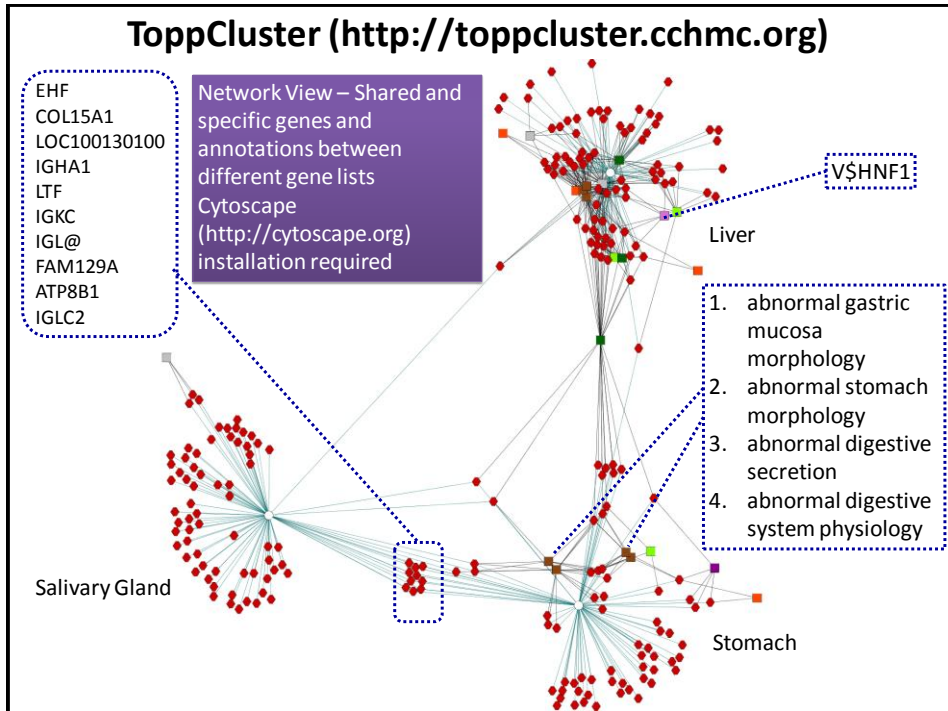
- GO: Molecular Function
- GO: Biological Process
- GO: Cellular Component
- Human Phenotype
- Mouse Phenotype
- Domain
- Pathway
- Pubmed
- Interaction
- Cytoband
- TFBS
- Coexpression
- Computational
- Drug
- Disease
- MicroRNA

Links: Back to Start, Shareable Link, Excel Version

Cytoscape: Include Orphaned Genes (XGMLL), Include Super Category

Re-Enrich: Build

Category	ID	Title (or Source)	liver_logP	salivary_gland_logP	stomach_cardiac_logP	pValues
GO: Molecular Function	GO:0016491	oxidoreductase activity	10.0000			
GO: Molecular Function	GO:0005201	extracellular matrix structural constituent			5.3780	
GO: Molecular Function	GO:0005506	iron ion binding	5.3161			
GO: Molecular Function	GO:0004497	monooxygenase activity	5.2535			
GO: Molecular Function	GO:0004022	alcohol dehydrogenase activity	4.8034			
GO: Molecular Function	GO:0046906	tetrapyrrole binding	4.5687			
GO: Molecular Function	GO:0020037	heme binding	4.5687			
GO: Molecular Function	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	4.4053			
GO: Molecular Function	GO:0019825	oxygen binding	3.9166			
GO: Molecular Function	GO:0004866	endopeptidase inhibitor activity	3.9162			
GO: Molecular Function	GO:0030414	peptidase inhibitor activity	3.8272			
GO: Molecular Function	GO:0004024	alcohol dehydrogenase activity, zinc-dependent	3.5535			
GO: Molecular Function	GO:0043499	euk-aryotic cell surface binding	3.4320			
GO: Molecular Function	GO:0030246	carbohydrate binding	2.0565		1.3457	
GO: Molecular Function	GO:0008289	lipid binding	3.1619			
GO: Molecular Function	GO:0004857	enzyme inhibitor activity	2.8756			
GO: Molecular Function	GO:0004867	serine-type endopeptidase inhibitor activity	2.8632			
GO: Molecular Function	GO:0048407	platelet-derived growth factor binding			2.7407	
GO: Molecular Function	GO:0008201	heparin binding	2.6553			



## DAVID (<http://david.abcc.ncifcrf.gov>)

### Database for Annotation, Visualization and Integrated Discovery

DAVID Bioinformatics Resources 2007  
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

