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

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 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…

Annotation Databases
Gene Ontology, Pathways

Genome-wide Promoters
Putative Regulatory
Signatures

DNA Repair
XRCC1
OGG1
ERCC1
MPG….

Angiogenesis
HIF1A
ANGPT1
VEGF
KLF5….

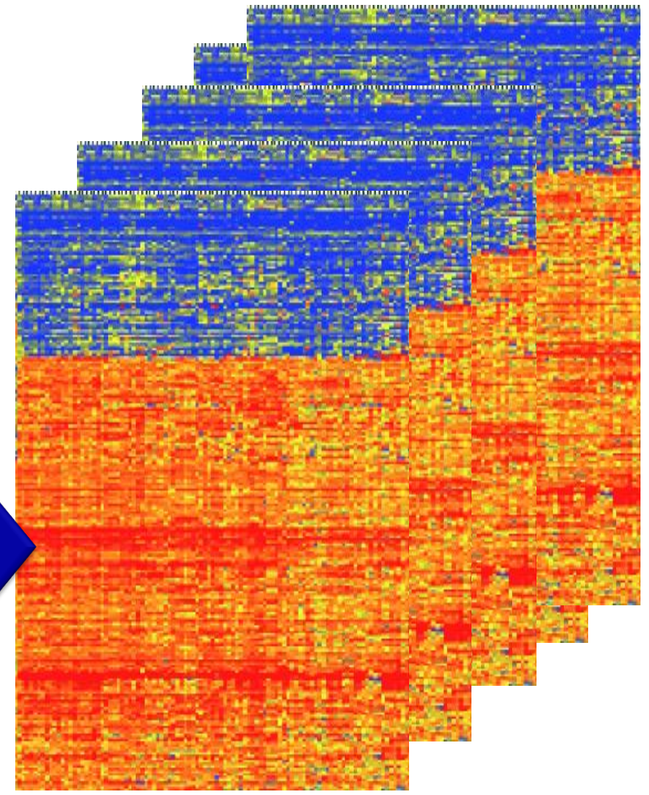
E2F
RB1
MCM4
FOS
SIVA…..

PDX1
GLUT2
PAX4
PDX1
IAPP….

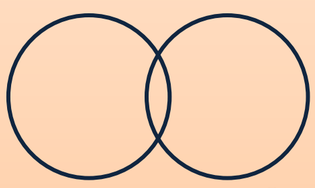
p53
CDKN1A
CTSD
CASP
DDB2….

Gene lists associated with similar
function/process/pathway

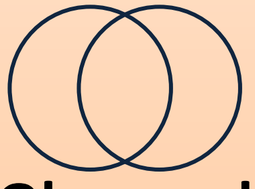
Expression Profile - Gene Lists



Enrichment Analysis



Expected

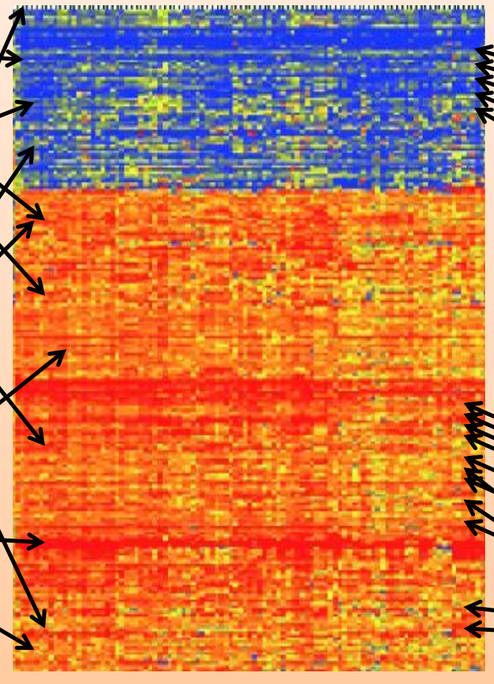


Observed

P53
CTSD
CASP
DDB2….

DNA Repair
XRCC1
OGG1
ERCC1
MPG….

Random Distribution



E2F
RB1
MCM4
FOS…

Angiogenesis
HIF1A
ANGPT1
VEGF…..

Significant Enrichment

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)

Supports human and mouse

oPOSSUM

About | Contact

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

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

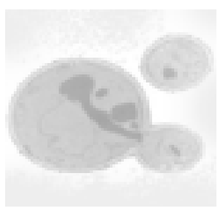
Human Combination Site Analysis (CSA) identifies over-represented **combinations** of conserved transcription factor binding sites in sets of human and mouse genes.

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.

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 | Clear

259
5265
350
335
335
1558

( ) **OR** upload a file containing a list of gene identifiers:

Browse...

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 8 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

JASPAR PhyloFACTS Profiles

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

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.

oPOSSUM (http://www.cisreg.ca/oPOSSUM)

STEP 3: Select parameters

Level of conservation:
Top 10% of conserved regions (min. conservation 70%)

Matrix match threshold:
80 %

Amount of upstream / downstream sequence:
2000 / 2000

Number of results to display:
Top 10 results
OR only results with **Z-score** >= 10 and **Fisher score** <= 0.01

Sort results by:
Z-score Fisher score

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

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.

oPOSSUM (http://www.cisreg.ca/oPOSSUM)

Analysis Results

Selected Parameters

Conservation level:	Top 10% of cons
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 5105 229 325 2168 2244 5053
Excluded: 5265 1558 125 3240 3827 5004

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
HNF1A	HOMEO	vertebrate	15.548	1466	13684	8	7	1860	0.0021	8	0.0136	22.69	2.692e-05
SRY	HMG	vertebrate	9.193	8624	6526	13	2	34149	0.0248	33	0.0361	6.567	1.552e-02
Fos	bZIP	vertebrate	10.670	7001	8149	11	4	16086	0.0104	16	0.0155	4.583	3.175e-02
HLF	bZIP	vertebrate	11.147	3376	11774	7	8	5014	0.0048	9	0.0131	10.72	3.196e-02
Foxq1	FORKHEAD	vertebrate	14.070	3533	11617	7	8	6047	0.0054	9	0.0120	8.207	4.026e-02
NKX3-1	HOMEO	vertebrate	11.127	5391	9759	9	6	12155	0.0069	13	0.0110	4.548	4.696e-02
FOXD1	FORKHEAD	vertebrate	11.926	5516	9634	9	6	11145	0.0072	15	0.0146	7.875	5.417e-02
Pdx1	HOMEO	vertebrate	9.040	9899	5251	13	2	54092	0.0261	47	0.0342	4.571	6.515e-02
Cebpa	bZIP	vertebrate	9.187	5863	9287	9	6	12176	0.0118	14	0.0204	7.21	7.844e-02
Nkx2-5	HOMEO	vertebrate	8.270	10169	4981	13	2	59121	0.0333	52	0.0442	5.46	8.496e-02

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	ENSG00000047457	3	-1	150422269	150420270	150424269	GGTTAATGTTTAAT	150421319	951	150421332	938	1	15.334
350	ENSG00000091583	17	-1	61655974	61653975	61657974	GGTTAATGTTTAAG	61656032	-58	61656045	-71	-1	13.479
3273	ENSG00000113905	3	1	187866487	187864487	187868486	TGTAAATGATTAGT	187866344	-143	187866357	-130	-1	9.708
1571	ENSG00000130649	10	1	135190857	135188857	135192856	GGTTTATTATTAGC	135190745	-112	135190758	-99	-1	14.409
5105	ENSG00000124253	20	1	55569543	55567543	55571542	AGATAATCATTGAA	55569396	-147	55569409	-134	-1	9.903
325	ENSG00000132703	1	1	157824239	157822239	157825284	AGTTATTTATTAGA	157824079	-160	157824092	-147	-1	12.759
2168	ENSG00000163586	2	-1	88208693	88206694	88210693	AGTTAATGTTTGAA	88208792	-99	88208805	-112	-1	12.830
2244	ENSG00000171564	4	1	155703596	155701596	155705595	AGTTAATATTTAAT	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	ENSG00000047457	3	-1	150422269	150420270	150424269	TTAAACATT	150421323	947	150421331	939	-1	6.961
				150422269	150420270	150424269	TGACACAAT	150422361	-92	150422369	-100	1	7.793
				150422269	150420270	150424269	TAAAACAAA	150423255	-986	150423263	-994	-1	9.474
350	ENSG00000091583	17	-1	61655974	61653975	61657974	TAATATAAT	61654150	1825	61654158	1817	1	5.862
				61655974	61653975	61657974	AAAAACAAA	61654256	1719	61654264	1711	-1	8.914
2158	ENSG00000101981	X	1	138440561	138438561	138442560	TTGGACAAA	138441494	934	138441502	942	1	6.016
383	ENSG00000118520	6	1	131936059	131934059	131938058	ATGAATAAT	131935824	-235	131935832	-227	1	5.865
3273	ENSG00000113905	3	1	187866487	187864487	187868486	TTAATCAAT	187866435	-52	187866443	-44	1	8.775
462	ENSG00000117601	1	-1	172153139	172151140	172155139	TTAAGCAAA	172153193	-54	172153201	-62	1	5.779
				172153139	172151140	172155139	TTAAACAAC	172153216	-77	172153224	-85	-1	7.440
1571	ENSG00000130649	10	1	135190857	135188857	135192856	GAAAATAAT	135188983	-1874	135188991	-1866	-1	8.003
				135190857	135188857	135192856	GCTAATAAT	135190745	-112	135190753	-104	1	6.366
				135200555	135198555	135202554	TAAAACATT	135199099	-1456	135199107	-1448	-1	6.342
5105	ENSG00000124253	20	1	55569543	55567543	55571542	GTACACAAA	55569204	-339	55569212	-331	1	8.214
				55569543	55567543	55571542	ATTAACAAC	55569381	-162	55569389	-154	1	6.352
229	ENSG00000136872	9	-1	103237926	103235927	103239926	TCTCACAAT	103237073	854	103237081	846	1	6.965
				103237926	103235927	103239926	GTAAATAAA	103237407	520	103237415	512	1	7.334

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	ENSG00000047457	3	-1	150422269	150420270	150424269	GGTTAATGTTTAAT	150421319	951	150421332	938	1	15.334
350	ENSG00000091583	17	-1	61655974	61653975	61657974	GGTTAATGTTTAAG	61656032	-58	61656045	-71	-1	13.479
3273	ENSG00000113905	3	1	187866487	187864487	187868486	TGTAAATGATTAGT	187866344	-143	187866357	-130	-1	9.708
1571	ENSG00000130649	10	1	135190857	135188857	135192856	GGTTTATTATTAGC	135190745	-112	135190758	-99	-1	14.409
5105	ENSG00000124253	20	1	55569543	55567543	55571542	AGATAATCATTGAA	55569396	-147	55569409	-134	-1	9.903
325	ENSG00000132703	1	1	157824239	157822239	157825284	AGTTATTTATTAGA	157824079	-160	157824092	-147	-1	12.759
2168	ENSG00000163586	2	-1	88208693	88206694	88210693	AGTTAATGTTTGAA	88208792	-99	88208805	-112	-1	12.830
2244	ENSG00000171564	4	1	155703596	155701596	155705595	AGTTAATATTTAAT	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	ENSG00000047457	3	-1	150422269	150420270	150424269	TTAAACATT	150421323	947	150421331	939	-1	6.961
				150422269	150420270	150424269	TGACACAAT	150422361	-92	150422369	-100	1	7.793
				150422269	150420270	150424269	TAAAACAAA	150423255	-986	150423263	-994	-1	9.474
350	ENSG00000091583	17	-1	61655974	61653975	61657974	TAATATAAT	61654150	1825	61654158	1817	1	5.862
				61655974	61653975	61657974	AAAAACAAA	61654256	1719	61654264	1711	-1	8.914
2158	ENSG00000101981	X	1	138440561	138438561	138442560	TTGGACAAA	138441494	934	138441502	942	1	6.016
383	ENSG00000118520	6	1	131936059	131934059	131938058	ATGAATAAT	131935824	-235	131935832	-227	1	5.865
3273	ENSG00000113905	3	1	187866487	187864487	187868486	TTAATCAAT	187866435	-52	187866443	-44	1	8.775
462	ENSG00000117601	1	-1	172153139	172151140	172155139	TTAAGCAAA	172153193	-54	172153201	-62	1	5.779
				172153139	172151140	172155139	TTAAACAAC	172153216	-77	172153224	-85	-1	7.440
1571	ENSG00000130649	10	1	135190857	135188857	135192856	GAAAATAAT	135188983	-1874	135188991	-1866	-1	8.003
				135190857	135188857	135192856	GCTAATAAT	135190745	-112	135190753	-104	1	6.366
				135200555	135198555	135202554	TAAAACATT	135199099	-1456	135199107	-1448	-1	6.342
5105	ENSG00000124253	20	1	55569543	55567543	55571542	GTACACAAA	55569204	-339	55569212	-331	1	8.214
				55569543	55567543	55571542	ATTAACAAC	55569381	-162	55569389	-154	1	6.352
229	ENSG00000136872	9	-1	103237926	103235927	103239926	TCTCACAAT	103237073	854	103237081	846	1	6.965
				103237926	103235927	103239926	GTAAATAAA	103237407	520	103237415	512	1	7.334

oPOSSUM (http://www.cisreg.ca/oPOSSUM)

Human SSA | Human CSA | Worm SSA | Yeast SSA

oPOSSUM Main | Home | Custom Analysis | Download API | Download DB

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):

[Use sample genes] [Clear]

```
259
5265
350
335
335
1558
```

( ) **OR** upload a file containing a list of gene identifiers:

[ ] [Browse...]

STEP 1b: Enter a background list of genes

( ) Use background set of 1000 random genes

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

[Clear]

```
1281
1281
1805
125
10551
```

( ) **OR** upload a file containing a list of gene identifiers:

[ ] [Browse...]

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
HNF1A	HOMEO	vertebrate	15.548	1	10	8	7	1	0.0025	8	0.0136	20.32	2.426e-02
HLF	bZIP	vertebrate	11.147	1	10	7	8	1	0.0021	9	0.0131	21.68	4.943e-02
NKX3-1	HOMEO	vertebrate	11.127	3	8	9	6	3	0.0037	13	0.0110	10.95	1.042e-01
Bapx1	HOMEO	vertebrate	8.542	4	7	10	5	5	0.0079	20	0.0218	14.26	1.286e-01
Lhx3	HOMEO	vertebrate	12.941	3	8	8	7	5	0.0079	11	0.0120	4.173	1.775e-01
Pdx1	HOMEO	vertebrate	9.040	7	4	13	2	28	0.0295	47	0.0342	2.533	1.826e-01
SRY	HMG	vertebrate	9.193	7	4	13	2	26	0.0410	33	0.0361	-2.303	1.826e-01
Nkx2-5	HOMEO	vertebrate	8.270	7	4	13	2	28	0.0344	52	0.0442	4.865	1.826e-01
FOXI1	FORKHEAD	vertebrate	13.183	4	7	9	6	8	0.0168	17	0.0248	5.555	2.142e-01
RORA_1	NUCLEAR RECEPTOR	vertebrate	13.190	1	10	4	11	1	0.0018	5	0.0061	9.233	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/)

DiRE

Distant Regulatory Elements of co-regulated genes

http://dire.dcode.org/

Home | Details | Output example | Screenshots | Return to submitted job... | Citing DiRE | Contact us

April 30, 2008. The DiRE tool operates with new ECR Browser alignments now. It is also possible to run DiRE on human (hg18), mouse (mm9), and rat (rn4) genomes.

Co-regulated genes:

Copy and paste **gene names** (or accession numbers) (example)

AMBP
SERPINA1
APOH
APOA1
APOA1
CYP2C8
CYP2E1
ALDOB
SERPINC1
ADH1B
HP
PCK1
SERPINA1
HRG
FGB
F9

Gene annotation: Gene symbols (GATA2, PAX6, etc)

Genome: human (hg18)

Background (control) genes:

Select the source of background genes:

- (•) random set of 100 genes
- ( ) copy and paste background genes

Target elements:

- (•) top 3 ECRs + promoter ECRs [default]
- ( ) UTR ECRs + promoter ECRs
- ( ) promoter ECRs only

Submit

job ID: 0926091038152309

22 signal genes; **100** background genes

Identifying distant regulatory elements...
parameter optimization: 82%

Automatic status update every 3 seconds.
Click here to manually refresh the page.

Request ID...

0926091006373607
perm link: http://dire.dcode.org/?id=0926091006373607

68 Potential Regulatory Elements...

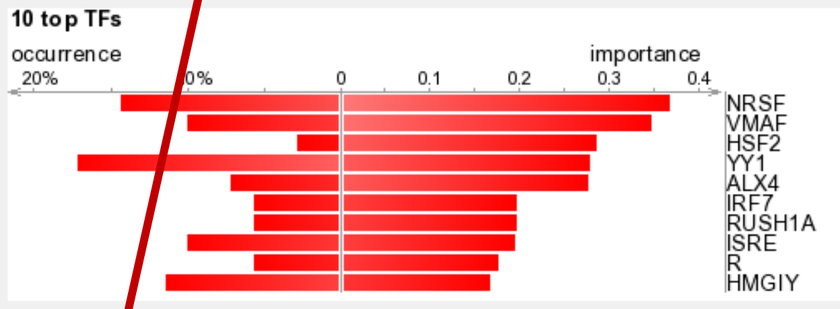
intergenic	40 (59%)
promoter	10 (15%)
utr	12 (18%)
intron	6 (9%)

Detailed description of regulatory elements (in tabulated textual format)

Chromosomal distribution

Candidate Transcription Factors...

10 top TFs



Full TF list

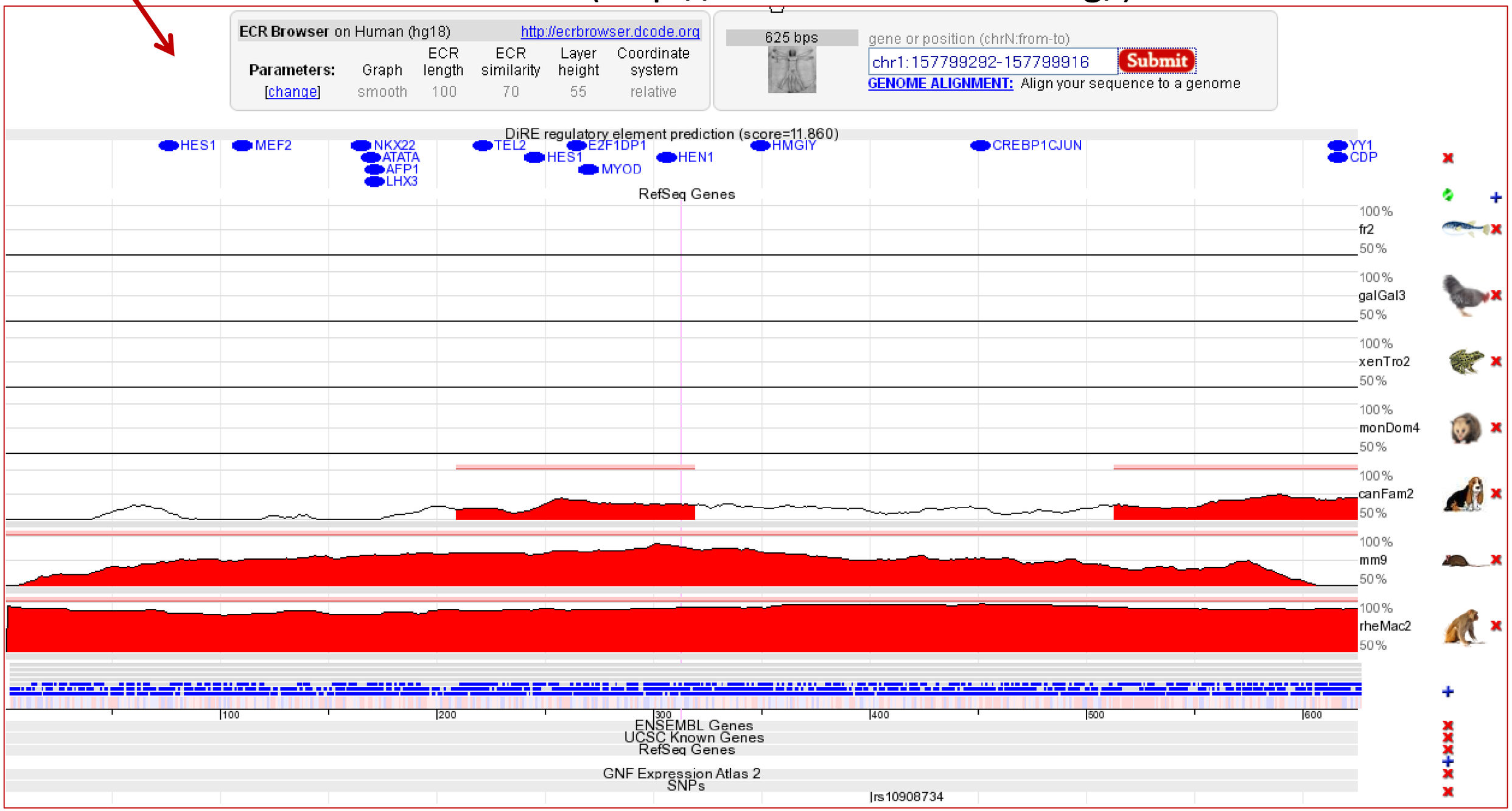
#	Transcription Factor	Occurrence	Importance
1	NRSF	14.29%	0.36786
2	VMAF	10.00%	0.34781
3	HSF2	2.86%	0.28571
4	YY1	17.14%	0.27750
5	ALX4	7.14%	0.27679
6	IRF7	5.71%	0.19643
7	RUSH1A	5.71%	0.19643
8	ISRE	10.00%	0.19375
9	R	5.71%	0.17661
10	HMGIY	11.43%	0.16571
11	CDP	8.57%	0.16286
12	LXR	5.71%	0.16161
13	POU6F1	5.71%	0.15732
14	PPARG	15.71%	0.15714
15	CEBP	21.43%	0.15000
16	STAT	11.43%	0.14857

#	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) NKX22(160) ATATA(164) AFP1(166) LHX3(166) TEL2(216) HES1(240) E2F1DP1(259) MYOD(265) HEN1(301) HMGIY(344) CREBP1CJUN(446) YY1(611) CDP(611)
2	chr1:157822774-157823321	promoter	1.327	chr1:157772437-157948651	APCS	4 :: AREB6(60) TBX5(62) CHX10(365) TBX5(520)
3	chr1:157823477-157823620	promoter	0.843	chr1:157772437-157948651	APCS	3 :: ARP1(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 :: PAX2(56) STAT3(65) OSF2(106) PEBP(107) AML1(107) RFX1(340) NFMUE1(396) TAL1BETAITF2(398) CHOP(403) YY1(507) GATA1(598) CDP(611) CLOX(611) NRSF(649)

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) NKX22(160) ATATA(164) AFP1(166) LHX3(166) TEL2(216) HES1(240) E2F1DP1(259) MYOD(265) HEN1(301) HMGIY(344) CREBP1CJUN(446) YY1(611) CDP(611)
2	chr1:157822774-157823321	promoter	1.327	chr1:157772437-157948651	APCS	4 :: AREB6(60) TBX5(62) CHX10(365) TBX5(520)
3	chr1:157823477-157823620	promoter	0.843	chr1:157772437-157948651	APCS	3 :: ARP1(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 :: PAX2(56) STAT3(65) OSF2(106) PEBP(107) AML1(107) RFX1(340) NFMUE1(396) TAL1BETAITF2(398) CHOP(403) YY1(507) GATA1(598) CDP(611) CLOX(611) NRSF(649)

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



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

<u>DiRE Summary</u>:

1. DiRE's unique feature is the detection of **<u>conserved</u>** 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)

Insert Gene/Sequence ID list: (help)

Select Organism: Homo sapiens

Select Region: -450 +50

Select Descriptors: Jaspar / Jaspar_Fam / Transfac / User Defined

Run! Undo changes Reset!

Messages:

Bioinformatics Evolution @nd COmparative GeNomics

Pscan Web Interface

Use the input form on the left to set up your query. The results will be displayed in this window.

If you need **HELP** please click here.

Source:
Download Pscan source code

Reference:
F.Zambelli, G.Pesole, G.Pavesi
Pscan: Finding Over-represented Transcription Factor Binding Site Motifs in Sequences from Co-Regulated or Co-Expressed Genes.
Nucleic Acids Research 2009 37(Web Server issue):W247-W252.

Contacts:
giulio.pavesi@unimi.it
federico.zambelli@unimi.it

Sample data

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.

MYC100 MYC90 MYC80 MYC75 MYC65 MYC55

List of NFkB target genes, collected from literature. NFKBxx should be read as in the MYC dataset.

NFkB100 NFkB90 NFkB80 NFkB70 NFkB60 NFkB50 NFkB40

List of NRF1 target genes. NRFxx 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).

