Making Sense Out of Transcriptome

Integrative Bioinformatic Approaches

Anil Jegga
Division of Biomedical Informatics, CCHMC
Department of Pediatrics, UC

Contact Information:
Anil Jegga
Biomedical Informatics, CCHMC
Room # 232, S Building 10th Floor
3333 Burnet Ave MLC 7024
Cincinnati OH-45229, USA
Tel: 513-636-0261
Fax: 513-636-2056
Homepage: http://anil.cchmc.org
Mail: anil.jegga@cchmc.org

Table of Contents

Chapter 1: Interpreting genome-wide expression profiles - A knowledge-based approach......3
Introduction......3
Chapter 2: Strategies for Identifying Putative Gene Regulatory Regions in a Group of Genes4
Objectives......4
Fetching the Promoter Sequence......4
Identification of Putative Common/Shared Cis-Elements......12
Known......12
Non-conserved......12
Conserved......14
Unknown/Novel......20
Non-conserved......24
Conserved......27
Chapter 3: Functional Enrichment Analysis of the Transcriptome......29
Objectives......29
Introduction......29
Tools and Servers......29
DAVID......29
MSigDB (Molecular Signatures Database)......31
Panther......33
L2L......36
ToppGene......36
Babelomics (FatioGO+)......38
OntoExpress......39
Chapter 4: Identification of Regulatory Regions: Using Trafac and Other Related Tools......42
Introduction......42
Method......42
Output......42
Utility......42
What do you need?......43
How to Use Trafac?......43
Uploading Sequences:......44
Input Files......45
Related Tools......46
Chapter 5: Annotation of Coding Single Nucleotide Polymorphisms: Using PolyDoms......47
APPENDIX 1: Other Useful bioinformatics resources and tools......52
APPENDIX 2: Co-expressed gene lists......53
APPENDIX 3: Querying NCBI - Example......56
Querying the NCBI's Entrez Gene:......56
Querying the NCBI's OMIM:......57
APPENDIX 4: Coping/Keeping up with literature (PubMed) searches......58

Chapter 1: Interpreting genome-wide expression profiles - A knowledge-based approach

Introduction

Genes typically operate in a sophisticated network of interactions and it is now well recognized that co-expressing genes tend to be playing some common roles in the cell. Recent evidences also suggest functionally related genes map close even in the eukaryotic genomes. Complex phenotypic traits, including diseases are now considered from a systems biology perspective. Thus, there is a clear necessity for methods and tools which can help to understand genome-scale experiments (for e.g. microarray-based gene expression) from a systems biology perspective. Genome-wide expression analysis with DNA microarrays has become a mainstay of genomics research. In fact, the challenge no longer lies in obtaining gene expression profiles, but rather in interpreting the results to gain insights into biological mechanisms *(Subramanian et al., PNAS, 102: 15545-15550)*. A typical experiment generates mRNA expression profiles for thousands of genes from a collection of samples belonging to different classes. The genes are ordered in a ranked list based on their differential expression between the classes. The proper interpretation of this data requires an integrative systems biology-based functional annotation wherein the collective properties of groups of genes are taken into account rather than individual genes. The principal challenge then is to extract meaning from this list(s) – what is the common or unifying theme(s)?

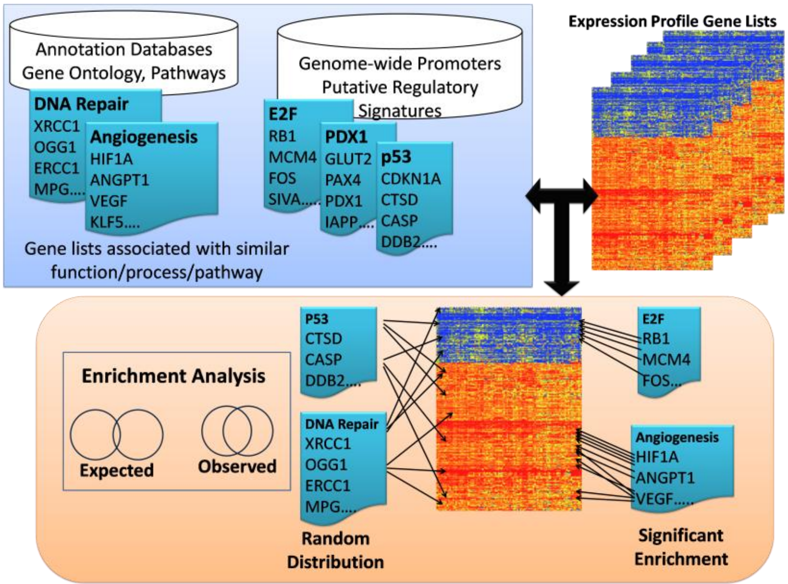


Figure 1: Enrichment analysis for functional and regulatory analysis aimed at identifying specific functions or processes or pathways (GO, KEGG) and transcription factor binding sites that are common for a group of coexpressed genes.

a. Common Regulation (shared cis regulatory elements or transcription factor binding sites)
b. Common Biological Function (common pathways or processes)
c. Chromosomal Location

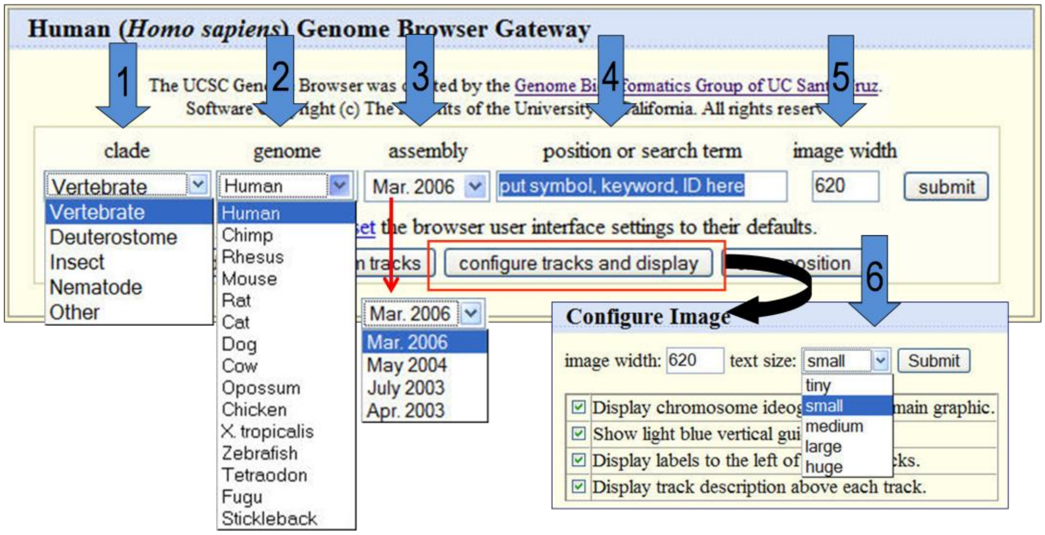
Chapter 2: Strategies for Identifying Putative Gene Regulatory Regions in a Group of Genes

Objectives

- i. Fetch the promoter sequences from a group of coexpressed genes
- ii. Identify the common/shared transcription factor binding sites (TFBSs) or *cis*-elements for this group of promoters
    - a. Known
        1. Non-conserved
        2. Conserved
    - b. Unknown/Novel
        1. Non-conserved
        2. Conserved

Fetching the Promoter Sequence

1. Go to the UCSC genome browser home page ((http://genome.ucsc.edu) and get the sequence for
    - a. Single gene:
        - i. Using gene symbols:
            - From the home page i.e. http://genome.ucsc.edu, click on the "Genomes" link (top navigation bar - left hand corner). Once you are on the Genome Browser Gateway page, select whichever genome you are interested in. In the box under "position or search term" enter the gene symbol and click submit.
            - The following pages list your query results under different categories (e.g. Known Genes, RefSeq Genes, etc.). What this means that your query has results from these different tables or databases.
            - Click on the entries below RefSeq Genes (wherever available; RefSeq or Reference Sequence is a database of curated mRNAs from NCBI) and this will take you to the Genome Browser page.
            - Click on the "DNA" (navigation bar on the top)
            - On the "Get DNA in Window" page, enter your sequence retrieval region options (for e.g. how many base pair upstream, etc.)
            - If you want to mask the repeat regions, check that option under "Sequence Formatting Options"
            - Finally click on "get DNA" to get the sequence in fasta format.
            - Explore the "extended case/color options"
        - ii. Using accession numbers
            - Same as above but using accession number instead of gene symbols.



Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Gen[...] Browser was [...]ted by the Genome Bi[...]formatics Group of UC San[...]ruz. Software [...]ight (c) The [...]ts of the University [...]alifornia. All rights reser[...]

clade: Vertebrate · genome: Human · assembly: Mar. 2006 · position or search term: put symbol, keyword, ID here · image width: 620

[...]et the browser user interface settings to their defaults.

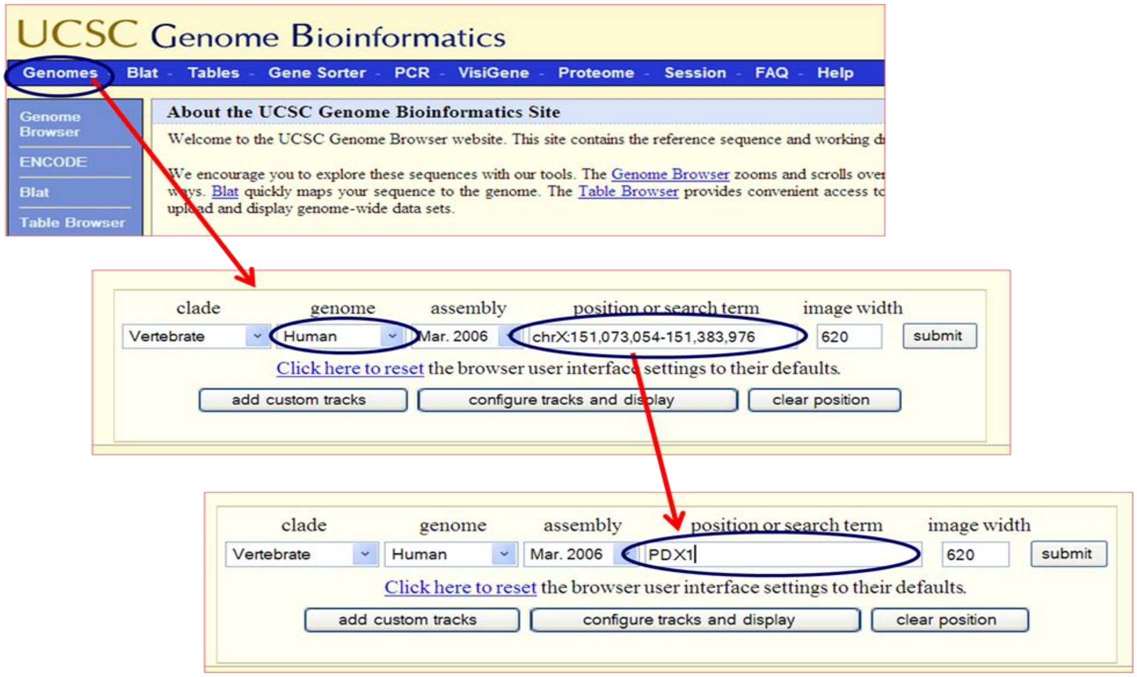
Configure Image

image width: 620 · text size: small

☑ Display chromosome ideog[...] main graphic.
☑ Show light blue vertical gui[...]
☑ Display labels to the left of [...]ks.
☑ Display track description above each track.

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



UCSC Genome Bioinformatics

About the UCSC Genome Bioinformatics Site

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

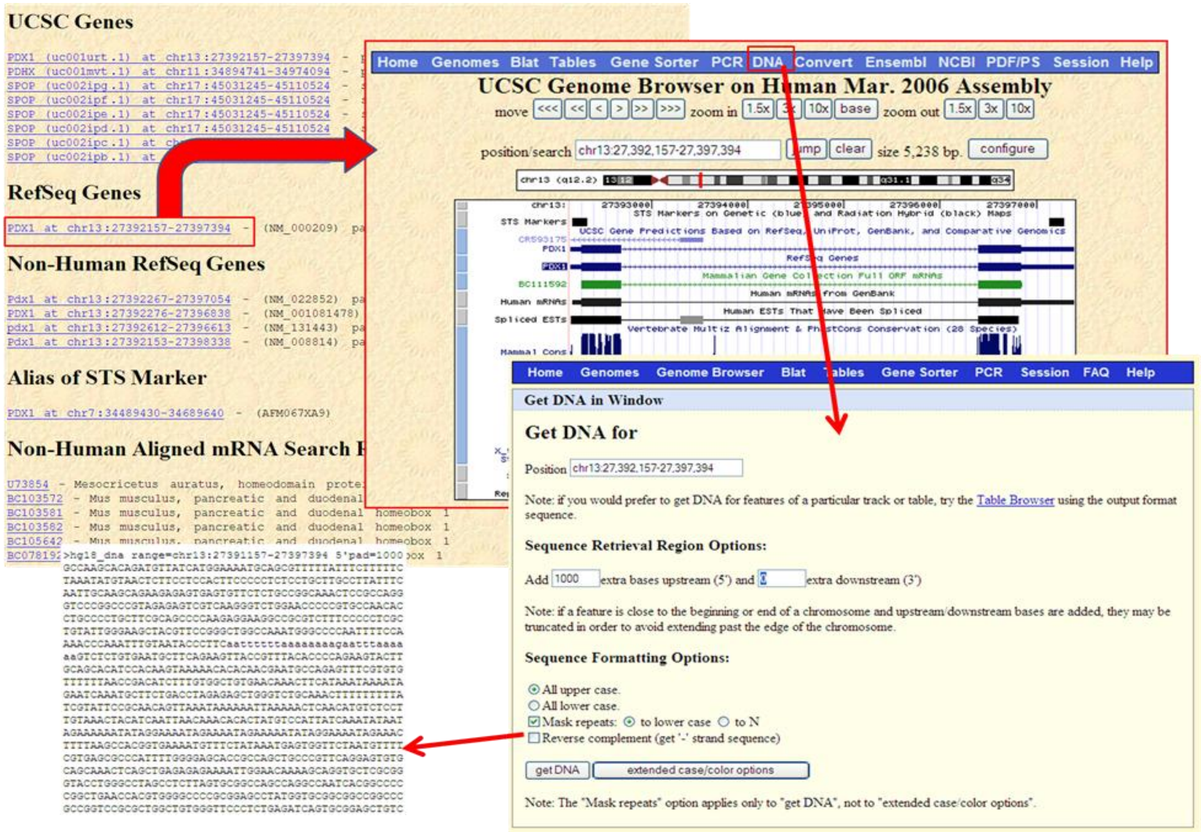
We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to upload and display genome-wide data sets.

clade: Vertebrate · genome: Human · assembly: Mar. 2006 · position or search term: chrX:151,073,054-151,383,976 · image width: 620

Click here to reset the browser user interface settings to their defaults.

clade: Vertebrate · genome: Human · assembly: Mar. 2006 · position or search term: PDX1 · image width: 620

Click here to reset the browser user interface settings to their defaults.



UCSC Genes

```
PDX1 (uc001urt.1) at chr13:27392157-27397394 -
PDHX (uc001mvt.1) at chr11:34894741-34974094 -
SPOP (uc002ipg.1) at chr17:45031245-45110524 -
SPOP (uc002ipf.1) at chr17:45031245-45110524 -
SPOP (uc002ipe.1) at chr17:45031245-45110524 -
SPOP (uc002ipd.1) at chr17:45031245-45110524
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 R

```
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
BC07819
```

UCSC Genome Browser on Human Mar. 2006 Assembly

position/search chr13:27,392,157-27,397,394 size 5,238 bp.

```
>hg18_dna range=chr13:27391157-27397394 5'pad=1000
GCCAAGCACAGATGTTATCATGGAAAATGCAGCGTTTTTATTTCTTTTTC
TAAATATGTAACTCTTCCTCCACTTCCCCCTCTCCTGCTTGCCTTATTTC
AATTGCAAGCAGAAGAGAGTGAGTGTTCTCTGCCGGCAAACTCCGCCAGG
GTCCCGGCCCGTAGAGAGTCGTCAAGGGTCTGGAACCCCCGTGCCAACAC
CTGCCCCTGCTTCGCAGCCCCAAGAGGAAGGCCGCGTCTTTCCCCCTCGC
TGTATTGGGAAGCTACGTTCCGGGCTGGCCAAATGGGCCCCAATTTTCCA
AAACCCAAATTTGTAATACCCTTCaatttttaaaaaaaagaatttaaaa
aaGTCTCTGTGAATGCTTCAGAAGTTACCGTTTACACCCCAGAAGTACTT
GCAGCACATCCACAAGTAAAAACACACAACGAATGCCAGAGTTTCGTGTG
TTTTTTAACCGACATCTTTGTGGCTGTGAACAAACTTCATAAATAAAATA
GAATCAAATGCTTCTGACCTAGAGAGCTGGGTCTGCAAACTTTTTTTTTA
TCGTATTCCGCAACAGTTAAATAAAAAATTAAAAACTCAACATGTCTCCT
TGTAAACTACATCAATTAACAAACACACTATGTCCATTATCAAATATAAT
AGAAAAAATATAGGAAAATAGAAAATAGAAAAATATAGGAAAATAGAAAC
TTTTAAGCCACGGTGAAAATGTTTCTATAAATGAGTGGTTCTAATGTTTT
CGTGAGCGCCCATTTTGGGGAGCACCGCCAGCTGCCCGTTCAGGAGTGTG
CAGCAAACTCAGCTGAGAGAGAAAATTGGAACAAAAGCAGGTGCTCGCGG
GTACCTGGGCCTAGCCTCTTAGTGCGGCCAGCCAGGCCAATCACGGCCCC
CGGCTGAACCACGTGGGGCCCCGCGGAGCCTATGGTGCGGCGGCCGGCCC
GCCGGTCCGCGCTGGCTGTGGGTTCCCTCTGAGATCAGTGCGGAGCTGTC
```

Get DNA in Window

Get DNA for

Position chr13:27,392,157-27,397,394

Note: if you would prefer to get DNA for features of a particular track or table, try the Table Browser using the output format sequence.

Sequence Retrieval Region Options:

Add 1000 extra bases upstream (5') and 0 extra downstream (3')

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:

● All upper case.
○ All lower case.
☑ Mask repeats: ● to lower case ○ to N
□ Reverse complement (get '-' strand sequence)

Note: The "Mask repeats" option applies only to "get DNA", not to "extended case/color options".

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump clear configure
chr13 (q12.2) 13q12 q31.1 q34
chr13: 27393000 27394000 27395000 27396000 27397000
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
CR593175
PDX1
RefSeq Genes
PDX1
Mammalian Gene Collection Full ORF mRNAs
BC111592
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help
get DNA
extended case/color options

b. Group of genes:

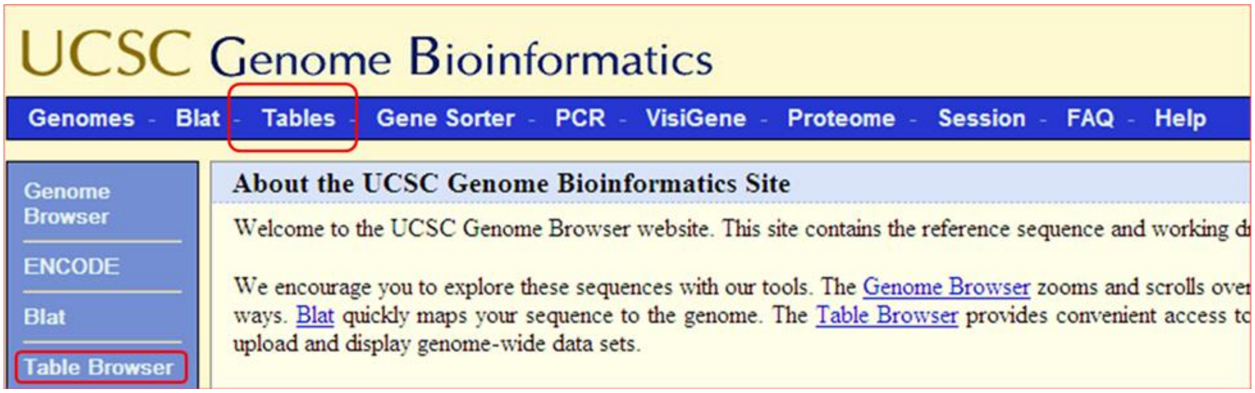
i. Using gene symbols

- Fetching upstream 1 kb sequences for a list of genes using gene symbols – Use "RefFlat" option.
- Human genes: Gene symbols should be in upper case and they should be approved gene symbols (for e.g. TP53 instead of p53 – although p53 is commonly used, the approved gene symbol is TP53)
- Mouse genes: Gene symbols should be lower case (the first letter should be upper case). For e.g. Trp53 for mouse p53. **Note:** the mouse orthologous gene for human TP53 gene is not Tp53 but Trp53!
- Where can I get a list of human and mouse approved symbols and the orthologous pairs?
  - NCBI's Homologene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene)
  - MGI's ortholog table (ftp://ftp.informatics.jax.org/pub/reports/index.html)

ii. Using accession numbers

- Same as above but with **two** differences:
  - Use RefSeq accession numbers (start with NM_XXXXXX
  - Use the option "RefGene" instead of "RefFlat"

2. **Advanced**: What if you want to download upstream 1 kb **plus** 200 bp downstream of the transcription start site? It can be done but you need to get the start positions of all your genes of interest and then (using excel) start subtracting 1000 from the start and also start adding 200. Thus, you will get a new start and end and you can use these to fetch the genomic sequences of desired length.
3. **Advanced**: How can I get a list of all SNPs occurring in the upstream 1 kb regions of genes of my interest?
    a. To do this you first need the coordinates or positions of the upstream regions (1 kb regions' start and end positions) (see 2 above)
    b. Then using these coordinates or regions as the base you can intersect with other features (for e.g. get me all SNPs that intersect or fall within these selected regions. Use the "intersection" feature from the table browser options.
4. **LIMITATIONS**:
    a. Transcription Start Sites (TSS) are still not well defined. Surprisingly the experimentally validated promoters for human and mouse are only about 1870 and 200 respectively! The **Eukaryotic Promoter Database** (EPD) is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally. EPD can be accessed at http://www.epd.isb-sib.ch/. For all practical purposes, we consider the region upstream to the NCBI's RefSeq mRNAs (http://www.ncbi.nlm.nih.gov/RefSeq/) as putative promoters. There is no single, perfect TSS or promoter prediction algorithms. Of the available ones, the firstEF (first exon finder; http://rulai.cshl.edu/tools/FirstEF/) does a reasonably good job (tested using experimentally validated promoters).
    b. Regulatory regions (especially enhancers and silencers) can occur just anywhere (intronic, downstream, farther upstream, UTRs).
    c. Regulatory mechanisms other than transcriptional: Posttranscriptional (e.g. miRNAs), posttranslational (protein modifications) or epigenetic mechanisms (differential effects of chromosome or chromatin packaging rather than by differences in DNA sequence).


UCSC Genome Bioinformatics
Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Proteome - Session - FAQ - Help
Genome Browser
ENCODE
Blat
Table Browser

About the UCSC Genome Bioinformatics Site

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

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls ove ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access tc upload and display genome-wide data sets.

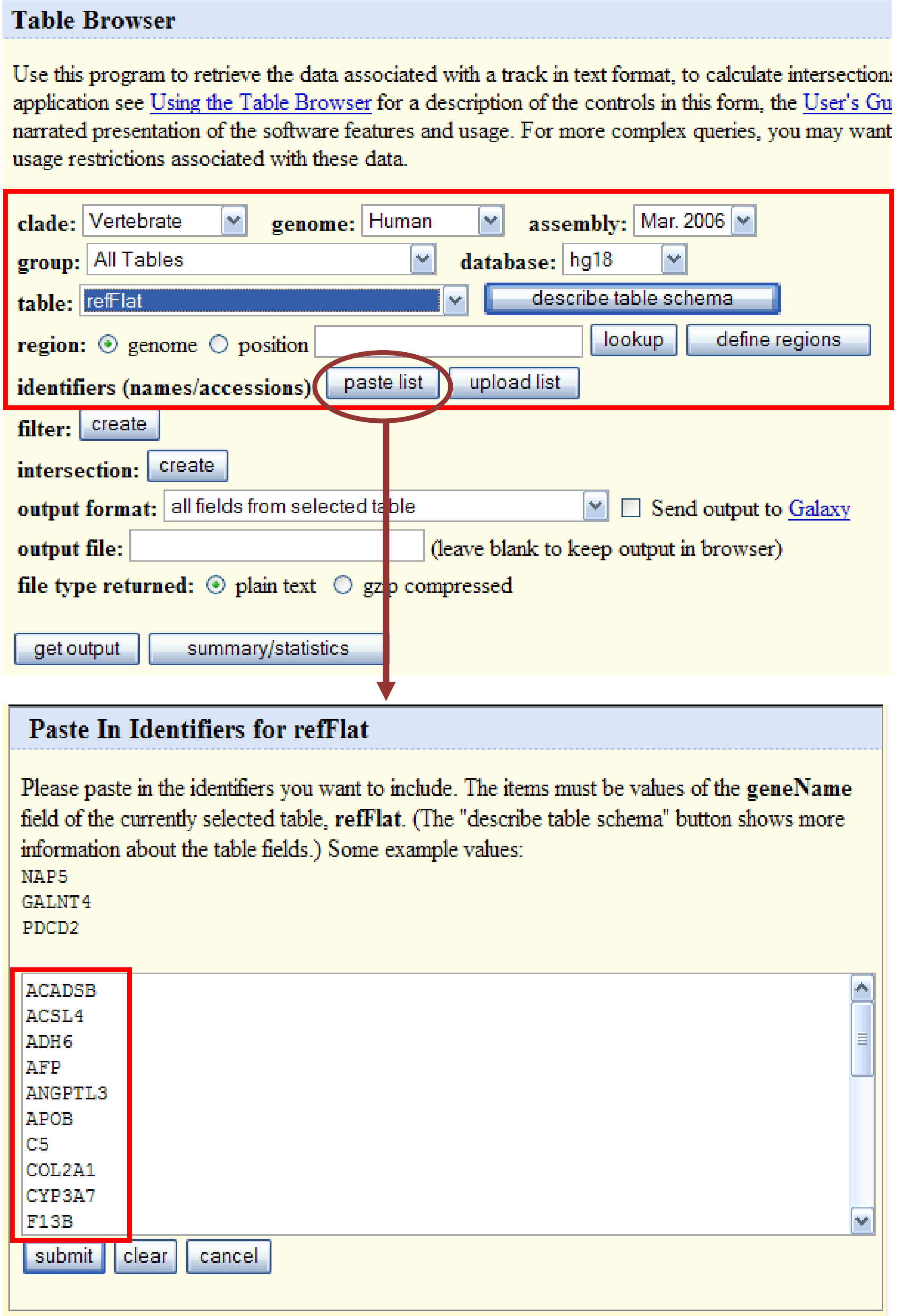


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersection: application see Using the Table Browser for a description of the controls in this form, the User's Gu narrated presentation of the software features and usage. For more complex queries, you may want usage restrictions associated with these data.

clade: Vertebrate **genome:** Human **assembly:** Mar. 2006
group: All Tables **database:** hg18
table: refFlat [describe table schema]
region: ● genome ○ position [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]

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:

```
NAP5
GALNT4
PDCD2
```

```
ACADSB
ACSL4
ADH6
AFP
ANGPTL3
APOB
C5
COL2A1
CYP3A7
F13B
```

[submit] [clear] [cancel]

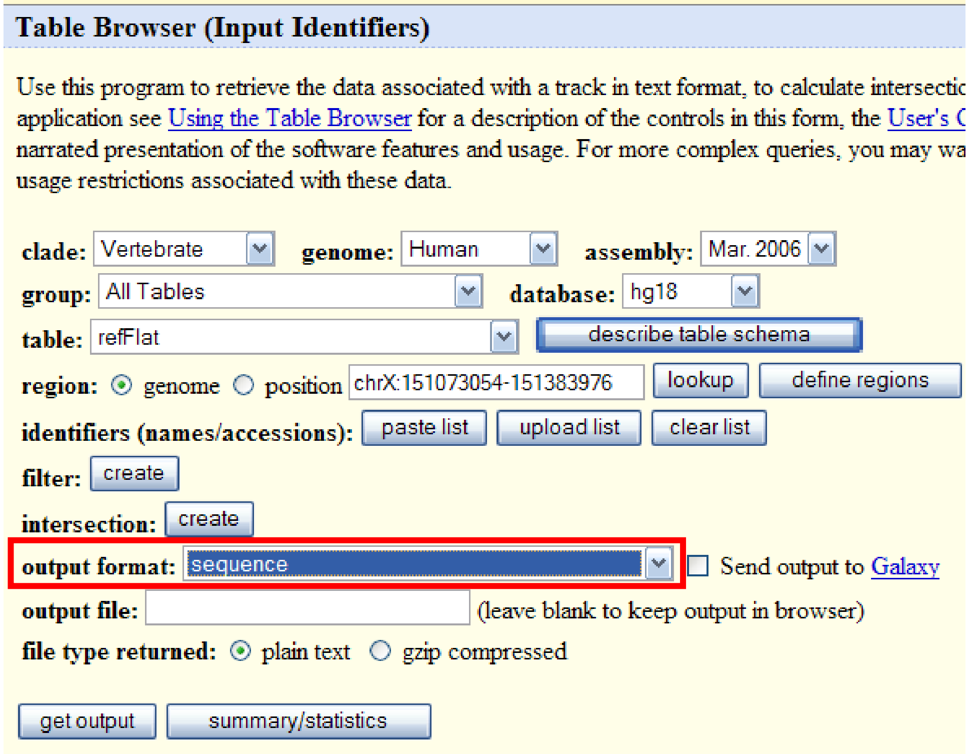


Table Browser (Input Identifiers)

Use this program to retrieve the data associated with a track in text format, to calculate intersectic
application see Using the Table Browser for a description of the controls in this form, the User's C
narrated presentation of the software features and usage. For more complex queries, you may wa
usage restrictions associated with these data.

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

group: All Tables **database:** hg18

table: refFlat

region: ● genome ○ position chrX:151073054-151383976

identifiers (names/accessions):

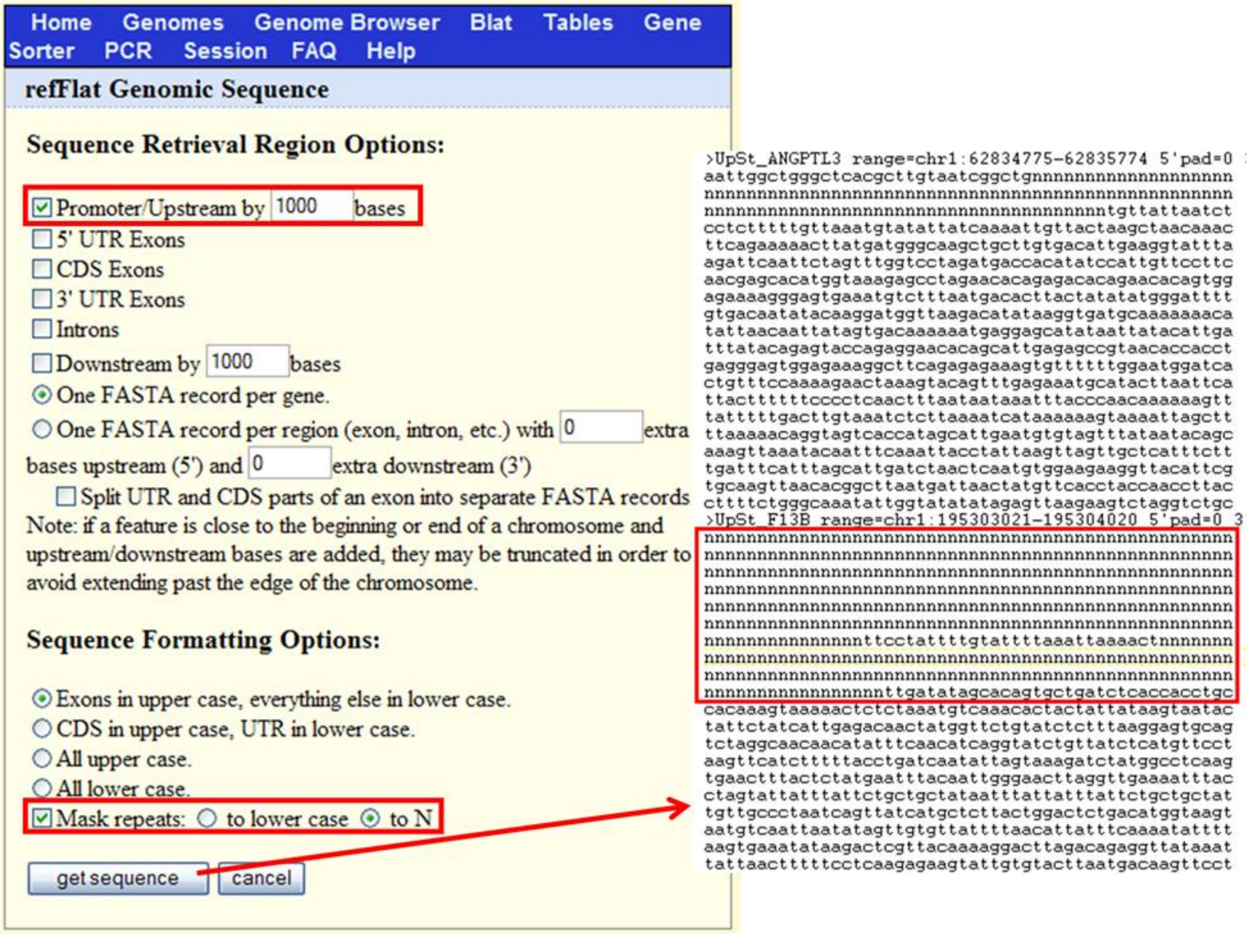
filter:

intersection:

output format: sequence □ Send output to Galaxy

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

file type returned: ● plain text ○ gzip compressed



refFlat Genomic Sequence

Sequence Retrieval Region Options:

☑ 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.
☑ Mask repeats: ○ to lower case ● to N

```
>UpSt_ANGPTL3 range=chr1:62834775-62835774 5'pad=0
aattggctgggctcacgcttgtaatcggctgnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnntgttattaatct
cctcttttttgttaaatgtatattatcaaaattgttactaagctaacaaac
ttcagaaaaacttatgatgggcaagctgcttgtgacattgaaggtattta
agattcaattctagtttggtcctagatgaccacatatccattgttccttc
aacgagcacatggtaaagagcctagaacacagagacacagaacacagtgg
agaaaagggagtgaaatgtctttaatgacacttactatatatgggatttt
gtgacaatatacaaggatggttaagacatataaggtgatgcaaaaaaaca
tattaacaattatagtgacaaaaaatgaggagcatataattatacattga
tttatacagagtaccagaggaacacagcattgagagccgtaacaccacct
gagggagtggagaaaggcttcagagagaaagtgtttttttggaatggatca
ctgtttccaaaagaactaaagtacagtttgagaaatgcatacttaattca
ttactttttttcccctcaactttaataataaatttacccaacaaaaaagtt
tatttttgacttgtaaatctcttaaaatcataaaaaagtaaaattagctt
ttaaaaacaggtagtcaccatagcattgaatgtgtagtttataatacagc
aaagttaaatacaatttcaaattacctattaagttagttgctcatttctt
tgatttcatttagcattgatctaactcaatgtggaagaaggttacattcg
tgcaagttaacacggcttaatgattaactatgttcacctaccaaccttac
cttttctgggcaaatattggtatatatagagttaagaagtctaggtctgc
>UpSt_F13B range=chr1:195303021-195304020 5'pad=0 3
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnttcctatttgtattttaaattaaaactnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnttgatatagcacagtgctgatctcaccacctgc
cacaaagtaaaaactctctaaatgtcaaacactactattataagtaatac
tattctatcattgagacaactatggttctgtatctctttaaggagtgcag
tctaggcaacaacatatttcaacatcaggtatctgttatctcatgttcct
aagttcatctttttacctgatcaatattagtaaagatctatggcctcaag
tgaactttactctatgaatttacaattgggaacttaggttgaaaatttac
ctagtattatttattctgctgctataatttattatttattctgctgctat
tgttgccctaatcagttatcatgctcttactggactctgacatggtaagt
aatgtcaattaatatagttgtgttattttaacattatttcaaaatatttt
aagtgaaatataagactcgttacaaaaggacttagacagaggttataaat
tattaactttttcctcaagagaagtattgtgtacttaatgacaagttcct
```

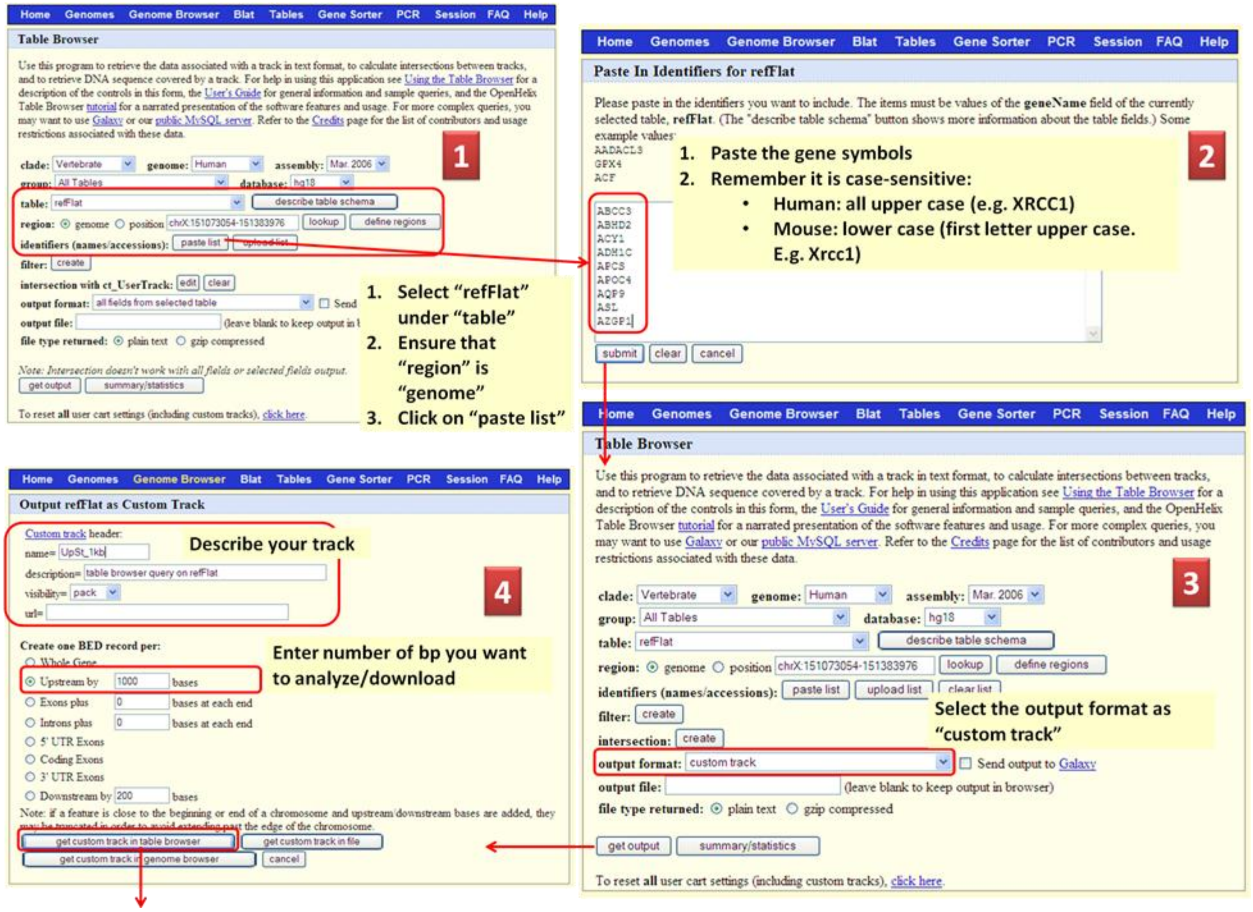


Table Browser

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: 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 b
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”

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

ABCC3
ABHD2
ACY1
ADH1C
APCS
APOC4
AQP9
ASL
AZGP1

[submit] [clear] [cancel]

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)

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: 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

[get output] [summary/statistics]

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

Select the output format as “custom track”

Output refFlat as Custom Track

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

Describe your track

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

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]

Enter number of bp you want to analyze/download

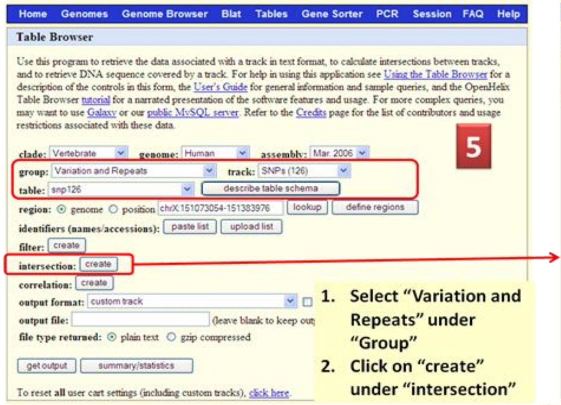


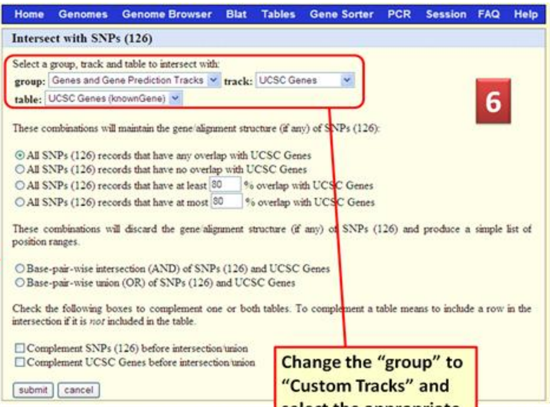
Table Browser

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
region: ● genome ○ position chrX:151073054-151383976
identifiers (names/accessions):
filter:
intersection:
correlation:
output format: custom track □
output file: (leave blank to keep out
file type returned: ● plain text ○ gzip compressed

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

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
6
submit
cancel

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)

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

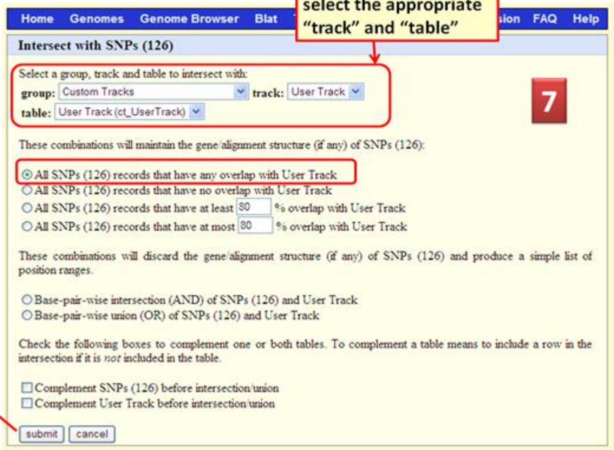
These combinations will discard the gene/alignment structure (if any) of SNPs (126) and produce a simple list of position ranges.

○ 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

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



Intersect with SNPs (126)

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

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

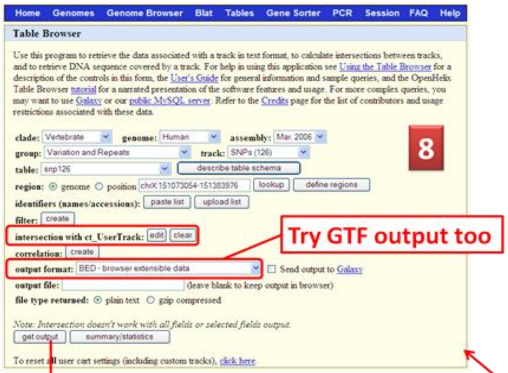


Table Browser

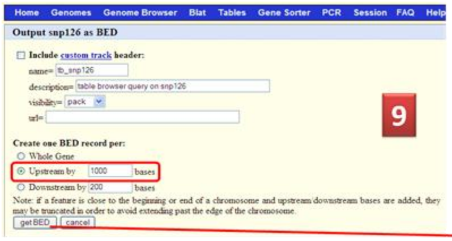
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for 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
region: ● genome ○ position chrX:151073054-151383976
identifiers (names/accessions):
filter:
intersection with ct_UserTrack:
correlation:
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.

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

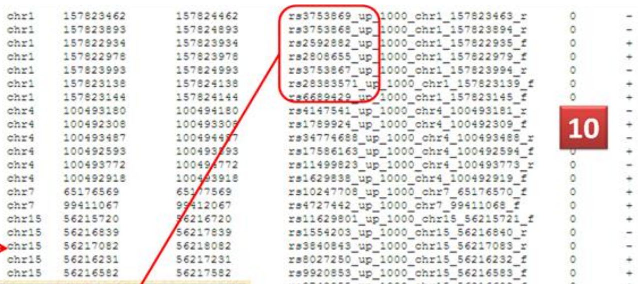
Try GTF output too


Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help
9
get BED
cancel

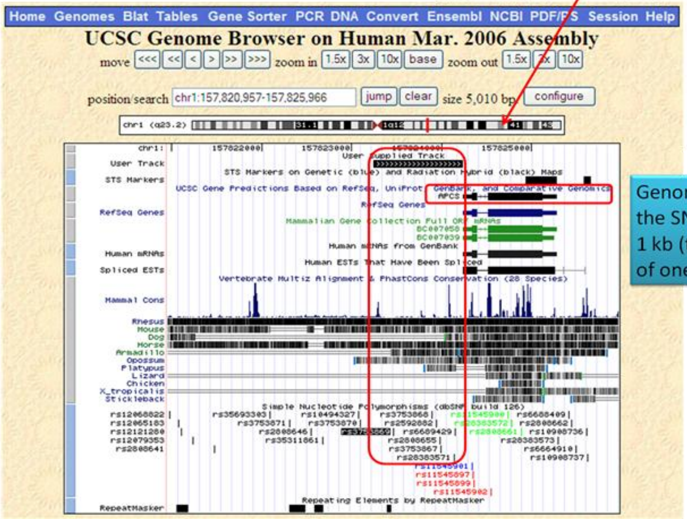
Output snp126 as BED

□ **Include custom track header:**
name= tb_snp126
description= table browser query on snp126
visibility= pack
url=

Create one BED record per:
○ Whole Gene
● Upstream by 1000 bases
○ Downstream by 200 bases
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.