**Shortcuts to DAVID Tools**

- ▶ **Functional Annotation**  
Gene-annotation enrichment analysis, functional annotation clustering (FAT), BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more
- ▶ **Gene Functional Classification**  
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological context captured by high throughput technologies. [More](#)
- ▶ **Gene ID Conversion**  
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. [More](#)
- ▶ **Gene Name Batch Viewer** beta!  
Display gene names for a given gene list. Search functionally related genes within your list or not in your list. Query links to enriched detailed information. [More](#)

Welcome to DAVID Bioinformatics Resources 2003 - 2007 [What's Special in DAVID 2007?](#)

The Database for Annotation, Visualization and Integrated Discovery (DAVID) 2007 is the fifth program of DAVID 2006 comprehensive set of functions. For a complete list of functions, see the DAVID 2007 user manual. For a complete list of gene list, DAVID tools are:

- Identify enriched biological terms
- Discover enriched functional gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more

Screen Shot 1



## DAVID (http://david.abcc.ncifcrf.gov)

**Annotation Summary Results**

Current Gene List: demolist1 171 DAVID IDs  
 Current Background: Homo sapiens Check Defaults [X] Clear All

Main Accessions (0 selected)  
 Other Accessions (0 selected)  
 Gene Ontology (1 selected)

Gene	Count	Chart
GOTERM_BP_1	79%	136
GOTERM_BP_2	76%	131
GOTERM_BP_3	74%	127
GOTERM_BP_4	69%	119
GOTERM_BP_5	66%	104
GOTERM_BP_ALL	79%	136
GOTERM_CC_1	79%	131
GOTERM_CC_2	61%	106
GOTERM_CC_3	55%	95
GOTERM_CC_4	50%	86
GOTERM_CC_5	38%	63
GOTERM_CC_ALL	79%	131
GOTERM_MF_1	75%	129
GOTERM_MF_2	69%	119
GOTERM_MF_3	60%	103
GOTERM_MF_4	56%	97
GOTERM_MF_5	45%	78
GOTERM_MF_ALL	75%	129

Protein Domains (3 selected)  
 Pathways (3 selected)  
 General Annotations (0 selected)  
 Functional Categories (3 selected)  
 Protein Interactions (0 selected)  
 Literature (0 selected)  
 Disease (1 selected)

GENETIC\_ASSOCIATION\_DB 13% 23  
 OMIM\_DISEASE 18% 32

Combined View for Selected Annotation

## DAVID (http://david.abcc.ncifcrf.gov)

**DAVID Bioinformatics Resources 2007**  
 National Institute of Allergy and Infectious Diseases (NIAID), NIH

Welcome to DAVID Bioinformatics Resources 2003 - 2007

**What's Special in DAVID 2007?**

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more

## DAVID (http://david.abcc.ncifcrf.gov)

### Convert NCBI Entrez Gene IDs to RefSeq Accession Numbers

**Upload** | **List** | **Background**

Upload Gene List

[Demolist 1](#) [Demolist 2](#)

[Upload Help](#)

Step 1: Enter Gene List

A: Paste a list

Clear

Or

B: Choose From a File

Browse...

Step 2: Select Identifier

AFFY\_ID

Step 3: List Type

Gene List

Background

Step 4: Submit List

Submit List

← Submit your gene list to start conversion!

**Gene ID Conversion Tool**

The Cross-Conversion of Gene ID Types:

- Entrez Gene ID
- Affy ID
- GenBank Accession
- Genpept Accession
- NCBI GI
- PIR Accession
- PIR ID
- PIR NREF ID
- RefSeq Genomic Accession
- RefSeq mRNA Accession
- RefSeq Protein Accession
- RefSeq RNA Accession
- Ungene
- UNIPROT Accession
- UNIPROT ID
- UNIREF100 ID
- Official Gene Symbol *new!*
- Not Sure *new!*

**Upload** | **List** | **Background**

Upload Gene List

[Demolist 1](#) [Demolist 2](#)

[Upload Help](#)

Step 1: Enter Gene List

A: Paste a list

3493

3512

10562

3535

5284

Clear

Or

B: Choose From a File

Browse...