NRF1_100 NRF1_90 NRF1_80 NRF1_70 NRF1_60 NRF1_50 NRF1_40

NRF1 Matrix

Pscan (http://159.149.109.9/pscan)

Insert Gene/Sequence ID list: (help)

PSCAN

```
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

Run! | Undo changes | Reset!

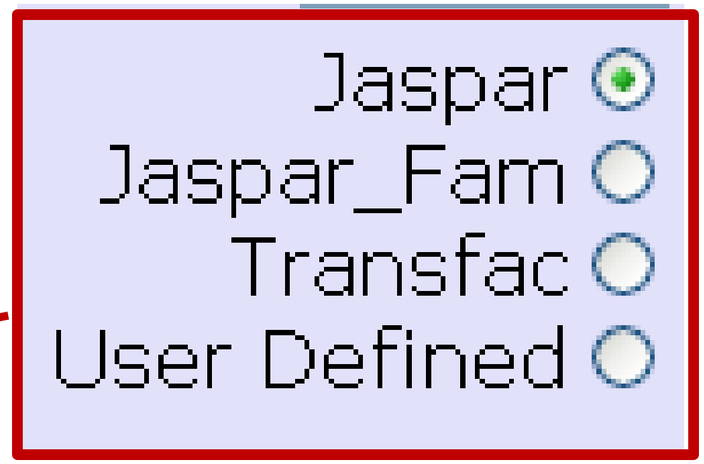
```
Messages:

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:

Select Descriptors:

- Homo sapiens
- Mus musculus
- Human and Mouse
- Drosophila melanogaster
- Arabidopsis thaliana
- Saccharomyces cerevisiae

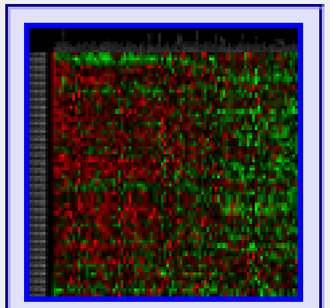


Pscan (http://159.149.109.9/pscan)

View Text Results

97 TF profiles used

Matrix Name	P-value
TBP	1.59074e-08
Foxa2	0.000274079
FOXL1	0.000657034
MEF2A	0.000657227
Hand1-Tcfe2a	0.000697277
Nobox	0.000790445
FOXI1	0.000804377
PBX1	0.00124224
SRF	0.00124647
Evi1	0.00128699
TEAD1	0.00212538
Lhx3	0.00303459
Foxq1	0.00355502
Prrx2	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	0.00940039



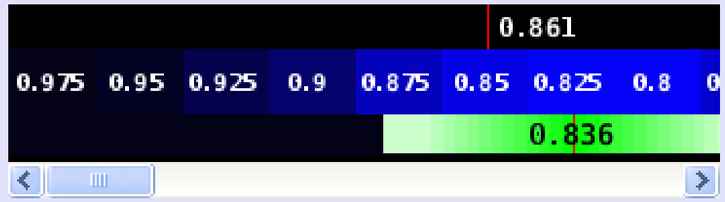
Matrix Info

ID	MA0047
Name	Foxa2
Class	FORKHEAD
Species	Rattus norvegicus
Inf. Content	12.43
SuperGroup	vertebrate
Protein Acc.	P32182
Type	COMPILED
PMID	8139574
Report Occurrences	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



Background Score Distribution

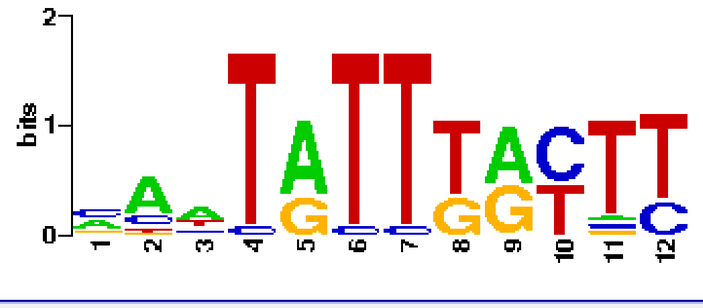
Sample Statistics

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	
Std Dev	
Size	

Go!



Pscan (http://159.149.109.9/pscan)

View Text Results

97 TF profiles used

Matrix Name	P-value
TBP	1.59074e-08
Foxa2	0.000274079
FOXL1	0.000657034
MEF2A	0.000657227
Hand1-Tcfe2a	0.000697277
Nobox	0.000790445
FOXI1	0.000804377
PBX1	0.00124224
SRF	0.00124647
Evi1	0.00128699
TEAD1	0.00212538
Lhx3	0.00309459
Foxg1	0.00355532
Prrx2	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	0.00940039

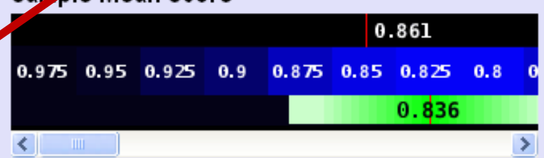
Matrix Info

ID	MA0047
Name	Foxa2
Class	FORKHEAD
Species	Rattus norvegicus
Inf. Content	12.43
SuperGroup	vertebrate
Protein Acc.	P32182
Type	COMPILED
PMID	8139574
Report Occurrences	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	0	0	0	6	8	0	1	0
T	1	2	5	19	0	16	16	11	0	8	14	12

Sample Mean Score



Background Score Distribution

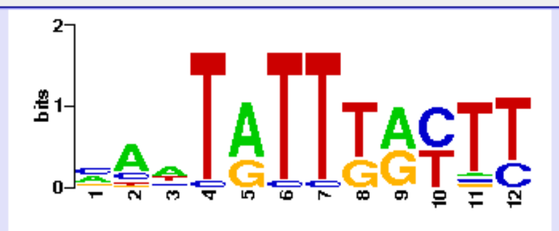
Sample Statistics

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	
Std Dev	
Size	

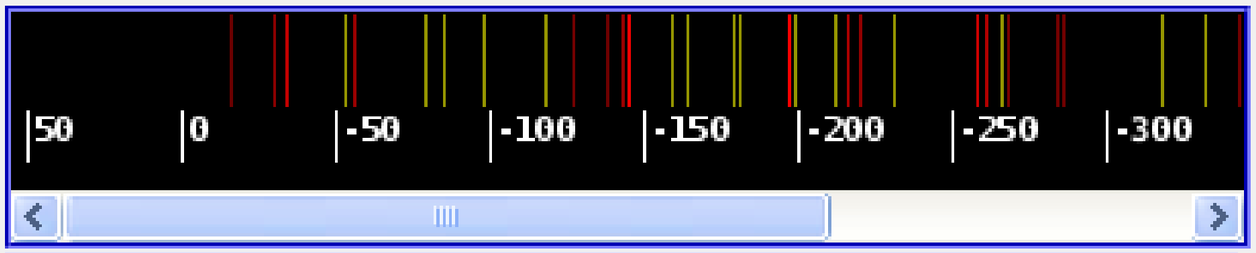
Go!



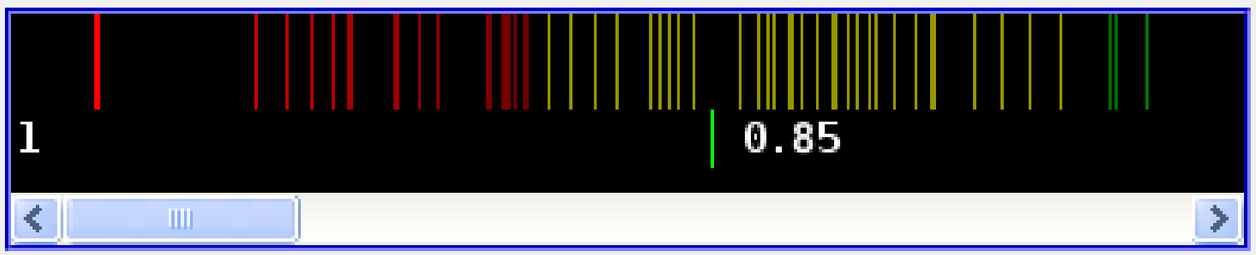
View Text Results

Name	Score	Position	Sequence	Strand
hg18_refGene_NM_002345	0.98343	-197	CAATATTGATTT	-
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	ACTTGTTTGCTT	+
hg18_refGene_NM_001033886	0.939791	-428	ACTTGTTTGCTT	+
hg18_refGene_NM_199168	0.939791	-428	ACTTGTTTGCTT	+
hg18_refGene_NM_021010	0.935506	-261	GAGTATTTACTT	-
hg18_refGene_NM_133477	0.932417	-356	AAACATTTATTT	+
hg18_refGene_NM_194435	0.931745	-216	CTTTGTTTGTTT	+
hg18_refGene_NM_003381	0.931745	-216	CTTTGTTTGTTT	+
hg18_refGene_NM_005603	0.922637	-143	GAATATTTACAT	+
hg18_refGene_NM_004616	0.9222	-56	ATCTGTTTACTT	+

Occurrences Position Distribution (score >=0.836)



Occurrences Score Distribution

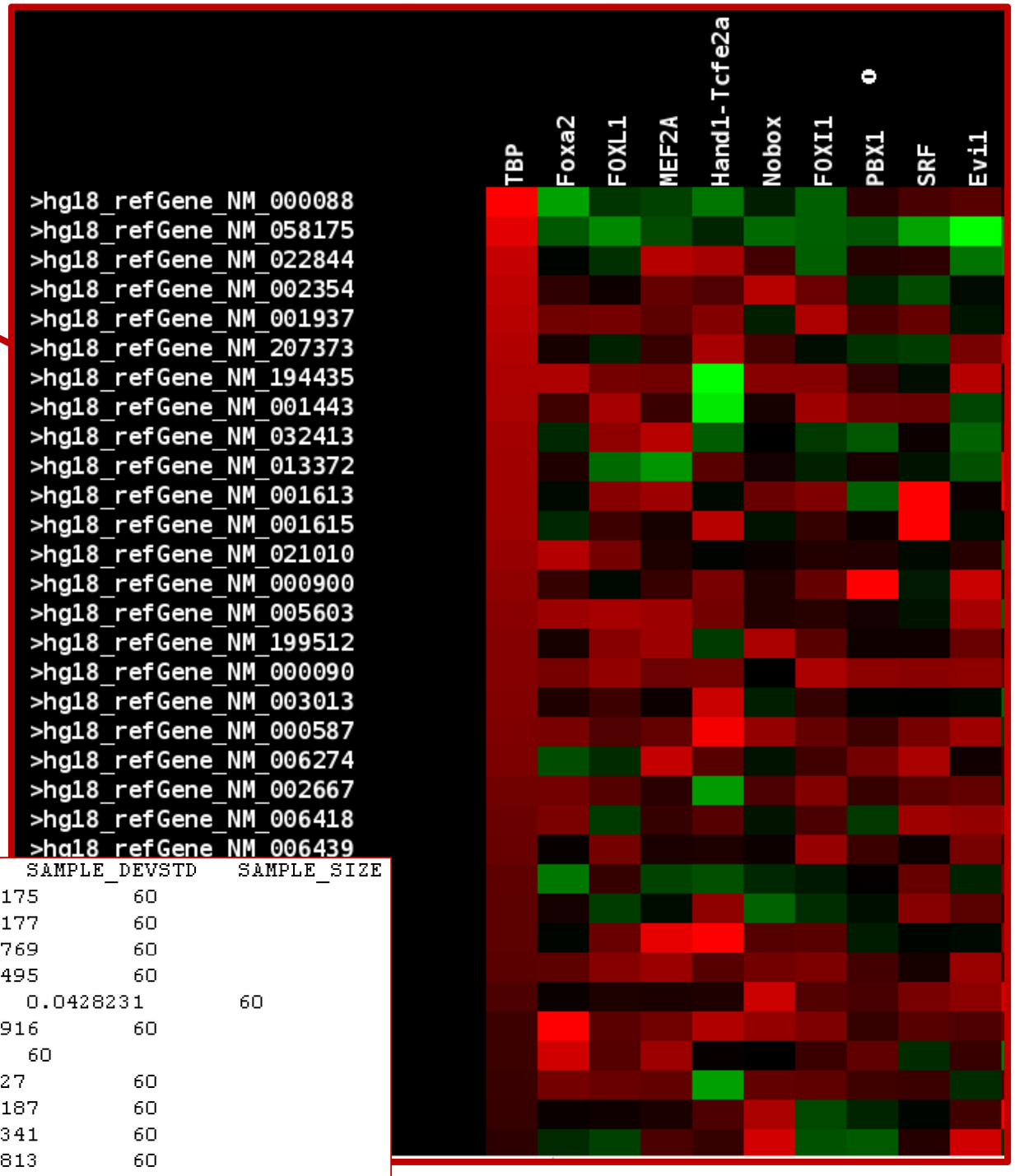


Pscan (http://159.149.109.9/pscan)

View Text Results

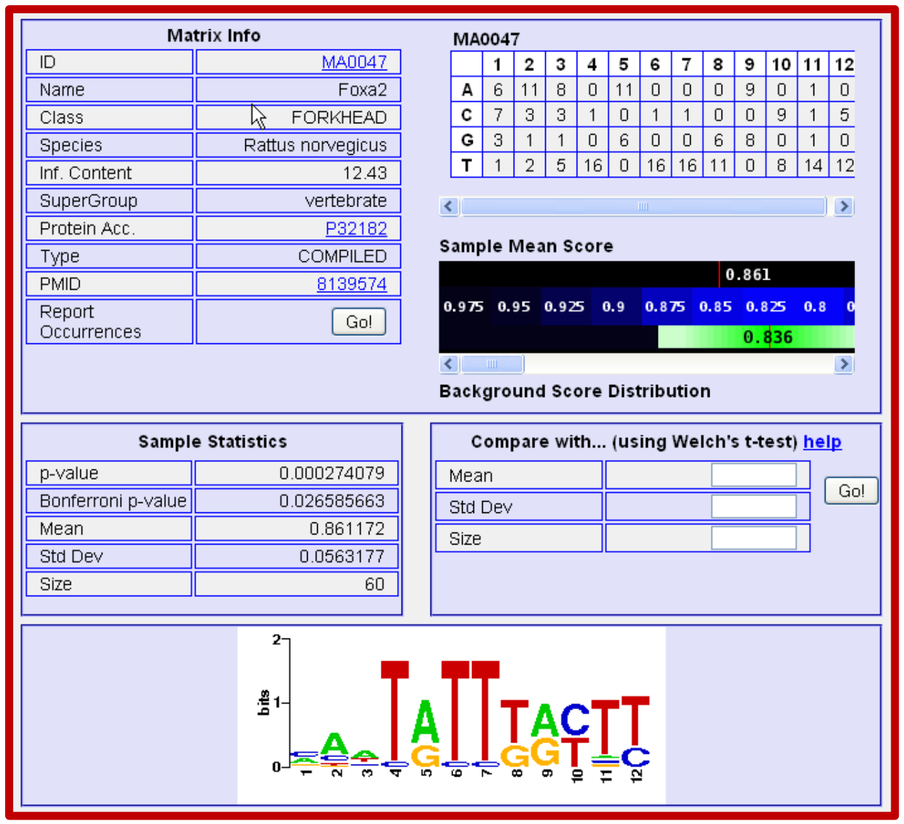
97 TF profiles used

Matrix Name	P-value
TBP	1.59074e-08
Foxa2	0.000274079
FOXL1	0.000657034
MEF2A	0.000657227
Hand1-Tcfe2a	0.000697277
Nobox	0.000790445
FOXI1	0.000804377
PBX1	0.00124224
SRF	0.00124647
Evi1	0.00128699
TEAD1	0.00212538
Lhx3	0.00303459
Foxq1	0.00355502
Prrx2	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	0.00940039



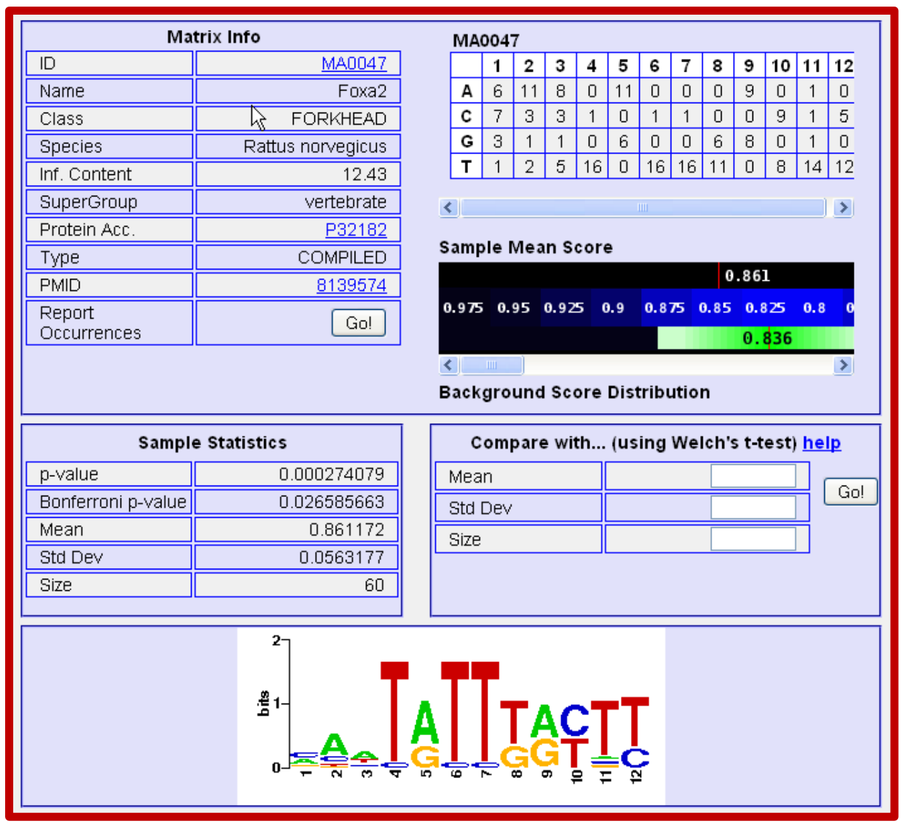
TF_NAME	MATRIX_ID	Z_SCORE	P_VALUE	SAMPLE_AVERAGE	BACKGROUND_AVERAGE	SAMPLE_DEVSTD	SAMPLE_SIZE
TBP	MA0108	5.52625	1.59074e-08	0.859494	0.816811	0.0493175	60
Foxa2	MA0047	3.45141	0.000274079	0.861172	0.836541	0.0563177	60
FOXL1	MA0033	3.21019	0.000657034	0.918922	0.891892	0.0523769	60
MEF2A	MA0052	3.20941	0.000657227	0.810217	0.777574	0.0710495	60
Hand1-Tcfe2a	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
FOXI1	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	3.01105	0.00128699	0.769179	0.746618	0.0546341	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
Prrx2	MA0075	2.58254	0.00486451	0.92434	0.89143	0.0897038	60
Lhx3	MA0135	2.55439	0.00527407	0.798641	0.773954	0.0704547	60
NKX3-1	MA0124	2.51437	0.00590862	0.853125	0.829214	0.0728452	60
NFIL3	MA0025	2.48463	0.00642618	0.80459	0.783263	0.0663844	60
REL	MA0101	2.46099	0.00685234	0.8843	0.869993	0.0500624	60
Pax6	MA0069	2.42112	0.00765503	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

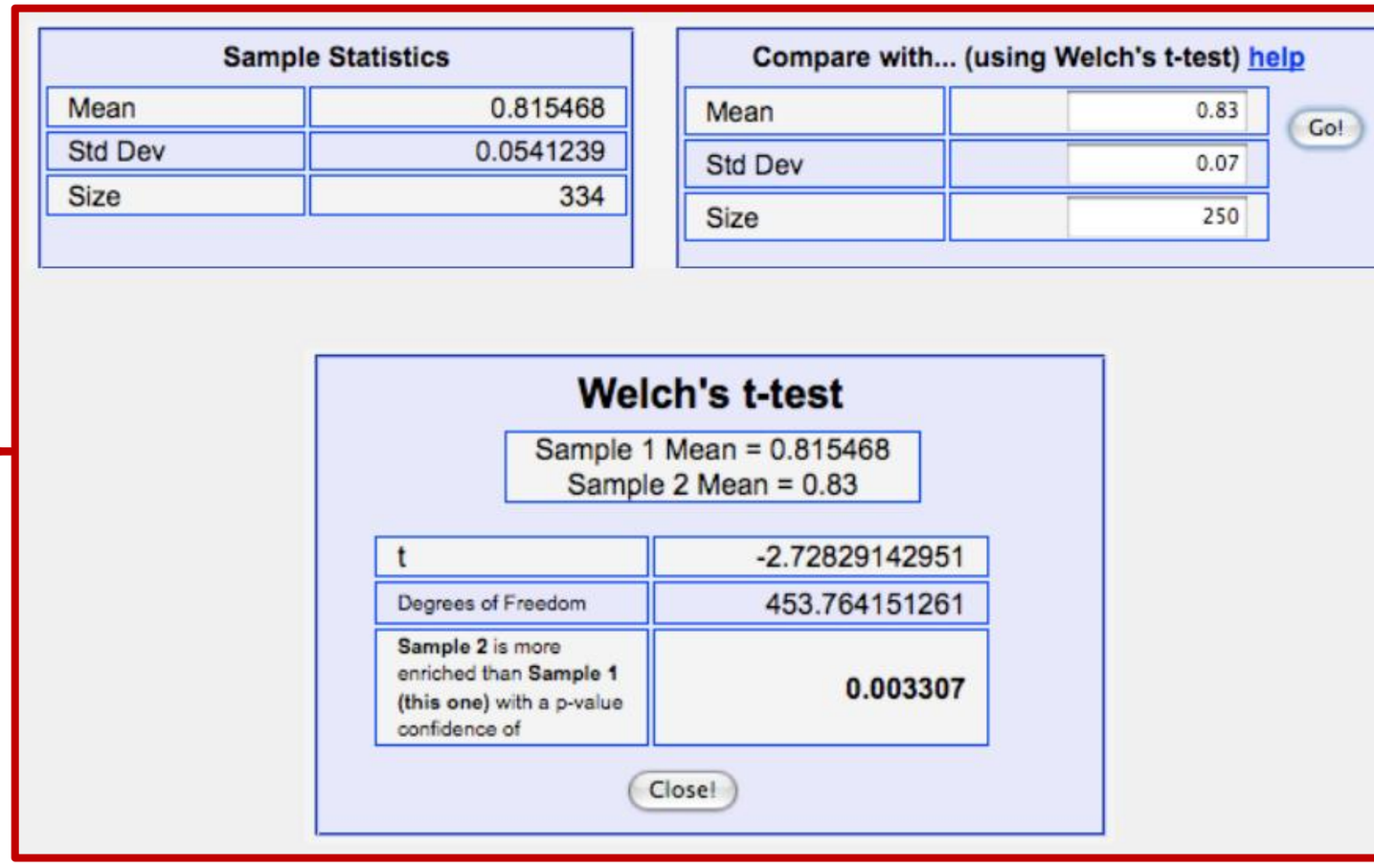
Pscan (http://159.149.109.9/pscan)



Comparing different input gene sets:

1. 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).
2. 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.
3. **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.





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

1. **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:

1. Pscan supports a variety of TFBS matrices (e.g. JASPAR, Transfac) including user input matrix.
2. Supports human, mouse, drosophila, and yeast
3. Limited options to select the regions for scanning
4. Cannot select the background set although comparisons can be computed
5. Does not support upload of your sequences; Input options are very limited
6. 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:

[Use sample genes] [Clear]

```
259
5265
350
335
335
1558
```

( ) **OR** upload a file containing a list of gene identifiers:

[ ] [Browse...]

oPOSSUM (http://www.cisreg.ca/oPOSSUM)

STEP 2: Select transcription factor binding site matrices

JASPAR CORE Profiles

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

oPOSSUM (http://www.cisreg.ca/oPOSSUM)

STEP 3: Select parameters

Level of conservation:
Top 10% of conserved regions (min. conservation 70%)

Matrix match threshold:
80 %

Amount of upstream / downstream sequence:
2000 / 2000

Number of results to display:
Top 10 results
OR only results with **Z-score** >= 10 and **Fisher score** <= 0.01

Sort results by:
Z-score Fisher score

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

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
RYTAAWNNNTGAY	Unknown	mammals	16.655	2636	12514	9	6	3909	0.0041	15	0.0237	27.75	2.676e-04
TAATTA	Unknown	mammals	12.000	7400	7750	13	2	27227	0.0132	31	0.0226	7.458	2.848e-03
TATAAA	Unknown	mammals	12.000	9219	5931	14	1	47951	0.0232	47	0.0342	6.638	6.209e-03
RGTTAMWNATT	Unknown	mammals	17.072	2277	12873	6	9	3189	0.0028	8	0.0107	13.33	1.705e-02
RTAAACA	Unknown	mammals	13.000	7918	7232	12	3	25209	0.0142	29	0.0246	7.953	2.670e-02
YATTNATC	Unknown	mammals	13.061	6858	8292	11	4	18528	0.0119	19	0.0185	5.394	2.682e-02
CTTTGA	Unknown	mammals	12.000	10591	4559	14	1	54148	0.0262	47	0.0342	4.553	3.478e-02
YCATTAA	Unknown	mammals	13.004	7484	7666	11	4	22958	0.0129	22	0.0187	4.57	5.404e-02
AACWWCAANK	Unknown	mammals	15.858	3060	12090	6	9	4631	0.0037	8	0.0097	8.817	6.381e-02
TGGAAA	Unknown	mammals	12.000	11182	3968	14	1	67892	0.0328	43	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
1356	ENSG00000047457	3	-1	150422269	150420270	150424269	ACTAAATTGTGTC	150422362	-93	150422375	-106	-1	8.802
383	ENSG00000118520	6	1	131936059	131934059	131938058	GTACAAGTTTGAC	131937518	1460	131937531	1473	1	6.292
3273	ENSG00000113905	3	1	187866487	187864487	187868486	ACTAATCATTTAC	187866344	-143	187866357	-130	1	8.969
462	ENSG00000117601	1	-1	172146654	172144655	172148654	CTTAATATCTGTC	172144991	1664	172145004	1651	1	6.144
				172153139	172151140	172155139	GTCAAAGGCTGAT	172153165	-26	172153178	-39	1	11.059
1571	ENSG00000130649	10	1	135190857	135188857	135192856	TTCAAAGGCTGAT	135190716	-141	135190729	-128	-1	6.038
5105	ENSG00000124253	20	1	55569543	55567543	55571542	ACTAAACCTTGAC	55569306	-237	55569319	-224	-1	13.603
				55569543	55567543	55571542	GTTAATGAATGCT	55569374	-169	55569387	-156	-1	8.174
				55569543	55567543	55571542	GATAATCATTGAA	55569396	-147	55569409	-134	-1	6.601
325	ENSG00000132703	1	1	157824239	157822239	157825284	ATTAAATACAGAC	157822921	-1318	157822934	-1305	-1	10.324
2168	ENSG00000163586	2	-1	88208693	88206694	88210693	GTTAATGTTTGAA	88208792	-99	88208805	-112	-1	12.880
				88208693	88206694	88210693	CTTTATCATTGAC	88208819	-126	88208832	-139	-1	6.066
				88208693	88206694	88210693	ATTAATGTTTGCT	88208867	-174	88208880	-187	-1	11.146
2244	ENSG00000171564	4	1	155703596	155701596	155705595	GTTAATATTTAAT	155703524	-72	155703537	-59	-1	11.267
				155703596	155701596	155705595	GCTAATGTAAGAT	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
ENCODE
Blat
Table Browser
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Proteome Browser
Utilities
Downloads
Release Log
Custom Tracks
Archaeal Genomes
Mirrors
Archives

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.