```
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    +
```


Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
move <<< << < > >> >>>
zoom in 1.5x 3x 10x base
zoom out 1.5x 3x 10x
jump
clear
configure
chr1 (q23.2)
31.1
1q12
41
chr1:
157822000
157823000
157824000
157825000
User Supplied Track
User Track
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
APCS
RefSeq Genes
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC007058
BC007039
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 126)
Repeating Elements by RepeatMasker
RepeatMasker

UCSC Genome Browser on Human Mar. 2006 Assembly

position/search chr1:157,820,957-157,825,966 size 5,010 bp

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.

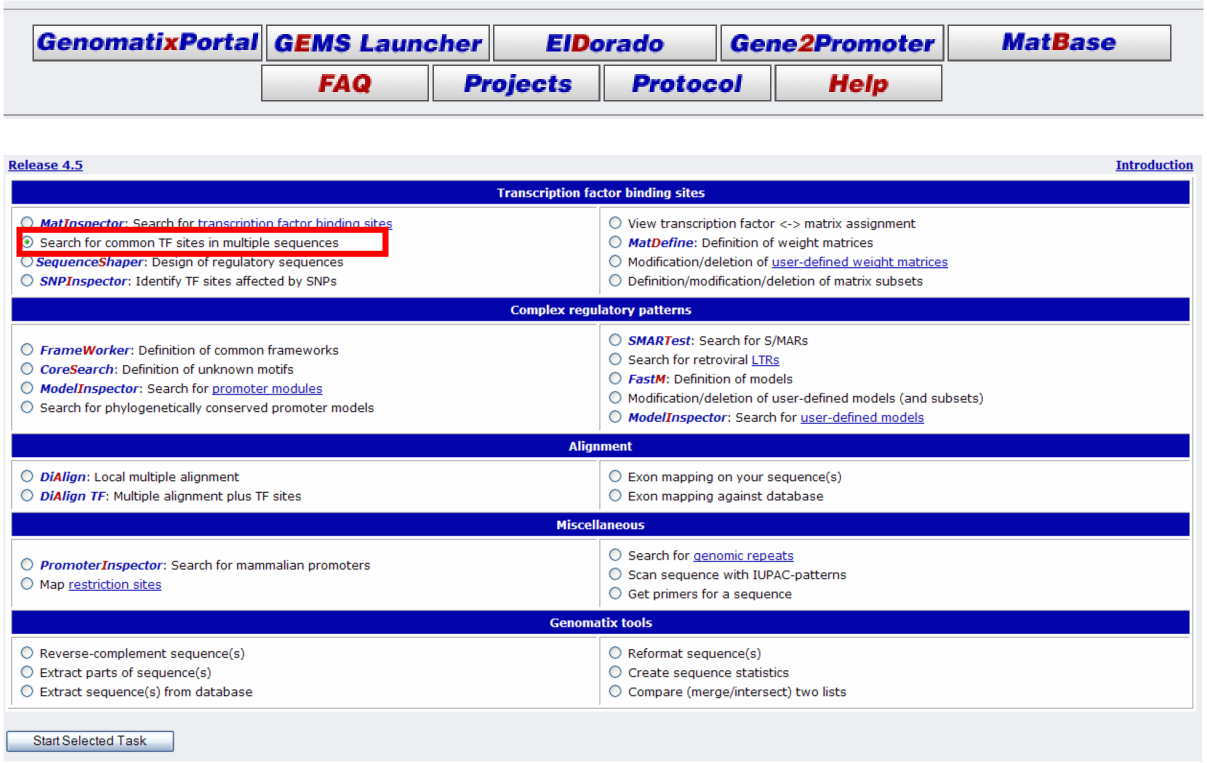
One drawback with this output is it doesn't tell you which SNPs are in the upstream region of which gene. However, since the positions of SNPs are included, you can compare them with the gene coordinates and figure it out .

Identification of Putative Common/Shared *Cis*-Elements

Known

Non-conserved

<u>GEMS Launcher (Genomatix) Search for common TF sites in multiple sequences</u>
URL: http://www.genomatix.de
Access: Licensed software (free for 20 analyses per month).
Description/Utility: You can search for transcription factor binding sites (TFBS) that are common to all or a subset of your input sequences. The results are displayed graphically. You can enter the percentage of sequences that have to contain the common sites. The search is based on known TFBSs that are stored as position weight matrices (PWMs) in the library.
Input: Fasta sequences (max 100 sequences) or any other format.
Output: Graphical and tab-delimited/spreadsheets
Other Comments: Genomatix has several other useful applications/tools that help to understand the molecular mechanisms of gene regulation as a central part of systems biology. The MatInspector tool of Genomatix is one of the widely used tools to identify potential binding sites in a DNA sequence.



Transcription factor binding sites

○ MatInspector: Search for transcription factor binding sites
● Search for common TF sites in multiple sequences
○ SequenceShaper: Design of regulatory sequences
○ SNPInspector: Identify TF sites affected by SNPs
○ View transcription factor <-> matrix assignment
○ MatDefine: Definition of weight matrices
○ Modification/deletion of user-defined weight matrices
○ Definition/modification/deletion of matrix subsets

Complex regulatory patterns

○ FrameWorker: Definition of common frameworks
○ CoreSearch: Definition of unknown motifs
○ ModelInspector: Search for promoter modules
○ Search for phylogenetically conserved promoter models
○ SMARTest: Search for S/MARs
○ Search for retroviral LTRs
○ FastM: Definition of models
○ Modification/deletion of user-defined models (and subsets)
○ ModelInspector: Search for user-defined models

Alignment

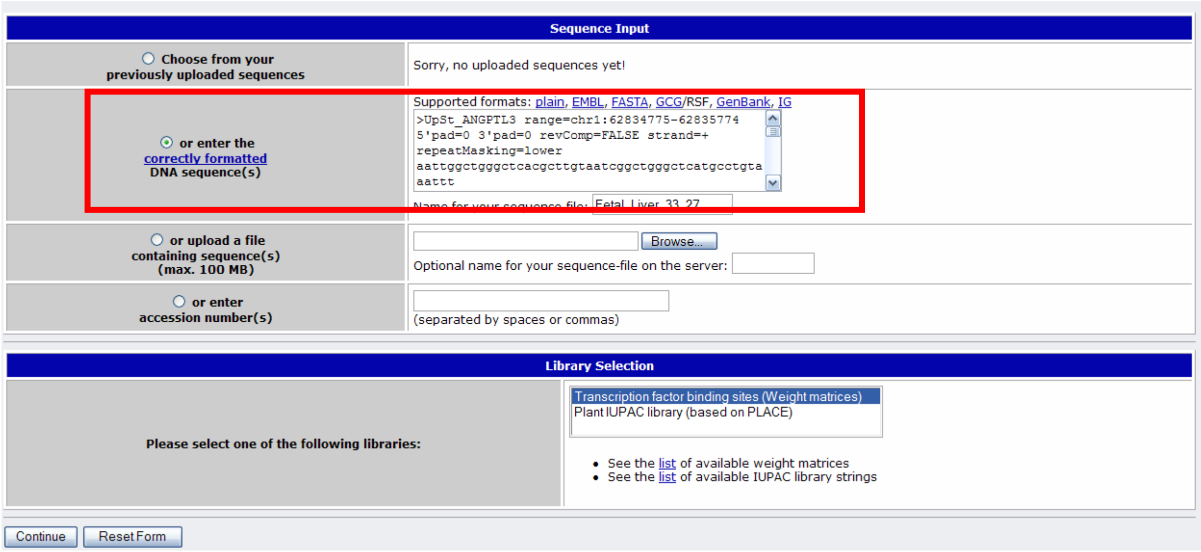
○ DiAlign: Local multiple alignment
○ DiAlign TF: Multiple alignment plus TF sites
○ Exon mapping on your sequence(s)
○ Exon mapping against database

Miscellaneous

○ PromoterInspector: Search for mammalian promoters
○ Map restriction sites
○ Search for genomic repeats
○ Scan sequence with IUPAC-patterns
○ Get primers for a sequence

Genomatix tools

○ Reverse-complement sequence(s)
○ Extract parts of sequence(s)
○ Extract sequence(s) from database
○ Reformat sequence(s)
○ Create sequence statistics
○ Compare (merge/intersect) two lists



Sequence Input

○ Choose from your previously uploaded sequences — Sorry, no uploaded sequences yet!

● or enter the correctly formatted DNA sequence(s) — Supported formats: plain, EMBL, FASTA, GCG/RSF, GenBank, IG