Step 2: Select Identifier

ENTREZ\_GENE\_ID

Step 3: List Type

Gene List

Background

Step 4: Submit List

Submit List

## DAVID (http://david.abcc.ncifcrf.gov)

**Upload** | **List** | **Background**

Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -

HOMO SAPIENS(88)

UNKNOWN(4)

UNIDENTIFIED(1)

Select

List Manager Help

Uploaded List: 1

Select list to:

Use Rename

Remove Combine

Show Gene List *new!*

[View Mapped IDs](#)

Convert the gene list being selected in left panel to REFSEQ\_MRNA

Submit

**Gene ID Conversion Tool Result**

[Right-click to Download the result](#) [Help](#)

[Submit Converted List to DAVID as a Gene List](#)
[Submit Converted List to DAVID as a Background](#)

Conversion Summary			From	To	Species	David Gene Name
ID Count	In DAVID DB	Conversion	202	NM_001624	HOMO SAPIENS	ABSENT IN MELANOMA 1
50	Yes	Successful	72	NM_001613	HOMO SAPIENS	ACTIN, ALPHA 2, SMOOTH MUSCLE, AORTA
8 IDs	Yes	None	72	NM_001615	HOMO SAPIENS	ACTIN, ALPHA 2, SMOOTH MUSCLE, AORTA
0 IDs	No	None	27299	NM_014479	HOMO SAPIENS	ADAM-LIKE, DECYSIN 1
0 IDs	Ambiguous	Pending	125	NM_000667	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS 1), ALPHA POLYPEPTIDE
<b>Total Unique User IDs: 68</b>			125	NM_000668	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS 1), ALPHA POLYPEPTIDE
<b>Summary of Ambiguous Gene IDs</b>			126	NM_000669	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS 1), ALPHA POLYPEPTIDE
ID Count	Possible Source	Convert All	126	NM_000668	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS 1), ALPHA POLYPEPTIDE
<b>All Possible Sources For Ambiguous IDs</b>			125	NM_000669	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS 1), ALPHA POLYPEPTIDE
Ambiguous ID	Possibility	Convert				

## Exercise 13: Convert affymetrix probeset IDs to gene symbols

## Exercise 14: What are the enriched pathways and diseases for this gene set?

From the same example data set ("Example-Set-1.xls"), use the probe set IDs (2<sup>nd</sup> column) and extract their RefSeq accession numbers

### PANTHER (<http://www.pantherdb.org/>) Protein Analysis Through Evolutionary Relationships

**Quick links**

- [Browse PANTHER](#)
- [Search PANTHER](#)
- [Batch search](#)
- [Browse pathways](#)
- [Community Curation](#)
- [My Workspace](#)
- [Gene expression tools](#)
- [HMM scoring](#)
- [cSNP analysis](#)
- [Downloads](#)
- [Site map](#)

**Batch ID Search**

Find PANTHER-classified genes, transcripts, and proteins by uploading a list of IDs

Enter IDs:

separate IDs by a space or comma - [supported IDs](#)

Upload IDs:

Select upload ID type:  Gene Symbol  ID List

Select File Type:  ID List  Previously exported text search results

Result page:  Genes  Transcripts/Proteins

Select datasets:

Celera:  H. sapiens  M. musculus  R. norvegicus

NCBI:  H. sapiens  M. musculus  R. norvegicus

FlyBase:  D. melanogaster

**GENE EXPRESSION DATA ANALYSIS**

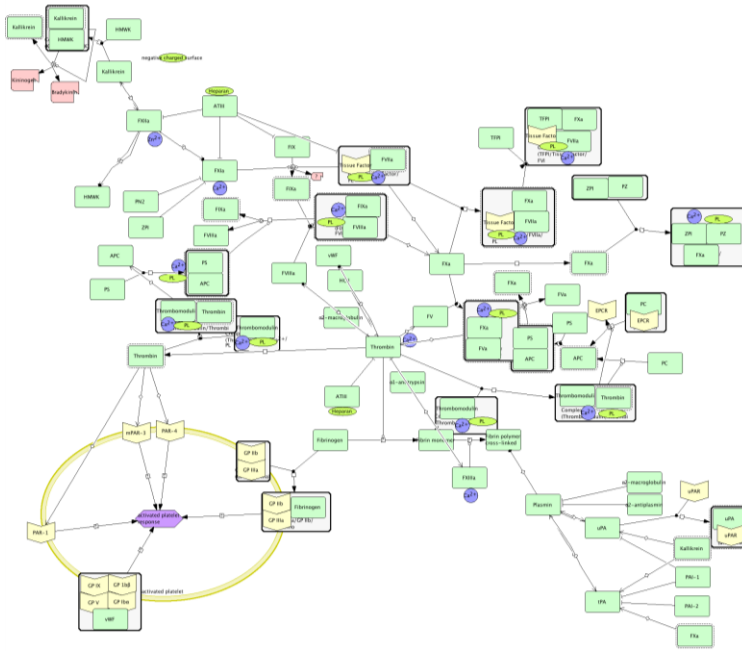
Our expression analysis tools can be used for microarray data interpretation. Multiple gene lists can be mapped to PANTHER molecular function and biological process categories, as well as to biological pathways. Our pathway visualization tool will display your experimental results on detailed diagrams of the relationships between genes/proteins in known pathways.