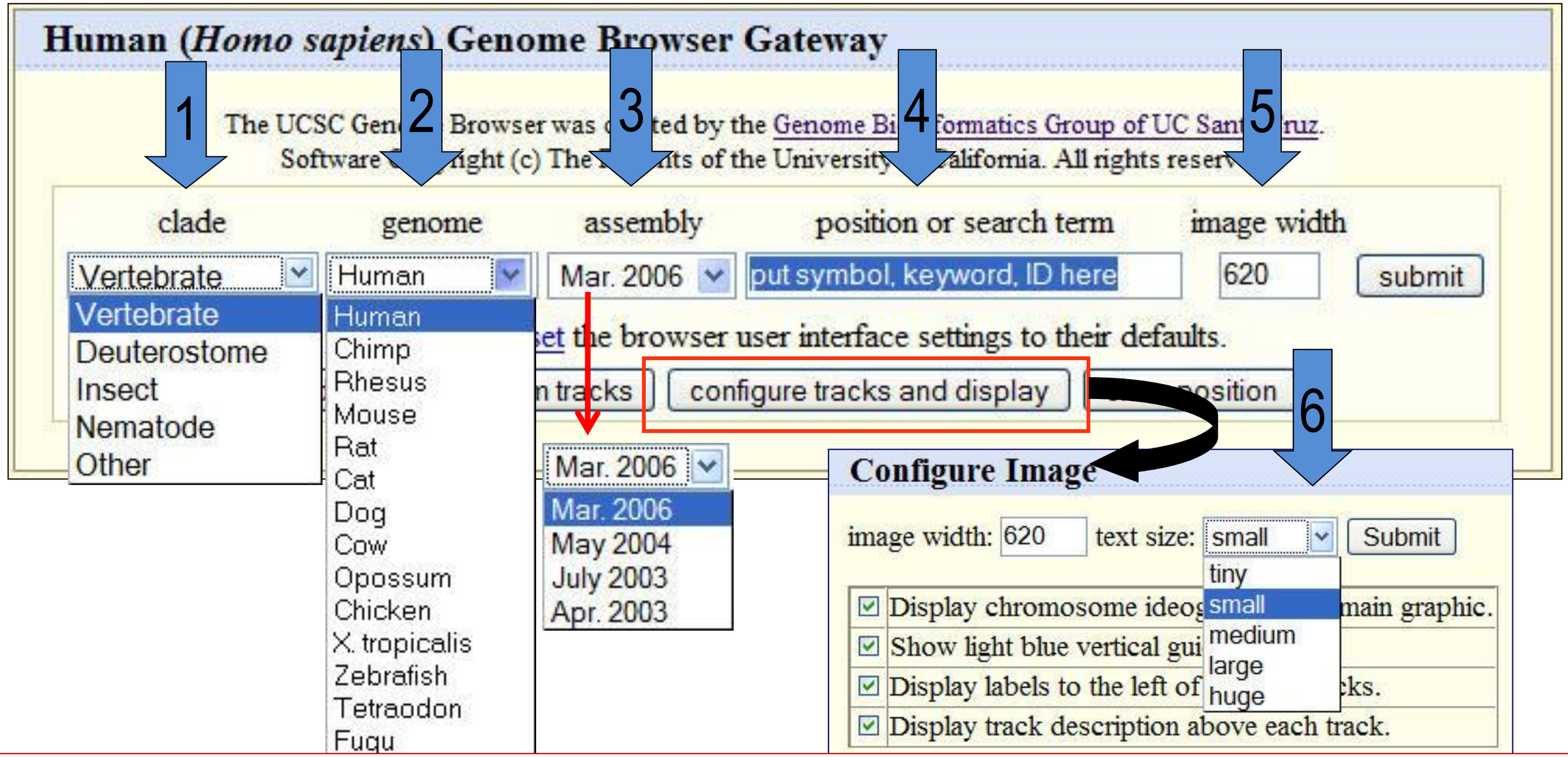
In conjunction with this change, there is also a change to the way you must specify your bigBed or bigWig Custom Track. When you specify the location of your local bigBed/bigWig file (on your web-accessible http, https, or ftp server), use this designation: bigDataUrl (instead of the old designation: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

e.g. `track type=bigBed name="My Big Bed" description="Some Data from My Lab" bigDataUrl=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)



Genome Browser Gateway choices:

1. Select Clade
2. Select genome/species: You can search only one species at a time
3. Assembly: the official backbone DNA sequence
4. Position: location in the genome to examine or search term (gene symbol, accession number, etc.)
5. Image width: how many pixels in display window; 5000 max
6. 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 dr

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over ways. 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.

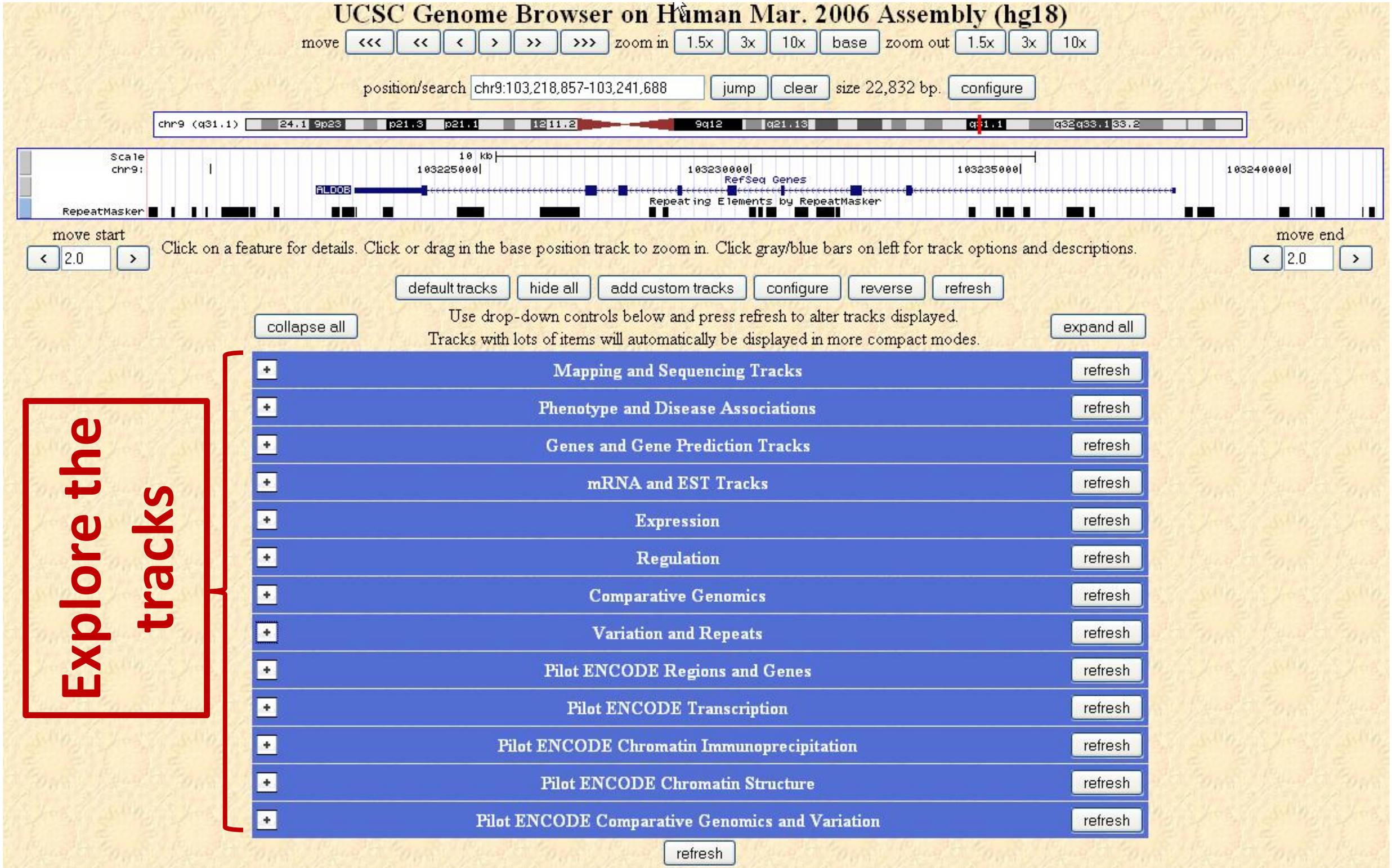
add custom tracks | configure tracks and display | clear position

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 Genes

PDX1 (uc001urt.1) at chr13:27392157-27397394 - p
PDHX (uc001mvt.1) at chr11:34894741-34974094 - p
SPOP (uc002ipg.1) at chr17:45031245-45110524 - s
SPOP (uc002ipf.1) at chr17:45031245-45110524 - s
SPOP (uc002ipe.1) at chr17:45031245-45110524 - s
SPOP (uc002ipd.1) at chr17:45031245-45110524 - s
SPOP (uc002ipc.1) at ch
SPOP (uc002ipb.1) at

RefSeq Genes

PDX1 at chr13:27392157-27397394 - (NM_000209) pa

Non-Human RefSeq Genes

Pdx1 at chr13:27392267-27397054 - (NM_022852) pa
PDX1 at chr13:27392276-27396838 - (NM_001081478)
pdx1 at chr13:27392612-27396613 - (NM_131443) pa
Pdx1 at chr13:27392153-27398338 - (NM_008814) pa

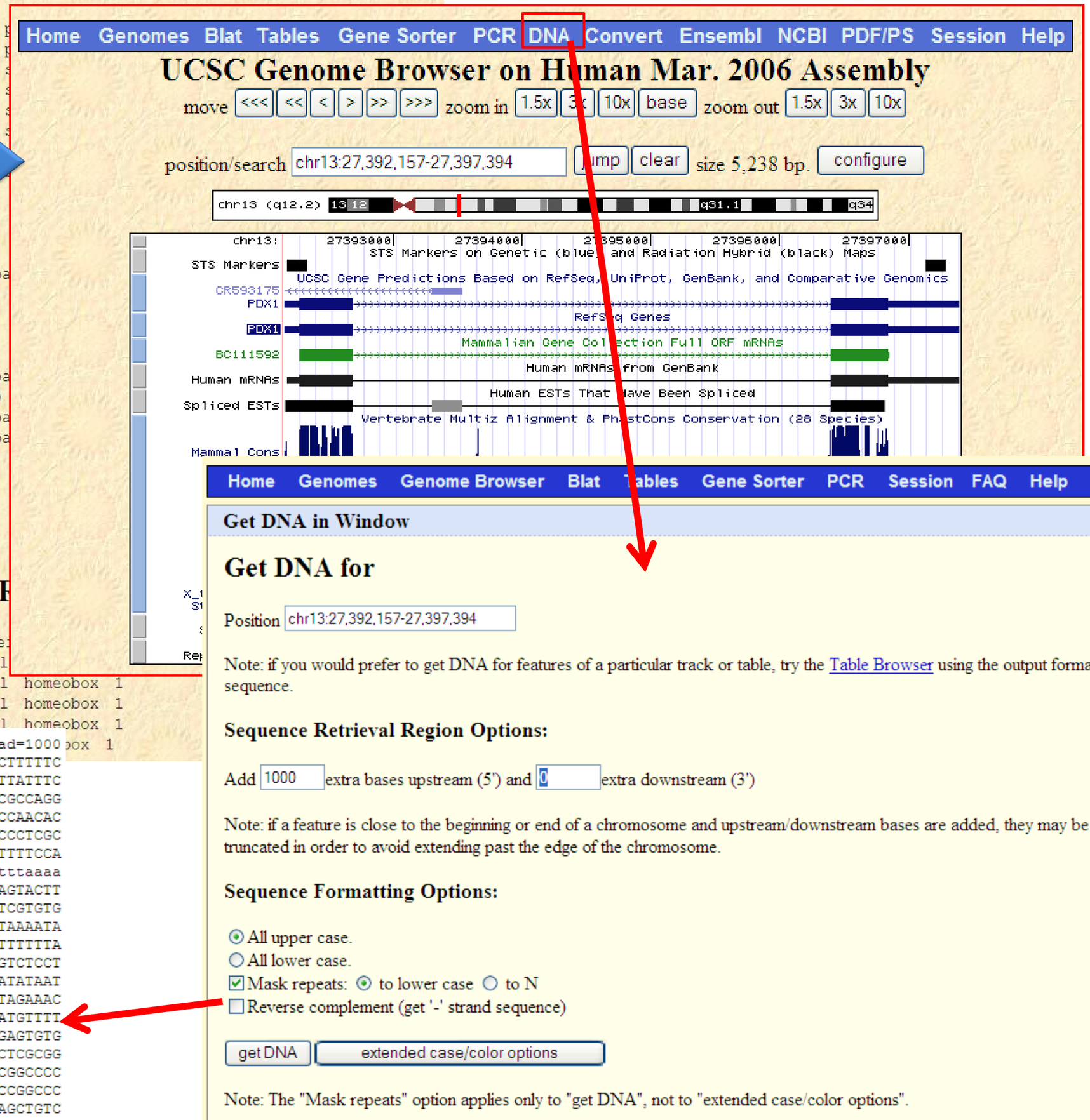
Alias of STS Marker

PDX1 at chr7:34489430-34689640 - (AFM067XA9)

Non-Human Aligned mRNA Search H

U73854 - Mesocricetus auratus, homeodomain prote
BC103572 - Mus musculus, pancreatic and duodenal
BC103581 - Mus musculus, pancreatic and duodenal homeobox 1
BC103582 - Mus musculus, pancreatic and duodenal homeobox 1
BC105642 - Mus musculus, pancreatic and duodenal homeobox 1
BC078192 >hg18_dna range=chr13:27391157-27397394 5'pad=1000 box 1

```
GCCAAGCACAGATGTTATCATGGAAAATGCAGCGTTTTTATTTCTTTTTC
TAAATATGTAACTCTTCCTCCACTTCCCCCTCTCCTGCTTGCCTTATTTC
AATTGCAAGCAGAAGAGAGTGAGTGTTCTCTGCCGGCAAACTCCGCCAGG
GTCCCGGCCCGTAGAGAGTCGTCAAGGGTCTGGAACCCCCGTGCCAACAC
CTGCCCCTGCTTCGCAGCCCCAAGAGGAAGGCCGCGTCTTTCCCCCTCGC
TGTATTGGGAAGCTACGTTCCGGGCTGGCCAAATGGGCCCCAATTTTCCA
AAACCCAAATTTGTAATACCCTTCaatttttttaaaaaaaagaatttaaaa
aaGTCTCTGTGAATGCTTCAGAAGTTACCGTTTACACCCCAGAAGTACTT
GCAGCACATCCACAAGTAAAAACACACAACGAATGCCAGAGTTTCGTGTG
TTTTTTAACCGACATCTTTGTGGCTGTGAACAAACTTCATAAATAAAATA
GAATCAAATGCTTCTGACCTAGAGAGCTGGGTCTGCAAACTTTTTTTTTA
TCGTATTCCGCAACAGTTAAATAAAAAATTAAAAACTCAACATGTCTCCT
TGTAAACTACATCAATTAACAAACACACTATGTCCATTATCAAATATAAT
AGAAAAAATATAGGAAAATAGAAAATAGAAAAATATAGGAAAATAGAAAC
TTTTAAGCCACGGTGAAAATGTTTCTATAAATGAGTGGTTCTAATGTTTT
CGTGAGCGCCCATTTTGGGGAGCACCGCCAGCTGCCCGTTCAGGAGTGTG
CAGCAAACTCAGCTGAGAGAGAAAATTGGAACAAAAGCAGGTGCTCGCGG
GTACCTGGGCCTAGCCTCTTAGTGCGGCCAGCCAGGCCAATCACGGCCCC
CGGCTGAACCACGTGGGGCCCCGCGGAGCCTATGGTGCGGCGGCCGGCCC
GCCGGTCCGCGCTGGCTGTGGGTTCCCTCTGAGATCAGTGCGGAGCTGTC
```



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
Gene Sorter
In Silico PCR
Genome Graphs
Galaxy
VisiGene
Proteome Browser
Utilities
Downloads
Release Log
Custom Tracks
Archaeal Genomes
Mirrors
Archives

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: bigDataUrl (instead of the old designation: dataUrl).

e.g. `track type=bigBed name="My Big Bed" description="Some Data from My Lab" bigDataUrl=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)

1

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to r
see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample
presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQ
restrictions associated with these data.

clade: Mammal **genome:** Human **assembly:** Mar. 2006

group: Genes and Gene Prediction Tracks **track:** RefSeq Genes add custom tracks

table: refFlat describe table schema

region: genome ENCODE position chrX:151073054-151383976 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

output format: all fields from selected table Send output to Galaxy

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

2

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to r
see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample
presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQ
restrictions associated with these data.

clade: Mammal **genome:** Human **assembly:** Mar. 2006

group: Genes and Gene Prediction Tracks **track:** RefSeq Genes add custom tracks

table: refFlat describe table schema

refGene
all_est
all_mrna
ccdsInfo
gbCdnaInfo
gbMiscDiff
gbSeq
gbStatus
gbWarn
imageClone
kgXref
knownToRefSeq
mrnaOrientInfo
mrnaRefseq
refFlat
refLink
refSeqAli
refSeqStatus
refSeqSummary
seq

position chrX:151073054-151383976 lookup define regions

list upload list

able Send output to Galaxy

(leave blank to keep output in browser)

zip compressed

ustom tracks), click here.

Using the Table Browser

3

Paste In Identifiers for refFlat

Please paste in the identifiers you want to include. The items must be values of the **geneName** fiel
schema" button shows more information about the table fields.) Some example values:
LOC133874
PLXDC1
SNORD115-17

AMBP
SERPINA1
APOH
APOA1
APOA1
CYP2C8
CYP2E1
ALDOB
SERPINC1

submit clear cancel

4

Table Browser (Input Identifiers)

Use this program to retrieve the data associated with a track in text f
see Using the Table Browser for a description of the controls in this
presentation of the software features and usage. For more complex
restrictions associated with these data.

clade: Mammal **genome:** Human **assembly:**

group: Genes and Gene Prediction Tracks

table: refFlat describe table schema

region: genome ENCODE position chrX:151073054-15

identifiers (names/accessions): paste list upload list cle

filter: create

intersection: create

output format: sequence

all fields from selected table
selected fields from primary and related tables
sequence
GTF - gene transfer format
BED - browser extensible data
custom track
hyperlinks to Genome Browser

output file:

file type return

get output

To reset **all** user cart settings (including custom tracks), click here.

5

Sequence Retrieval Region Options:

Promoter/Upstream by 500 bases
5' UTR Exons
CDS Exons
3' UTR Exons
Introns
Downstream by 1000 bases
One FASTA record per gene.
One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra dow
Split UTR and CDS parts of an exon into separate FASTA records

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they m

Sequence Formatting Options:

Exons in upper case, everything else in lower case.
CDS in upper case, UTR in lower case.
All upper case.
All lower case.
Mask repeats: to lower case to N

get sequence cancel

6

```
>hg18_refFlat_SERPINC1 range=chr1:172153140-172153639 5
gaaacgcgcttggcatgcaccccgaggccctgctcttctctccctgtccc
accacttcagggctgctgggaatgggtctctctgtgggccacaggtgta
accattgtgttttccttgtctgtgccagggacaccttggcactcagatgc
ctgaaggtagcagcttgtccctctttgccttctctaattagatatttctc
tctctctctccctctctccataaagaaaactatgagagagggtgggtatg
aaccaagtttgtttccttggttagtttcctaaccaagtttgagggtatga
acatactctccttttccttttctataaagctgaggagaagagtgagggag
tgtgggcaagagaggtggctcaggctttccctgggcctgattgaacttta
aaacttctctactaattaaacaacactgggctctacactttgcttaaccc
tgggaactggtcatcagcctttgacctcagttccccctcctgaccagctc
>hg18_refFlat_ALDOB range=chr9:103237884-103238383 5'pad=0 3'
ttcttgaatgcctactgcgtacagacactatacaactcacttaatgctcc
tgtgaaattgagatcccacttacagagaagaagaaaataacggctcagacaa
gtaaaaaacaaaaccaaaaacccgtcaccaagttcatgtgttttgtaagt
gttacagctacactaacattcttcctagtatcatttaatggaccctagag
ctccatcattgactagctgtatgacttaagggaacctccttatcttcctt
gccttagtttcctcatctgtgaaatggagggtctggattaggaactccac
ggtctaaagacattccttgcagctttaacattctatgagcataaggagta
tatgggcatgattttaaggactggttgtttatgagccaatcagaggtgtt
gaataaacacctccctactaggtcaaggtagaaaggggagggcaaatatt
ggaaaaaaaaaacatgatgagaagtctataaaaattgtgtgctaccaaag
>hg18_refFlat_AMBP range=chr9:115880574-115881073 5'pad=0 3'p
tctgggcaagtcacttcacttcttggagactcaactcctctgcctgtaaa
ggggacaaatagcacctacttcgtaagcttgttgtggggatttactggac
acctgtggcacgtgtaaagtgcctggtacacagccagcacatagcaggtg
ctcagtgaacggaagctcgtaacagtcatcattgttaatgtctctattat
ttgcaaaggctagggcaagcctgggggctggagaaggaactctggagaca
ggatacctgggttctggtcttggctctgcccatagacttccctgaggtct
ctggcaagtcacttcctctccttgggcctcagtttttcctgtgagaaat
gggcccaatcaggcttgacctgctggctcacagagctgtgaggtccagat
ggcatgactcagcagaaaggccaggtcgctccccatcctcgcatgccctc
tgtgggggacttttgggggggcggagcccaatcctggtgctccagggccct
>hg18_refFlat_CYP2C8 range=chr10:96819245-96819744 5'pad=0 3'
tatacatttattagctaataatttgatatatgtatggtaattcaacatgt
atgagttatattcactatttcatgtttaggcagctgtattttaagtgaac
tatactaaatatttgaaaggcttttgttatcaagggctaagtctcctatt
ttttgatatagcattacaatgtacatttttatacacaaaatatagaata
cactgatttccctcaaggtcataaattcccaactggtcattaatctgaga
atattgaattttgagtatattctaacatagaatcatttacttcagtgttt
ctccatcatcacagcacattggaacaaccagggacttttaattaaaaata
cctgggctccaatccaatacaattaaaccagaatctcctagattggcact
ggaaagaaggagtaggacaaaagaacatttttatttctatccatgggccaa
agtccactcagaaaaaaagtataaattggatctaggtgattgtttacttt
```

Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

refFlat Genomic Sequence

Sequence Retrieval Region Options:

- [x] Promoter/Upstream by 1000 bases
- [ ] 5' UTR Exons
- [ ] CDS Exons
- [ ] 3' UTR Exons
- [ ] Introns
- [ ] Downstream by 1000 bases
- (•) One FASTA record per gene.
- ( ) One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra downstream (3')
- [ ] Split UTR and CDS parts of an exon into separate FASTA records

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

Sequence Formatting Options:

- (•) Exons in upper case, everything else in lower case.
- ( ) CDS in upper case, UTR in lower case.
- ( ) All upper case.
- ( ) All lower case.
- [x] Mask repeats: ( ) to lower case (•) to N

get sequence cancel

```
>UpSt_ANGPTL3 range=chr1:62834775-62835774 5'pad=0
aattggctgggctcacgcttgtaatcggctgnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnntgttattaatct
cctctttttgttaaatgtatattatcaaaattgttactaagctaacaaac
ttcagaaaaacttatgatgggcaagctgcttgtgacattgaaggtattta
agattcaattctagtttggtcctagatgaccacatatccattgttccttc
aacgagcacatggtaaagagcctagaacacagagacacagaacacagtgg
agaaaagggagtgaaatgtctttaatgacacttactatatatgggatttt
gtgacaatatacaaggatggttaagacatataaggtgatgcaaaaaaaca
tattaacaattatagtgacaaaaaatgaggagcatataattatacattga
tttatacagagtaccagaggaacacagcattgagagccgtaacaccacct
gagggagtggagaaaggcttcagagagaaagtgtttttggaatggatca
ctgtttccaaaagaactaaagtacagtttgagaaatgcatacttaattca
ttacttttttcccctcaactttaataataaatttacccaacaaaaaagtt
tatttttgacttgtaaatctcttaaaatcataaaaaagtaaaattagctt
ttaaaaacaggtagtcaccatagcattgaatgtgtagtttataatacagc
aaagttaaatacaatttcaaattacctattaagttagttgctcatttctt
tgatttcatttagcattgatctaactcaatgtggaagaaggttacattcg
tgcaagttaacacggcttaatgattaactatgttcacctaccaaccttac
ctttctgggcaaatattggtatatatagagttaagaagtctaggtctgc
>UpSt_F13B range=chr1:195303021-195304020 5'pad=0 3
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnttcctattttgtattttaaattaaaactnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnttgatatagcacagtgctgatctcaccacctgc
cacaaagtaaaaactctctaaatgtcaaacactactattataagtaatac
tattctatcattgagacaactatggttctgtatctctttaaggagtgcag
tctaggcaacaacatatttcaacatcaggtatctgttatctcatgttcct
aagttcatctttttacctgatcaatattagtaaagatctatggcctcaag
tgaactttactctatgaatttacaattgggaacttaggttgaaaatttac
ctagtattatttattctgctgctataatttattatttattctgctgctat
tgttgccctaatcagttatcatgctcttactggactctgacatggtaagt
aatgtcaattaatatagttgtgttatttttaacattatttcaaaatatttt
aagtgaaatataagactcgttacaaaaggacttagacagaggttataaat
tattaacttttttcctcaagagaagtattgtgtacttaatgacaagttcct
```

Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.