```
>UpSt_ANGPTL3 range=chr1:62834775-62835774
5'pad=0 3'pad=0 revComp=FALSE strand=+
repeatMasking=lower
aattggctgggctcacgcttgtaatcggctgggctcatgcctgta
aattt
```

Name for your sequence file: Fetal_Liver_33_27

○ or upload a file containing sequence(s) (max. 100 MB) — Browse... Optional name for your sequence-file on the server:

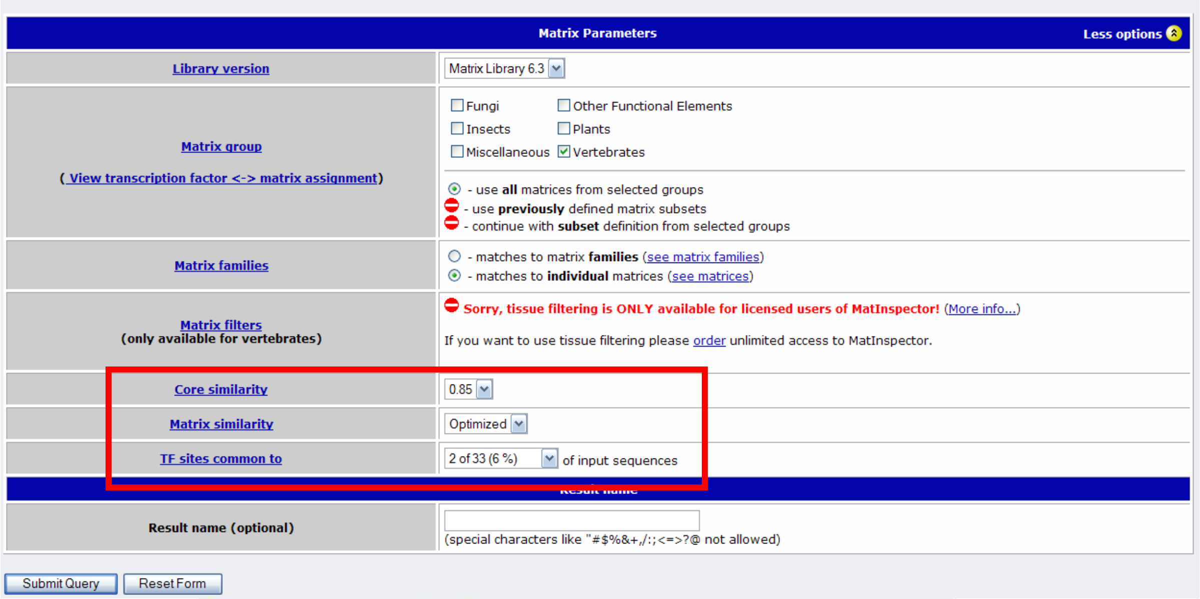
○ or enter accession number(s) — (separated by spaces or commas)

Library Selection

Please select one of the following libraries:

Transcription factor binding sites (Weight matrices)
Plant IUPAC library (based on PLACE)

- See the list of available weight matrices
- See the list of available IUPAC library strings



Matrix Parameters Less options

Library version: Matrix Library 6.3

Matrix group (View transcription factor <-> matrix assignment):

☐ Fungi ☐ Other Functional Elements
☐ Insects ☐ Plants
☐ Miscellaneous ☑ Vertebrates

● - use **all** matrices from selected groups
- use **previously** defined matrix subsets
- continue with **subset** definition from selected groups

Matrix families:
○ - matches to matrix **families** (see matrix families)
● - matches to **individual** matrices (see matrices)

Matrix filters (only available for vertebrates): Sorry, tissue filtering is ONLY available for licensed users of MatInspector! (More info...)

If you want to use tissue filtering please order unlimited access to MatInspector.

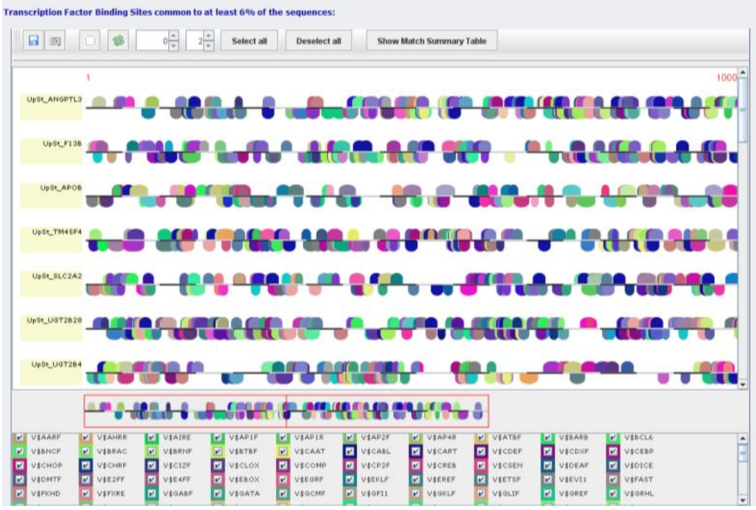
Core similarity: 0.85

Matrix similarity: Optimized

TF sites common to: 2 of 33 (6 %) of input sequences

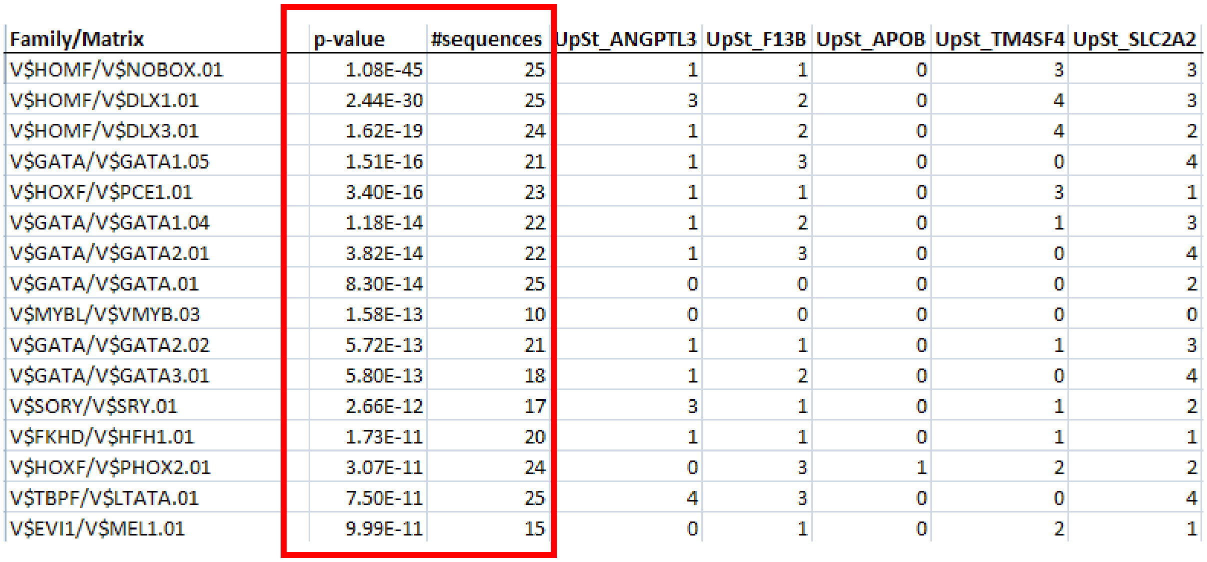
Result name

Result name (optional) — (special characters like "#$%&+,/:;<=>?@ not allowed)



Transcription Factor Binding Sites common to at least 6% of the sequences:

☑ V$AARF ☑ V$AHRR ☑ V$AIRE ☑ V$AP1F ☑ V$AP1R ☑ V$AP2F ☑ V$AP4R ☑ V$ATBF ☑ V$BARB ☑ V$BCL6
☑ V$BNCF ☑ V$BRAC ☑ V$BRNF ☑ V$BTBF ☑ V$CAAT ☑ V$CABL ☑ V$CART ☑ V$CDEF ☑ V$CDXF ☑ V$CEBP
☑ V$CHOP ☑ V$CHRF ☑ V$CIZF ☑ V$CLOX ☑ V$COMP ☑ V$CP2F ☑ V$CREB ☑ V$CSEN ☑ V$DEAF ☑ V$DICE
☑ V$DMTF ☑ V$E2FF ☑ V$E4FF ☑ V$EBOX ☑ V$EGRF ☑ V$EKLF ☑ V$EREF ☑ V$ETSF ☑ V$EVI1 ☑ V$FAST
☑ V$FKHD ☑ V$FXRE ☑ V$GABF ☑ V$GATA ☑ V$GCMF ☑ V$GFI1 ☑ V$GKLF ☑ V$GLIF ☑ V$GREF ☑ V$GRHL

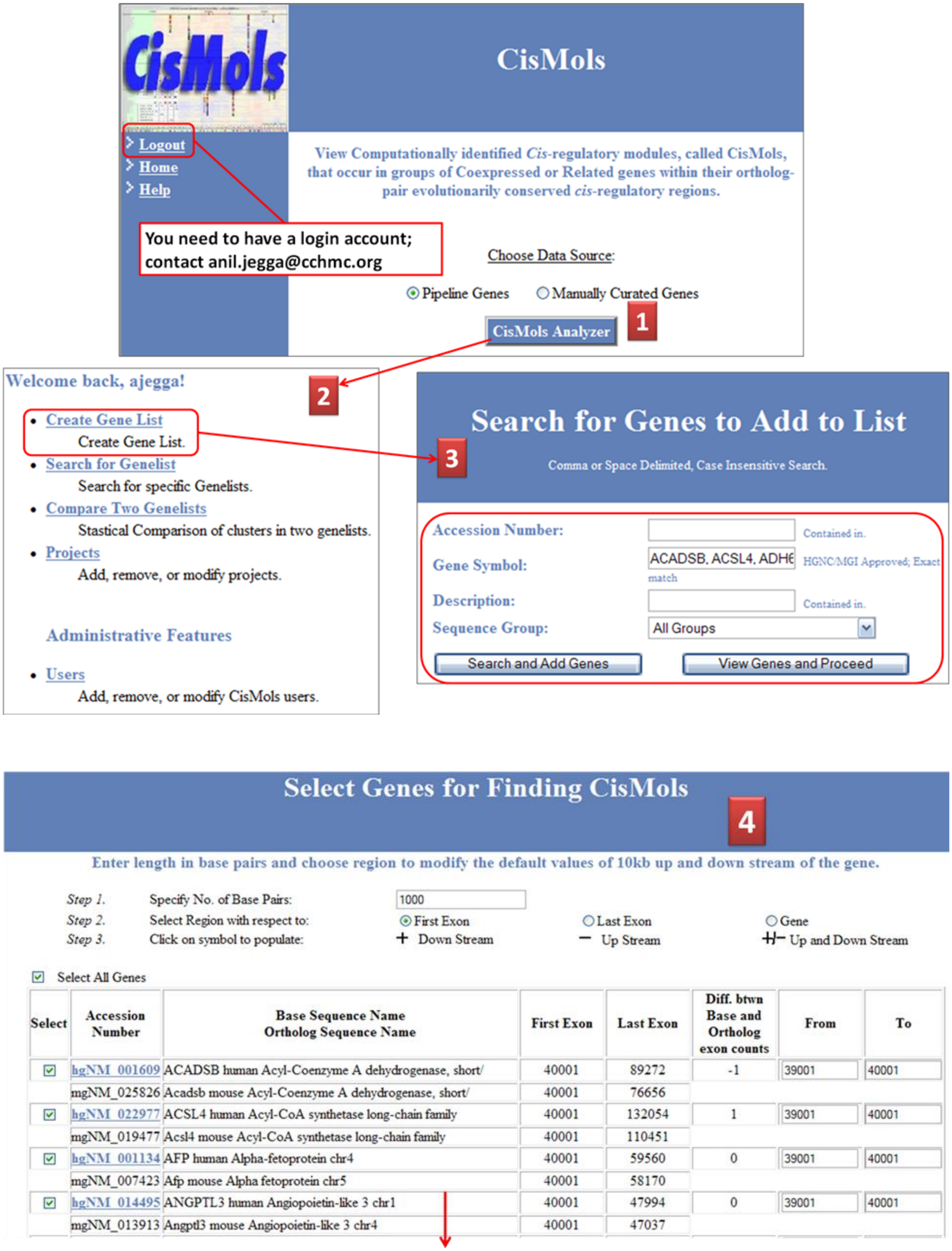


Family/Matrix	p-value	#sequences	UpSt_ANGPTL3	UpSt_F13B	UpSt_APOB	UpSt_TM4SF4	UpSt_SLC2A2
V$HOMF/V$NOBOX.01	1.08E-45	25	1	1	0	3	3
V$HOMF/V$DLX1.01	2.44E-30	25	3	2	0	4	3
V$HOMF/V$DLX3.01	1.62E-19	24	1	2	0	4	2
V$GATA/V$GATA1.05	1.51E-16	21	1	3	0	0	4
V$HOXF/V$PCE1.01	3.40E-16	23	1	1	0	3	1
V$GATA/V$GATA1.04	1.18E-14	22	1	2	0	1	3
V$GATA/V$GATA2.01	3.82E-14	22	1	3	0	0	4
V$GATA/V$GATA.01	8.30E-14	25	0	0	0	0	2
V$MYBL/V$VMYB.03	1.58E-13	10	0	0	0	0	0
V$GATA/V$GATA2.02	5.72E-13	21	1	1	0	1	3
V$GATA/V$GATA3.01	5.80E-13	18	1	2	0	0	4
V$SORY/V$SRY.01	2.66E-12	17	3	1	0	1	2
V$FKHD/V$HFH1.01	1.73E-11	20	1	1	0	1	1
V$HOXF/V$PHOX2.01	3.07E-11	24	0	3	1	2	2
V$TBPF/V$LTATA.01	7.50E-11	25	4	3	0	0	4
V$EVI1/V$MEL1.01	9.99E-11	15	0	1	0	2	1

Conserved

CisMols

URL: http://cismols.cchmc.org
Access: Free; Web-based
Description/Utility: Filtering candidate transcription factor binding site clusters (cis-regulatory element clusters) based on sequence conservation is helpful for an individual ortholog gene pair, but combining data from cis-conservation and coordinate expression across multiple genes is a more difficult problem. To approach this, we have extended an ortholog gene pair database with additional analytical architecture to allow for the analysis and identification of maximal numbers of compositionally similar and phylogenetically conserved cis-regulatory element clusters from a list of user-selected genes. Starting with identification of cis-clusters in phylogenetic footprints, we intend to extend the query to identify compositionally similar cis regulatory element clusters that occur in groups of co-regulated genes within each of their ortholog-pair evolutionarily conserved cis-regulatory regions. These computationally predicted cis-clusters, which we call as cismols, could serve as valuable probes for genome wide identification of regulatory regions and novel gene targets.
Support: Please refer to the “Help” section (**http://info.cchmc.org/help/cismols/index.html**) on CisMols home page (http://cismols.cchmc.org). For accounts, any problems or questions or analysis, send a mail to anil.jegga@cchmc.org.
Input: Human or mouse gene symbols or RefSeq accession numbers (hint: start with NM_).
Output: Graphical output (can be stored as pdf or any other image formats). Data can be downloaded in a spreadsheet.



CisMols

View Computationally identified *Cis*-regulatory modules, called CisMols, that occur in groups of Coexpressed or Related genes within their ortholog-pair evolutionarily conserved *cis*-regulatory regions.

You need to have a login account; contact anil.jegga@cchmc.org

Choose Data Source:

● Pipeline Genes ○ Manually Curated Genes

Welcome back, ajegga!

- Create Gene List
  Create Gene List.
- Search for Genelist
  Search for specific Genelists.
- Compare Two Genelists
  Stastical Comparison of clusters in two genelists.
- Projects
  Add, remove, or modify projects.

Administrative Features

- Users
  Add, remove, or modify CisMols users.

Search for Genes to Add to List

Comma or Space Delimited, Case Insensitive Search.

Accession Number:		Contained in.
Gene Symbol:	ACADSB, ACSL4, ADHE	HGNC/MGI Approved; Exact match
Description:		Contained in.
Sequence Group:	All Groups	

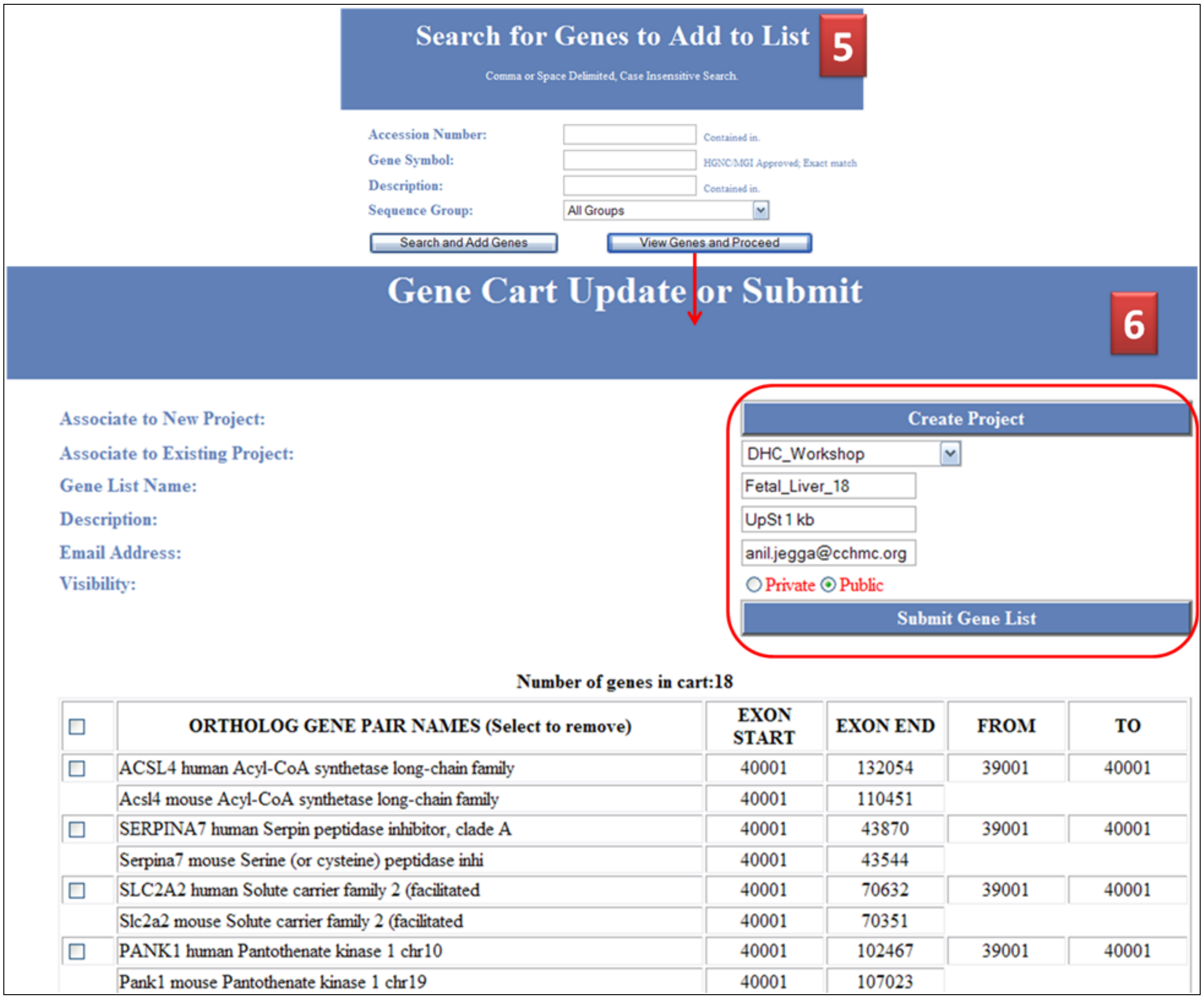
Select Genes for Finding CisMols

Enter length in base pairs and choose region to modify the default values of 10kb up and down stream of the gene.

Step 1.	Specify No. of Base Pairs:	1000		
Step 2.	Select Region with respect to:	● First Exon	○ Last Exon	○ Gene
Step 3.	Click on symbol to populate:	+ Down Stream	− Up Stream	+/− Up and Down Stream

☑ Select All Genes

Select	Accession Number	Base Sequence Name Ortholog Sequence Name	First Exon	Last Exon	Diff. btwn Base and Ortholog exon counts	From	To
☑	hgNM_001609	ACADSB human Acyl-Coenzyme A dehydrogenase, short/	40001	89272	-1	39001	40001
	mgNM_025826	Acadsb mouse Acyl-Coenzyme A dehydrogenase, short/	40001	76656			
☑	hgNM_022977	ACSL4 human Acyl-CoA synthetase long-chain family	40001	132054	1	39001	40001
	mgNM_019477	Acsl4 mouse Acyl-CoA synthetase long-chain family	40001	110451			
☑	hgNM_001134	AFP human Alpha-fetoprotein chr4	40001	59560	0	39001	40001
	mgNM_007423	Afp mouse Alpha fetoprotein chr5	40001	58170			
☑	hgNM_014495	ANGPTL3 human Angiopoietin-like 3 chr1	40001	47994	0	39001	40001
	mgNM_013913	Angptl3 mouse Angiopoietin-like 3 chr4	40001	47037			



Search for Genes to Add to List

Comma or Space Delimited, Case Insensitive Search.

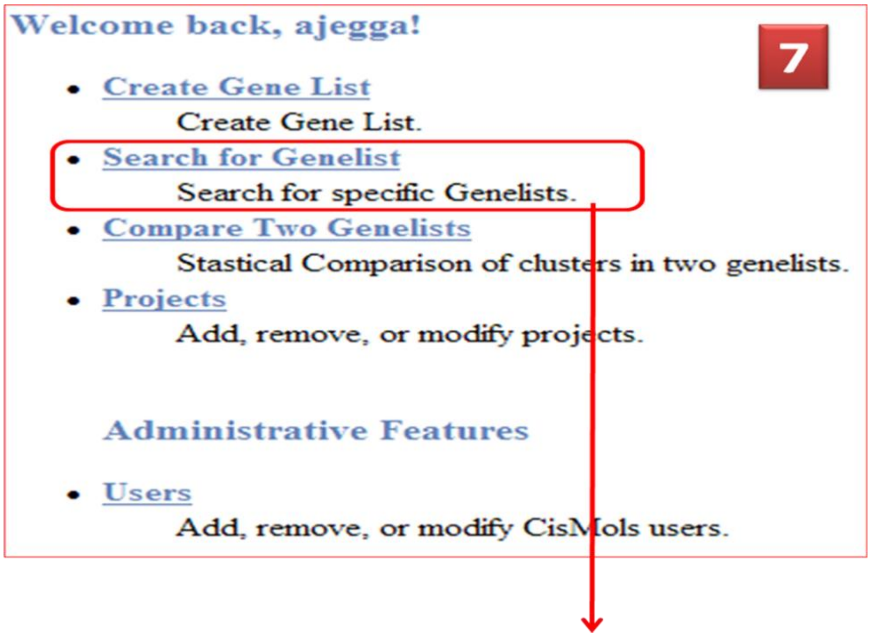
Accession Number:		Contained in.
Gene Symbol:		HGNC/MGI Approved; Exact match
Description:		Contained in.
Sequence Group:	All Groups	

Gene Cart Update or Submit

Associate to New Project:	Create Project
Associate to Existing Project:	DHC_Workshop
Gene List Name:	Fetal_Liver_18
Description:	UpSt 1 kb
Email Address:	anil.jegga@cchmc.org
Visibility:	○ Private ● Public

Number of genes in cart:18

□	ORTHOLOG GENE PAIR NAMES (Select to remove)	EXON START	EXON END	FROM	TO
□	ACSL4 human Acyl-CoA synthetase long-chain family	40001	132054	39001	40001
	Acsl4 mouse Acyl-CoA synthetase long-chain family	40001	110451		
□	SERPINA7 human Serpin peptidase inhibitor, clade A	40001	43870	39001	40001
	Serpina7 mouse Serine (or cysteine) peptidase inhi	40001	43544		
□	SLC2A2 human Solute carrier family 2 (facilitated	40001	70632	39001	40001
	Slc2a2 mouse Solute carrier family 2 (facilitated	40001	70351		
□	PANK1 human Pantothenate kinase 1 chr10	40001	102467	39001	40001
	Pank1 mouse Pantothenate kinase 1 chr19	40001	107023		

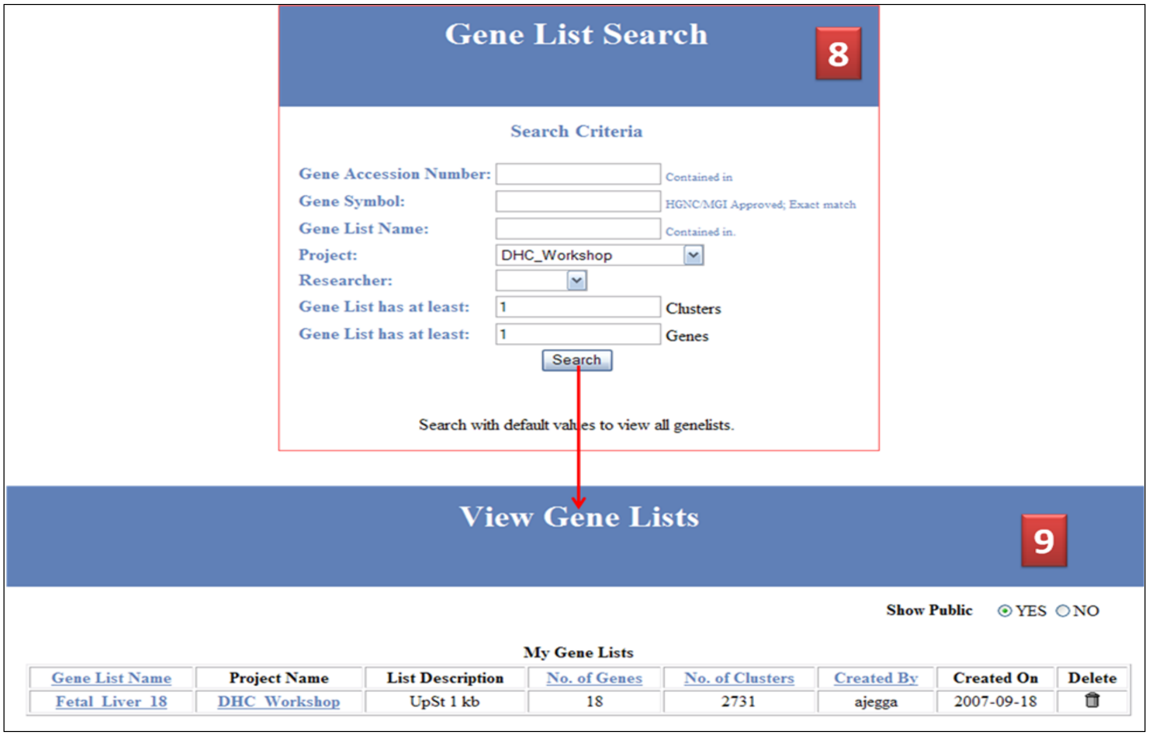


Welcome back, ajegga!

- Create Gene List
  Create Gene List.
- Search for Genelist
  Search for specific Genelists.
- Compare Two Genelists
  Stastical Comparison of clusters in two genelists.
- Projects
  Add, remove, or modify projects.

Administrative Features

- Users
  Add, remove, or modify CisMols users.



Gene List Search

Search Criteria

Gene Accession Number: Contained in

Gene Symbol: HGNC/MGI Approved; Exact match

Gene List Name: Contained in.

Project: DHC_Workshop

Researcher:

Gene List has at least: 1 Clusters

Gene List has at least: 1 Genes

Search

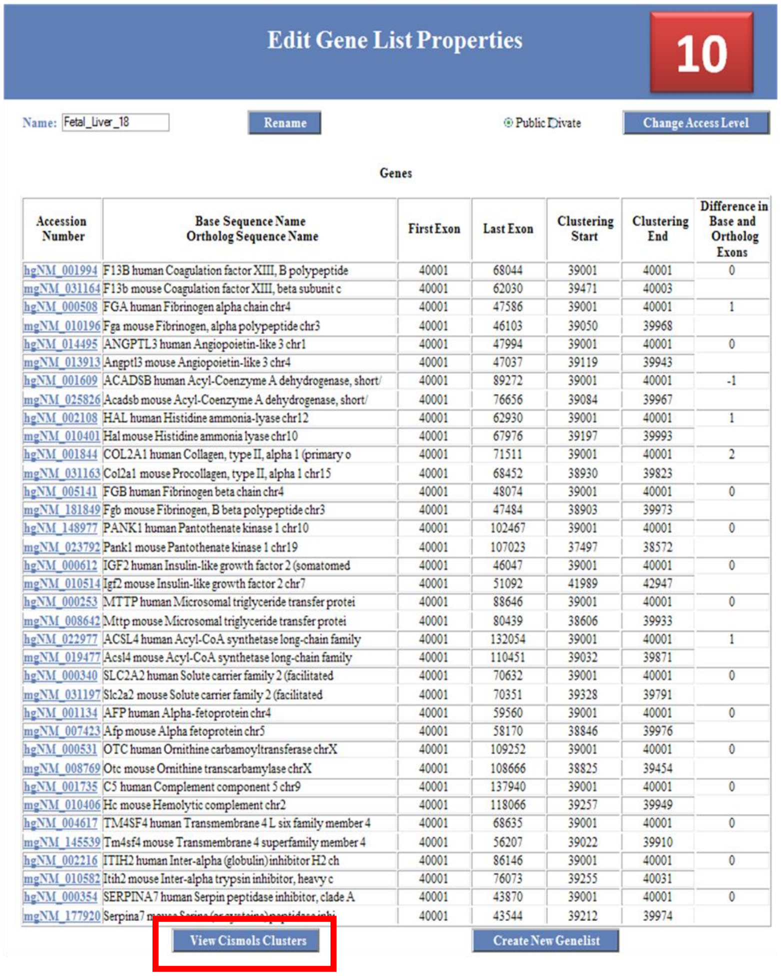
Search with default values to view all genelists.

View Gene Lists

Show Public ◉ YES ○ NO

My Gene Lists

Gene List Name	Project Name	List Description	No. of Genes	No. of Clusters	Created By	Created On	Delete
Fetal_Liver_18	DHC_Workshop	UpSt 1 kb	18	2731	ajegga	2007-09-18	🗑



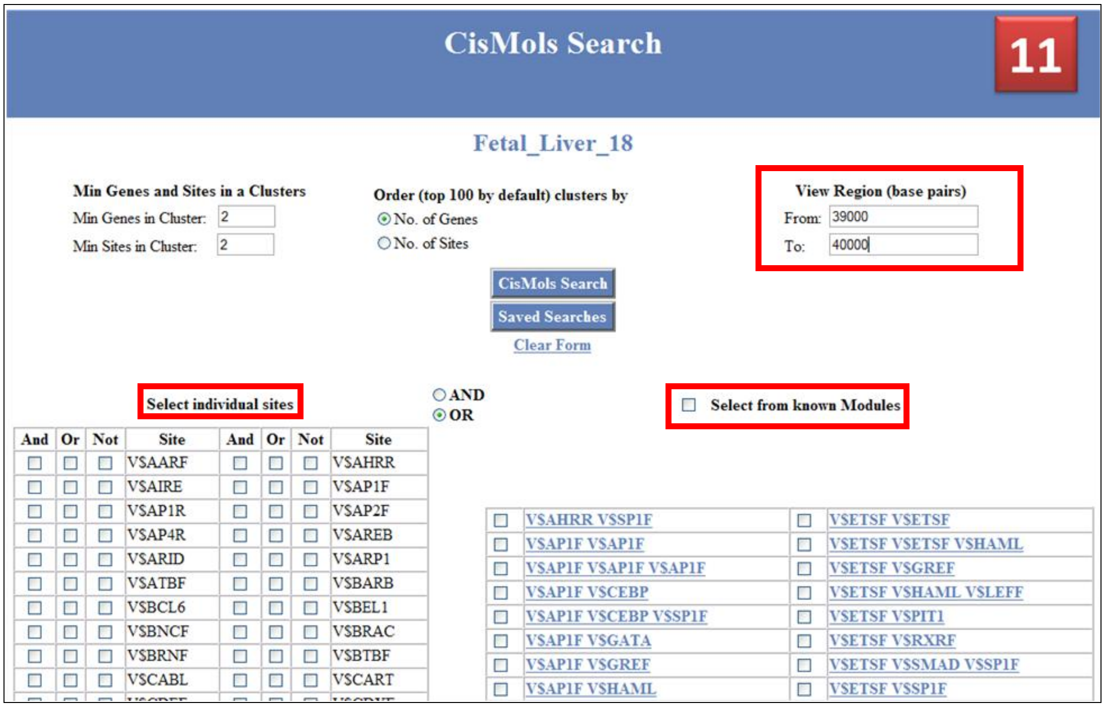
Edit Gene List Properties

Name: Fetal_Liver_18

◉ Public ○ Private

Genes

Accession Number	Base Sequence Name Ortholog Sequence Name	First Exon	Last Exon	Clustering Start	Clustering End	Difference in Base and Ortholog Exons
hgNM_001994	F13B human Coagulation factor XIII, B polypeptide	40001	68044	39001	40001	0
mgNM_031164	F13b mouse Coagulation factor XIII, beta subunit c	40001	62030	39471	40003	
hgNM_000508	FGA human Fibrinogen alpha chain chr4	40001	47586	39001	40001	1
mgNM_010196	Fga mouse Fibrinogen, alpha polypeptide chr3	40001	46103	39050	39968	
hgNM_014495	ANGPTL3 human Angiopoietin-like 3 chr1	40001	47994	39001	40001	0
mgNM_013913	Angptl3 mouse Angiopoietin-like 3 chr4	40001	47037	39119	39943	
hgNM_001609	ACADSB human Acyl-Coenzyme A dehydrogenase, short/	40001	89272	39001	40001	-1
mgNM_025826	Acadsb mouse Acyl-Coenzyme A dehydrogenase, short/	40001	76656	39084	39967	
hgNM_002108	HAL human Histidine ammonia-lyase chr12	40001	62930	39001	40001	1
mgNM_010401	Hal mouse Histidine ammonia lyase chr10	40001	67976	39197	39993	
hgNM_001844	COL2A1 human Collagen, type II, alpha 1 (primary o	40001	71511	39001	40001	2
mgNM_031163	Col2a1 mouse Procollagen, type II, alpha 1 chr15	40001	68452	38930	39823	
hgNM_005141	FGB human Fibrinogen beta chain chr4	40001	48074	39001	40001	0
mgNM_181849	Fgb mouse Fibrinogen, B beta polypeptide chr3	40001	47484	38903	39973	
hgNM_148977	PANK1 human Pantothenate kinase 1 chr10	40001	102467	39001	40001	0
mgNM_023792	Pank1 mouse Pantothenate kinase 1 chr19	40001	107023	37497	38572	
hgNM_000612	IGF2 human Insulin-like growth factor 2 (somatomed	40001	46047	39001	40001	0
mgNM_010514	Igf2 mouse Insulin-like growth factor 2 chr7	40001	51092	41989	42947	
hgNM_000253	MTTP human Microsomal triglyceride transfer protei	40001	88646	39001	40001	0
mgNM_008642	Mttp mouse Microsomal triglyceride transfer protei	40001	80439	38606	39933	
hgNM_022977	ACSL4 human Acyl-CoA synthetase long-chain family	40001	132054	39001	40001	1
mgNM_019477	Acsl4 mouse Acyl-CoA synthetase long-chain family	40001	110451	39032	39871	
hgNM_000340	SLC2A2 human Solute carrier family 2 (facilitated	40001	70632	39001	40001	0
mgNM_031197	Slc2a2 mouse Solute carrier family 2 (facilitated	40001	70351	39328	39791	
hgNM_001134	AFP human Alpha-fetoprotein chr4	40001	59560	39001	40001	0
mgNM_007423	Afp mouse Alpha fetoprotein chr5	40001	58170	38846	39976	
hgNM_000531	OTC human Ornithine carbamoyltransferase chrX	40001	109252	39001	40001	0
mgNM_008769	Otc mouse Ornithine transcarbamylase chrX	40001	108666	38825	39454	
hgNM_001735	C5 human Complement component 5 chr9	40001	137940	39001	40001	0
mgNM_010406	Hc mouse Hemolytic complement chr2	40001	118066	39257	39949	
hgNM_004617	TM4SF4 human Transmembrane 4 L six family member 4	40001	68635	39001	40001	0
mgNM_145539	Tm4sf4 mouse Transmembrane 4 superfamily member 4	40001	56207	39022	39910	
hgNM_002216	ITIH2 human Inter-alpha (globulin) inhibitor H2 ch	40001	86146	39001	40001	0
mgNM_010582	Itih2 mouse Inter-alpha trypsin inhibitor, heavy c	40001	76073	39255	40031	
hgNM_000354	SERPINA7 human Serpin peptidase inhibitor, clade A	40001	43870	39001	40001	0
mgNM_177920	Serpina7 mouse Serine (or cysteine) peptidase inhi	40001	43544	39212	39974	



CisMols Search

Fetal_Liver_18

Min Genes and Sites in a Clusters

Min Genes in Cluster: 2

Min Sites in Cluster: 2

Order (top 100 by default) clusters by

● No. of Genes

○ No. of Sites

View Region (base pairs)

From: 39000

To: 40000

Select individual sites

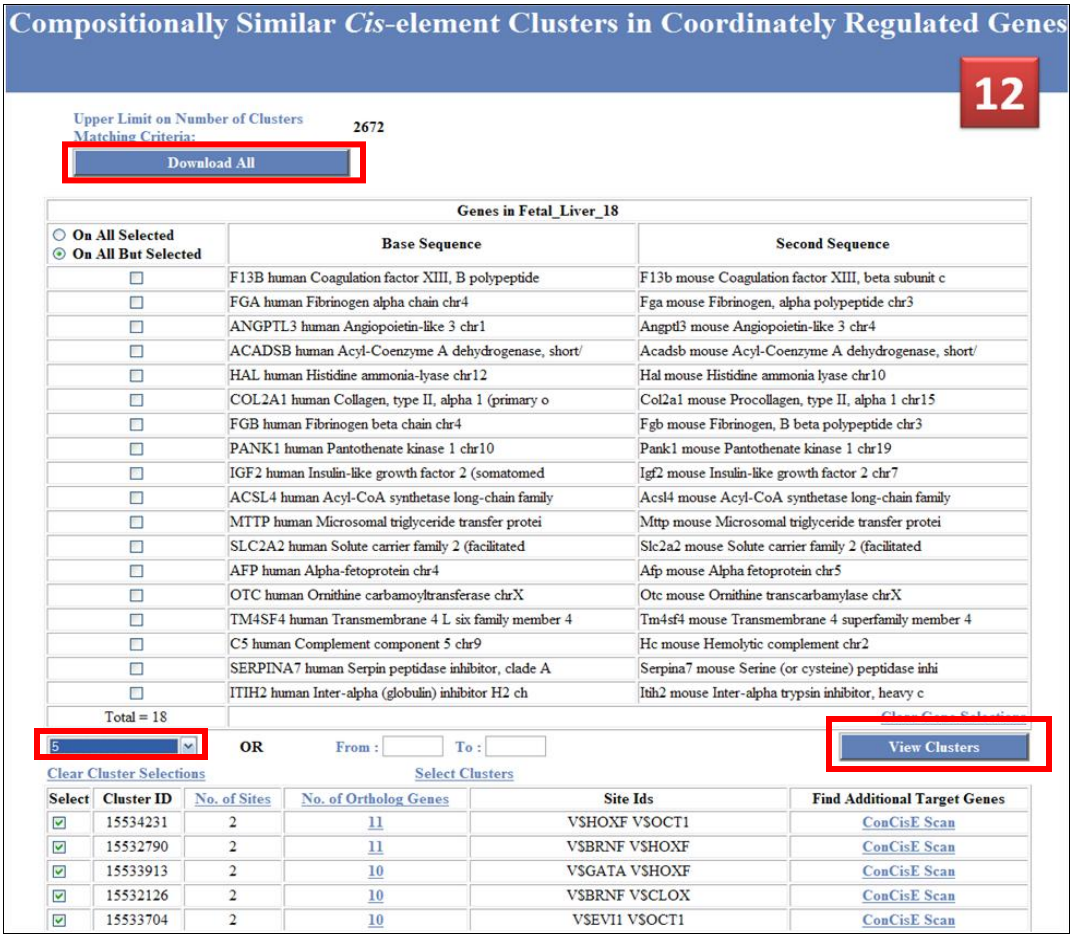
○ AND

● OR

□ **Select from known Modules**

And	Or	Not	Site	And	Or	Not	Site
□	□	□	V$AARF	□	□	□	V$AHRR
□	□	□	V$AIRE	□	□	□	V$AP1F
□	□	□	V$AP1R	□	□	□	V$AP2F
□	□	□	V$AP4R	□	□	□	V$AREB
□	□	□	V$ARID	□	□	□	V$ARP1
□	□	□	V$ATBF	□	□	□	V$BARB
□	□	□	V$BCL6	□	□	□	V$BEL1
□	□	□	V$BNCF	□	□	□	V$BRAC
□	□	□	V$BRNF	□	□	□	V$BTBF
□	□	□	V$CABL	□	□	□	V$CART

□	V$AHRR V$SP1F	□	V$ETSF V$ETSF
□	V$AP1F V$AP1F	□	V$ETSF V$ETSF V$HAML
□	V$AP1F V$AP1F V$AP1F	□	V$ETSF V$GREF
□	V$AP1F V$CEBP	□	V$ETSF V$HAML V$LEFF
□	V$AP1F V$CEBP V$SP1F	□	V$ETSF V$PIT1
□	V$AP1F V$GATA	□	V$ETSF V$RXRF
□	V$AP1F V$GREF	□	V$ETSF V$SMAD V$SP1F
□	V$AP1F V$HAML	□	V$ETSF V$SP1F



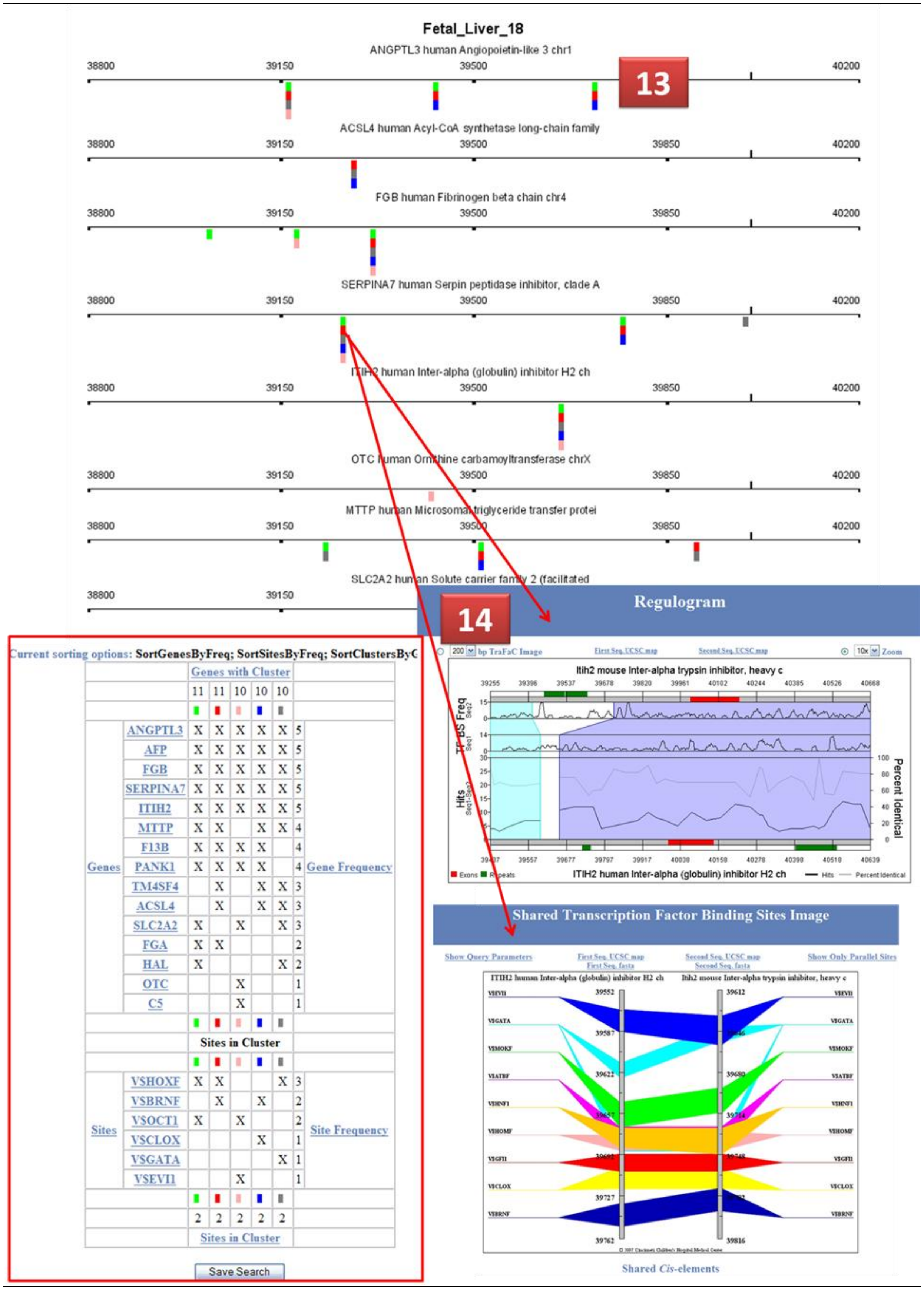
Compositionally Similar *Cis*-element Clusters in Coordinately Regulated Genes

Upper Limit on Number of Clusters Matching Criteria: 2672

Genes in Fetal_Liver_18		
○ On All Selected ● On All But Selected	**Base Sequence**	**Second Sequence**
□	F13B human Coagulation factor XIII, B polypeptide	F13b mouse Coagulation factor XIII, beta subunit c
□	FGA human Fibrinogen alpha chain chr4	Fga mouse Fibrinogen, alpha polypeptide chr3
□	ANGPTL3 human Angiopoietin-like 3 chr1	Angptl3 mouse Angiopoietin-like 3 chr4
□	ACADSB human Acyl-Coenzyme A dehydrogenase, short/	Acadsb mouse Acyl-Coenzyme A dehydrogenase, short/
□	HAL human Histidine ammonia-lyase chr12	Hal mouse Histidine ammonia lyase chr10
□	COL2A1 human Collagen, type II, alpha 1 (primary o	Col2a1 mouse Procollagen, type II, alpha 1 chr15
□	FGB human Fibrinogen beta chain chr4	Fgb mouse Fibrinogen, B beta polypeptide chr3
□	PANK1 human Pantothenate kinase 1 chr10	Pank1 mouse Pantothenate kinase 1 chr19
□	IGF2 human Insulin-like growth factor 2 (somatomed	Igf2 mouse Insulin-like growth factor 2 chr7
□	ACSL4 human Acyl-CoA synthetase long-chain family	Acsl4 mouse Acyl-CoA synthetase long-chain family
□	MTTP human Microsomal triglyceride transfer protei	Mttp mouse Microsomal triglyceride transfer protei
□	SLC2A2 human Solute carrier family 2 (facilitated	Slc2a2 mouse Solute carrier family 2 (facilitated
□	AFP human Alpha-fetoprotein chr4	Afp mouse Alpha fetoprotein chr5
□	OTC human Ornithine carbamoyltransferase chrX	Otc mouse Ornithine transcarbamylase chrX
□	TM4SF4 human Transmembrane 4 L six family member 4	Tm4sf4 mouse Transmembrane 4 superfamily member 4
□	C5 human Complement component 5 chr9	Hc mouse Hemolytic complement chr2
□	SERPINA7 human Serpin peptidase inhibitor, clade A	Serpina7 mouse Serine (or cysteine) peptidase inhi
□	ITIH2 human Inter-alpha (globulin) inhibitor H2 ch	Itih2 mouse Inter-alpha trypsin inhibitor, heavy c
Total = 18		Clear Gene Selections

5 OR From : To :

Select	Cluster ID	No. of Sites	No. of Ortholog Genes	Site Ids	Find Additional Target Genes
☑	15534231	2	11	V$HOXF V$OCT1	ConCisE Scan
☑	15532790	2	11	V$BRNF V$HOXF	ConCisE Scan
☑	15533913	2	10	V$GATA V$HOXF	ConCisE Scan
☑	15532126	2	10	V$BRNF V$CLOX	ConCisE Scan
☑	15533704	2	10	V$EVI1 V$OCT1	ConCisE Scan



Current sorting options: **SortGenesByFreq; SortSitesByFreq; SortClustersByC**

		Genes with Cluster						
		11	11	10	10	10		
		🟩	🟥	▫	🟦	▪		
Genes	ANGPTL3	X	X	X	X	X	5	Gene Frequency
	AFP	X	X	X	X	X	5	
	FGB	X	X	X	X	X	5	
	SERPINA7	X	X	X	X	X	5	
	ITIH2	X	X	X	X	X	5	
	MTTP	X	X		X	X	4	
	F13B	X	X	X	X		4	
	PANK1	X	X	X	X		4	
	TM4SF4		X		X	X	3	
	ACSL4		X		X	X	3	
	SLC2A2	X		X		X	3	
	FGA	X	X				2	
	HAL	X				X	2	
	OTC			X			1	
	C5			X			1	
		🟩	🟥	▫	🟦	▪		
		Sites in Cluster						
		🟩	🟥	▫	🟦	▪		
Sites	V$HOXF	X	X			X	3	Site Frequency
	V$BRNF		X		X		2	
	V$OCT1	X		X			2	
	V$CLOX				X		1	
	V$GATA					X	1	
	V$EVI1			X			1	
		🟩	🟥	▫	🟦	▪		
		2	2	2	2	2		
		Sites in Cluster						

DiRE (Distant Regulatory Elements of co-regulated genes)

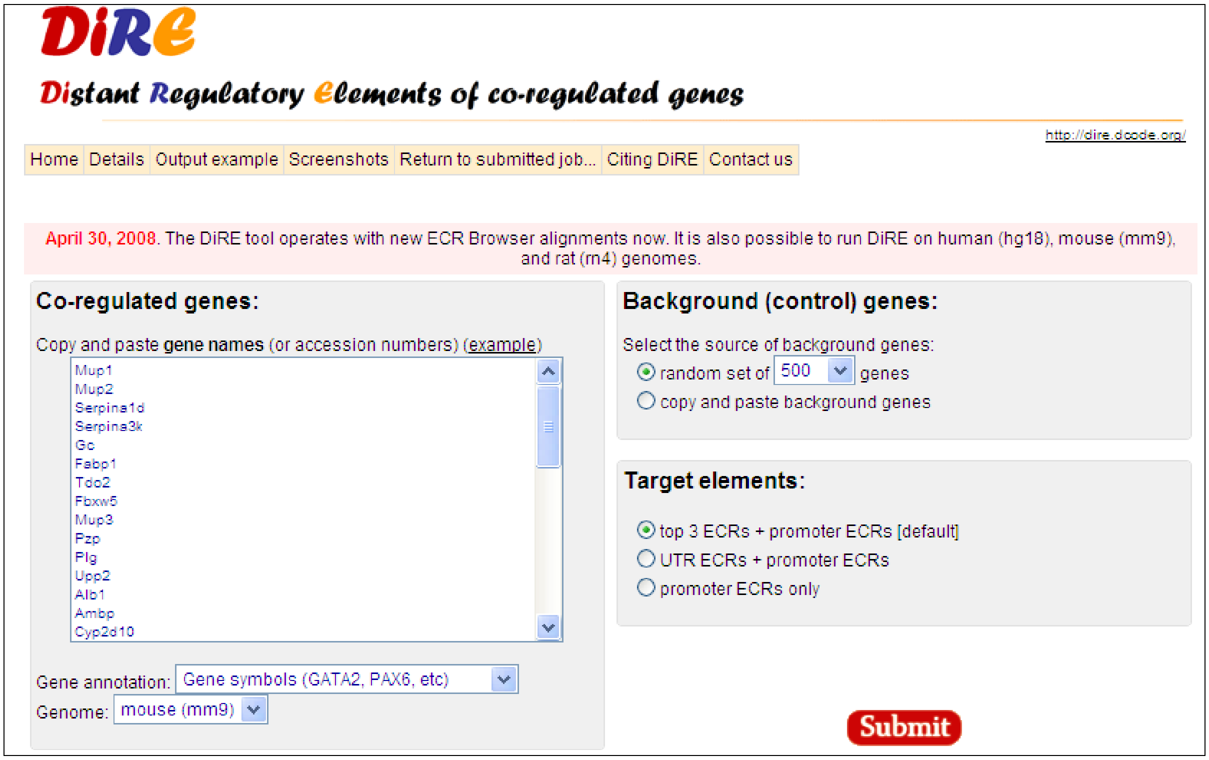
URL: http://dire.dcode.org

Access: Free; Web-based

Description/Utility: DiRE uses gene co-expression data, comparative genomics, and combinatorics of transcription factor binding sites (TFBSs) to find TFBS-association signatures that can be used for discriminating specific regulatory functions. DiRE's unique feature is the detection of REs outside of proximal promoter regions, as it takes advantage of the full gene locus to conduct the search. DiRE can predict common REs for any set of input genes for which the user has prior knowledge of co-expression, co-function, or other biologically meaningful grouping.

Input: Human or mouse or rat gene symbols or RefSeq accession numbers or chromosomal location.

Output: Graphical output and table.



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)

Mup1
Mup2
Serpina1d
Serpina3k
Gc
Fabp1
Tdo2
Fbxw5
Mup3
Fzp
Plg
Upp2
Alb1
Ambp
Cyp2d10

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

Genome: mouse (mm9)

Background (control) genes:

Select the source of background genes:

● random set of 500 genes

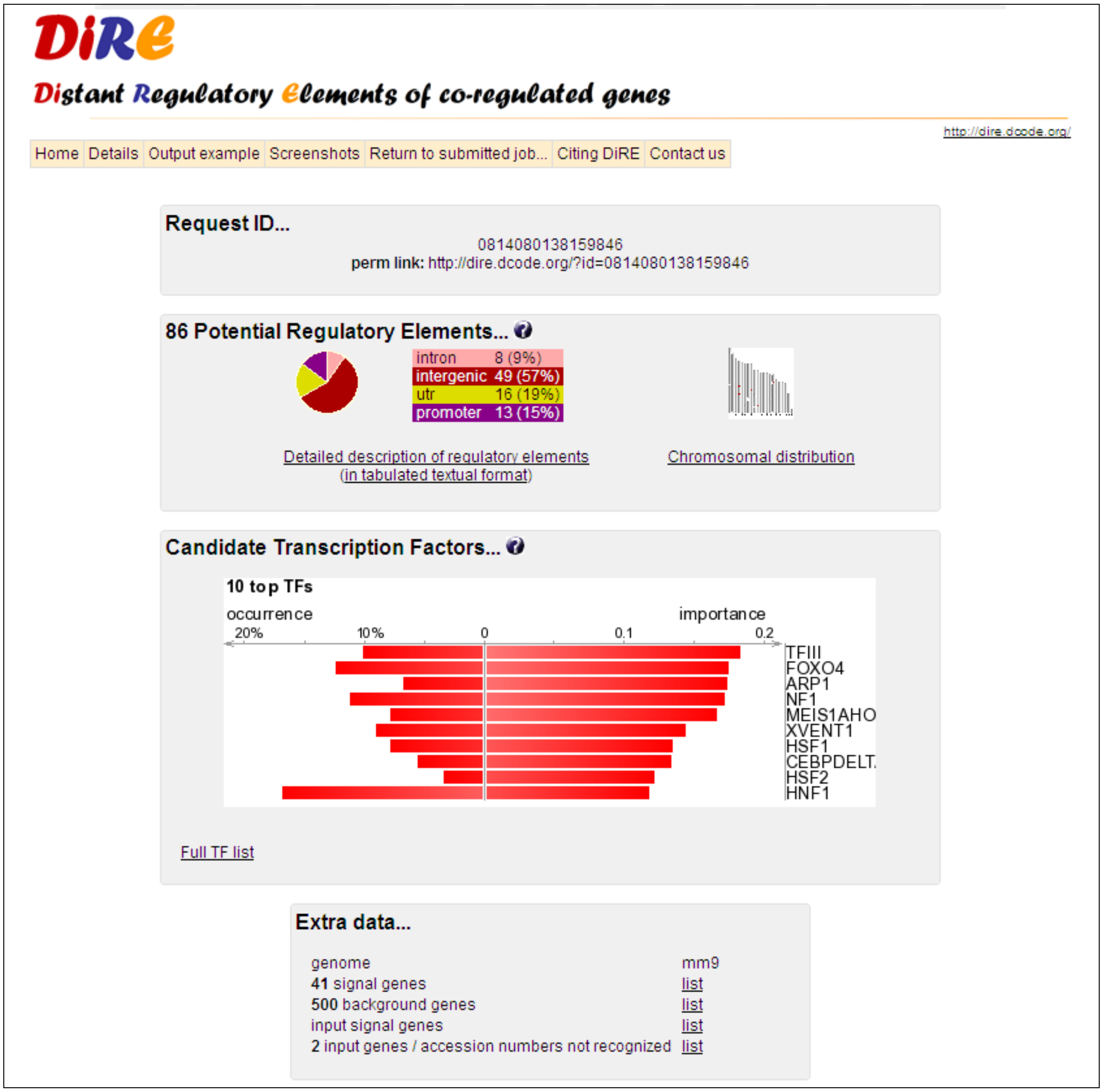
○ copy and paste background genes

Target elements:

● top 3 ECRs + promoter ECRs [default]

○ UTR ECRs + promoter ECRs

○ promoter ECRs only



DiRE

Distant Regulatory Elements of co-regulated genes

Request ID...

0814080138159846

perm link: http://dire.dcode.org/?id=0814080138159846

86 Potential Regulatory Elements...

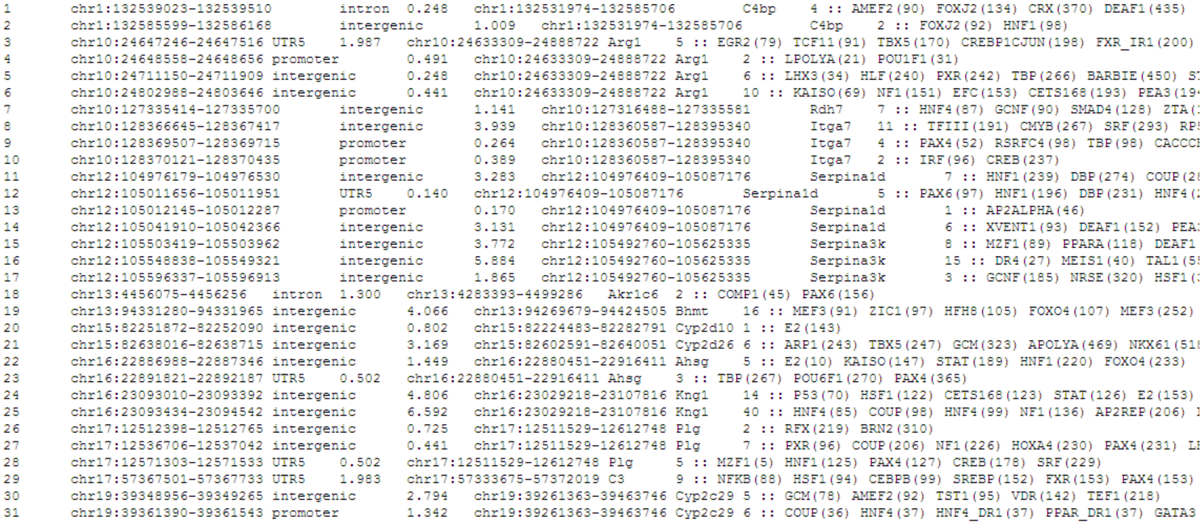
intron	8	(9%)
intergenic	49	(57%)
utr	16	(19%)
promoter	13	(15%)

Detailed description of regulatory elements
(in tabulated textual format)
Chromosomal distribution

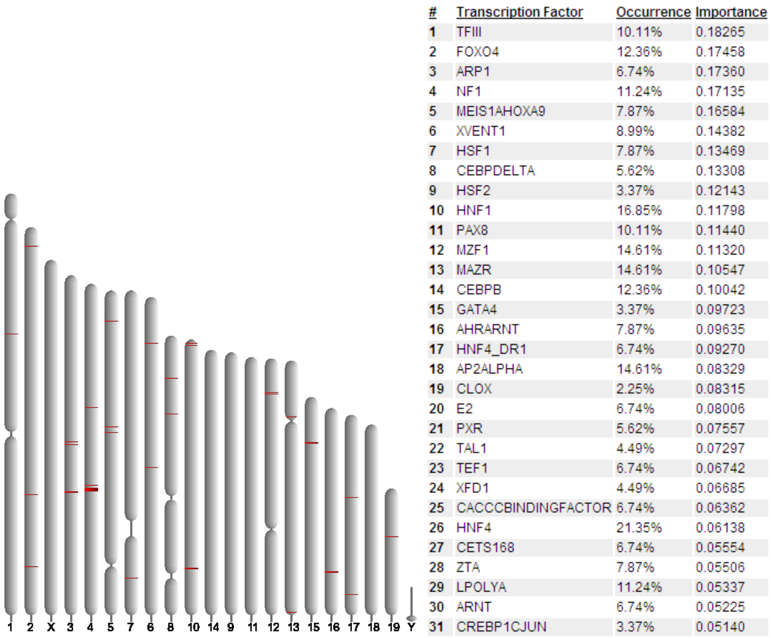
Candidate Transcription Factors...

Extra data...

genome	mm9
41 signal genes	list
500 background genes	list
input signal genes	list
2 input genes / accession numbers not recognized	list



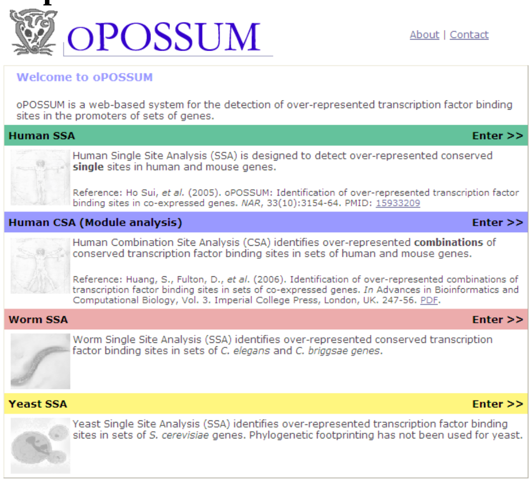
```
1       chr1:132539023-132539510        intron  0.248   chr1:132531974-132585706        C4bp    4 :: AMEF2(90) FOXJ2(134) CRX(370) DEAF1(435)
2       chr1:132585599-132586168        intergenic      1.009   chr1:132531974-132585706        C4bp    2 :: FOXJ2(92) HNF1(98)
3       chr10:24647246-24647516 UTR5    1.987   chr10:24633309-24888722 Arg1    5 :: EGR2(79) TCF11(91) TBX5(170) CREBP1CJUN(198) FXR_IR1(200)
4       chr10:24648558-24648656 promoter        0.491   chr10:24633309-24888722 Arg1    2 :: LPOLYA(21) POU1F1(31)
5       chr10:24711150-24711909 intergenic      0.248   chr10:24633309-24888722 Arg1    6 :: LHX3(34) HLF(240) PXR(242) TBP(266) BARBIE(450) S
6       chr10:24802988-24803646 intergenic      0.441   chr10:24633309-24888722 Arg1    10 :: KAISO(69) NF1(151) EFC(153) CETS168(193) PEA3(19
7       chr10:127335414-127335700       intergenic      1.141   chr10:127316488-127335581       Rdh7    7 :: HNF4(87) GCNF(90) SMAD4(128) ZTA(
8       chr10:128366645-128367417       intergenic      3.939   chr10:128360587-128395340       Itga7   11 :: TFIII(191) CMYB(267) SRF(293) RP
9       chr10:128369507-128369715       promoter        0.264   chr10:128360587-128395340       Itga7   4 :: PAX4(52) RSRFC4(98) TBP(98) CACCC
10      chr10:128370121-128370435       promoter        0.389   chr10:128360587-128395340       Itga7   2 :: IRF(96) CREB(237)
11      chr12:104976179-104976530       intergenic      3.283   chr12:104976409-105087176       Serpina1d       7 :: HNF1(239) DBP(274) COUP(2
12      chr12:105011656-105011951       UTR5    0.140   chr12:104976409-105087176       Serpina1d       5 :: PAX6(97) HNF1(196) DBP(231) HNF4(
13      chr12:105012145-105012287       promoter        0.170   chr12:104976409-105087176       Serpina1d       1 :: AP2ALPHA(46)
14      chr12:105041910-105042366       intergenic      3.131   chr12:104976409-105087176       Serpina1d       6 :: XVENT1(93) DEAF1(152) PEA
15      chr12:105503419-105503962       intergenic      3.772   chr12:105492760-105625335       Serpina3k       8 :: MZF1(89) PPARA(118) DEAF1
16      chr12:105548838-105549321       intergenic      5.884   chr12:105492760-105625335       Serpina3k       15 :: DR4(27) MEIS1(40) TAL1(5
17      chr12:105596337-105596913       intergenic      1.865   chr12:105492760-105625335       Serpina3k       3 :: GCNF(185) NRSE(320) HSF1(
18      chr13:4456075-4456256   intron  1.300   chr13:4283393-4499286   Akr1c6  2 :: COMP1(45) PAX6(156)
19      chr13:94331280-94331965 intergenic      4.066   chr13:94269679-94424505 Bhmt    16 :: MEF3(91) ZIC1(97) HFH8(105) FOXO4(107) MEF3(252)
20      chr15:82251872-82252090 intergenic      0.802   chr15:82224483-82282791 Cyp2d10 1 :: E2(143)
21      chr15:82638016-82638715 intergenic      3.169   chr15:82602591-82640051 Cyp2d26 6 :: ARP1(243) TBX5(247) GCM(323) APOLYA(469) NKX61(51
22      chr16:22886988-22887346 intergenic      1.449   chr16:22880451-22916411 Ahsg    5 :: E2(10) KAISO(147) STAT(189) HNF1(220) FOXO4(233)
23      chr16:22891821-22892187 UTR5    0.502   chr16:22880451-22916411 Ahsg    3 :: TBP(267) POU6F1(270) PAX4(365)
24      chr16:23093010-23093392 intergenic      4.806   chr16:23029218-23107816 Kng1    14 :: P53(70) HSF1(122) CETS168(123) STAT(126) E2(153)
25      chr16:23093434-23094542 intergenic      6.592   chr16:23029218-23107816 Kng1    40 :: HNF4(85) COUP(98) HNF4(99) NF1(136) AP2REP(206) I
26      chr17:12512398-12512765 intergenic      0.725   chr17:12511529-12612748 Plg     2 :: RFX(219) BRN2(310)
27      chr17:12536706-12537042 intergenic      0.441   chr17:12511529-12612748 Plg     7 :: PXR(96) COUP(206) NF1(226) HOXA4(230) PAX4(231) L
28      chr17:12571303-12571533 UTR5    0.502   chr17:12511529-12612748 Plg     5 :: MZF1(5) HNF1(125) PAX4(127) CREB(178) SRF(229)
29      chr17:57367501-57367733 UTR5    1.983   chr17:57333675-57372019 C3      9 :: NFKB(88) HSF1(94) CEBPB(99) SREBP(152) FXR(153) PAX4(153)
30      chr19:39348956-39349265 intergenic      2.794   chr19:39261363-39463746 Cyp2c29 5 :: GCM(78) AMEF2(92) TST1(95) VDR(142) TEF1(218)
31      chr19:39361390-39361543 promoter        1.342   chr19:39261363-39463746 Cyp2c29 6 :: COUP(36) HNF4(37) HNF4_DR1(37) PPAR_DR1(37) GATA3
```



#	Transcription Factor	Occurrence	Importance
1	TFIII	10.11%	0.18265
2	FOXO4	12.36%	0.17458
3	ARP1	6.74%	0.17360
4	NF1	11.24%	0.17135
5	MEIS1AHOXA9	7.87%	0.16584
6	XVENT1	8.99%	0.14382
7	HSF1	7.87%	0.13469
8	CEBPDELTA	5.62%	0.13308
9	HSF2	3.37%	0.12143
10	HNF1	16.85%	0.11798
11	PAX8	10.11%	0.11440
12	MZF1	14.61%	0.11320
13	MAZR	14.61%	0.10547
14	CEBPB	12.36%	0.10042
15	GATA4	3.37%	0.09723
16	AHRARNT	7.87%	0.09635
17	HNF4_DR1	6.74%	0.09270
18	AP2ALPHA	14.61%	0.08329
19	CLOX	2.25%	0.08315
20	E2	6.74%	0.08006
21	PXR	5.62%	0.07557
22	TAL1	4.49%	0.07297
23	TEF1	6.74%	0.06742
24	XFD1	4.49%	0.06685
25	CACCCBINDINGFACTOR	6.74%	0.06362
26	HNF4	21.35%	0.06138
27	CETS168	6.74%	0.05554
28	ZTA	7.87%	0.05506
29	LPOLYA	11.24%	0.05337
30	ARNT	6.74%	0.05225
31	CREBP1CJUN	3.37%	0.05140

oPOSSUM (Distant Regulatory Elements of co-regulated genes)

URL: http://burgundy.cmmt.ubc.ca/oPOSSUM/
Access: Free; Web-based
Description/Utility: oPOSSUM is a system for determing the over-representation of transcription factor binding sites (TFBS) within a set of (co-expressed) genes as compared with a pre-compiled background set. The input is a set of gene identifiers. Analysis parameters are chosen. The system then compares the number of hits for each selected TFBS in the target gene set against the background set. Two different measures of statistical significance are applied to determine which sites are over-represented in the target set. The results of the analysis are displayed in tabular form.
Input: Human or mouse gene symbols or RefSeq accession numbers or Ensembl ID, or Entrez Gene IDs.
Output: Tab-delimited file.



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.

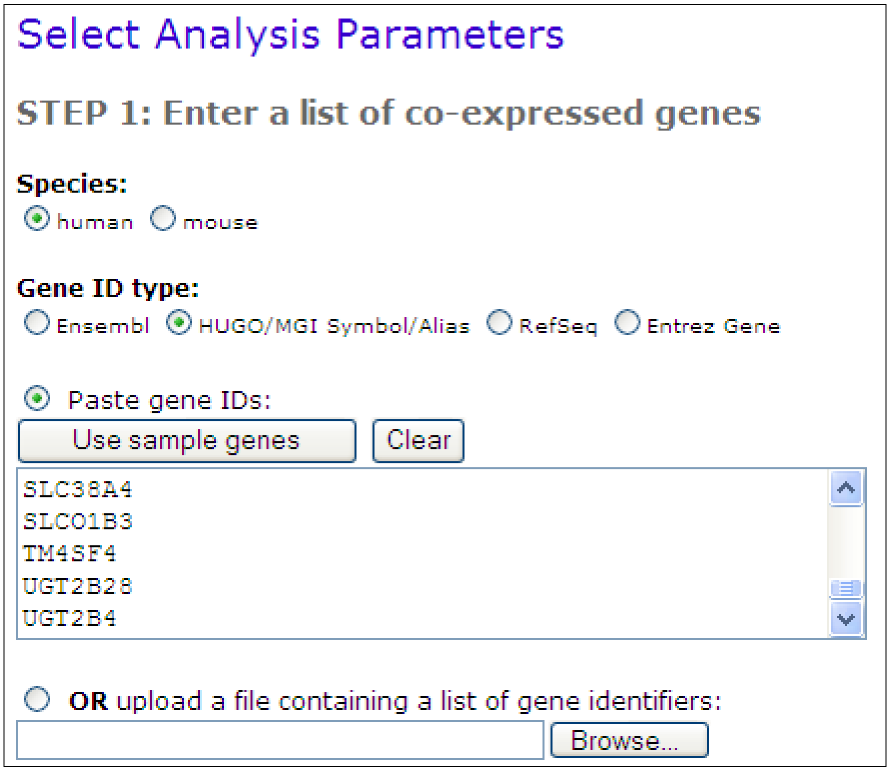
Reference: Huang, S., Fulton, D., *et al.* (2006). Identification of over-represented combinations of transcription factor binding sites in sets of co-expressed genes. *In* Advances in Bioinformatics and Computational Biology, Vol. 3. Imperial College Press, London, UK. 247-56. PDF.

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.



Select Analysis Parameters

STEP 1: Enter a list of co-expressed genes

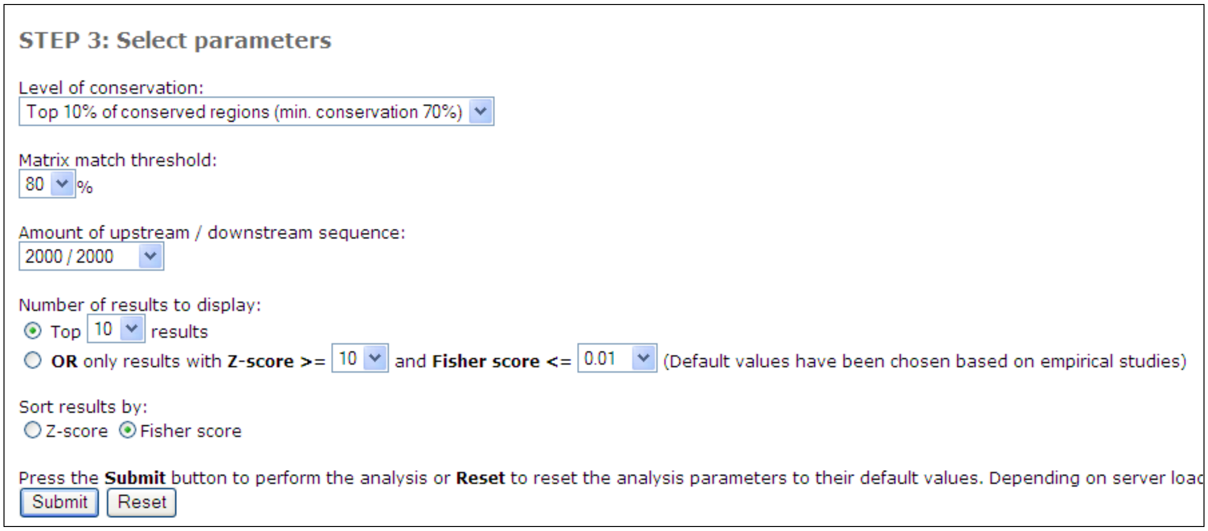
Species:
● human ○ mouse

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

● Paste gene IDs:

```
SLC38A4
SLCO1B3
TM4SF4
UGT2B28
UGT2B4
```

○ **OR** upload a file containing a list of gene identifiers:



STEP 3: Select parameters

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

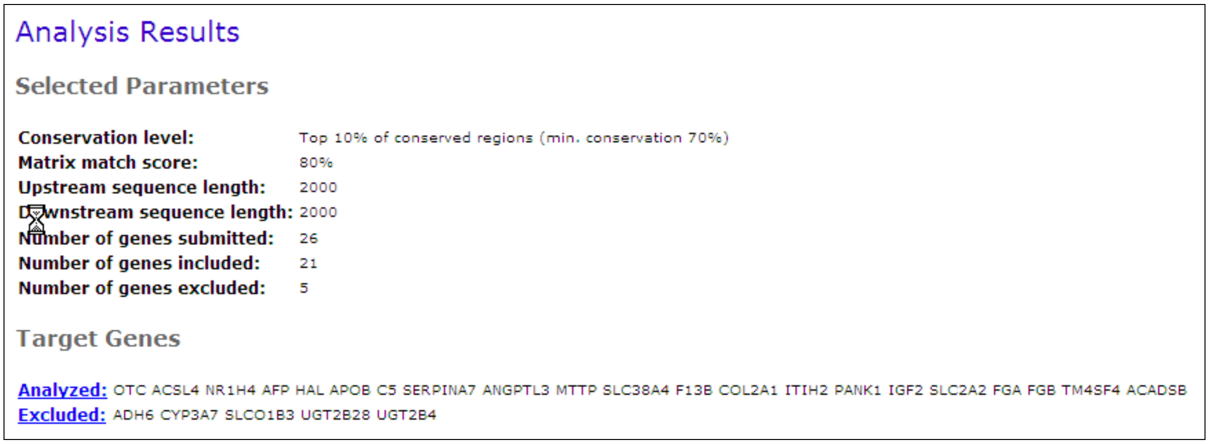
Matrix match threshold:
80 %

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 (Default values have been chosen based on empirical studies)

Sort results by:
○ Z-score ● Fisher score

Press the **Submit** button to perform the analysis or **Reset** to reset the analysis parameters to their default values. Depending on server load



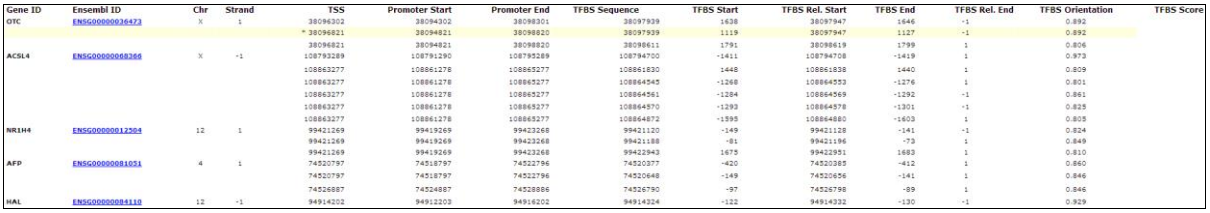
Analysis Results

Selected Parameters

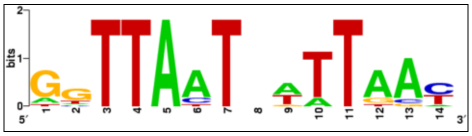
Conservation level:	Top 10% of conserved regions (min. conservation 70%)
Matrix match score:	80%
Upstream sequence length:	2000
Downstream sequence length:	2000
Number of genes submitted:	26
Number of genes included:	21
Number of genes excluded:	5

Target Genes

Analyzed: OTC ACSL4 NR1H4 AFP HAL APOB C5 SERPINA7 ANGPTL3 MTTP SLC38A4 F13B COL2A1 ITIH2 PANK1 IGF2 SLC2A2 FGA FGB TM4SF4 ACADSB
Excluded: ADH6 CYP3A7 SLCO1B3 UGT2B28 UGT2B4



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
OTC	ENSG00000036473	X	1	38096302	38094302	38098301	38097939	1638	38097947	1646	-1	0.892	
				* 38096821	38094821	38098820	38097939	1119	38097947	1127	-1	0.892	
				38096821	38094821	38098820	38098611	1791	38098619	1799	1	0.806	
ACSL4	ENSG00000068366	X	-1	108793289	108791290	108795289	108794700	-1411	108794708	-1419	1	0.973	
				108863277	108861278	108865277	108861830	1448	108861838	1440	1	0.809	
				108863277	108861278	108865277	108864545	-1268	108864553	-1276	1	0.801	
				108863277	108861278	108865277	108864561	-1284	108864569	-1292	-1	0.861	
				108863277	108861278	108865277	108864570	-1293	108864578	-1301	-1	0.825	
				108863277	108861278	108865277	108864872	-1595	108864880	-1603	1	0.805	
NR1H4	ENSG00000012504	12	1	99421269	99419269	99423268	99421120	-149	99421128	-141	-1	0.824	
				99421269	99419269	99423268	99421188	-81	99421196	-73	1	0.849	
				99421269	99419269	99423268	99422943	1675	99422951	1683	1	0.810	
AFP	ENSG00000081051	4	1	74520797	74518797	74522796	74520377	-420	74520385	-412	1	0.860	
				74520797	74518797	74522796	74520648	-149	74520656	-141	1	0.846	
				74526887	74524887	74528886	74526790	-97	74526798	-89	1	0.846	
HAL	ENSG00000084110	12	-1	94914202	94912203	94916202	94914324	-122	94914332	-130	-1	0.929	



Unknown/Novel

Non-conserved

MEME – Multiple Em for Motif Elicitation
URL: http://meme.sdsc.edu/meme/meme.html
Access: Free; Web-based; The web-based version has limit of number of base pairs (should not be more than 60,000 bp). Alternately, you can download the software and install it on your own computer. This will allow you to use many features that are not available with the interactive version.
Description/Utility: MEME is a tool for discovering motifs in a group of related DNA or protein sequences. A motif is defined as a sequence pattern that occurs repeatedly in a group of related protein or DNA sequences. MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs.

MEME takes as input a group of DNA or protein sequences (the training set) 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.
MEME sends you three e-mail messages:

- a confirmation message,

- the MEME results, and
- the MAST results of searching the training set for the motifs found by MEME using MAST.

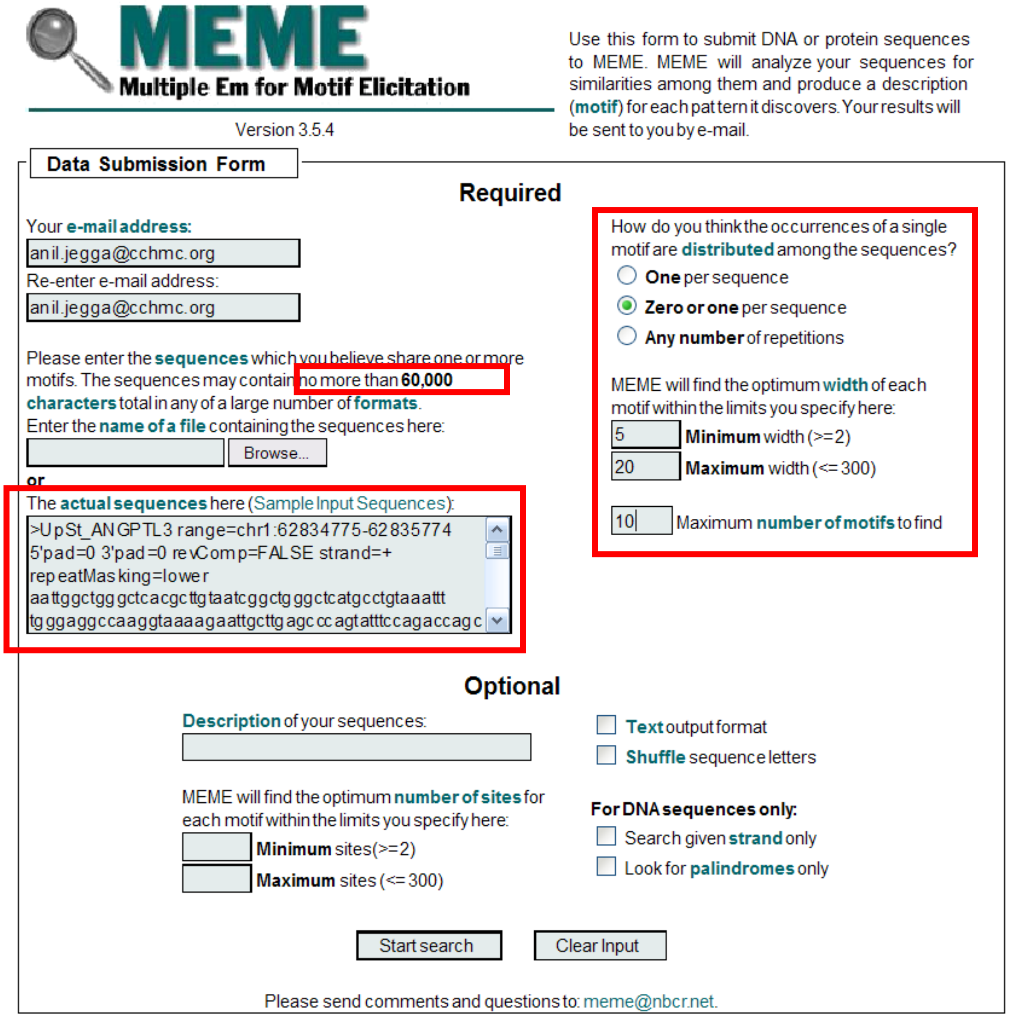
One of the features of MEME that come handy is it facilitates you to compare the identified motifs to known TFBSs. This will help you to find out potentially “real” novel motifs in the sequences of coexpressed genes.

Input: Multi-fasta files (not more than 60,000 bp)

Output: There is no good way of using/storing the graphical output (especially if you want to use it in a presentation or publication) except for taking a screen capture.

Drawbacks/Limitations:

- Larger sequences can take considerable amount of time (more than a day). Explore the mirror sites OR download the application and use it locally.
- Results can be difficult to interpret.
- Always try to mask the repeat elements otherwise your top scored motifs could be repeat elements!



Use this form to submit DNA or protein sequences to MEME. MEME will analyze your sequences for similarities among them and produce a description (**motif**) for each pattern it discovers. Your results will be sent to you by e-mail.

Data Submission Form

Required

Your **e-mail address:** anil.jegga@cchmc.org

Re-enter e-mail address: anil.jegga@cchmc.org

Please enter the **sequences** which you believe share one or more motifs. The sequences may contain no more than **60,000** **characters** total in any of a large number of **formats**.

Enter the **name of a file** containing the sequences here:

or

The **actual sequences** here (Sample Input Sequences):