- Compare gene lists [?](#)  
Upload lists of genes or gene products and statistically compare them to a reference list to look for under- and over-represented functional categories.
- Analyze a list of genes with expression values [?](#)  
Upload a list of genes and their corresponding fold-change values from a differential expression experiment.

**You can compare multiple lists!**



## PANTHER (<http://www.pantherdb.org/>)



## Summary

### Cis-Element Finding Matrix

	CONSERVED	NON-CONSERVED
KNOWN TFBS	oPOSSUM DiRE	Pscan MatInspector*
NOVEL/UNKNOWN TFBS OR MOTIFS	oPOSSUM WEEDER-H	MEME WEEDER

## RESOURCES - URLs: Summary

Application/Resource	URL
oPOSSUM	<a href="http://burgundy.cmmmt.ubc.ca/oPOSSUM/">http://burgundy.cmmmt.ubc.ca/oPOSSUM/</a>
DiRE	<a href="http://dire.dcode.org/">http://dire.dcode.org/</a>
Weeder-H	<a href="http://159.149.109.9/modtools/">http://159.149.109.9/modtools/</a>
Weeder	<a href="http://159.149.109.9/modtools/">http://159.149.109.9/modtools/</a>
Pscan	<a href="http://159.149.109.9/modtools/">http://159.149.109.9/modtools/</a>
MEME	<a href="http://meme.sdsc.edu/">http://meme.sdsc.edu/</a>
MatInspector	<a href="http://www.genomatix.de/">http://www.genomatix.de/</a>
GenomeTrafac	<a href="http://genometrafac.cchmc.org">http://genometrafac.cchmc.org</a>
ToppGene	<a href="http://toppgene.cchmc.org">http://toppgene.cchmc.org</a>
ToppCluster	<a href="http://toppcluster.cchmc.org">http://toppcluster.cchmc.org</a>
DAVID	<a href="http://david.abcc.ncifcrf.gov">http://david.abcc.ncifcrf.gov</a>
PANTHER	<a href="http://www.pantherdb.org">http://www.pantherdb.org</a>
Genome Browser	<a href="http://genome.ucsc.edu">http://genome.ucsc.edu</a>
ECR Browser	<a href="http://ecrbrowser.dcode.org">http://ecrbrowser.dcode.org</a>
Slides/Exercises	<a href="http://anil.cchmc.org/dhc.html">http://anil.cchmc.org/dhc.html</a>

## Exercises - Summary

1. **Exercise 1:** Use oPOSSUM to find shared conserved cis-elements in a group of co-expressed genes
2. **Exercise 2:** Use DiRE to find shared conserved cis-elements in a group of co-expressed genes
3. **Exercise 3:** Use Pscan to find shared cis-elements (Transfac) in a group of co-expressed genes
4. **Exercise 4:** Download upstream 500 bp sequence for a list of genes
5. **Exercise 5:** Download all SNPs overlapping with these genes
6. **Exercise 6:** Download the orthologous promoter sequences (human, mouse, and rat) for the gene SLC7A1
7. **Exercise 7:** Are there any putative microRNA regulators for SLC7A1? If yes, download all of them using table browser
8. **Exercise 8:** Use the downloaded SLC7A1 ortholog promoter sequences to find out common motifs using WeederH
9. **Exercise 9:** Use the downloaded promoter sequences to find out common motifs using Weeder and MEME
10. **Exercise 10:** Does any of the motifs found by Meme match known TFBS?
11. **Exercise 11:** Use the gene list from the downloaded file ("Example-Set-2") and find out:
  - How many of these genes are transcription factors?
  - What are the enriched TFBSs and miRNAs?
  - What gene families are enriched in this list?
  - Are there any salivary gland development associated genes present in this list?
  - How many and which genes from this list are associated with non-insulin dependent diabetes mellitus (NIDDM)?
12. **Exercise 12:** Prioritize the 721 genes ("Example-Set-2") using "stomach genes" from the "Example-Set-1".
  - What are the top 10 ranked genes using ToppGene and ToppNet?
  - Why is TFF3 ranked among the top 5 in ToppGene prioritization? What is its rank in ToppNet?
13. **Exercise 13:** Convert Affymetrix probeset IDs to gene symbols
14. **Exercise 14:** What are the enriched pathways and diseases for this gene set?

For additional exercises, see <http://anil.cchmc.org/dhc.html>