1

clade: Vertebrate genome: Human assembly: Mar. 2006
group: All Tables database: hg18
table: refFlat describe table schema
region: genome position chrX:151073054-151383976 lookup define regions
identifiers (names/accessions): paste list upload list
filter: create
intersection with ct_UserTrack: edit clear
output format: all fields from selected table Send
output file: (leave blank to keep output in t
file type returned: plain text gzip compressed

Note: Intersection doesn't work with all fields or selected fields output.
get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

1. Select "refFlat" under "table"
2. Ensure that "region" is "genome"
3. Click on "paste list"

Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Paste In Identifiers for refFlat

Please paste in the identifiers you want to include. The items must be values of the **geneName** field of the currently selected table, **refFlat**. (The "describe table schema" button shows more information about the table fields.) Some example values:
AADACL3
GPX4
ACF

2

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)

ABCC3
ABHD2
ACY1
ADH1C
APCS
APOC4
AQP9
ASL
AZGP1

submit clear cancel

Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.

3

clade: Vertebrate genome: Human assembly: Mar. 2006
group: All Tables database: hg18
table: refFlat describe table schema
region: genome position chrX:151073054-151383976 lookup define regions
identifiers (names/accessions): paste list upload list clear list
filter: create
intersection: create
output format: custom track Send output to Galaxy
output file: (leave blank to keep output in browser)
file type returned: plain text gzip compressed

Select the output format as "custom track"

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Output refFlat as Custom Track

Custom track header:
name= UpSt_1kb
description= table browser query on refFlat
visibility= pack
url=

Describe your track

4

Create one BED record per:
- Whole Gene
- Upstream by 1000 bases
- Exons plus 0 bases at each end
- Introns plus 0 bases at each end
- 5' UTR Exons
- Coding Exons
- 3' UTR Exons
- Downstream by 200 bases

Enter number of bp you want to analyze/download

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

get custom track in table browser | get custom track in file
get custom track in genome browser | cancel

Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.

clade: Vertebrate **genome:** Human **assembly:** Mar. 2006
group: Variation and Repeats **track:** SNPs (126)
table: snp126 describe table schema
region: genome position chrX:151073054-151383976 lookup define regions
identifiers (names/accessions): paste list upload list
filter: create
intersection: create
correlation: create
output format: custom track
output file: (leave blank to keep out
file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

5

1. Select "Variation and Repeats" under "Group"
2. Click on "create" under "intersection"

Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Intersect with SNPs (126)

Select a group, track and table to intersect with:
group: Genes and Gene Prediction Tracks **track:** UCSC Genes
table: UCSC Genes (knownGene)

6

These combinations will maintain the gene/alignment structure (if any) of SNPs (126):

- All SNPs (126) records that have any overlap with UCSC Genes
- All SNPs (126) records that have no overlap with UCSC Genes
- All SNPs (126) records that have at least 80 % overlap with UCSC Genes
- All SNPs (126) records that have at most 80 % overlap with UCSC Genes

These combinations will discard the gene/alignment structure (if any) of SNPs (126) and produce a simple list of position ranges.

- Base-pair-wise intersection (AND) of SNPs (126) and UCSC Genes
- Base-pair-wise union (OR) of SNPs (126) and UCSC Genes

Check the following boxes to complement one or both tables. To complement a table means to include a row in the intersection if it is *not* included in the table.

- Complement SNPs (126) before intersection/union
- Complement UCSC Genes before intersection/union

submit cancel

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

Home Genomes Genome Browser Blat ... sion FAQ Help

Intersect with SNPs (126)

Select a group, track and table to intersect with:
group: Custom Tracks **track:** User Track
table: User Track (ct_UserTrack)

7

These combinations will maintain the gene/alignment structure (if any) of SNPs (126):

- All SNPs (126) records that have any overlap with User Track
- All SNPs (126) records that have no overlap with User Track
- All SNPs (126) records that have at least 80 % overlap with User Track
- All SNPs (126) records that have at most 80 % overlap with User Track

These combinations will discard the gene/alignment structure (if any) of SNPs (126) and produce a simple list of position ranges.

- Base-pair-wise intersection (AND) of SNPs (126) and User Track
- Base-pair-wise union (OR) of SNPs (126) and User Track

Check the following boxes to complement one or both tables. To complement a table means to include a row in the intersection if it is *not* included in the table.

- Complement SNPs (126) before intersection/union
- Complement User Track before intersection/union

submit cancel

Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.

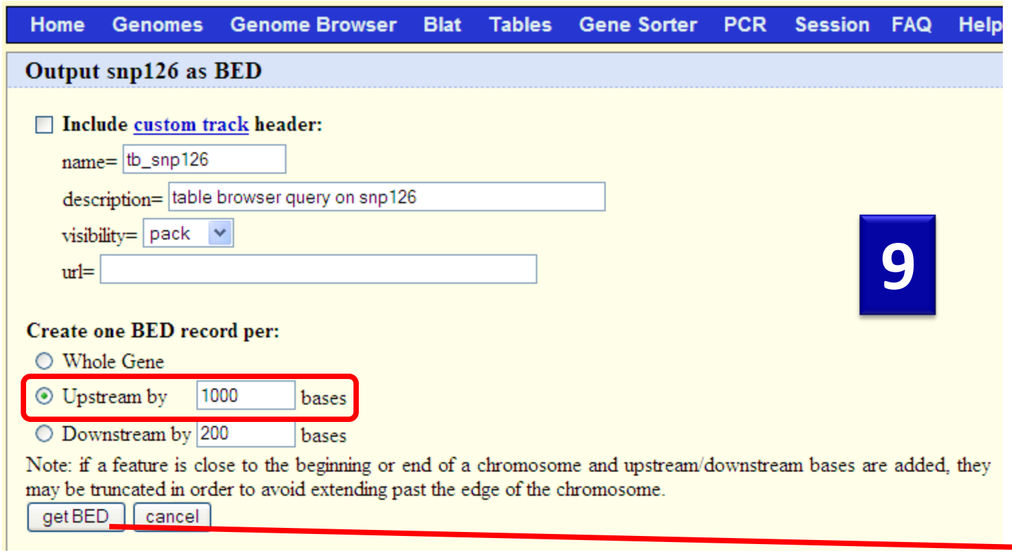
clade: Vertebrate **genome:** Human **assembly:** Mar. 2006
group: Variation and Repeats **track:** SNPs (126)
table: snp126 describe table schema
region: genome position chrX:151073054-151383976 lookup define regions
identifiers (names/accessions): paste list upload list
filter: create
intersection with ct_UserTrack: edit clear
correlation: create
output format: BED - browser extensible data Send output to Galaxy
output file: (leave blank to keep output in browser)
file type returned: plain text gzip compressed

Note: Intersection doesn't work with all fields or selected fields output.
get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

8

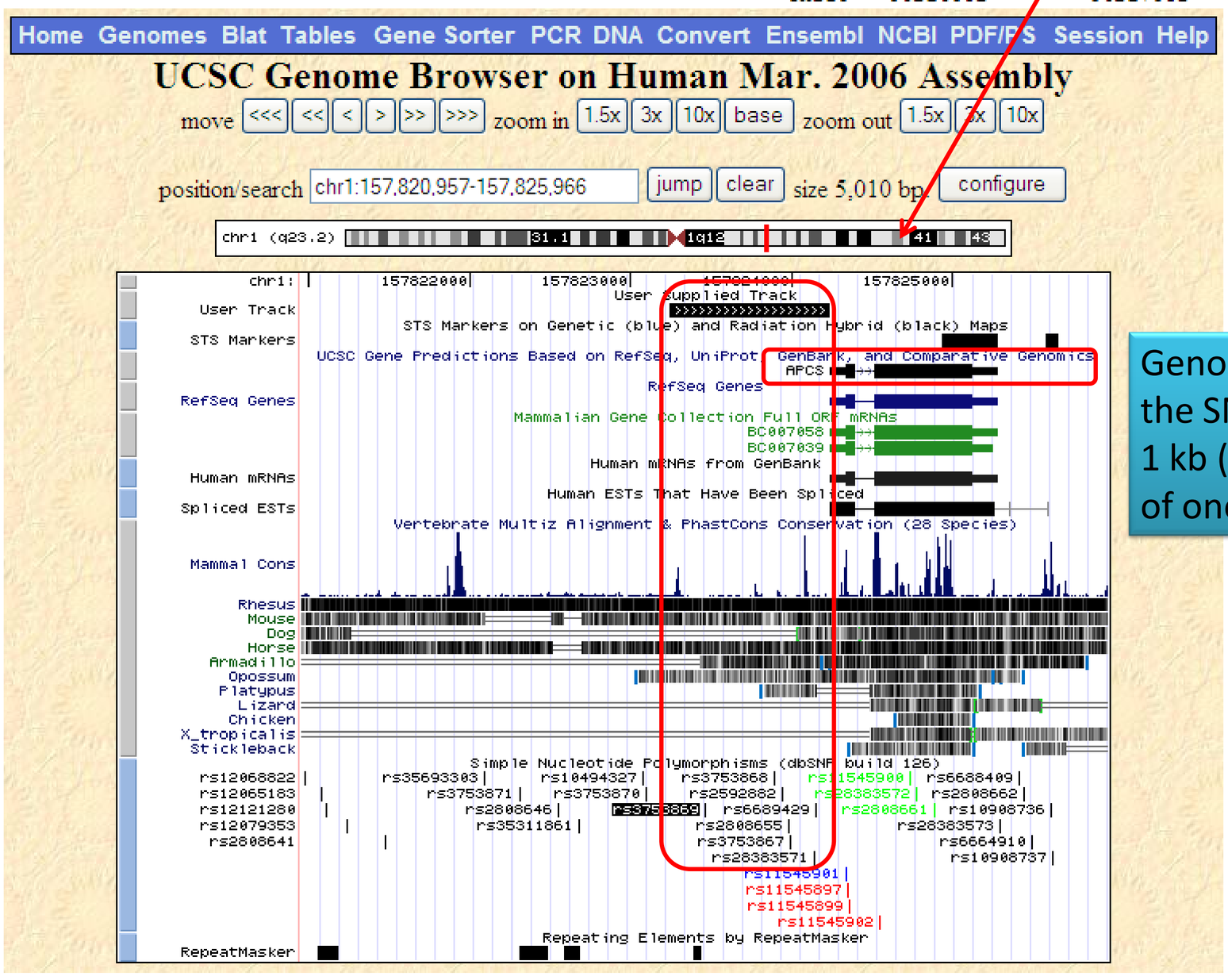
Try GTF output too



9

```
chr1    157823462    157824462    rs3753869_up_1000_chr1_157823463_r    0    -
chr1    157823893    157824893    rs3753868_up_1000_chr1_157823894_r    0    -
chr1    157822934    157823934    rs2592882_up_1000_chr1_157822935_f    0    +
chr1    157822978    157823978    rs2808655_up_1000_chr1_157822979_f    0    +
chr1    157823993    157824993    rs3753867_up_1000_chr1_157823994_r    0    -
chr1    157823138    157824138    rs28383571_up_1000_chr1_157823139_f   0    +
chr1    157823144    157824144    rs6689429_up_1000_chr1_157823145_f    0    +
chr4    100493180    100494180    rs4147541_up_1000_chr4_100493181_r    0    -
chr4    100492308    100493308    rs1789924_up_1000_chr4_100492309_f    0    +
chr4    100493487    100494487    rs34774688_up_1000_chr4_100493488_r   0    -
chr4    100492593    100493593    rs17586163_up_1000_chr4_100492594_f   0    +
chr4    100493772    100494772    rs11499823_up_1000_chr4_100493773_r   0    -
chr4    100492918    100493918    rs1629838_up_1000_chr4_100492919_f    0    +
chr7    65176569     65177569     rs10247708_up_1000_chr7_65176570_f    0    +
chr7    99411067     99412067     rs4727442_up_1000_chr7_99411068_f     0    +
chr15   56215720     56216720     rs11629801_up_1000_chr15_56215721_f   0    +
chr15   56216839     56217839     rs1554203_up_1000_chr15_56216840_r    0    -
chr15   56217082     56218082     rs3840843_up_1000_chr15_56217083_r    0    -
chr15   56216231     56217231     rs8027250_up_1000_chr15_56216232_f    0    +
chr15   56216582     56217582     rs9920853_up_1000_chr15_56216583_f    0    +
                                  rs3742955_up_1000_chr15_56216602_f    0    +
                                  rs293370_up_1000_chr15_87431481_r     0    -
                                  rs293369_up_1000_chr15_87432137_r     0    -
                                  rs4148404_up_1000_chr17_46065400_f    0    +
                                  rs35467079_up_1000_chr17_46065407_f   0    +
                                  rs9895420_up_1000_chr17_46066037_f    0    +
                                  rs4793665_up_1000_chr17_46066086_f    0    +
                                  rs12721101_up_1000_chr19_50135470_f   0    +
                                  rs35366732_up_1000_chr19_50135470_f   0    +
                                  rs4803773_up_1000_chr19_50135582_f    0    +
                                  rs12721105_up_1000_chr19_50136135_f   0    +
```

10



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 their any putative microRNA regulators for SLC7A1? If yes, download all of them using table browser.

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

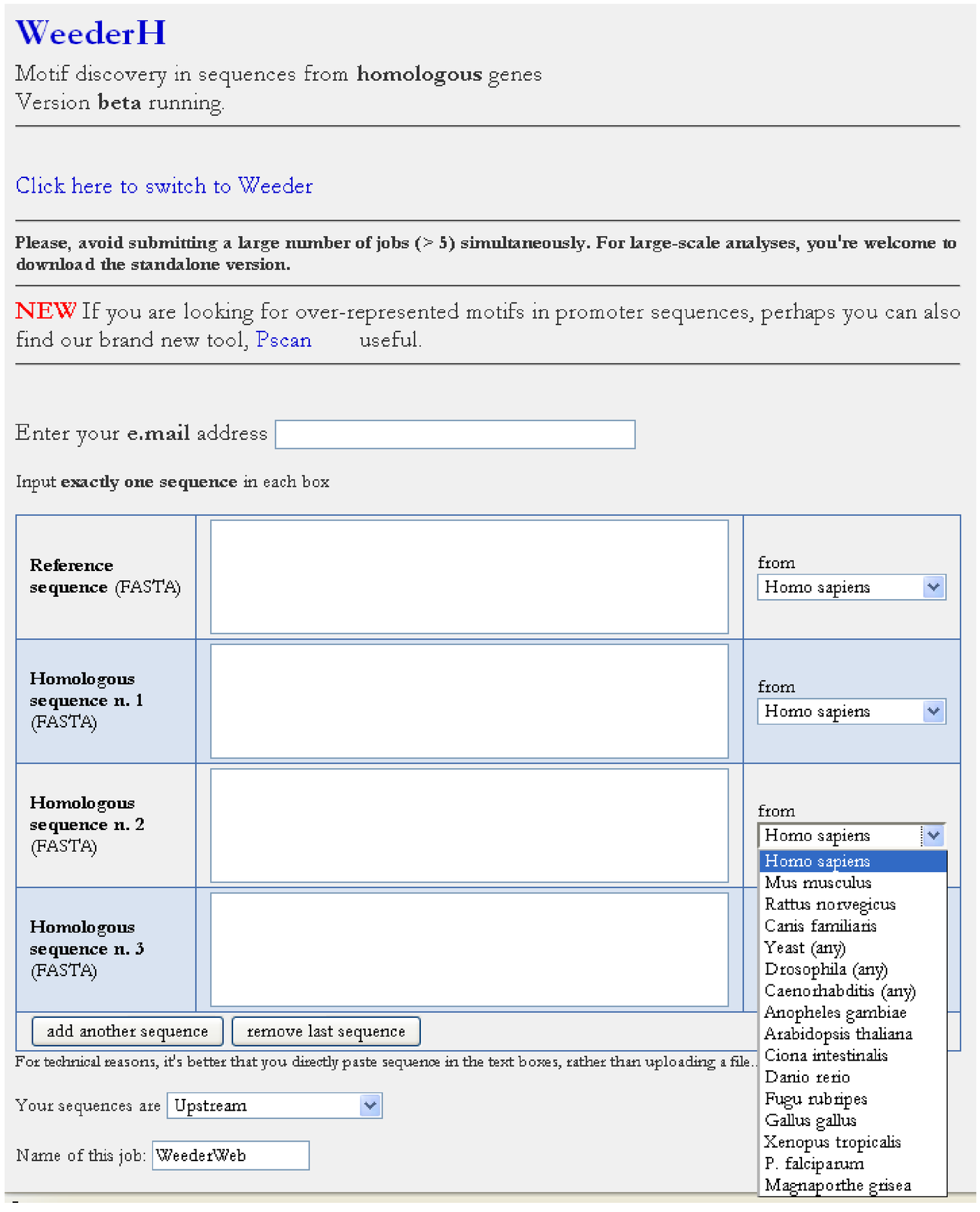
I want to find the shared cis-elements – Known and Novel

- Known transcription factor binding sites (TFBS)
  - Conserved
    - oPOSSUM
    - DiRE
  - Non-conserved
    - Pscan
    - MatInspector (*Licensed)
- 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.

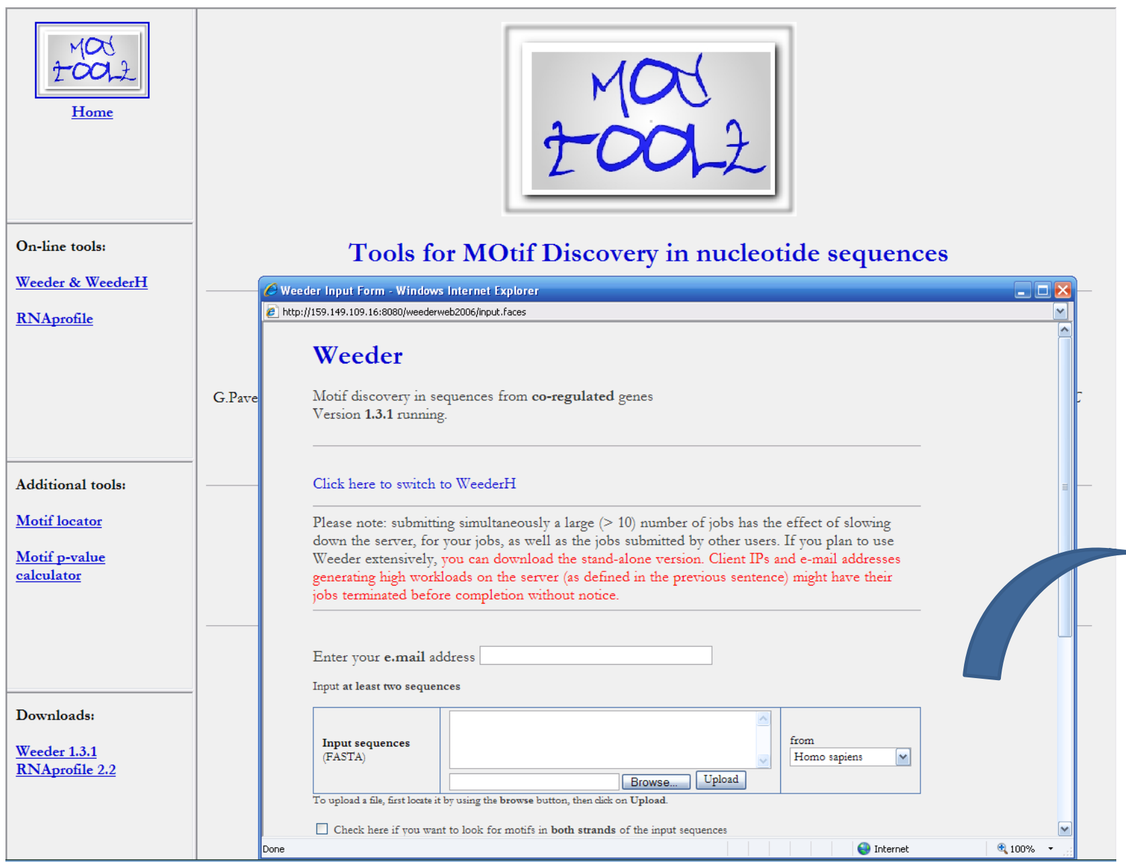
Use the fetched promoter/upstream sequences for the following analyses

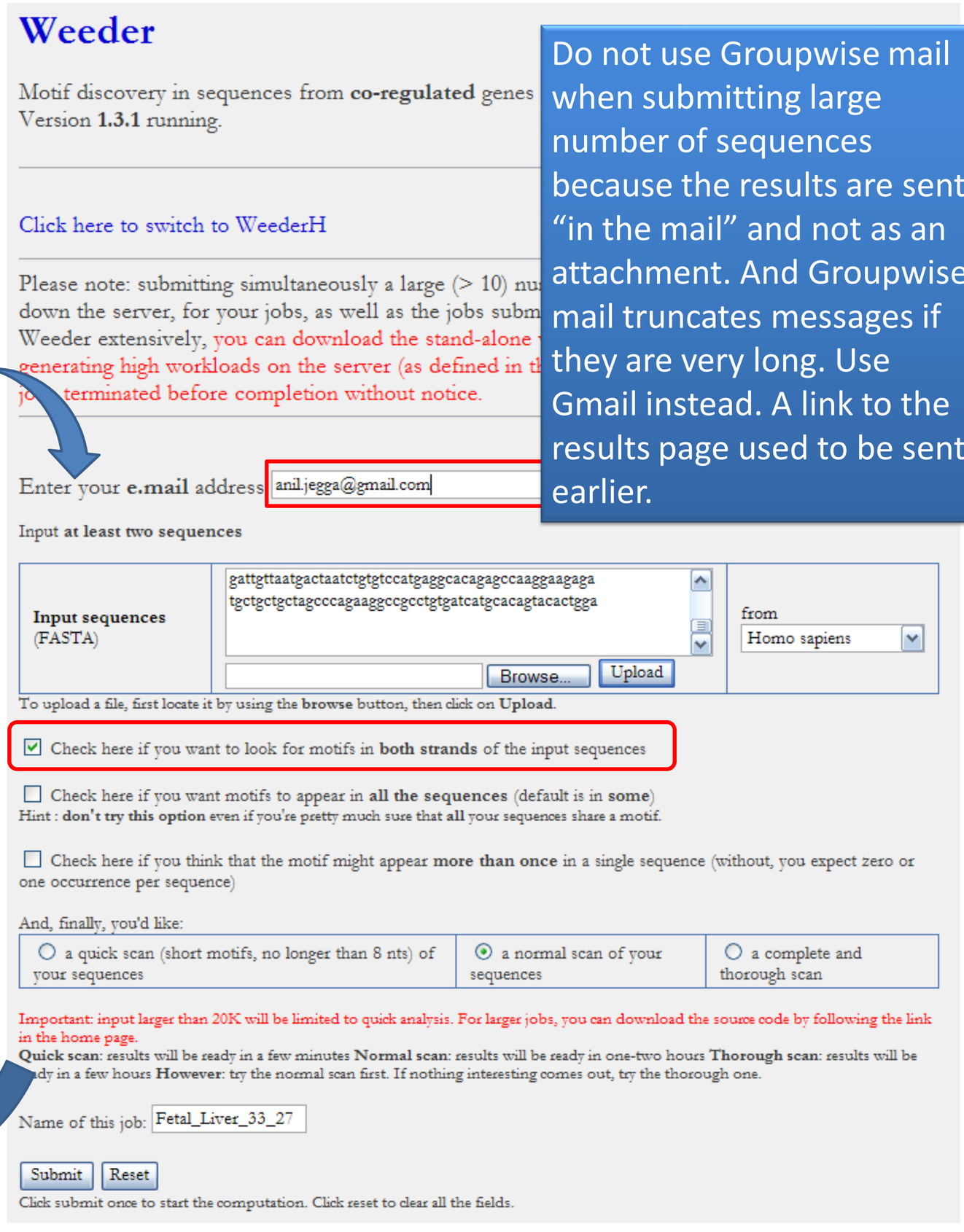
WeederH (http://159.149.109.9/pscan)



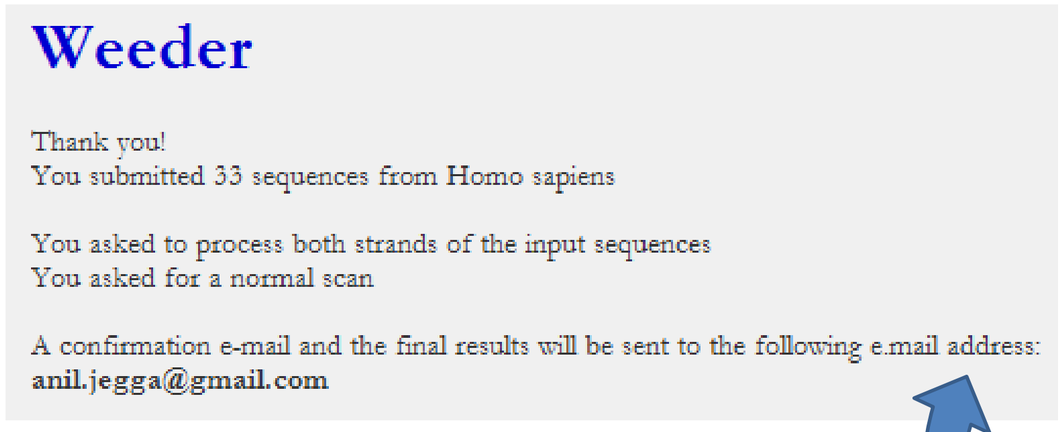
1. Supports large number of species.
2. Does not support multiple sequences (multifasta) input. You have to enter each sequence separately.
3. Good for small number of sequences where you expect a potential novel (or not included in the TFBS libraries) conserved motif.