```
>UpSt_ANGPTL3 range=chr1:62834775-62835774
5'pad=0 3'pad=0 revComp=FALSE strand=+
repeatMasking=lower
aattggctggctcacgcttgtaatcggctgggctcatgcctgtaaattt
tgggaggccaaggtaaaagaattgcttgagcccagtatttccagaccagc
```

How do you think the occurrences of a single motif are **distributed** among the sequences?

○ **One** per sequence
● **Zero or one** per sequence
○ **Any number** of repetitions

MEME will find the optimum **width** of each motif within the limits you specify here:

5 **Minimum** width (>=2)
20 **Maximum** width (<=300)

10 Maximum **number of motifs** to find

Optional

Description of your sequences:

MEME will find the optimum **number of sites** for each motif within the limits you specify here:

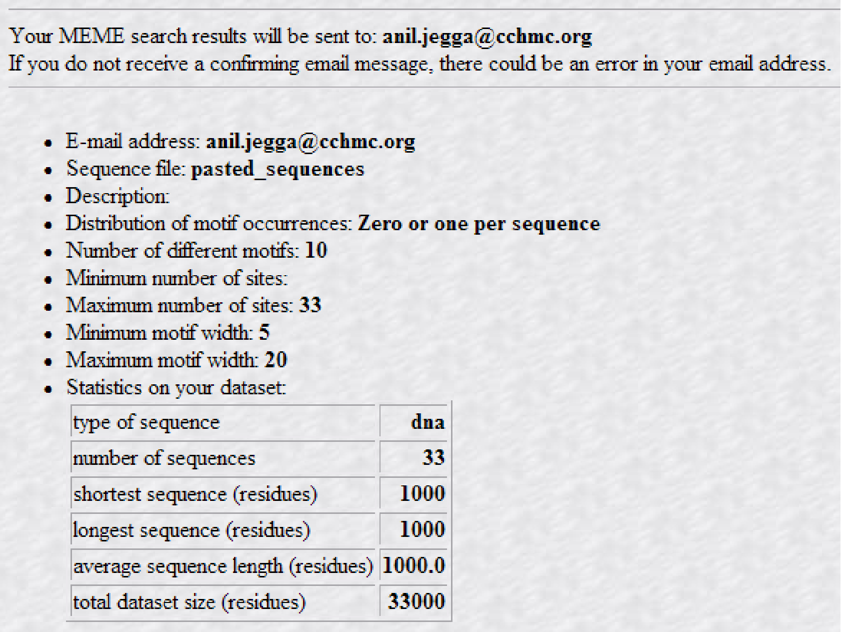
Minimum sites (>=2)
Maximum sites (<=300)

□ **Text** output format
□ **Shuffle** sequence letters

For DNA sequences only:

□ Search given **strand** only
□ Look for **palindromes** only

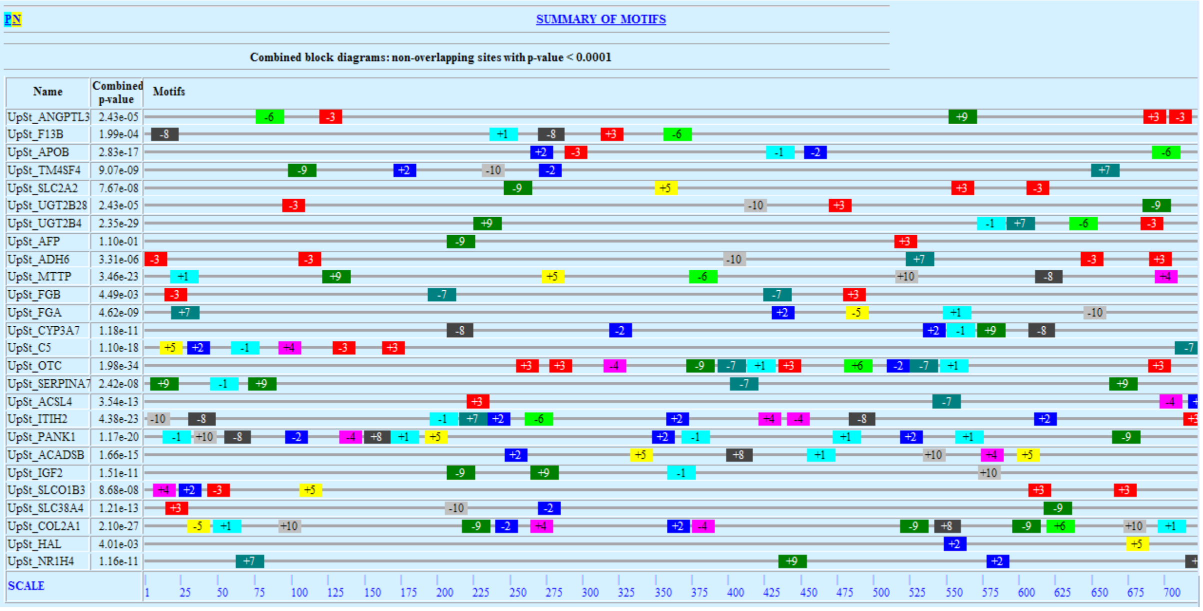
Please send comments and questions to: meme@nbcr.net.



Your MEME search results will be sent to: **anil.jegga@cchmc.org**
If you do not receive a confirming email message, there could be an error in your email address.

- E-mail address: **anil.jegga@cchmc.org**
- Sequence file: **pasted_sequences**
- Description:
- Distribution of motif occurrences: **Zero or one per sequence**
- Number of different motifs: **10**
- Minimum number of sites:
- Maximum number of sites: **33**
- Minimum motif width: **5**
- Maximum motif width: **20**
- Statistics on your dataset:

type of sequence	**dna**
number of sequences	**33**
shortest sequence (residues)	**1000**
longest sequence (residues)	**1000**
average sequence length (residues)	**1000.0**
total dataset size (residues)	**33000**



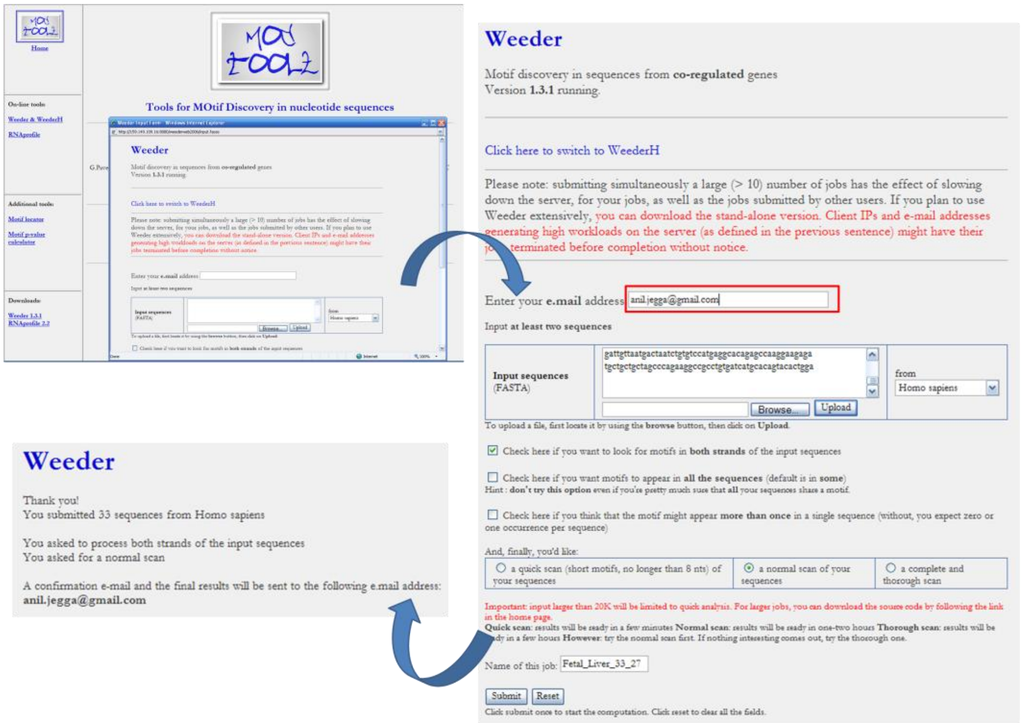
PN

SUMMARY OF MOTIFS

Combined block diagrams: non-overlapping sites with p-value < 0.0001

Name	Combined p-value	Motifs
UpSt_ANGPTL3	2.43e-05	
UpSt_F13B	1.99e-04	
UpSt_APOB	2.83e-17	
UpSt_TM4SF4	9.07e-09	
UpSt_SLC2A2	7.67e-08	
UpSt_UGT2B28	2.43e-05	
UpSt_UGT2B4	2.35e-29	
UpSt_AFP	1.10e-01	
UpSt_ADH6	3.31e-06	
UpSt_MTTP	3.46e-23	
UpSt_FGB	4.49e-03	
UpSt_FGA	4.62e-09	
UpSt_CYP3A7	1.18e-11	
UpSt_C5	1.10e-18	
UpSt_OTC	1.98e-34	
UpSt_SERPINA7	2.42e-08	
UpSt_ACSL4	3.54e-13	
UpSt_ITIH2	4.38e-23	
UpSt_PANK1	1.17e-20	
UpSt_ACADSB	1.66e-15	
UpSt_IGF2	1.51e-11	
UpSt_SLCO1B3	8.68e-08	
UpSt_SLC38A4	1.21e-13	
UpSt_COL2A1	2.10e-27	
UpSt_HAL	4.01e-03	
UpSt_NR1H4	1.16e-11	
SCALE		

Weeder
Similar to MEME but comparative studies have shown it be better than MEME (speed, output, results, robustness)
Drawbacks/Limitations: Results are sent in the mail message. Some email clients (e.g. Groupwise mail) truncate longer messages and therefore you may not see complete results (especially if you are using several sequences). I use Gmail and that works fine.



Weeder

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

Click here to switch to WeederH

Please note: submitting simultaneously a large (> 10) number of jobs has the effect of slowing down the server, for your jobs, as well as the jobs submitted by other users. If you plan to use Weeder extensively, you can download the stand-alone version. Client IPs and e-mail addresses generating high workloads on the server (as defined in the previous sentence) might have their jobs terminated before completion without notice.

Enter your **e.mail** address anil.jegga@gmail.com

Input **at least two sequences**

Input sequences (FASTA)
gattgttaatgactaatctgtgtccatgaggcacagagccaaggaagaga
tgctgctgctagcccagaaggccgcctgtgatcatgcacagtacactgga

from Homo sapiens

Browse... Upload

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

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

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

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

And, finally, you'd like:
○ a quick scan (short motifs, no longer than 8 nts) of your sequences ● a normal scan of your sequences ○ a complete and thorough scan

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

Name of this job: Fetal_Liver_33_27

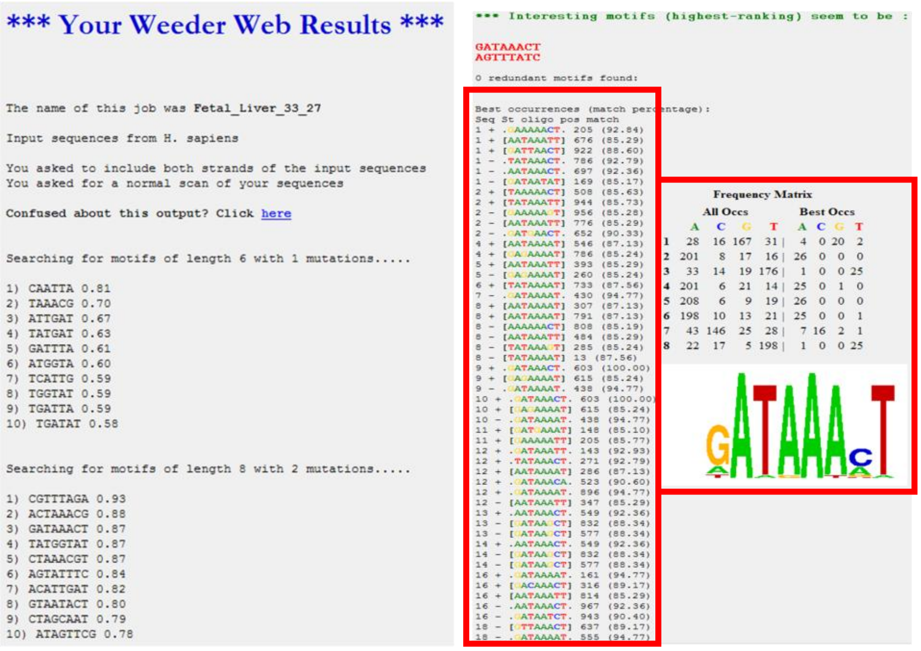
Submit Reset
Click submit once to start the computation. Click reset to clear all the fields.

Weeder

Thank you!
You submitted 33 sequences from Homo sapiens

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

A confirmation e-mail and the final results will be sent to the following e.mail address:
anil.jegga@gmail.com



*** 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)
```

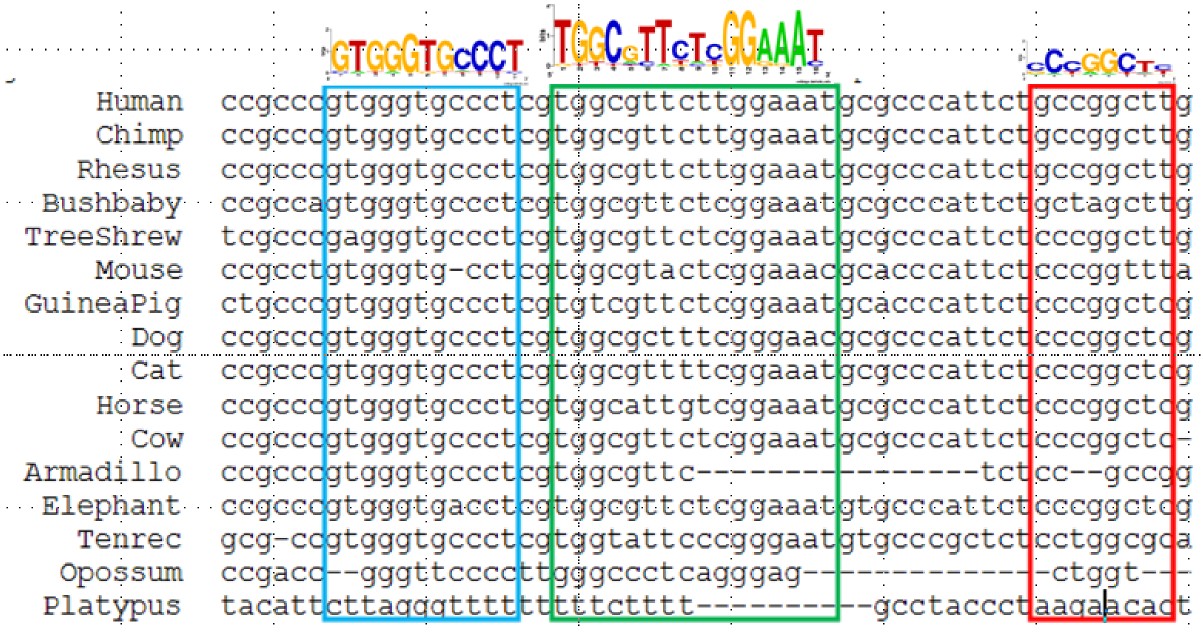
Frequency Matrix

	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

Conserved

Currently, there are no publicly available tools that can look for conserved motifs across multiple sequences. You can use MEME or Weeder by using orthologous promoters. But this can sometimes lead to erroneous results (for e.g. the high scoring motifs can be the same ones just coming from ortholog copy of the same gene).