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





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 (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) ATTGAT 0.67
4) TATGAT 0.63
5) GATTTA 0.61
6) ATGGTA 0.60
7) TCATTG 0.59
8) TGGTAT 0.59
9) TGATTA 0.59
10) TGATAT 0.58

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

1) CGTTTAGA 0.93
2) ACTAAACG 0.88
3) GATAAACT 0.87
4) TATGGTAT 0.87
5) CTAAACGT 0.87
6) AGTATTTC 0.84
7) ACATTGAT 0.82
8) GTAATACT 0.80
9) CTAGCAAT 0.79
10) ATAGTTCG 0.78

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

GATAAACT
AGTTTATC

0 redundant motifs found:

Best occurrences (match percentage):

Seq	St	oligo	pos	match
1	+	.GAAAAACT.	205	(92.84)
1	+	[AATAAATT]	676	(85.29)
1	+	[GATTAACT]	922	(88.60)
1	-	.TATAAACT.	786	(92.79)
1	-	.AATAAACT.	697	(92.36)
1	-	[GATAATAT]	169	(85.17)
2	+	[TAAAAACT]	508	(85.63)
2	+	[TATAAATT]	944	(85.73)
2	-	[GAAAAAGT]	956	(85.28)
2	-	[AATAAATT]	776	(85.29)
2	-	.GATGAACT.	652	(90.33)
4	+	[AATAAAAT]	546	(87.13)
4	+	[GAGAAAAT]	786	(85.24)
5	+	[AATAAATT]	393	(85.29)
5	-	[GAGAAAAT]	260	(85.24)
6	+	[TATAAAAT]	733	(87.56)
7	-	.GATAAAAT.	430	(94.77)
8	+	[AATAAAAT]	307	(87.13)
8	+	[AATAAAAT]	791	(87.13)
8	-	[AAAAAACT]	808	(85.19)
8	-	[AATAAATT]	484	(85.29)
8	-	[TATAAAGT]	285	(85.24)
8	-	[TATAAAAT]	13	(87.56)
9	+	.GATAAACT.	603	(100.00)
9	+	[GAGAAAAT]	615	(85.24)
9	-	.GATAAAAT.	438	(94.77)
10	+	.GATAAACT.	603	(100.00)
10	+	[GAGAAAAT]	615	(85.24)
10	-	.GATAAAAT.	438	(94.77)
11	+	[GATGAAAT]	148	(85.10)
11	+	[GAAAAATT]	205	(85.77)
12	+	.GATAAATT.	143	(92.93)
12	+	.TATAAACT.	271	(92.79)
12	+	[AATAAAAT]	286	(87.13)
12	+	.GATAAACA.	523	(90.60)
12	+	.GATAAAAT.	896	(94.77)
12	-	[AATAAATT]	347	(85.29)
13	+	.AATAAACT.	549	(92.36)
13	-	[GATAAGCT]	832	(88.34)
13	-	[GATAAGCT]	577	(88.34)
14	+	.AATAAACT.	549	(92.36)
14	-	[GATAAGCT]	832	(88.34)
14	-	[GATAAGCT]	577	(88.34)
16	+	.GATAAAAT.	161	(94.77)
16	+	[GACAAACT]	316	(89.17)
16	+	[AATAAATT]	814	(85.29)
16	-	.AATAAACT.	967	(92.36)
16	-	.GATAATCT.	943	(90.40)
18	-	[GTTAAACT]	637	(89.17)
18	-	.GATAAAAT.	555	(94.77)

Frequency Matrix

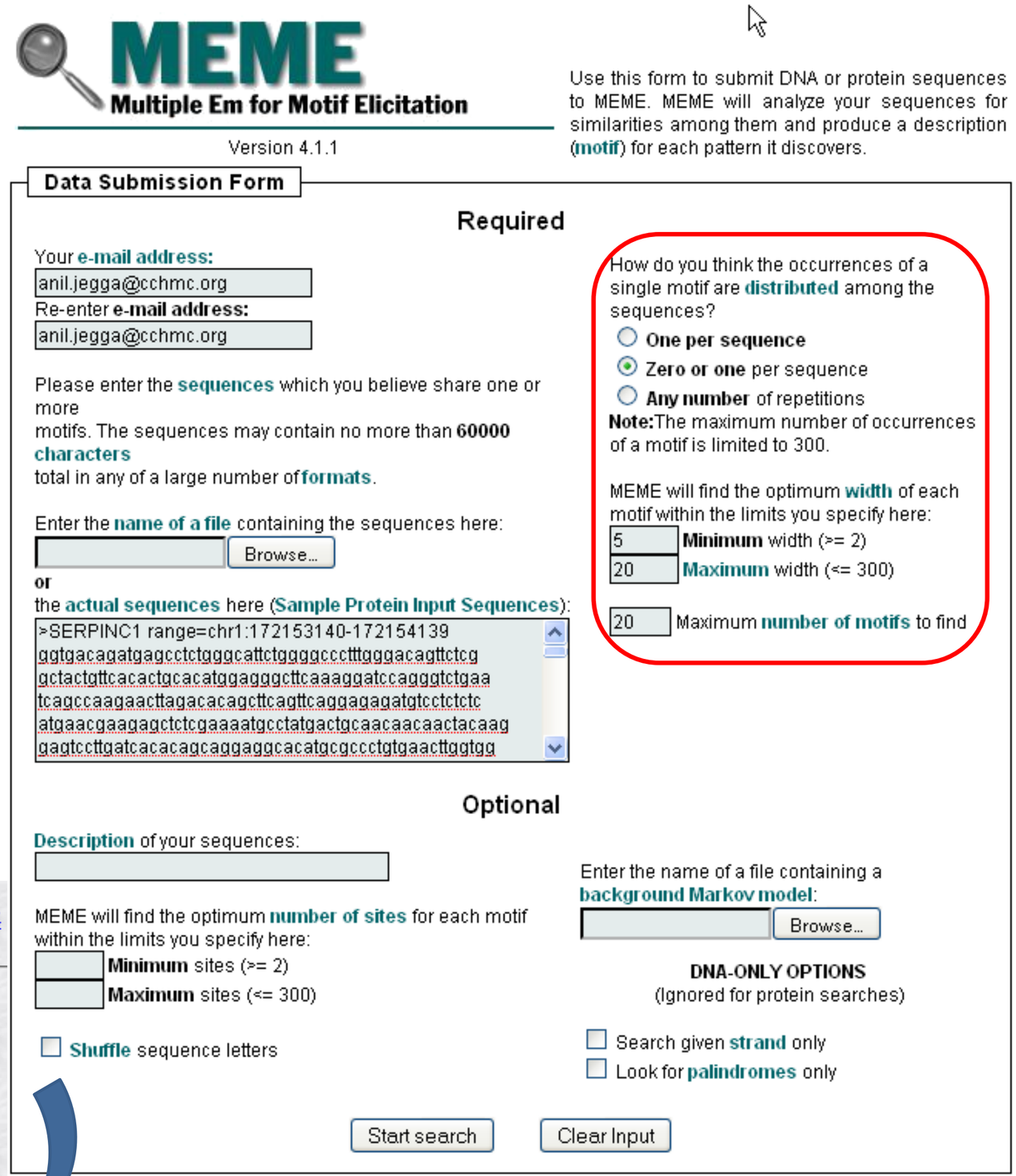
	All Occs				Best Occs			
	A	C	G	T	A	C	G	T
1	28	16	167	31	4	0	20	2
2	201	8	17	16	26	0	0	0
3	33	14	19	176	1	0	0	25
4	201	6	21	14	25	0	1	0
5	208	6	9	19	26	0	0	0
6	198	10	13	21	25	0	0	1
7	43	146	25	28	7	16	2	1
8	22	17	5	198	1	0	0	25

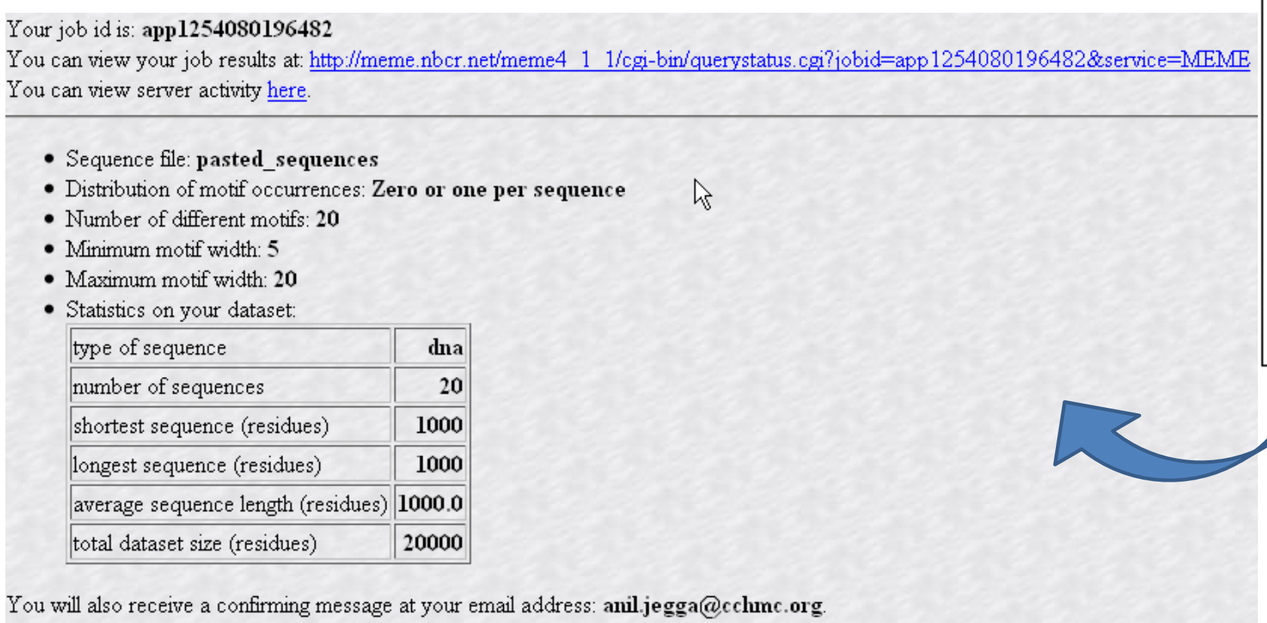
MEME (http://meme.sdsc.edu)

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.nbcr.net/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: **anil.jegga@cchmc.org**.

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

MEME Job app1254080196482

- /home/install/usr/apps/meme_4.1.1/bin/meme sequences -sf past

Results

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

Messages

- Processing Messages
- Error Messages

P N

MOTIF 1 width = 20 sites = 7 llr = 125 E-value = 1.9e+000

SEQUENCE LOGO

Information Content

24.4 (bits)

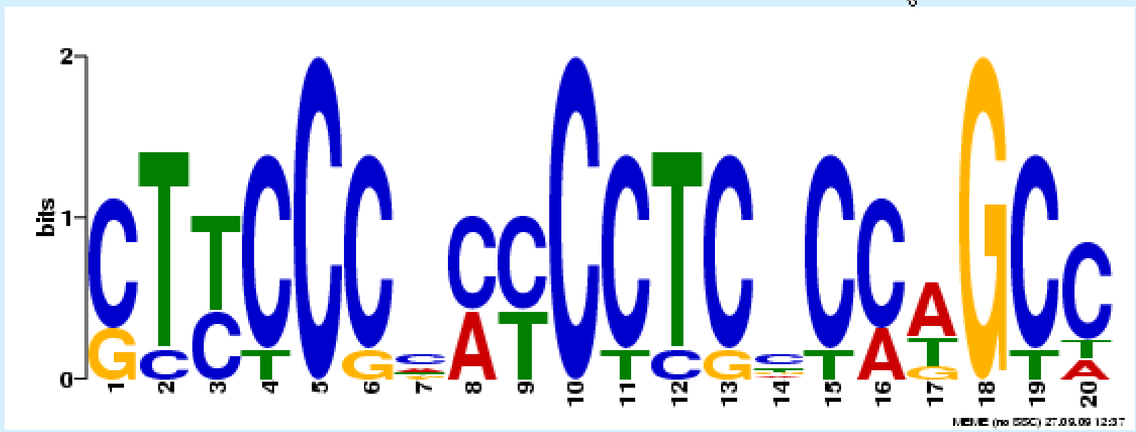
Relative Entropy

25.8 (bits)

Download LOGO

Without SSC: [EPS][PNG]

With SSC: [EPS][PNG]

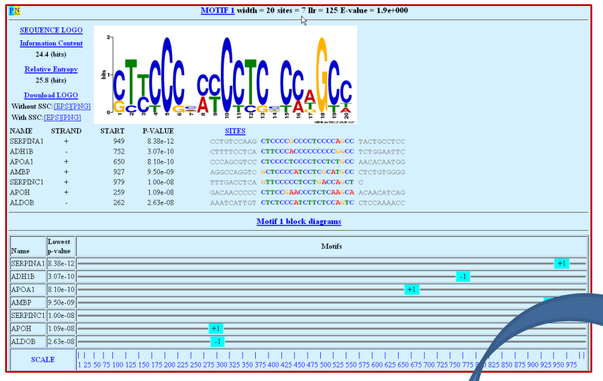


NAME	STRAND	START	P-VALUE	SITES
SERPINA1	+	949	8.38e-12	CCTGTCCAAG CTCCCCGCCCCTCCCCAGCC TACTGCCTCC
ADH1B	-	752	3.07e-10	CTTTTCCTCA CTTCCCACCCCCCCCCGGCC TCTGGAATTC
APOA1	+	650	8.10e-10	CCCAGCGTCC CTCCCCTCCCCTCCTCTGCC AACACAATGG
AMBP	+	927	9.50e-09	AGGCCAGGTC GCTCCCCATCCTCGCATGCC CTCTGTGGGG
SERPINC1	+	979	1.00e-08	TTTGACCTCA GTTCCCCCTCCTGACCAGCT C
APOH	+	259	1.09e-08	GACAACCCCC CTTCCGAACCCTCTCAAGCA ACAACATCAG
ALDOB	-	262	2.63e-08	AAATCATTGT CTCTCCCATCTTCTCCAGTC 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 (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
Motif Comparison Tool
Version 4.1.1

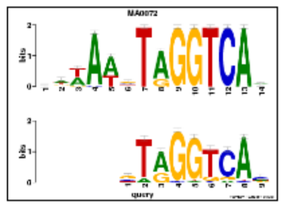
Use this form to use TOMTOM to compare your DNA motif against a database of known motifs (e.g., JASPAR or Transfac).
TOMTOM will rank the motifs in the target database by the *E*-value of the similarity score. TOMTOM outputs the *E*-value of each match, and an alignment of LOGOs representing your query and the target motif it matches. For JASPAR and Transfac searches, you can click on the motif ID to get more information. (A Transfac license is required for viewing Transfac details.)

Data Submission Form

Required

TOMTOM will compare your uploaded motif with each motif in a DNA motif database.

Example Alignment



Select the DNA Motif Database to search:
TRANSFAC (COMMERCIAL)

Select the Motif Column Comparison Function:
Pearson correlation coefficient (pearson)

Optional

Description of your motif:

Start search | Clear Input

TOMTOM OUTPUT

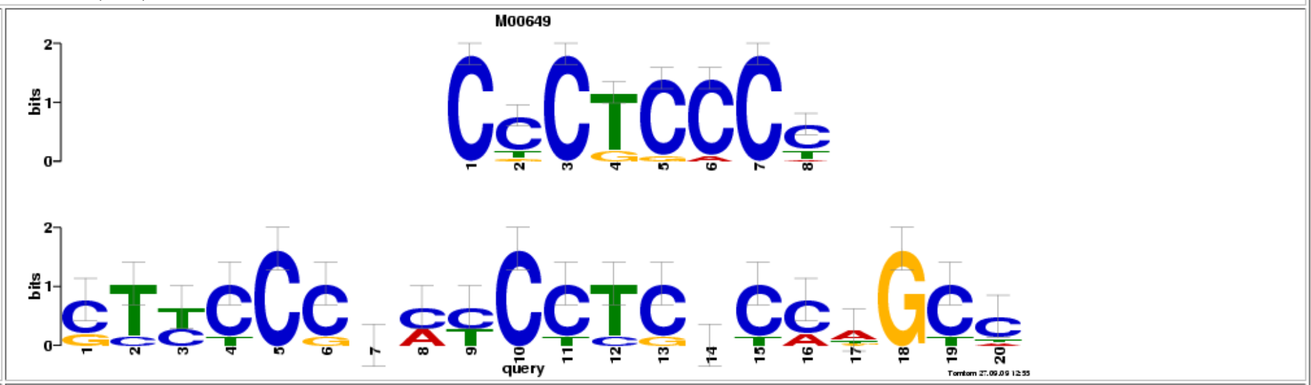
Query File: ../output/tomtom_64592614_99603686/query.meme

Target File: /home/meme/meme_4.1.1/bin//db/motif_databases/transfac_matrix.meme

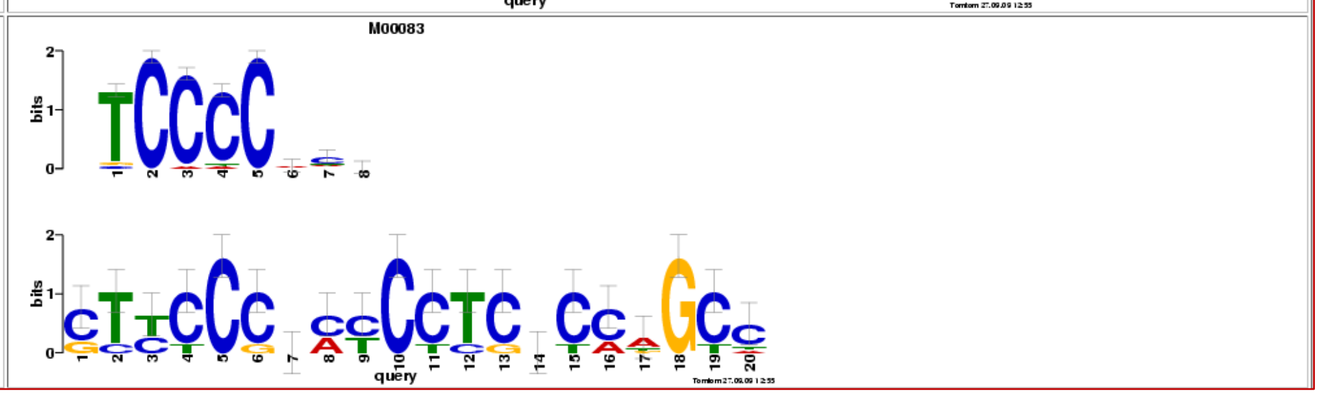
Distance Measure: pearson

All Motif Matches with *q* value at most: 0.5. The q-value is the estimated false discovery rate if 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*

Target Motif:	M00649
Target Description:	MAZ
Query Motif:	query
Query Description:	
p-value:	0.00015
q-value:	0.24
Overlap:	8
Query Offset:	-8
Orientation:	-
Figures:	[EPS][PNG]



Target Motif:	M00083
Target Description:	MZF1
Query Motif:	query
Query Description:	
p-value:	0.00034
q-value:	0.27
Overlap:	8
Query Offset:	-1
Orientation:	-
Figures:	[EPS][PNG]



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

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

Select	Query
◉	Disease *(Always use **Disease Selector**)*
○	Pathway *(Always use **Pathway Selector**)*
○	Gene ontology *(Always use **Ontology Selector**)*
○	Mammalian phenotype *(Always use **Phenotype Selector**)*
○	Select gene family from the list
○	Select custom group from the list

Search

Query took 1.514 s

(2 genes meet the search criteria)

Submit Reset

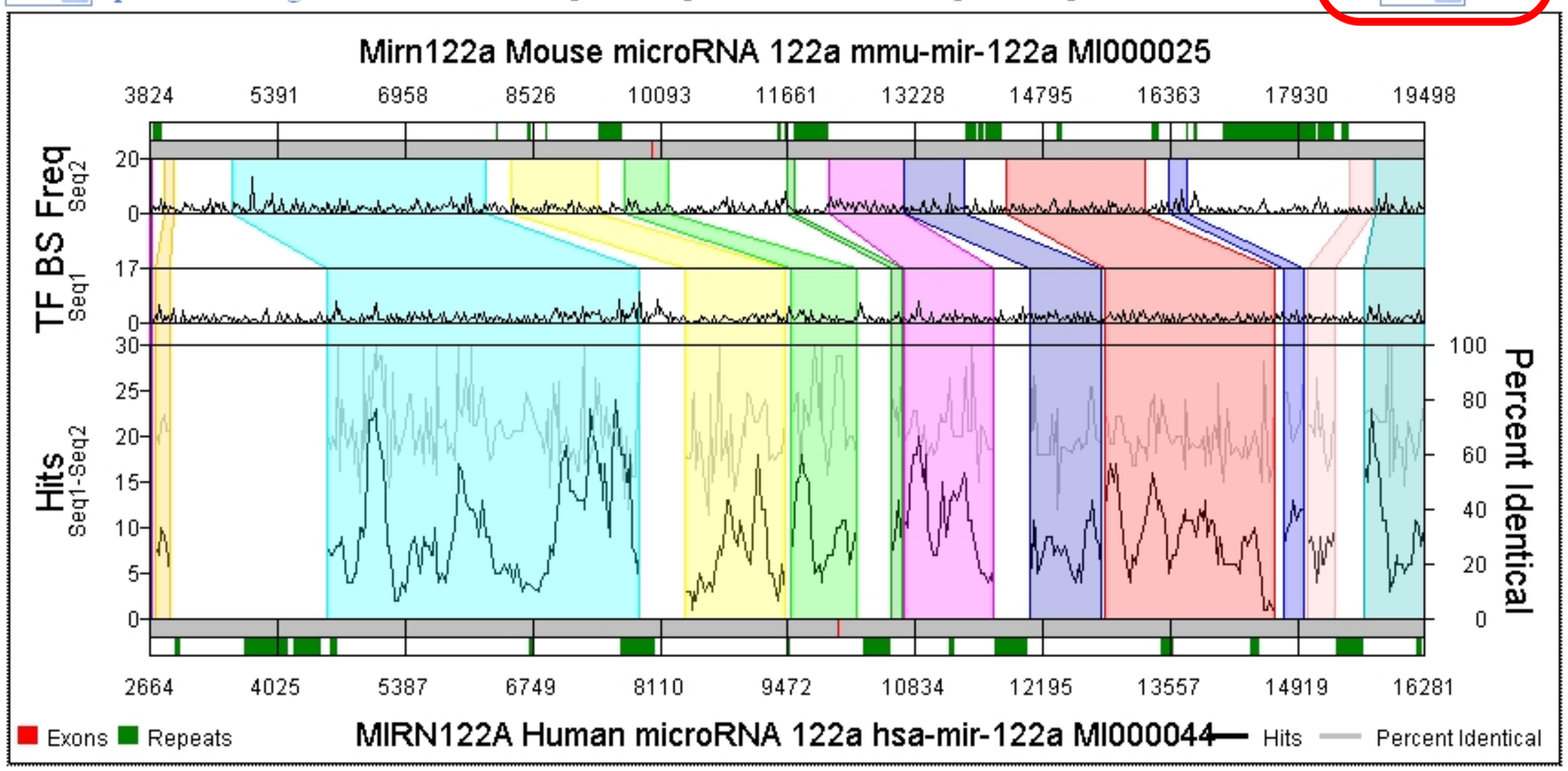
✓	Query Term	Accession Number	Name
☑	MIR-122A	hgMIRN122A	MIRN122A Human microRNA 122a hsa-mir-122a MI0000044
☑	MIR-122A	mgMirn122a	Mirn122a Mouse microRNA 122a mmu-mir-122a MI0000025

GenomeTrafac: http://genometrafac.cchmc.org

First Sequence	Second Sequence	Timestamp	Action
hgMIRN122A, MIRN122A Human microRNA 122a hsa-mir-122a MI000044	mgMirn122a, Mirn122a Mouse microRNA 122a mmu-mir-122a MI000025	07/25/2006 12:00	View Regulogram

Regulogram

200 bp TraFaC Image First Seq. UCSC map Second Seq. UCSC map 10x Zoom



Size: 850 X 412

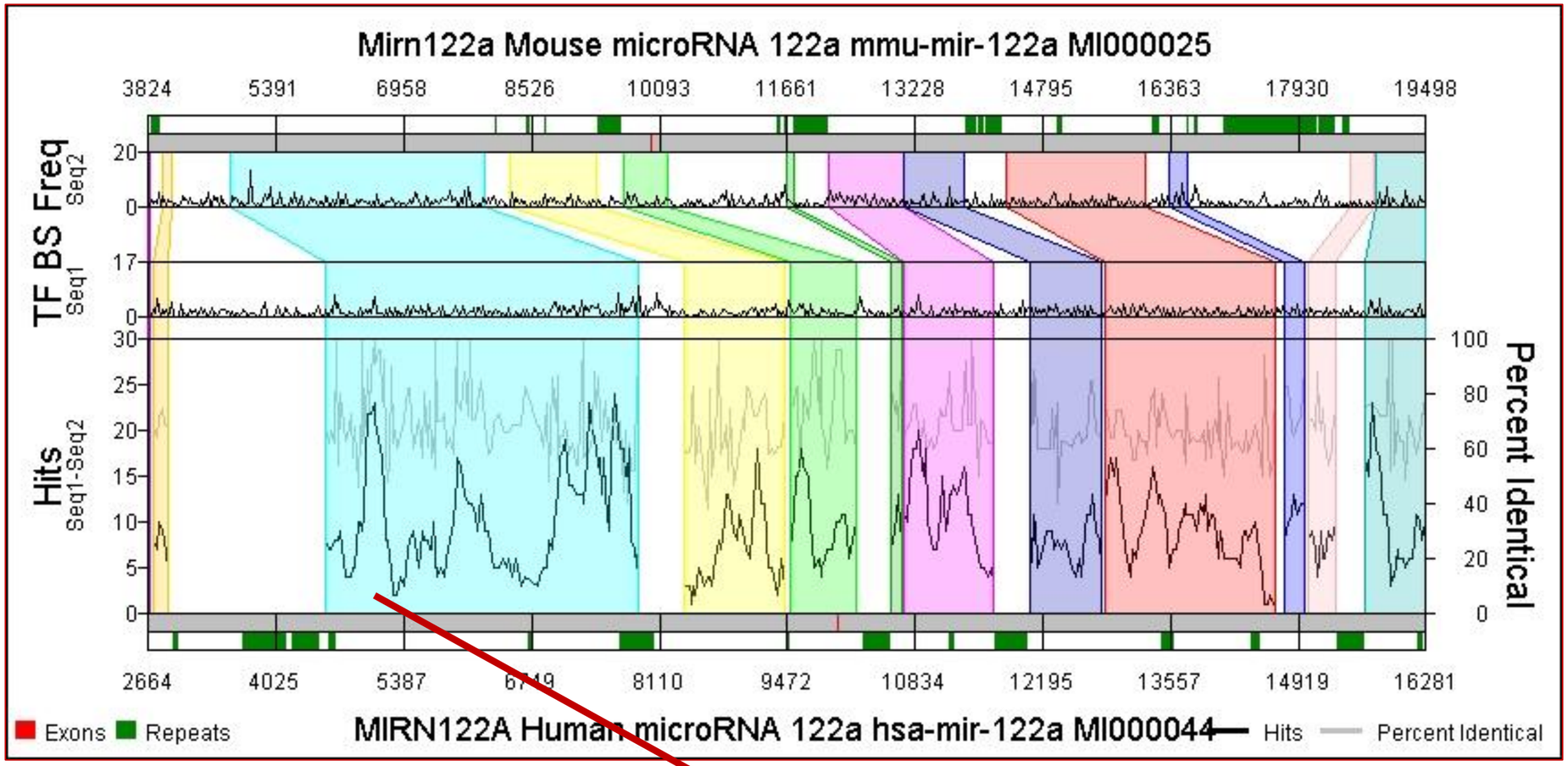
Show ordered same-family hits also

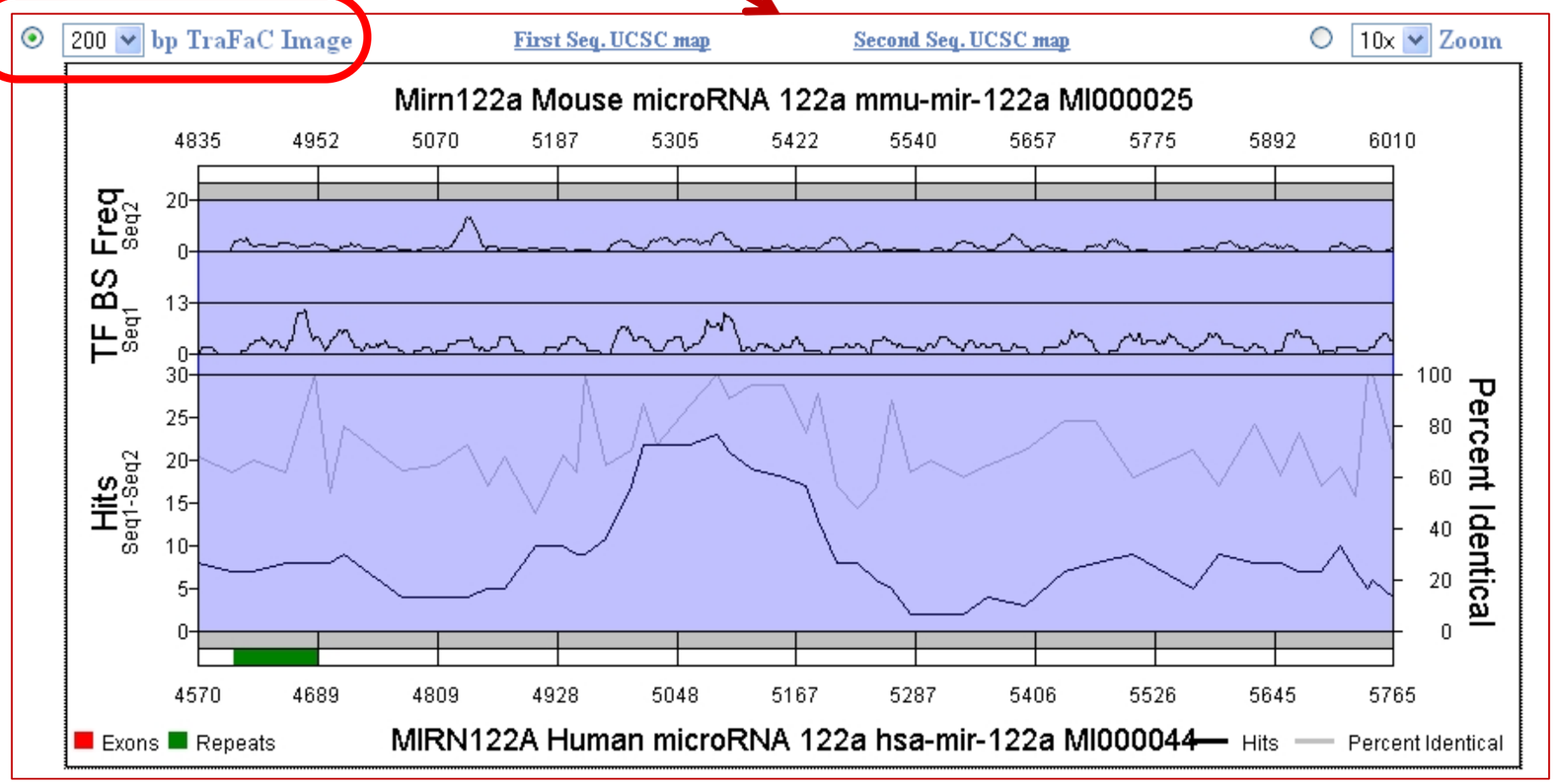
Combine unordered same-family matrices

Show only ordered same-family hits

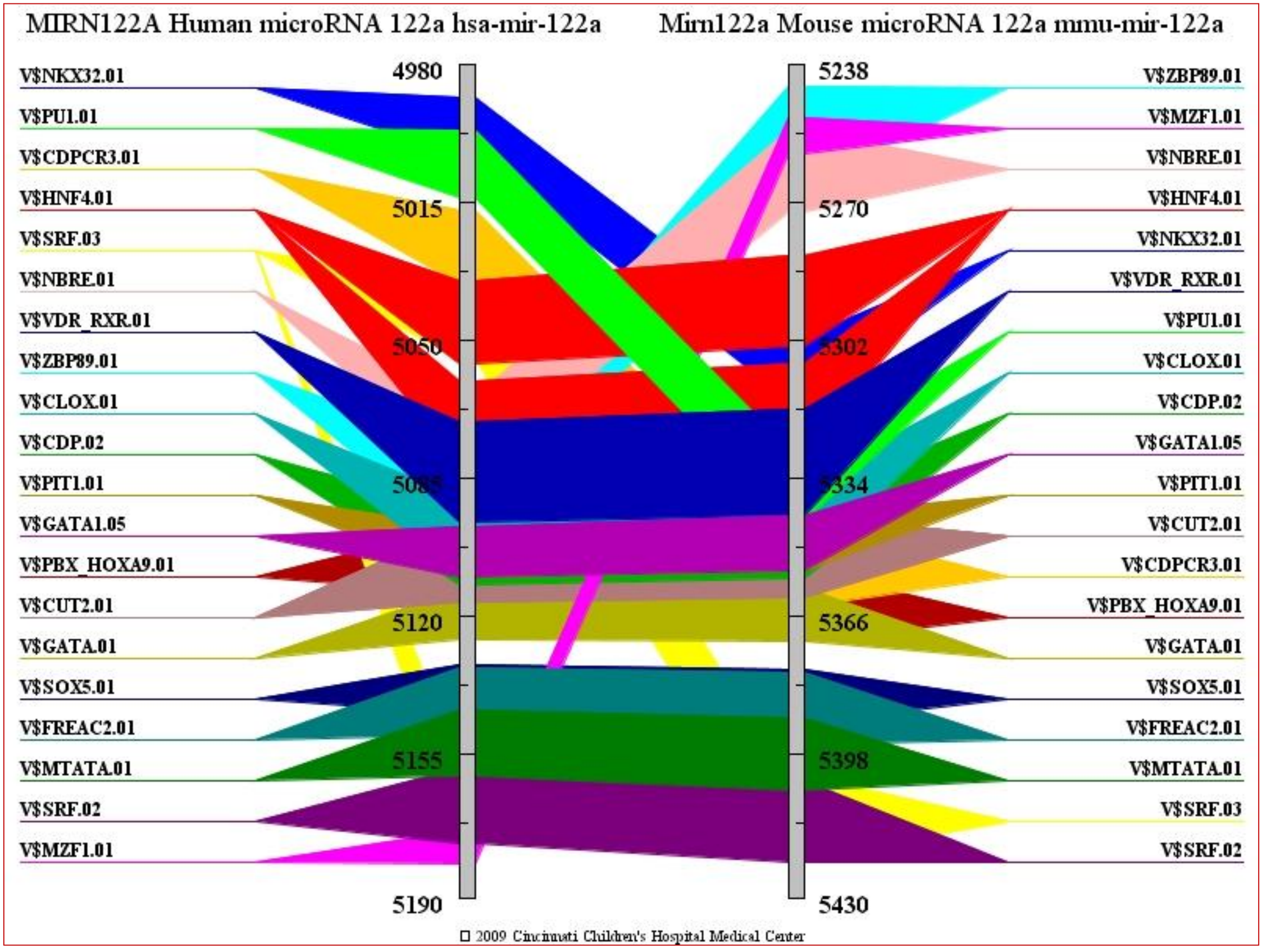
Mirn122a Mouse microRNA 122a mmu-mir-122a MI000025 Export mouse GFF Open mouse GFF_UCSC	From: 3824 To: 19498
MIRN122A Human microRNA 122a hsa-mir-122a MI000044 Export human GFF Open human GFF_UCSC	From: 2664 To: 16281

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)

Family/Matrix	Description	hgMIRN122A				mgMirn122a			
		Begin	End	Sequence		Begin	End	Sequence	
V$NKXH/V$NKX32.01	Homeodomain protein NKX3.2 (BAPX1, NKX3B, Bagpipe homolog)	4993	5007	CCCCCACTCAGCAGA	-	5301	5315	CTGACTTAGTGGACT	+
V$ETSF/V$PU1.01	Pu.1 (Pu120) Ets-like transcription factor identified in lymphoid B-cells	5001	5017	CAGCAGAGGAATGGACT	+	5326	5342	CCTCTCTTCCCCCACAA	-
V$CLOX/V$CDPCR3.01	Cut-like homeodomain protein	5020	5038	CCAATCTTGCTGAGTGTGT	-	5343	5361	TCGATAATTTAATGTGACT	-
V$HNF4/V$HNF4.01	Hepatic nuclear factor 4	5037	5057	GTTTGACCAAAGGTGGTGCTG	+	5283	5303	GTTTGACCAAAGGTGACTCTG	+
V$SRFF/V$SRF.03	Serum responsive factor	5038	5056	TTTGACCAAAGGTGGTGCT	-	5399	5417	GGATCCCATAAAGGGAGAG	-
V$HNF4/V$HNF4.01	Hepatic nuclear factor 4	5061	5081	TAGTGGCCTAAGGTCGTGCCC	+	5307	5327	TAGTGGACTAAGGTCATGCCC	+
V$RORA/V$NBRE.01	Monomers of the nur subfamily of nuclear receptors (nur77, nurr1, nor-1)	5065	5083	GGCCTAAGGTCGTGCCCTC	+	5255	5273	GGGAGCTGGACCTTCGGTT	-
V$RXRF/V$VDR_RXR.01	VDR/RXR Vitamin D receptor RXR heterodimer site	5071	5095	AGGTCGTGCCCTCCCTCCCCCACTG	-	5317	5341	AGGTCATGCCCTCTCTTCCCCCACA	-
V$ZBPF/V$ZBP89.01	Zinc finger transcription factor ZBP-89	5077	5099	TGCCCTCCCTCCCCCACTGAATC	+	5245	5267	GGGGCATGGGGGGAGCTGGACCT	-
V$CLOX/V$CLOX.01	Clox	5089	5107	CCCACTGAATCGATAAATA	+	5334	5352	CCCCCACAATCGATAATTT	+

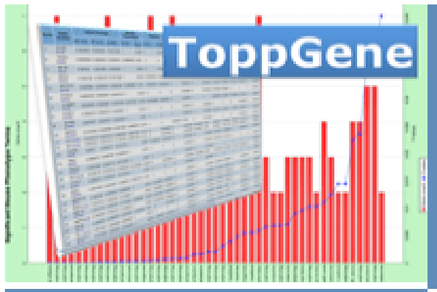
I have a list of co-expressed mRNAs (Transcriptome)….

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)



- Home
- Links
- Database details
- Supplementary
- Help
- Publications
- Terms of Use
- Contacts

Supported by:

Computational Medicine Center

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)

ToppFun: 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: Entrez ID

Example gene sets: HGNC Symbol Entrez ID
(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

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

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	FGB	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
6822	Duplicated
2328	Duplicated

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

Calculations

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

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

GO Biological Process

Annotations:	16,372
Genes:	15,079
	Updated Aug 26, 2009

GO Cellular Component

Annotations:	2,335
Genes:	16,728
	Updated Aug 26, 2009

GO Molecular Function

Annotations:	8,583
Genes:	15,948
	Updated Aug 26, 2009

Pathways

Annotations:		1,672
	BioCyc	164
Aug 25, 2009	CGAP BioCarta	314
	GenMAPP	67
Jun 15, 2009	KEGG pathway	202
May 10, 2009	MSigDB	431
	PantherDB	150
Aug 25, 2009	Pathway Ontology	306
	Reactome	25
	SigmaAldrich	2
	Signalling Transduction KE	11
Genes:		6,697

Human Phenotype

Annotations:	9,551
Genes:	2,531
	Updated Sep 10, 2009

Mouse Phenotype

Annotations:	6,203
Genes:	5,590
	Updated Aug 25, 2009

Domains

Annotations:		10,223
	Gene3D	285
	InterPro	4,859
	PROSITE	1,351
	Pfam	2,774
	ProDom	385
	SMART	569
Genes:		12,430

Pubmed

Annotations:	221,282
Genes:	22,176
	Updated Aug 25, 2009

Interactions

Annotations:		18,047
	BIND	4,370
	BioGRID	7,602
	HPRD	6,075
Genes:		5,541

Cytoband

Annotations:	382
Genes:	29,821

TFBS

Annotations:	615
Genes:	9,770

miRNA

Annotations:		740
	MSigDB	313
	PicTar	178
	TargetScan	249
Genes:		11,618

Gene Families

Annotations:	151
Genes:	6,098

Coexpression

Annotations:		1,203
	Body Map	23
	mSigDB	1,180
Genes:		12,694

Computational Gene Set

Annotations:	427
Genes:	4,712

Drugs

Annotations:		13,141
Aug 28, 2009	CTD	4,977
Aug 25, 2009	Drug Bank	2,009
Aug 25, 2009	Stitch	6,155
Genes:		14,836

Disease

Annotations:		3,789
Aug 28, 2009	CTD	1,006
	GWAS	291
	OMIM	2,492
Genes:		4,385

Master Gene Info File

For All Annotations	35,449
	Updated Aug 28, 2009

1. Database updated regularly
2. Exhaustive collection of annotations

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

Results

Go To Start Page

Input Parameters [Show Detail]

Training Results [Show All] [Download All] [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
Correction and Cutoff:	

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
Minimun feature count in test set:	2
Analysis took:	2 seconds
Analysis finished at:	Sun Sep 27 16:45:06 EDT 2009

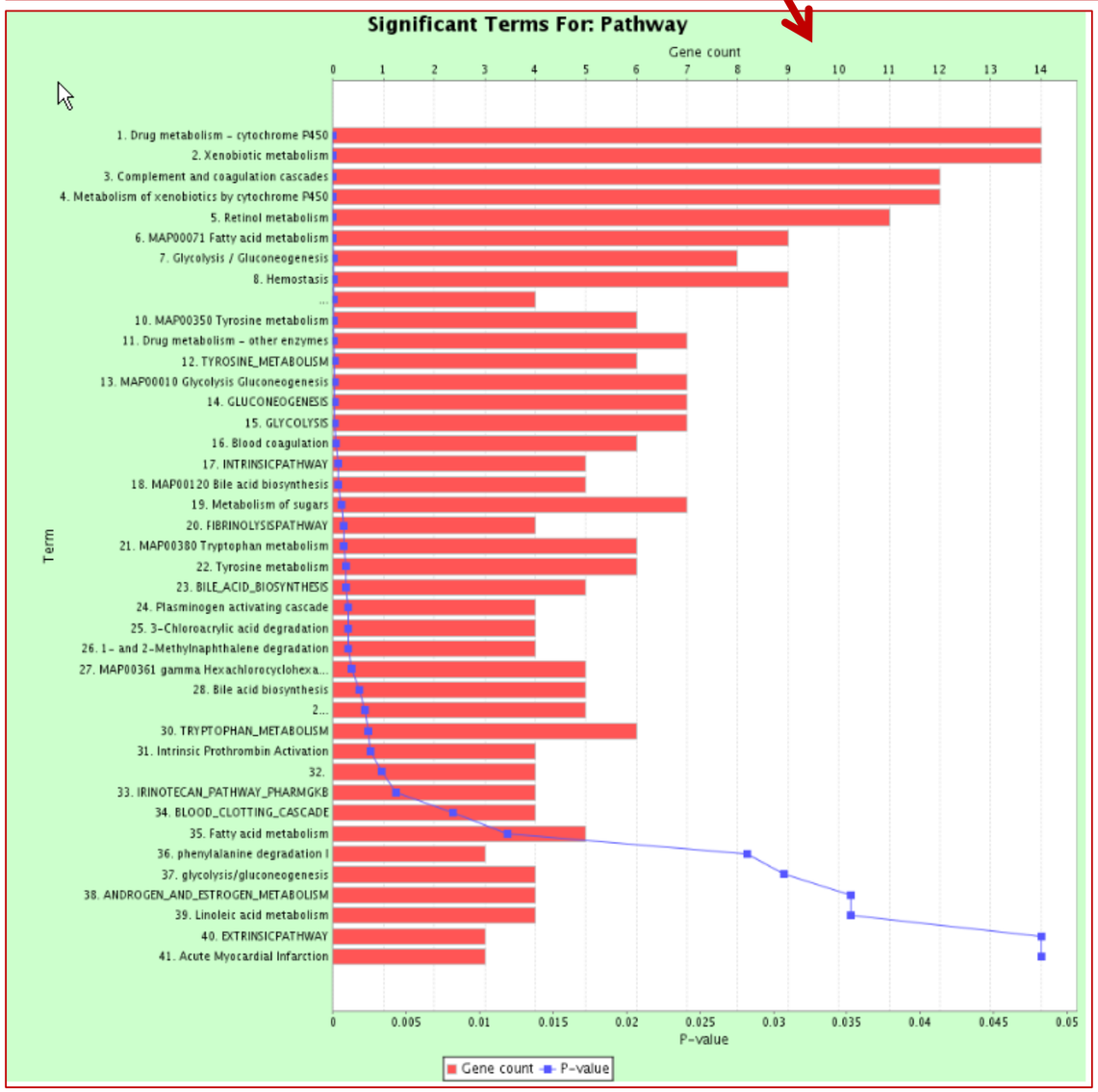
2: GO: Biological Process [Display Chart] [Hide Detail]

	ID	Name	Source	P-value	Term in Query	Term in Genome
1	GO:0009605	response to external stimulus		0	27	893
2	GO:0007596	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	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:0009605	response to external stimulus		0	27	893
2	GO:0007596	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	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



response to external stimulus; GO:0009605

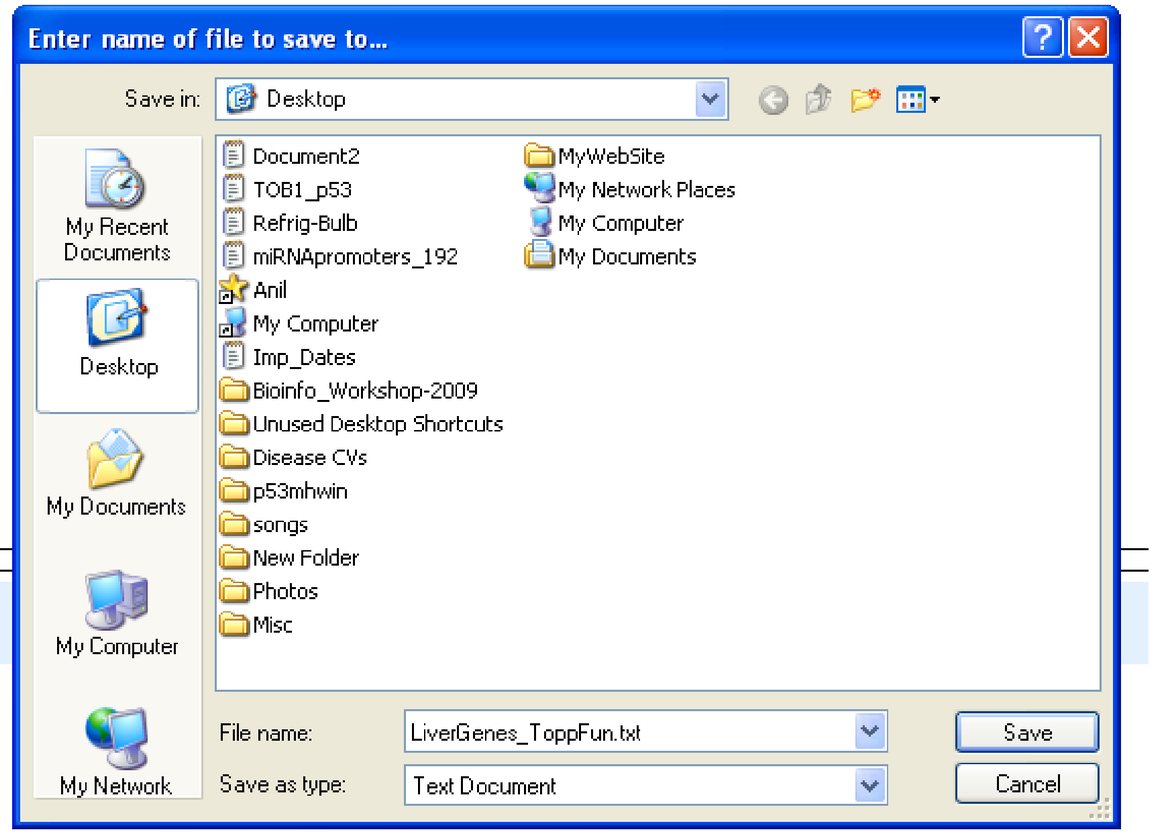
	Entrez Gene ID	Gene Symbol	Gene Name	Original Symbol
1	126	ADH1C	alcohol dehydrogenase 1C (class I), gamma polypeptide	126
2	335	APOA1	apolipoprotein A-I	335
3	350	APOH	apolipoprotein H (beta-2-glycoprotein I)	350
4	2158	F9	coagulation factor IX	2158
5	5950	RBP4	retinol binding protein 4, plasma	5950
6	197	AHSG	alpha-2-HS-glycoprotein	197
7	2243	FGA	fibrinogen alpha chain	2243
8	213	ALB	albumin	213
9	2244	FGB	fibrinogen beta chain	2244
10	629	CFB	complement factor B	629
11	3158	HMGCS2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)	3158
12	5444	PON1	paraoxonase 1	5444
13	1361	CPB2	carboxypeptidase B2 (plasma)	1361
14	3078	CFHR1	complement factor H-related 1	3078
15	5265	SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	5265
16	3827	KNG1	kininogen 1	3827
17	325	APCS	amyloid P component, serum	325
18	2538	G6PC	glucose-6-phosphatase, catalytic subunit	2538
19	4153	MBL2	mannose-binding lectin (protein C) 2, soluble (opsonic defect)	4153
20	735	C9	complement component 9	735
21	462	SERPINC1	serpin peptidase inhibitor, clade C (antithrombin), member 1	462
22	3273	HRG	histidine-rich glycoprotein	3273
23	5340	PLG	plasminogen	5340
24	5004	ORM1	orosomucoid 1	5004
25	316	AOX1	aldehyde oxidase 1	316
26	3053	SERPIND1	serpin peptidase inhibitor, clade D (heparin cofactor), member 1	3053
27	1356	CP	ceruloplasmin (ferroxidase)	1356