In case anyone needs assistance in this type of analysis (for e.g. a genome-wide scan for two or multispecies conserved motifs), we can assist (the tool/application we are working on is in development stage).



```
    Human  ccgcccgtgggtgccctcgtggcgttcttggaaatgcgcccattctgccggcttg
    Chimp  ccgcccgtgggtgccctcgtggcgttcttggaaatgcgcccattctgccggcttg
   Rhesus  ccgcccgtgggtgccctcgtggcgttcttggaaatgcgcccattctgccggcttg
 Bushbaby  ccgccagtgggtgccctcgtggcgttctcggaaatgcgcccattctgctagcttg
TreeShrew  tcgcccgagggtgccctcgtggcgttctcggaaatgcgcccattctcccggcttg
    Mouse  ccgcctgtgggtg-cctcgtggcgtactcggaaacgcacccattctcccggttta
GuineaPig  ctgcccgtgggtgccctcgtgtcgttctcggaaatgcacccattctcccggctcg
      Dog  ccgcccgtgggtgccctcgtggcgctttcgggaacgcgcccattctcccggctcg
      Cat  ccgcccgtgggtgccctcgtggcgttttcggaaatgcgcccattctcccggctcg
    Horse  ccgcccgtgggtgccctcgtggcattgtcggaaatgcgcccattctcccggctcg
      Cow  ccgcccgtgggtgccctcgtggcgttctcggaaatgcgcccattctcccggctc-
Armadillo  ccgcccgtgggtgccctcgtggcgttc----------------tctcc--gccgg
 Elephant  ccgcccgtgggtgacctcgtggcgttctcggaaatgtgcccattctcccggctcg
   Tenrec  gcg-ccgtgggtgccctcgtggtattcccgggaatgtgcccgctctcctggcgca
  Opossum  ccgacc--gggttcccctttgggccctcagggag-------------ctggt---
 Platypus  tacattcttagggtttttttttctttt-----------gcctaccctaagaacact
```

Chapter 3: Functional Enrichment Analysis of the Transcriptome

Objectives

1. What is the functional enrichment for a given set of genes
    a. Gene Ontology (Biological Process, Molecular Function and Cellular Component)
    b. Pathways (Kegg, Biocarta, etc.)
    c. Protein Domains (Interpro, PFAM, etc.)
    d. Functional Keywords (SwissProt keywords – genes associated with controlled vocabulary terms e.g. all genes associated with the word "apoptosis")
2. What is the expression pattern of these genes in other conditions (or how do my genes score in other microarray expression experiments)
3. How to prioritize/rank the genes in my gene list so that I can select a handful for further experimental validation

Introduction

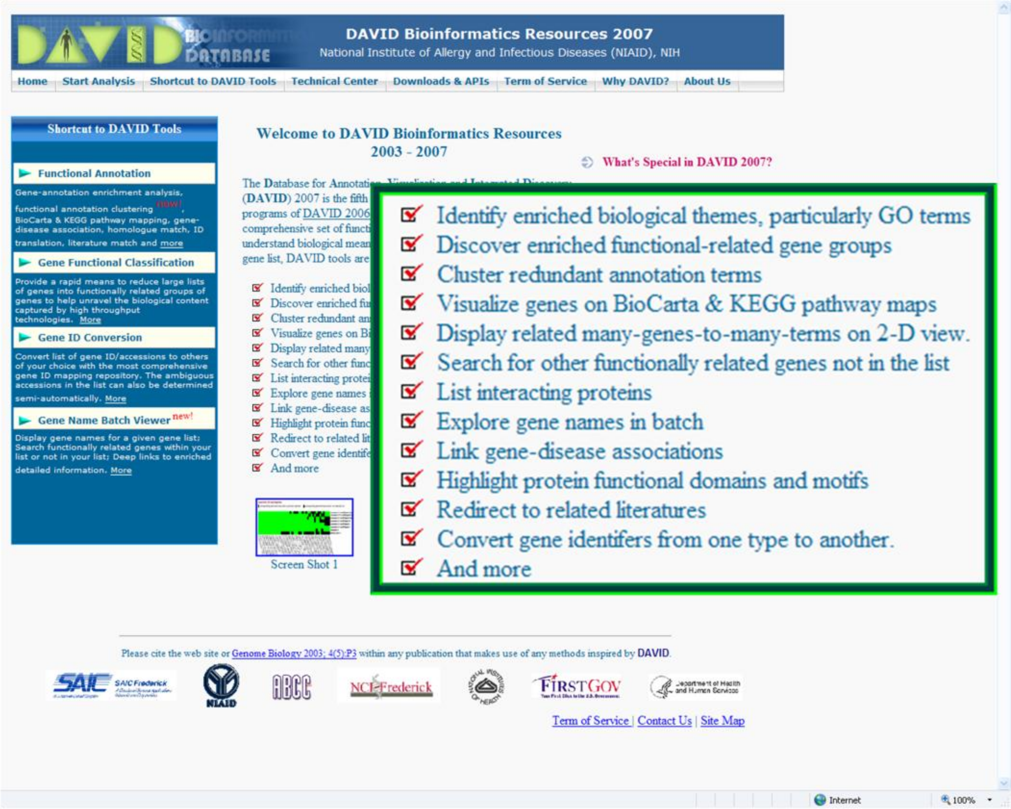
The pathway, ontology data sources and analysis tools establish a basis for finding links between lists of genes in their associated biological network context. Several tools/servers are available that effectively utilize these various resources and help in understanding of the unifying biological themes underlying one's data. Here I list some of the "popular" and useful ones. **My favorites** are:

1. For single gene list enrichment analysis: DAVID, MSigDB, FatiGO+ and Panther (not necessarily in that order)
2. For comparison:
    a. Two gene lists: FatiGO+ (extensive parameter coverage including miRNAs!)
    b. Multiple gene lists: Panther (but the parameters are limited)
3. For comparison of a gene list with other published gene sets based on different experiments: L2L and MSigDB (for cancer related datasets Oncomine is probably the best)
4. For both enrichment analysis and also candidate gene prioritization: ToppGene. It's probably the only one database currently that takes into account literature co-citations and also mouse phenotype data for functional enrichment analysis.

Tools and Servers

DAVID

Description: **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (DAVID) provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes.
URL: http://david.abcc.ncifcrf.gov/



BIOINFORMATICS DATABASE
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 functi
understand biological mean
gene list, DAVID tools are

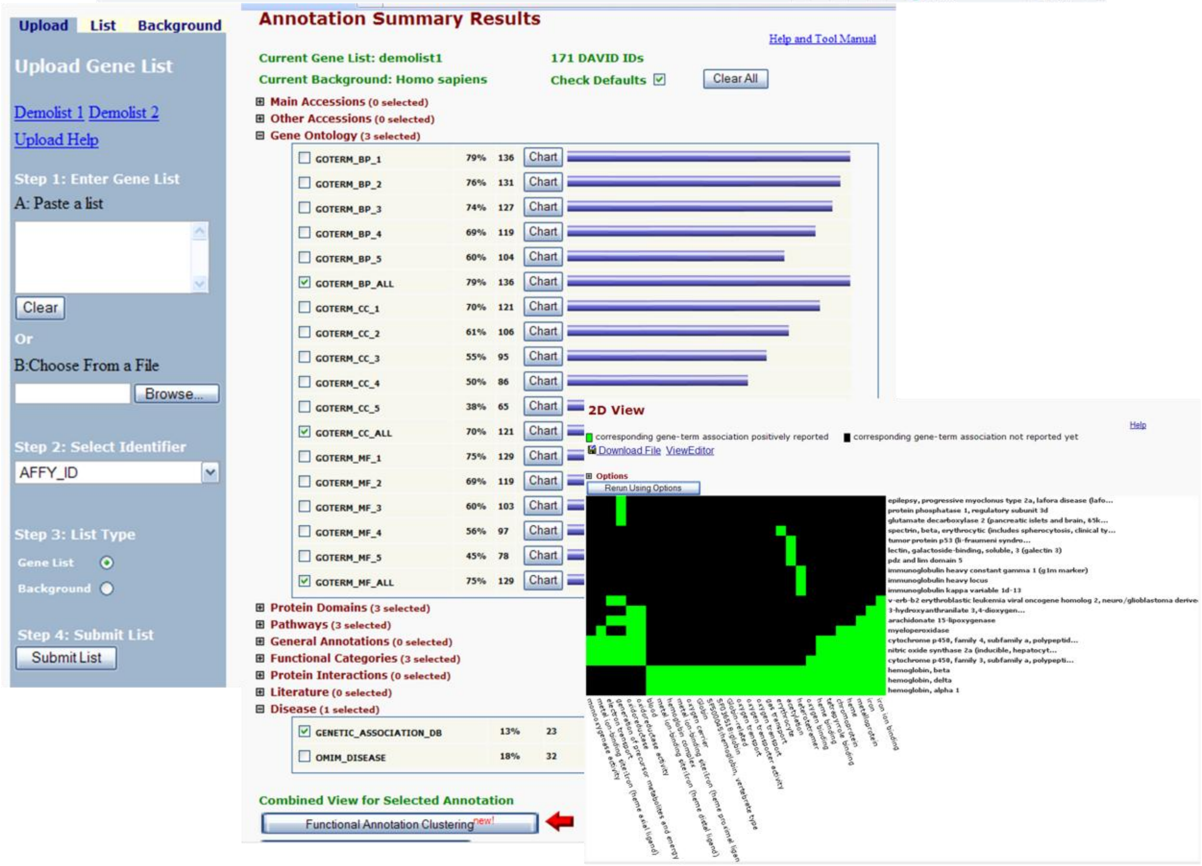
- ☑ Identify enriched biological themes, particularly GO terms
- ☑ Discover enriched functional-related gene groups
- ☑ Cluster redundant annotation terms
- ☑ Visualize genes on BioCarta & KEGG pathway maps
- ☑ Display related many-genes-to-many-terms on 2-D view.
- ☑ Search for other functionally related 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

Screen Shot 1

Please cite the web site or Genome Biology 2003; 4(5):P3 within any publication that makes use of any methods inspired by DAVID.

SAIC SAIC Frederick
NIAID
ABCC
NCI-Frederick
National Institutes of Health
FIRSTGOV
Department of Health and Human Services

Term of Service | Contact Us | Site Map



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

epilepsy, progressive myoclonus type 2a, lafora disease (lafo...
protein phosphatase 1, regulatory subunit 3d
glutamate decarboxylase 2 (pancreatic islets and brain, 65k...
spectrin, beta, erythrocytic (includes spherocytosis, clinical ty...
tumor protein p53 (li-fraumeni syndro...
lectin, galactoside-binding, soluble, 3 (galectin 3)
pdz and lim domain 5
immunoglobulin heavy constant gamma 1 (g1m marker)
immunoglobulin heavy locus
immunoglobulin kappa variable 1d-13
v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derive
3-hydroxyanthranilate 3,4-dioxygen...
arachidonate 15-lipoxygenase
myeloperoxidase
cytochrome p450, family 4, subfamily a, polypeptid...
nitric oxide synthase 2a (inducible, hepatocyt...
cytochrome p450, family 3, subfamily a, polypepti...
hemoglobin, beta
hemoglobin, delta
hemoglobin, alpha 1
monooxygenase activity
metal ion-binding site:iron (heme axial ligand)
electron transport
generation of precursor metabolites and energy
oxidoreductase
oxidoreductase activity
blood
metal ion-binding site:iron (heme distal ligand)
hemoglobin complex
metal ion-binding site:iron (heme proximal ligan
oxygen carrier
Globin
SF00045:hemoglobin, vertebrate type
SF036518:globin
Globin-related
oxygen transport
oxygen transporter activity
gas transport
erythrocyte
acetylation
heterotetramer
oxygen binding
heme binding
tetrapyrrole binding
chromoprotein
heme
metalloprotein
iron
iron ion binding

MSigDB (Molecular Signatures Database)

Description: The “annotation” features helps you to compute overlaps between your gene set and other gene sets in MSigDB. Additionally, you can also categorize members of the gene set by gene families and display the gene set expression profile based on a selected compendium of expression profiles. The analysis results include:

Statistics:

- overlaps shown lists the number of overlapping gene sets displayed in the report: By default, the report displays the 10 gene sets in the collection that best overlap with your gene set. If you compute overlaps from the Annotations page, you can choose the number of overlapping gene sets to display in the report.
- gene sets in collection lists the total number of gene sets being analyzed
- genes in comparison lists the number of genes in your gene set
- genes in collection lists the number of unique genes in the gene sets being analyzed

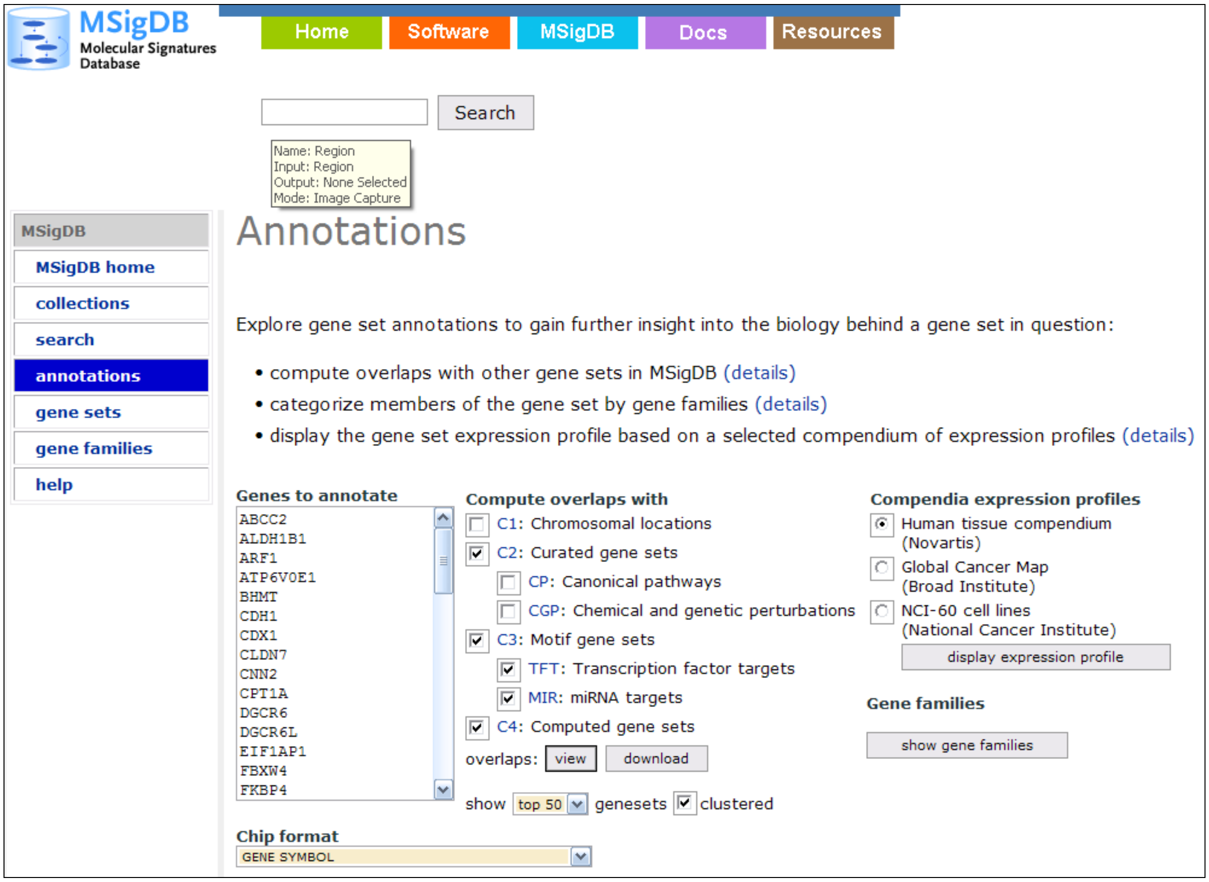
Descriptions of the overlapping gene sets, including

- Link to the gene set card
- Number of genes in the gene set
- Description of the gene set
- Number of genes in the overlap between this gene set and your gene set
- *p value* indicating the significance of the overlap
- Color bar shading from light green to black, where **lighter colors indicate more significant p values (< 0.05) and black indicates less significant p values (≥ 0.05)**.

Overlap matrix showing the genes in the overlapping gene sets

- Rows list the genes in your gene set, with gene descriptions and links to gene annotations
- Columns list the overlapping gene sets, with links to the gene set cards
- Overlaps are computed using HUGO gene symbols. In rare instances, a gene set may contain a gene symbol that is not in the GENE_SYMBOL chip annotation file. Such gene symbols are ignored when overlaps are computed and appear crossed out in the matrix. If a gene set has a source platform other than GENE_SYMBOL, each gene symbol in the gene set is translated to its probe identifier(s) on the source platform. The matrix lists the probe identifier(s) in parentheses following the gene symbol. If a gene symbol cannot be translated, a question mark (?) appears in place of the probe identifier(s).

URL: http://www.broad.mit.edu/gsea/msigdb/annotate.jsp
Access: Free for academics (need to register)
Others: Explore the GSEA (Gene Set Enrichment Analysis) and also Gene Pattern



Annotations

Explore gene set annotations to gain further insight into the biology behind a gene set in question:

- compute overlaps with other gene sets in MSigDB (details)
- categorize members of the gene set by gene families (details)
- display the gene set expression profile based on a selected compendium of expression profiles (details)

Genes to annotate

```
ABCC2
ALDH1B1
ARF1
ATP6V0E1
BHMT
CDH1
CDX1
CLDN7
CNN2
CPT1A
DGCR6
DGCR6L
EIF1AP1
FBXW4
FKBP4
```

Compute overlaps with

- □ C1: Chromosomal locations
- ☑ C2: Curated gene sets
  - □ CP: Canonical pathways
  - □ CGP: Chemical and genetic perturbations
- ☑ C3: Motif gene sets
  - ☑ TFT: Transcription factor targets
  - ☑ MIR: miRNA targets
- ☑ C4: Computed gene sets

overlaps: view download

show top 50 genesets ☑ clustered

Compendia expression profiles

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

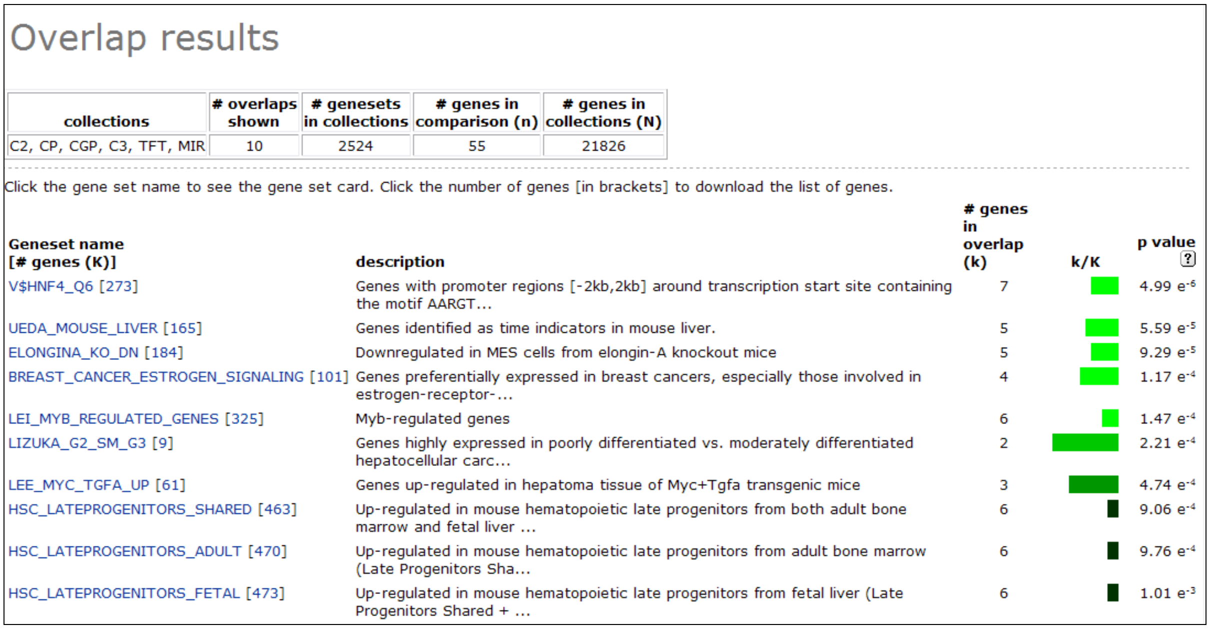
display expression profile

Gene families

show gene families

Chip format

GENE SYMBOL

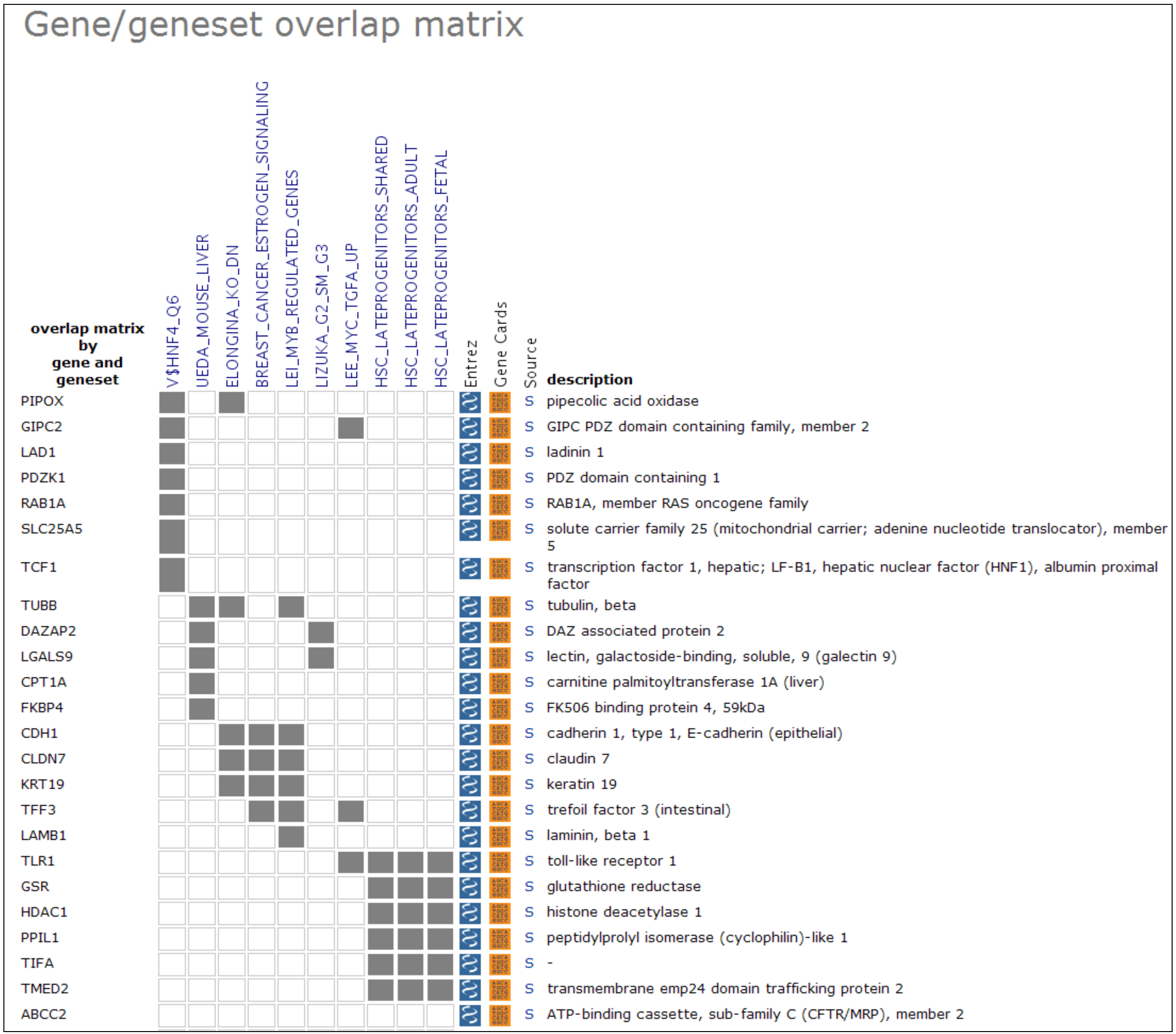


Overlap results

collections	# overlaps shown	# genesets in collections	# genes in comparison (n)	# genes in collections (N)
C2, CP, CGP, C3, TFT, MIR	10	2524	55	21826