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

ToppGene Result Page

Number of genes in training set:	81
Number of genes in test set:	0
Correction and Cutoff:	(see table below)
Random sampling size in analysis:	0
Minimun feature count in test set:	2
Analysis took:	2 seconds
Analysis finished at:	Sun Sep 27 16:45:06 EDT 2009

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



Training Results [Show All] [Download All] [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]

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

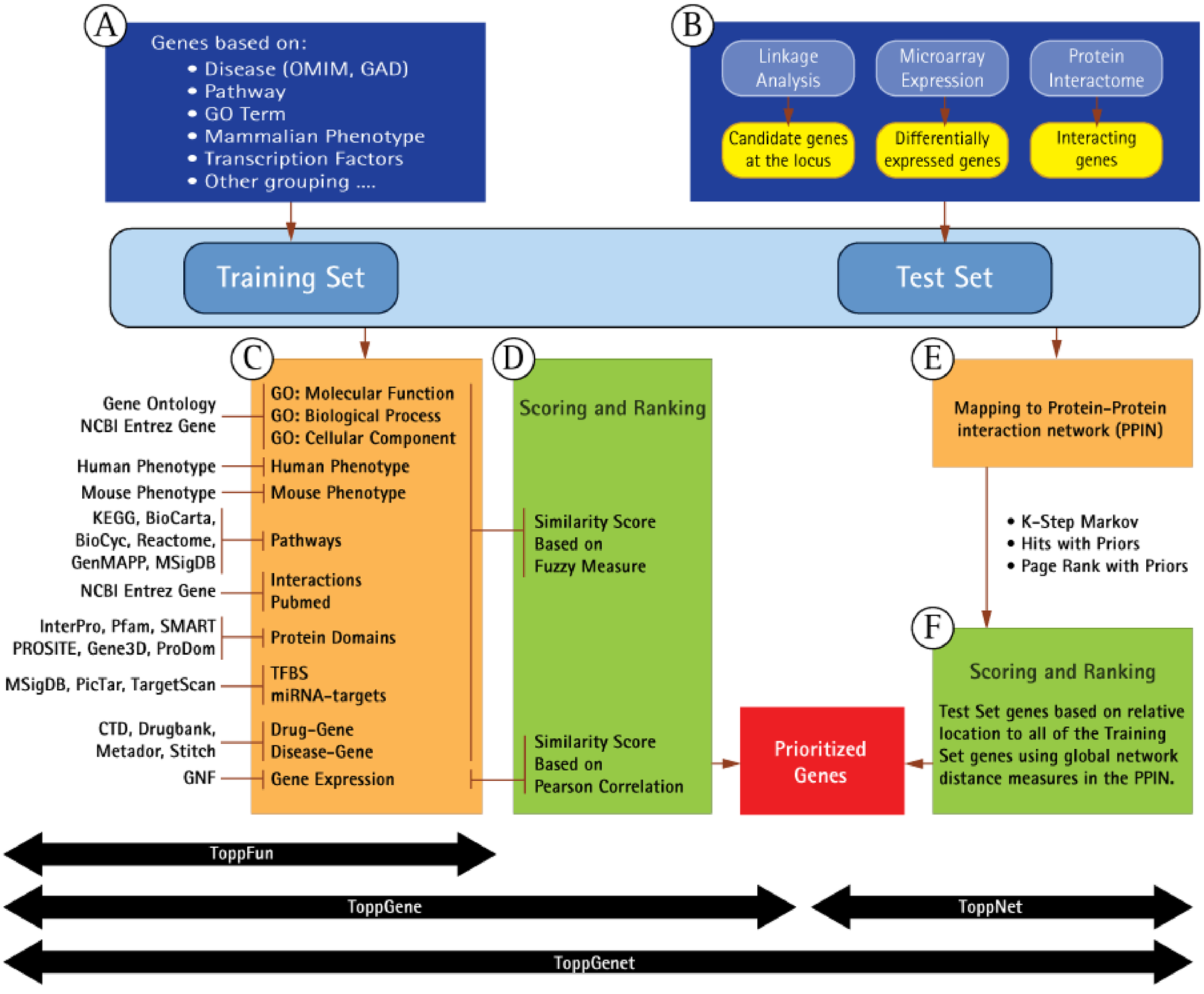
  Prioritize or rank candidate genes based on topological features in protein-protein interaction network.

- **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)…..



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

ToppGene: 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

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

Training set (7 / 7)

Entered	Human Symbol	Gene ID
NKX2-5	NKX2-5	1482
MEF2A	MEF2A	4205
GATA4	GATA4	2626
HAND1	HAND1	9421
HAND2	HAND2	9464
TBX5	TBX5	6910
SRF	SRF	6722

Test set (146 / 158)

Entered	Human Symbol	Gene ID
ACVR1	ACVR1	90
ACVR2B	ACVR2B	93
ADAM19	ADAM19	8728
ADM	ADM	133
ADRA1A	ADRA1A	148
ADRA1B	ADRA1B	147
ADRBK1	ADRBK1	156
ALDH1A2	ALDH1A2	8854
ALPK3	ALPK3	57538
ATP6V0A1	ATP6V0A1	535
BMP10	BMP10	27302
BMP2	BMP2	650
BMP4	BMP4	652
BMPR1A	BMPR1A	657
CALCRL	CALCRL	10203
CASP3	CASP3	836
CASP7	CASP7	840
CASP8	CASP8	841
CASQ2	CASQ2	845
CHD7	CHD7	55636
CITED2	CITED2	10370

Entered	Suggestions
CENTA2	☑ ADAP2 - ArfGAP with dual PH domains 2 Human Synonym
CMYA1	☑ XIRP1 - xin actin-binding repeat ontaining 1 Human Synonym
GJA7	☑ GJC1 - gap junction protein, gamma 1, 45kDa Human Synonym
HOP	☑ HOPX - HOP homeobox Human Synonym
	☐ ST13 - suppression of tumorigenicity 13 (colon carcinoma) (Hsp70 interacting protein) Human Synonym
	☐ STIP1 - stress-induced-phosphoprotein 1 Human Synonym
PPARBP	☑ MED1 - mediator complex subunit 1 Human Synonym
RBPSUH	RBPJ Duplicated

Update ☐ Check All

Ignored

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
All	Bonferroni	0.05	1 ≤ n ≤ 1500
GO: Molecular Function	Bonferroni	0.05	1 ≤ n ≤ 1500
GO: Biological Process	Bonferroni	0.05	1 ≤ n ≤ 1500
GO: Cellular Component	Bonferroni	0.05	1 ≤ n ≤ 1500
Human Phenotype	Bonferroni	0.05	1 ≤ n ≤ 1500
Mouse Phenotype	Bonferroni	0.05	1 ≤ n ≤ 1500
Domain	Bonferroni	0.05	1 ≤ n ≤ 1500
Pathway	Bonferroni	0.05	1 ≤ n ≤ 1500
Pubmed	Bonferroni	0.05	1 ≤ n ≤ 1500
Interaction	Bonferroni	0.05	1 ≤ n ≤ 1500
Cytoband	Bonferroni	0.05	1 ≤ n ≤ 1500
TFBS	Bonferroni	0.05	1 ≤ n ≤ 1500
Gene Family	Bonferroni	0.05	1 ≤ n ≤ 1500
Coexpression	Bonferroni	0.05	1 ≤ n ≤ 1500
Computational	Bonferroni	0.05	1 ≤ n ≤ 1500
MicroRNA	Bonferroni	0.05	1 ≤ n ≤ 1500
Drug	Bonferroni	0.05	1 ≤ n ≤ 1500
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)

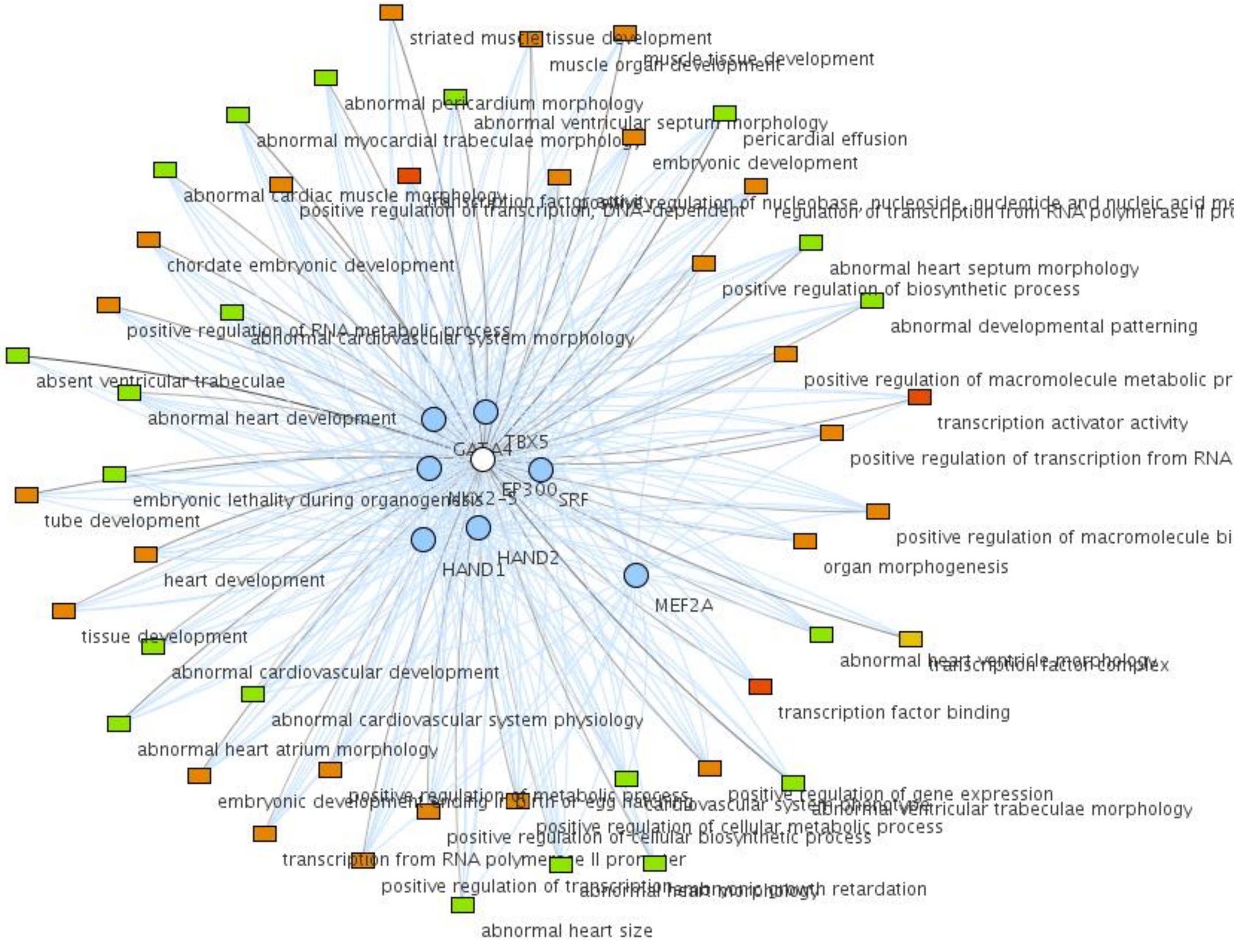
Rank	Gene Symbol	Gene ID	GO: Molecular Function		GO: Biological Process		GO: Cellular Component		Human Phenotype		Mouse Phenotype		Domain		Pathway		Pubmed		Interaction		Cytoband	
			Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue
1	EP300	2033	0.7136999	0.0235641	0.9999726	0.0029455	0.4305814	0.001	0	0.5	0.9999881	0.005			0	0.5049834	0.8410056	0.001	0.7991753	0.001	0	0.5004907
2	TEAD1	7003	0.5804123	0.0309278	0.997771	0.0103093	0.4305814	0.001	0	0.5	0.9989337	0.035	0	0.5			0.7993714	0.001	0.7160863	0.001	0	0.5004907
3	HIF1A	3091	0.9391513	0.0029455	1	0.001	0.4305814	0.001			0.9997067	0.02	0.8885697	0.001	0	0.5049834	0	0.4995093	0	0.505	0	0.5004907
4	CTNNB1	1499	0.7136999	0.0235641	1	0.001	0.4305814	0.001	0	0.5	0.9529218	0.06			0.6371253	0.001	0	0.4995093	0	0.505	0	0.5004907
5	TBX20	57057	0.5804123	0.0309278	0.9999964	0.0014728	0	0.5022091			0.9999902	0.005	0	0.5			0	0.4995093			0	0.5004907
6	ZFPM2	23414	0.6308709	0.0250368	0.9999978	0.001	0	0.5022091	0	0.5	1	0.001	0	0.5			0	0.4995093			0	0.5004907
7	BMP4	652	0	0.5493373	1	0.001	0	0.5022091	0	0.5	0.9999435	0.01	0	0.5	0.6478057	0.001	0.7993714	0.001	0	0.505	0	0.5004907
8	TBX1	6899	0.9660807	0.001	0.999997	0.001	0	0.5022091	0	0.5	0.9996966	0.02	0	0.5			0	0.4995093			0	0.5004907
9	TBX2	6909	0.5418852	0.0397644	0.9943991	0.0162003	0.4305814	0.001			0.9993508	0.035	0	0.5	0	0.5049834	0	0.4995093			0	0.5004907
10	TGFB2	7042	0.8603852	0.005891	1	0.001	0	0.5022091			0.9999998	0.005	0	0.5	0.3937178	0.0033223	0	0.4995093	0	0.505	0	0.5004907

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
0.229808	0.0001787
0.2395215	0.0002528
0.2566618	0.0002615
0.2503156	0.0002619
0.3307561	0.0002975

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

Why is a test set gene ranked higher?



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)

Graph prioritization parameters

Prioritization method: k-Step Markov

Step Size(normally 4-8): 6

Training gene neighborhood subnetwork visualization parameters

Neighborhood distance: 1

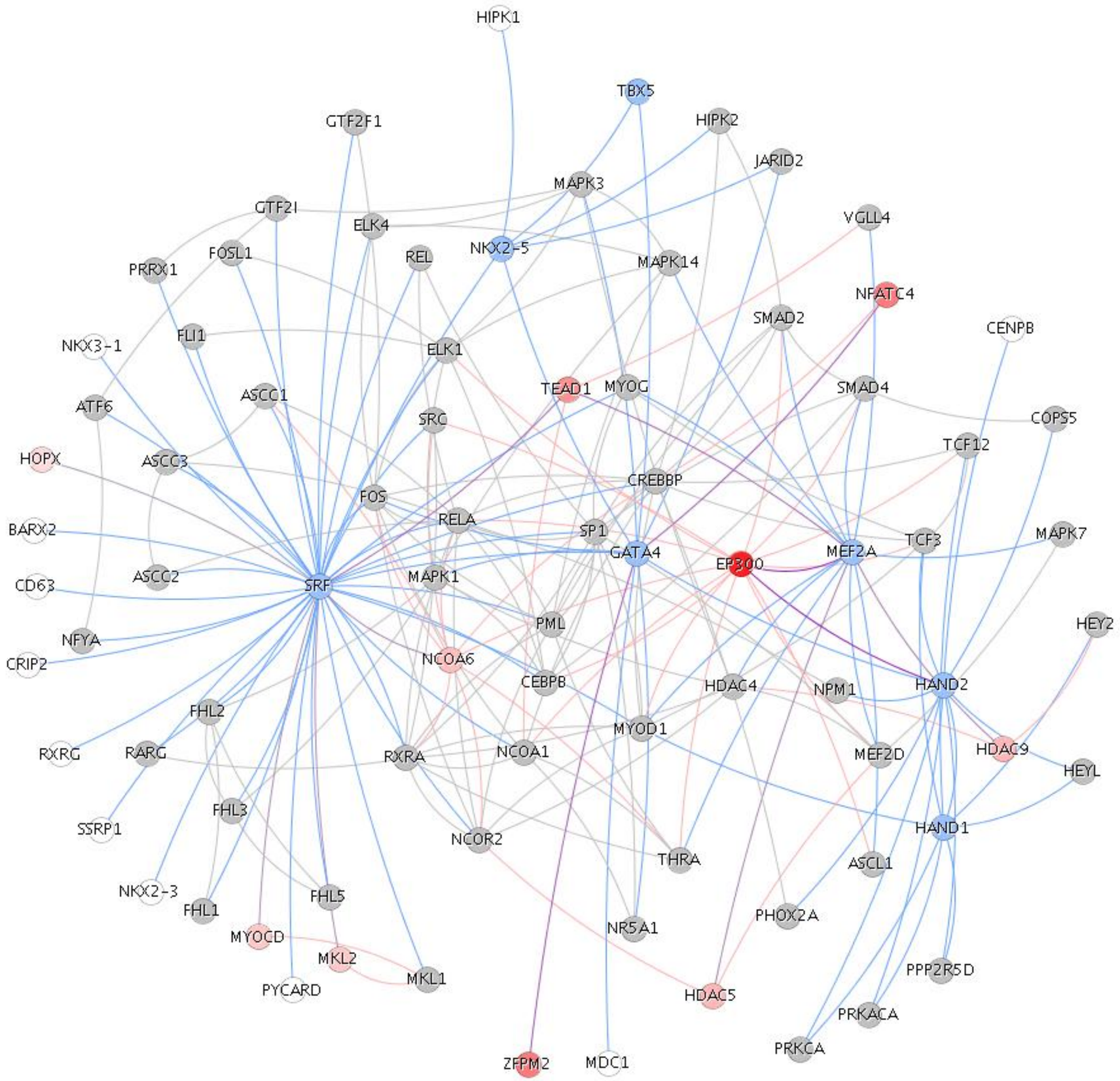
When training set is big, the training gene neighborhood subnetwork can be huge.

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	CTNNB1	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:

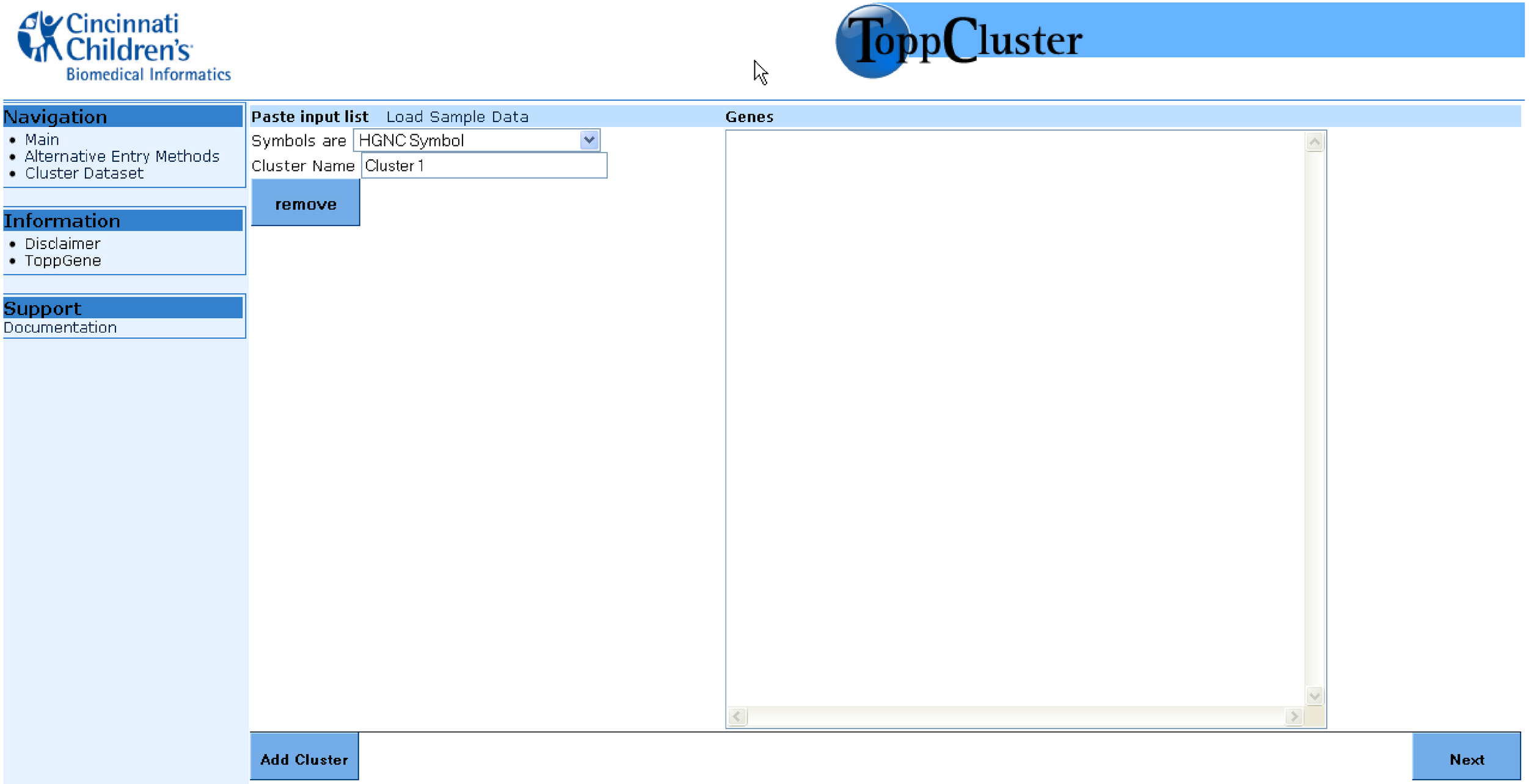
a. How many of these genes are transcription factors?
b. What are the enriched TFBSs and miRNAs?
c. What gene families are enriched in this list?
d. Are there are salivary gland development associated genes present in this list?
e. How many and which genes from this list are associated with non-insulin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

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

a. What are the top 10 ranked genes using ToppGene and ToppNet?
b. 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)



ToppCluster (http://toppcluster.cchmc.org)

Paste input list Load Sample Data

Symbols are: Entrez ID

Cluster Name: Liver

remove

Genes
259
5265
350
335
335
1558
1571
229
462
125
3240
5105
5265
3273
2244
2158
5053
125
1356
3827
383
5004
2168
1571
325
5950
127
1373
213
735

Symbols are: Entrez ID

Cluster Name: Salivary_Glands

remove

Genes
64065
3855
100129410
8842
3514
55504
26298
64065
245973

Options

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

Annotations must have at least 2 gene(s)

Chose Toppcluster output format: Extended HTML Table (Cytoscape Link)

Gene Sets

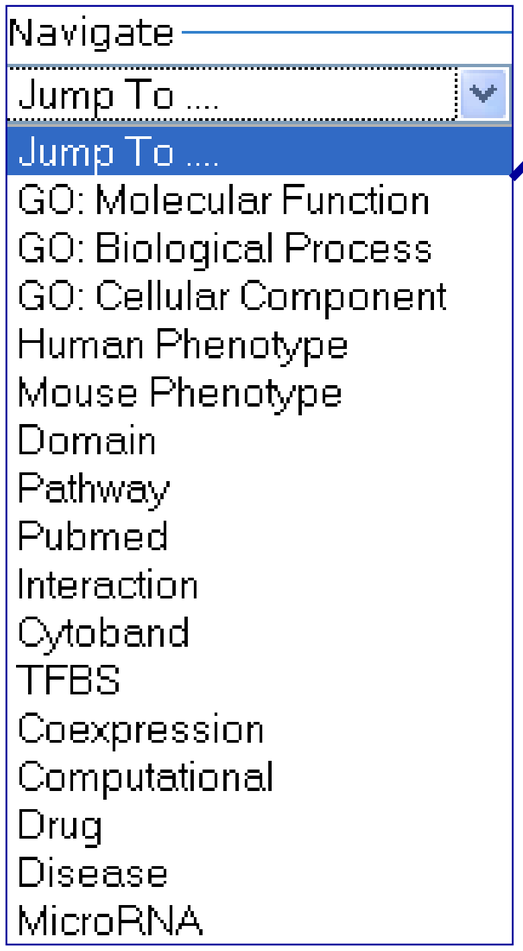
Liver
44 known - 0 unknown

Original	Human Symbol	Entrez ID
1576	CYP3A4	1576
1577	CYP3A5	1577
7036	TFR2	7036
229	ALDOB	229
341	APOC1	341
126	ADH1C	126
125	ADH1B	125
7448	VTN	7448
5053	PAH	5053
3240	HP	3240
197	AHSG	197
3078	CFHR1	3078
383	ARG1	383

Salivary_Glands
46 known - 0 unknown

Original	Human Symbol	Entrez ID
2591	GALNT3	2591
9073	CLDN8	9073
486	FXYD2	486
54959	ODAM	54959
5349	FXYD3	5349
155006	TMEM213	155006
100129410	LOC100129410	100129410
54097	FAM3B	54097
352999	C6orf58	352999
57535	KIAA1324	57535
26298	EHF	26298
999	CDH1	999
360	AQP3	360

ToppCluster (http://toppcluster.cchmc.org)



Navigate: Jump To ....

Links: Back to Start, Shareable Link, Excel Version

Cytoscape: Include Orphaned Genes, Include Super Category, XGMML, Build

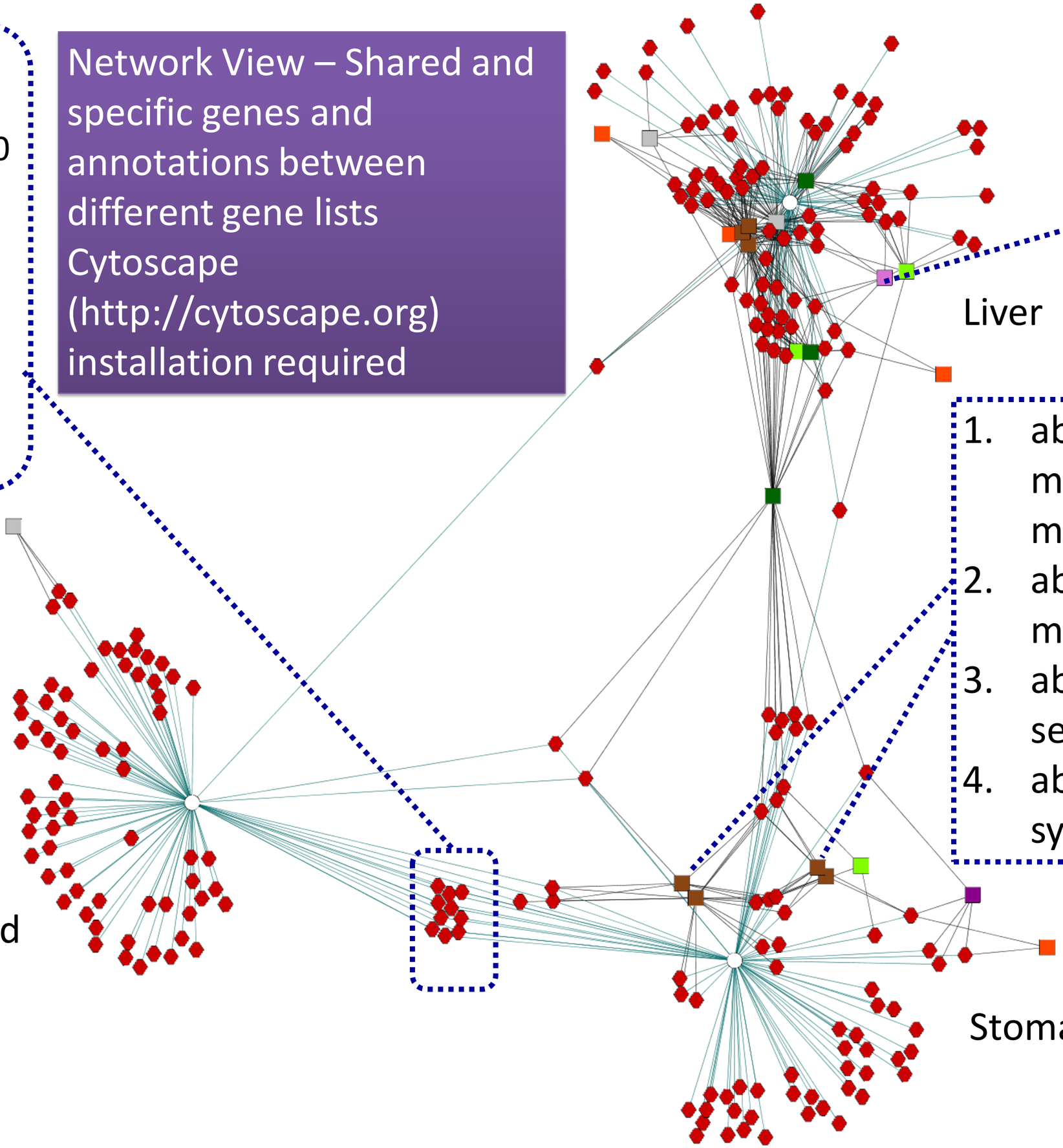
Re-Enrich

Category	ID	Title (or Source)			liver_logP	salivary gland_logP	stomach cardiac_logP
GO: Molecular Function					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	eukaryotic 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		

ToppCluster (http://toppcluster.cchmc.org)

EHF
COL15A1
LOC100130100
IGHA1
LTF
IGKC
IGL@
FAM129A
ATP8B1
IGLC2

Network View – Shared and specific genes and annotations between different gene lists Cytoscape (http://cytoscape.org) installation required



V$HNF1

Liver

1. abnormal gastric mucosa morphology
2. abnormal stomach morphology
3. abnormal digestive secretion
4. abnormal digestive system physiology

Salivary Gland

Stomach

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

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering new!, 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 content 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 new!

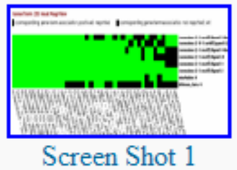
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Welcome to DAVID Bioinformatics Resources 2003 - 2007

What's Special in DAVID 2007?

The Database for Annotati... (DAVID) 2007 is the fifth ... programs of DAVID 2006 ... comprehensive set of funct... understand biological mean... gene list, DAVID tools are...

- Identify enriched biol...
- Discover enriched fu...
- Cluster redundant an...
- Visualize genes on B...
- Display related many...
- Search for other func...
- List interacting protei...
- Explore gene names ...
- Link gene-disease as...
- Highlight protein func...
- Redirect to related lit...
- Convert gene identife...
- And more



Screen Shot 1

- 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 identifers from one type to another.
- And more

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

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

Annotation Summary Results

Help and Tool Manual

Current Gene List: demolist1 — 171 DAVID IDs

Current Background: Homo sapiens — Check Defaults ☑ Clear All

⊞ Main Accessions (0 selected)

⊞ Other Accessions (0 selected)

⊟ Gene Ontology (3 selected)

☐ GOTERM_BP_1	79%	136	Chart
☐ GOTERM_BP_2	76%	131	Chart
☐ GOTERM_BP_3	74%	127	Chart
☐ GOTERM_BP_4	69%	119	Chart
☐ GOTERM_BP_5	60%	104	Chart
☑ GOTERM_BP_ALL	79%	136	Chart
☐ GOTERM_CC_1	70%	121	Chart
☐ GOTERM_CC_2	61%	106	Chart
☐ GOTERM_CC_3	55%	95	Chart
☐ GOTERM_CC_4	50%	86	Chart
☐ GOTERM_CC_5	38%	65	Chart
☑ GOTERM_CC_ALL	70%	121	Chart
☐ GOTERM_MF_1	75%	129	Chart
☐ GOTERM_MF_2	69%	119	Chart
☐ GOTERM_MF_3	60%	103	Chart
☐ GOTERM_MF_4	56%	97	Chart
☐ GOTERM_MF_5	45%	78	Chart
☑ GOTERM_MF_ALL	75%	129	Chart

⊞ 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

Functional Annotation Clustering new!

2D View

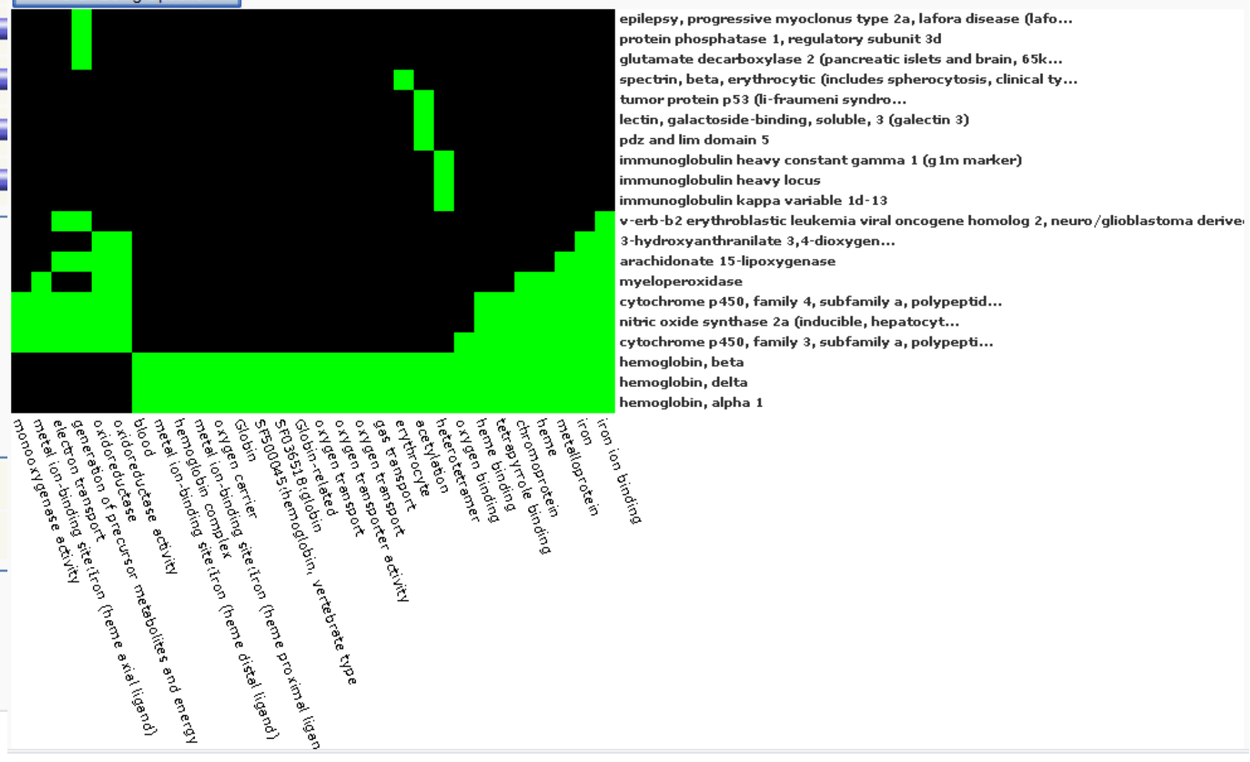
Help

corresponding gene-term association positively reported ■ corresponding gene-term association not reported yet

Download File ViewEditor

⊞ Options

Rerun Using Options



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

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

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering new!, 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 content 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 new!

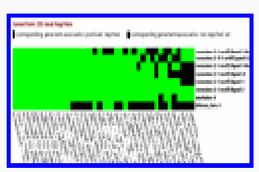
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Welcome to DAVID Bioinformatics Resources 2003 - 2007

What's Special in DAVID 2007?

The Database for Annotati… (DAVID) 2007 is the fifth … programs of DAVID 2006 … comprehensive set of funct… understand biological mean… gene list, DAVID tools are…

- Identify enriched biol…
- Discover enriched fu…
- Cluster redundant an…
- Visualize genes on Bi…
- Display related many…
- Search for other func…
- List interacting protei…
- Explore gene names …
- Link gene-disease as…
- Highlight protein func…
- Redirect to related lit…
- Convert gene identife…
- And more



Screen Shot 1

- 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 identifers 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

Gene ID Conversion Tool

Submit your gene list to start conversion!

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
- Unigene
- 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

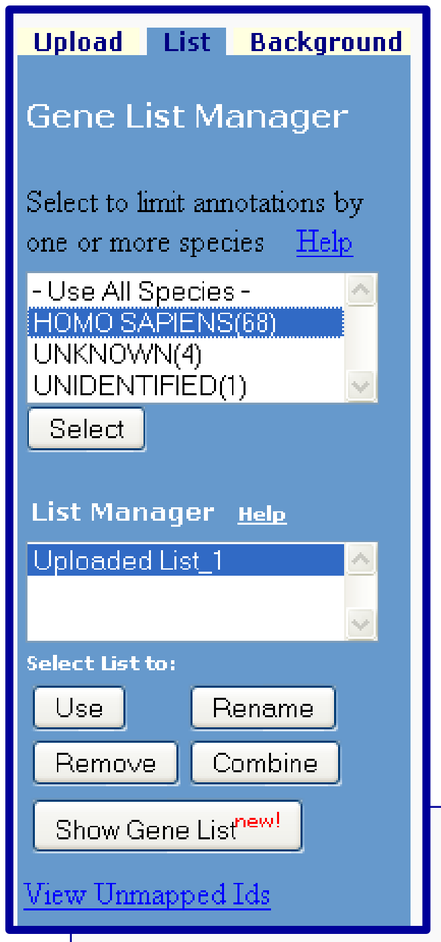
Submit List

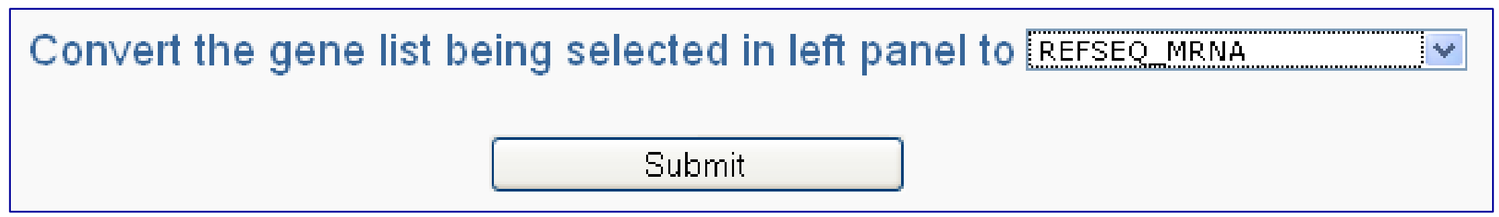
us how you like the tool

chnical notes of the tool

Contact us for questions

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





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

Gene Accession Conversion Statistics Help

Conversion Summary		
ID Count	**In DAVID DB**	**Conversion**
60	Yes	Successful
8 IDs	Yes	None
0 IDs	No	None
0 IDs	Ambiguous	Pending
Total Unique User IDs: 68		

Summary of Ambiguous Gene IDs		
ID Count	**Possible Source**	**Convert All**

All Possible Sources For Ambiguous IDs		
Ambiguous ID	**Possibility**	**Convert**

From	To	Species	David Gene Name
202	NM_001624	HOMO SAPIENS	ABSENT IN MELANOMA 1
72	NM_001613	HOMO SAPIENS	ACTIN, ALPHA 2, SMOOTH MUSCLE, AORTA
72	NM_001615	HOMO SAPIENS	ACTIN, ALPHA 2, SMOOTH MUSCLE, AORTA
27299	NM_014479	HOMO SAPIENS	ADAM-LIKE, DECYSIN 1
125	NM_000667	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS I), ALPHA POLYPEPTIDE
125	NM_000668	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS I), ALPHA POLYPEPTIDE
126	NM_000669	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS I), ALPHA POLYPEPTIDE
126	NM_000668	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS I), ALPHA POLYPEPTIDE
125	NM_000669	HOMO SAPIENS	ALCOHOL DEHYDROGENASE 1A (CLASS I), ALPHA POLYPEPTIDE

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 (2nd column) and extract their RefSeq accession numbers

PANTHER (http://www.pantherdb.org/)

Protein ANalysis THrough Evolutionary Relationships

PANTHER Classification System

LOGIN REGISTER CONTACT US HELP

Home | Browse | Genes | Families and HMMs | Pathways | Ontologies | Tools | Workspace

Quick links

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

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

Batch ID Search

Enter IDs: ACADSB, ACSL4, ADH6

separate IDs by a space or comma - supported IDs

Upload IDs: Browse...

- file format

Select upload ID type: Gene Symbol

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		

Search

GENE EXPRESSION DATA ANALYSIS

Our expression analysis tools can be used for microarray data intrepretation. 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!

Compare Classifications of Lists

Map lists of genes to a PANTHER ontology.For pathways, you can then view the gene expression values overlaid on top of a pathway diagram, where genes will be colored differently for different clusters of genes.

Use the binomial statistics tool to compare classifications of multiple clusters of lists to a reference list to statistically determine over- or under- representation of PANTHER classification categories. Each list is compared to the reference list using the binomial test (Cho & Campbell, TIGs 2000) for each molecular function, biological process, or pathway term in PANTHER.

Steps:

1. Select list(s) to analyze
2. Select reference list

1. Select Lists to Compare to a Reference List

For example, each selected list may be a cluster of co-expressed genes under a particular set of conditions.

Select list(s)

selected: FetalLiverSpecific.txt
FetalBrainSpecific.txt
AdultHeartSpecific.txt

2. Select Reference List

For example, the reference list may be the set of all genes in the experiment, or the set of all genes in the genome being analyzed.

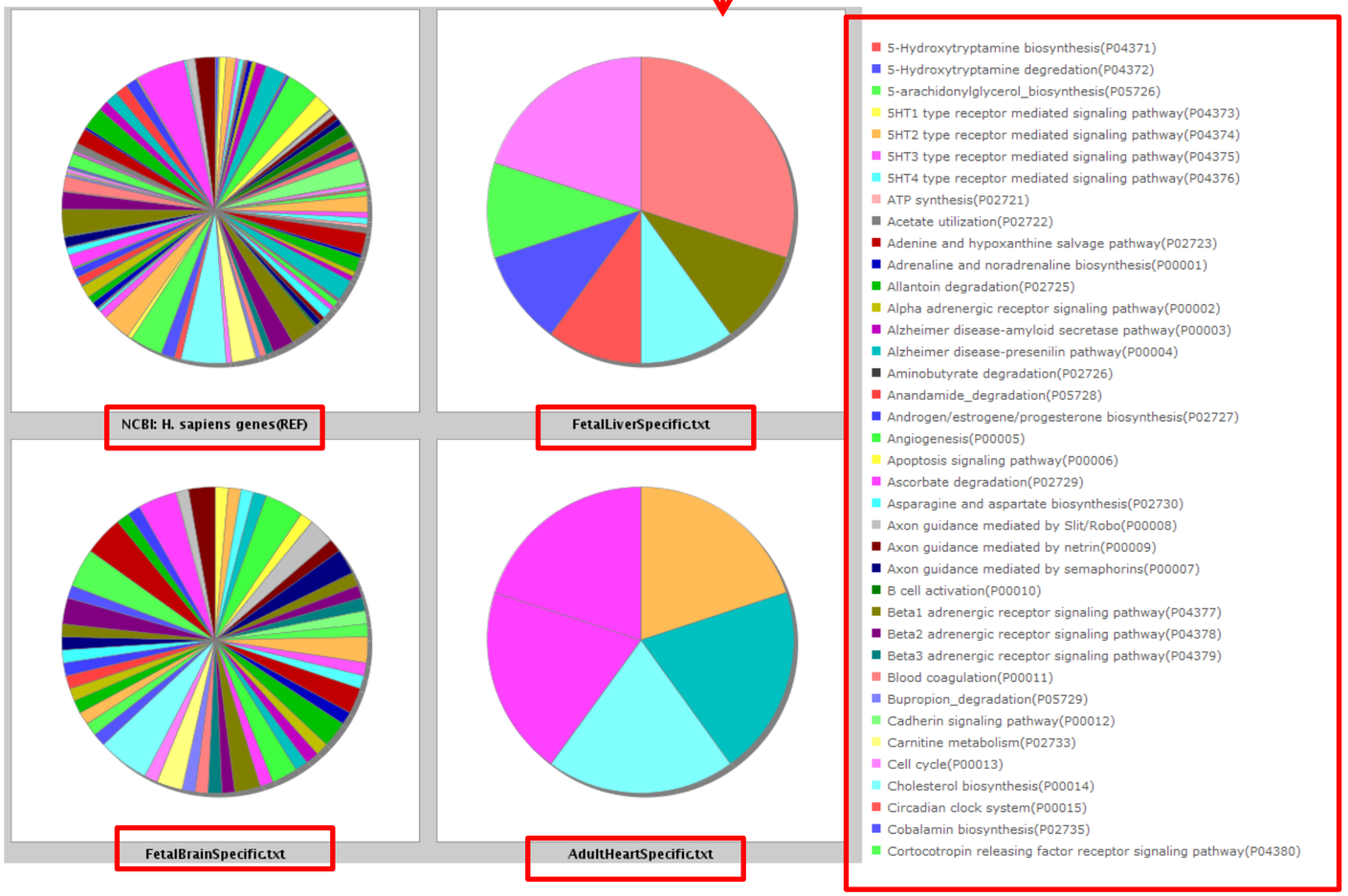
Select reference list default: NCBI: H. sapiens genes

Search options

PANTHER Ontology:

- (•) Pathways
- ( ) Biological Process
- ( ) Molecular Function

[✓] Use the Bonferroni correction for multiple testing



PANTHER (http://www.pantherdb.org/)

Protein ANalysis THrough Evolutionary Relationships

Results

Colors for viewing genes in pathway diagrams:

Example-Set-3.txt: red

gray: components only in the reference list

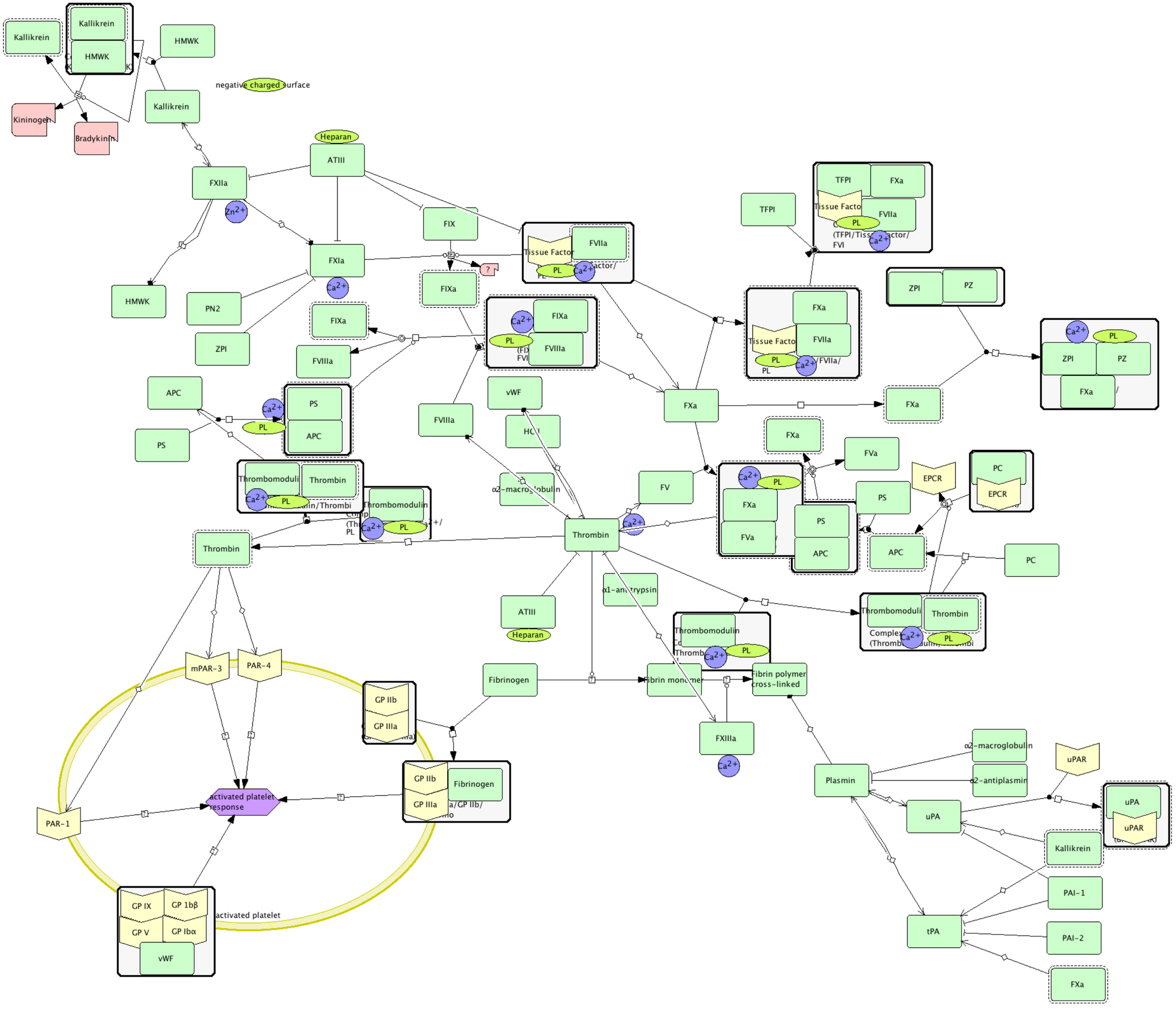
	Reference list	Example-Set-3.txt
Mapped IDs:	25431	135
Unmapped IDs:	0	15

Click on pathway name to see genes highlighted on pathway diagram

Export results View: Bar Chart of Gene Count

Pathways	NCBI: H. sapiens genes (REF) #	Example-Set-3.txt #	expected	+/-	P value
Blood coagulation	55	8	.29	+	1.38E-07
Plasminogen activating cascade	21	4	.11	+	9.37E-04

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

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 their 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 are 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