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

Geneset name [# genes (K)]	description	# genes in overlap (k)	k/K	p value ?
V$HNF4_Q6 [273]	Genes with promoter regions [-2kb,2kb] around transcription start site containing the motif AARGT...	7	🟩	4.99 e^{-6}
UEDA_MOUSE_LIVER [165]	Genes identified as time indicators in mouse liver.	5	🟩	5.59 e^{-5}
ELONGINA_KO_DN [184]	Downregulated in MES cells from elongin-A knockout mice	5	🟩	9.29 e^{-5}
BREAST_CANCER_ESTROGEN_SIGNALING [101]	Genes preferentially expressed in breast cancers, especially those involved in estrogen-receptor-...	4	🟩	1.17 e^{-4}
LEI_MYB_REGULATED_GENES [325]	Myb-regulated genes	6	🟩	1.47 e^{-4}
LIZUKA_G2_SM_G3 [9]	Genes highly expressed in poorly differentiated vs. moderately differentiated hepatocellular carc...	2	🟩	2.21 e^{-4}
LEE_MYC_TGFA_UP [61]	Genes up-regulated in hepatoma tissue of Myc+Tgfa transgenic mice	3	🟩	4.74 e^{-4}
HSC_LATEPROGENITORS_SHARED [463]	Up-regulated in mouse hematopoietic late progenitors from both adult bone marrow and fetal liver ...	6	🟩	9.06 e^{-4}
HSC_LATEPROGENITORS_ADULT [470]	Up-regulated in mouse hematopoietic late progenitors from adult bone marrow (Late Progenitors Sha...	6	🟩	9.76 e^{-4}
HSC_LATEPROGENITORS_FETAL [473]	Up-regulated in mouse hematopoietic late progenitors from fetal liver (Late Progenitors Shared + ...	6	🟩	1.01 e^{-3}

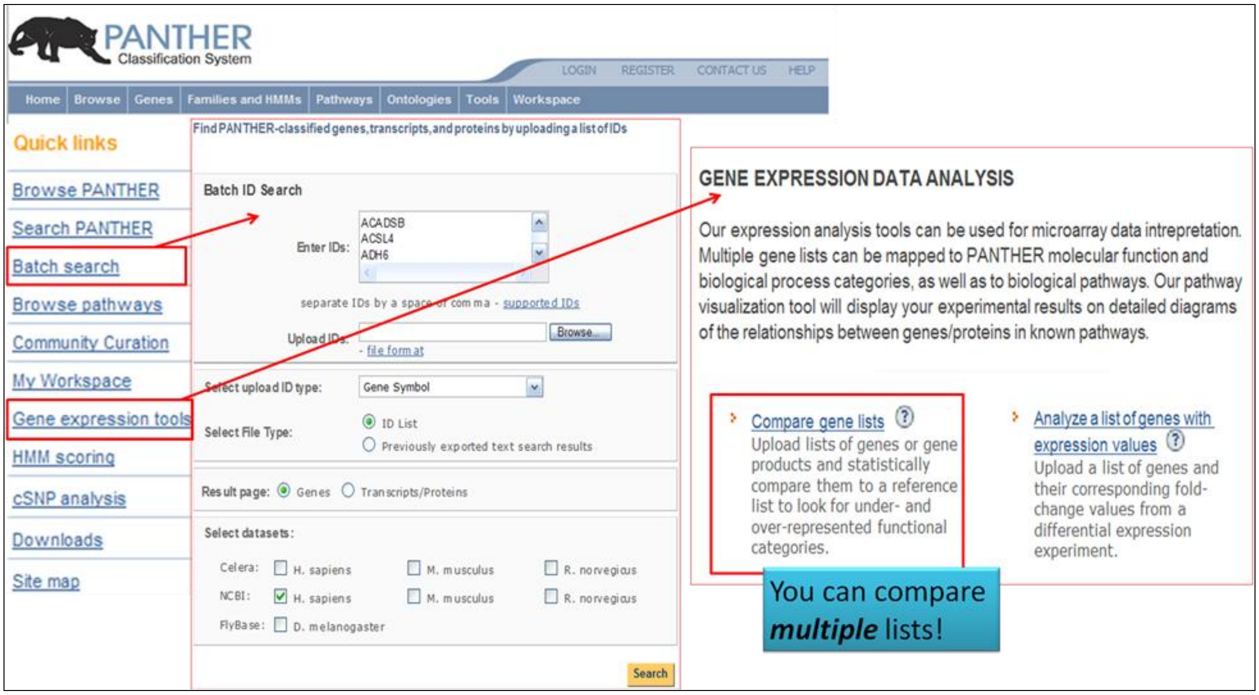


Gene/geneset overlap matrix

overlap matrix by gene and geneset	V$HNF4_Q6	UEDA_MOUSE_LIVER	ELONGINA_KO_DN	BREAST_CANCER_ESTROGEN_SIGNALING	LEI_MYB_REGULATED_GENES	LIZUKA_G2_SM_G3	LEE_MYC_TGFA_UP	HSC_LATEPROGENITORS_SHARED	HSC_LATEPROGENITORS_ADULT	HSC_LATEPROGENITORS_FETAL	Entrez	Gene Cards	Source	description
PIPOX	■		■										S	pipecolic acid oxidase
GIPC2	■						■						S	GIPC PDZ domain containing family, member 2
LAD1	■												S	ladinin 1
PDZK1	■												S	PDZ domain containing 1
RAB1A	■												S	RAB1A, member RAS oncogene family
SLC25A5	■												S	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5
TCF1	■												S	transcription factor 1, hepatic; LF-B1, hepatic nuclear factor (HNF1), albumin proximal factor
TUBB		■	■		■								S	tubulin, beta
DAZAP2		■				■							S	DAZ associated protein 2
LGALS9		■				■							S	lectin, galactoside-binding, soluble, 9 (galectin 9)
CPT1A		■											S	carnitine palmitoyltransferase 1A (liver)
FKBP4		■											S	FK506 binding protein 4, 59kDa
CDH1			■	■	■								S	cadherin 1, type 1, E-cadherin (epithelial)
CLDN7			■	■	■								S	claudin 7
KRT19			■	■	■								S	keratin 19
TFF3				■	■		■						S	trefoil factor 3 (intestinal)
LAMB1					■								S	laminin, beta 1
TLR1							■	■	■	■			S	toll-like receptor 1
GSR								■	■	■			S	glutathione reductase
HDAC1								■	■	■			S	histone deacetylase 1
PPIL1								■	■	■			S	peptidylprolyl isomerase (cyclophilin)-like 1
TIFA								■	■	■			S	-
TMED2								■	■	■			S	transmembrane emp24 domain trafficking protein 2
ABCC2													S	ATP-binding cassette, sub-family C (CFTR/MRP), member 2

Panther

Description: The PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System is a unique resource that classifies genes by their functions, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence. Proteins are classified by expert biologists into families and subfamilies of shared function, which are then categorized by molecular function and biological process ontology terms. For an increasing number of proteins, detailed biochemical interactions in canonical pathways are captured and can be viewed interactively.
URL: http://www.pantherdb.org
Access: Free web-based. By registering, you can store you gene lists and results (“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

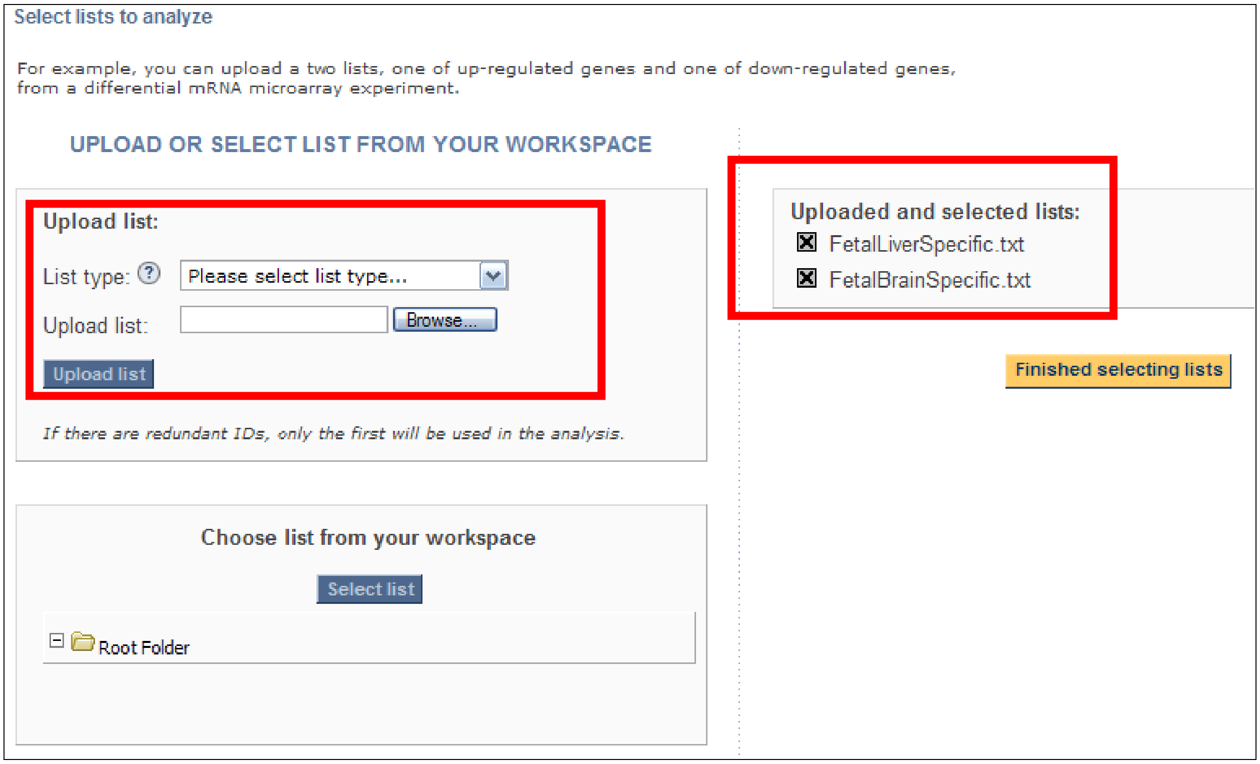
Search

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!



Select lists to analyze

For example, you can upload a two lists, one of up-regulated genes and one of down-regulated genes, from a differential mRNA microarray experiment.

UPLOAD OR SELECT LIST FROM YOUR WORKSPACE

Upload list:

List type: ? Please select list type...

Upload list: Browse...

Upload list

If there are redundant IDs, only the first will be used in the analysis.

Choose list from your workspace

Select list

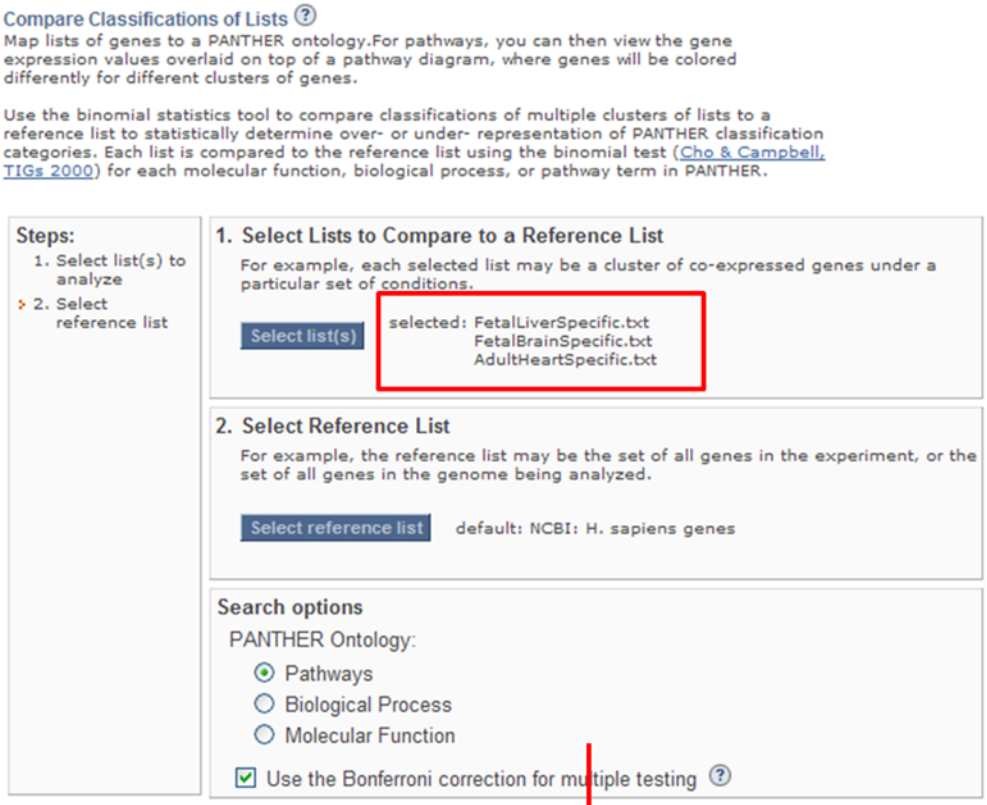
⊟ Root Folder

Uploaded and selected lists:

☒ FetalLiverSpecific.txt

☒ FetalBrainSpecific.txt

Finished selecting 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

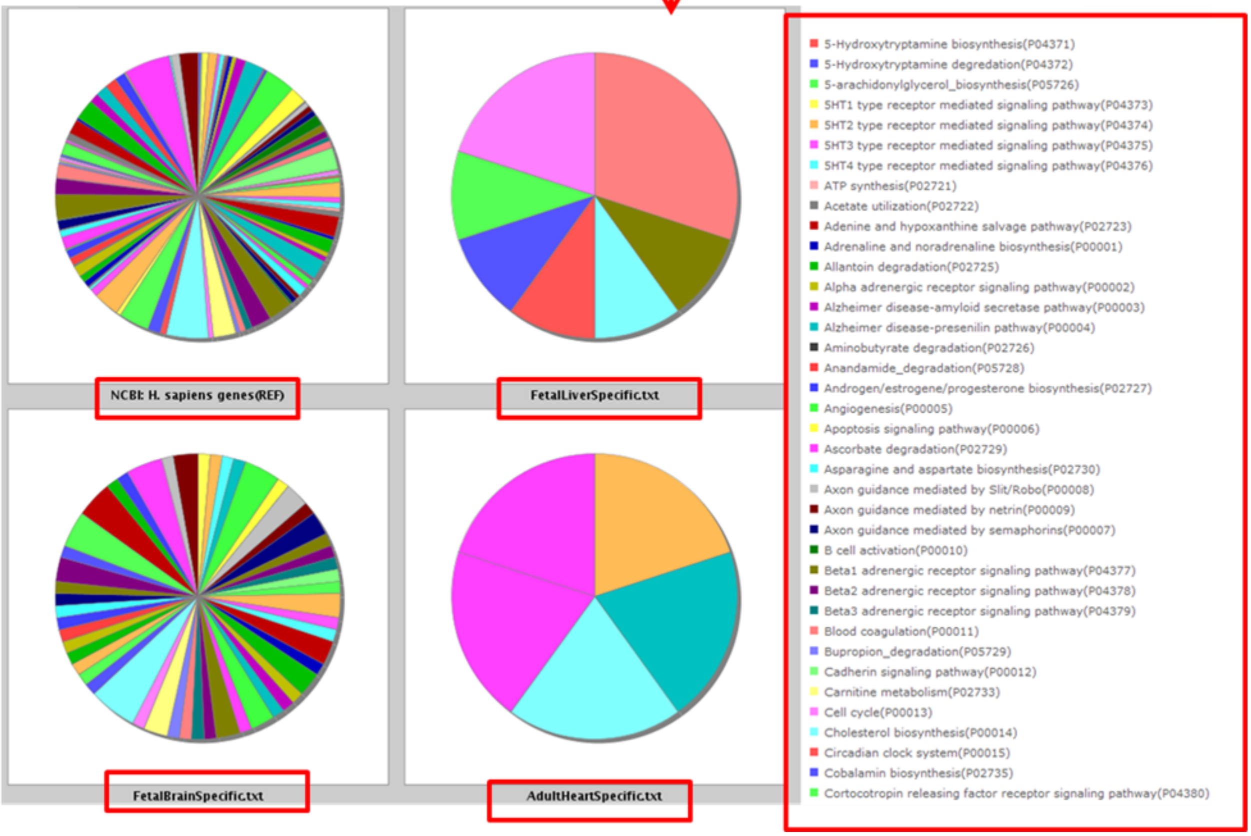
default: NCBI: H. sapiens genes

Search options

PANTHER Ontology:

● Pathways
○ Biological Process
○ Molecular Function

☑ Use the Bonferroni correction for multiple testing ?

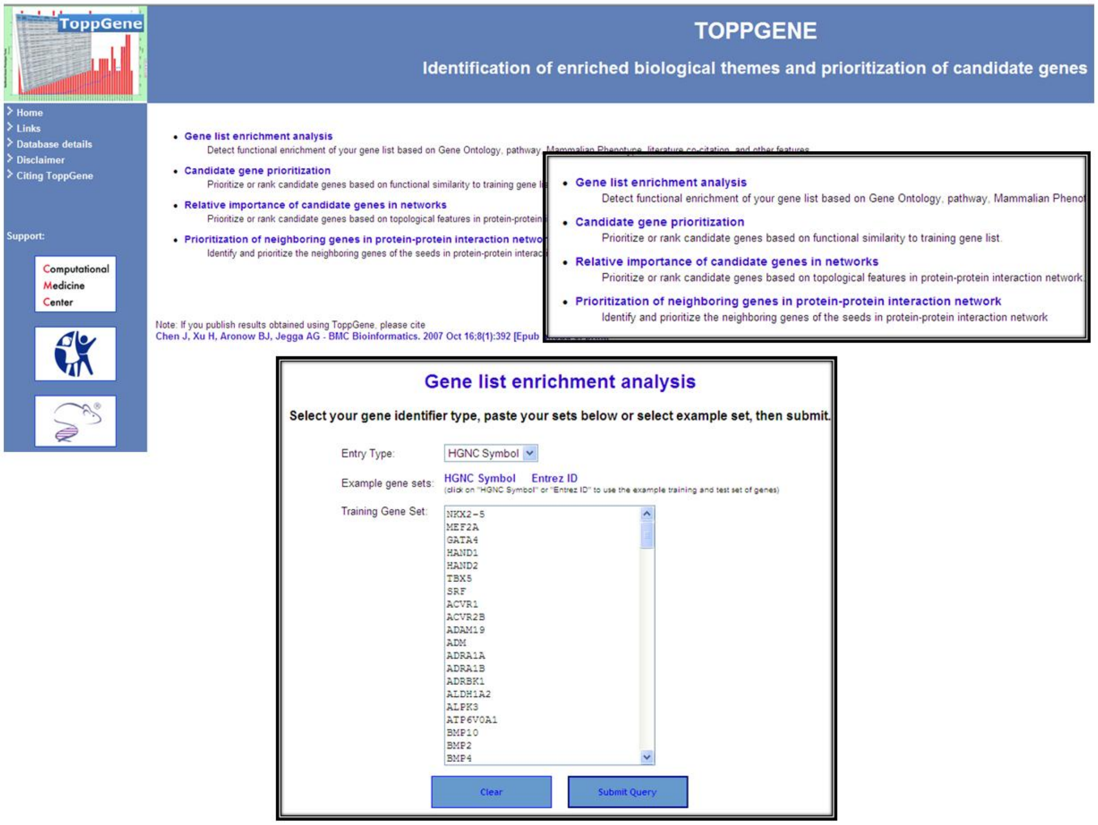


L2L

Description: L2L is a database of published microarray gene expression data, and a software tool for comparing the published data to a user's own microarray results.
URL: http://depts.washington.edu/l2l/

ToppGene

Description: The majority of common diseases are multi-factorial and modified by genetically and mechanistically complex polygenic interactions and environmental factors. High-throughput genome-wide studies like linkage analysis and gene expression profiling, tend to be most useful for classification and characterization but do not provide sufficient information to identify or prioritize specific disease causal genes. Hypothesizing that the majority of genes that impact or cause disease share membership in any of several functional relationships ToppGene integrates several data sources for disease candidate gene prioritization. ToppGene for the first times uses mouse phenotype data as one of the features for gene prioritization and we have observed that using mouse phenotype data greatly improves the human disease candidate gene analysis and prioritization
URL: http://toppgene.cchmc.org
Access: Free, web-based
Utility: Can be used for both functional enrichment in gene lists and also to prioritize candidate genes.
Input: Human gene symbols or gene IDs (NCBI's Entrez gene IDs)
Output: Graphical and results are downloadable as tab-delimited text files.
Limitations: Currently works only for human genes.


ToppGene
TOPPGENE
Identification of enriched biological themes and prioritization of candidate genes
Home
Links
Database details
Disclaimer
Citing ToppGene
Support:
Computational Medicine Center

- **Gene list enrichment analysis**
  Detect functional enrichment of your gene list based on Gene Ontology, pathway, Mammalian Phenotype, literature co-citation, and other features
- **Candidate gene prioritization**
  Prioritize or rank candidate genes based on functional similarity to training gene list.
- **Relative importance of candidate genes in networks**
  Prioritize or rank candidate genes based on topological features in protein-protein interaction network
- **Prioritization of neighboring genes in protein-protein interaction network**
  Identify and prioritize the neighboring genes of the seeds in protein-protein interaction network

Note: If you publish results obtained using ToppGene, please cite
Chen J, Xu H, Aronow BJ, Jegga AG - BMC Bioinformatics. 2007 Oct 16;8(1):392 [Epub

Gene list enrichment analysis

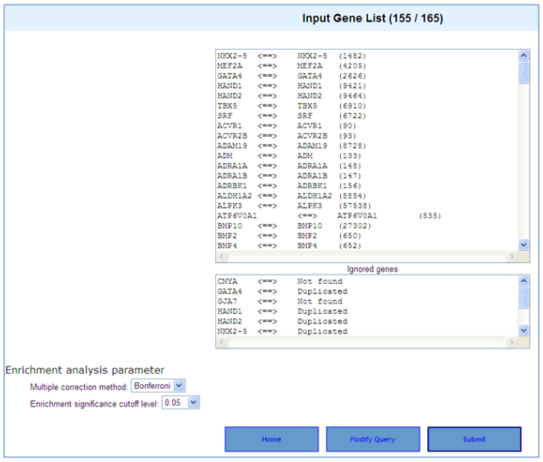
Select your gene identifier type, paste your sets below or select example set, then submit.

Entry Type: HGNC Symbol

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:

```
NKX2-5
MEF2A
GATA4
HAND1
HAND2
TBX5
SRF
ACVR1
ACVR2B
ADAM19
ADM
ADRA1A
ADRA1B
ADRBK1
ALDH1A2
ALPK3
ATP6V0A1
BMP10
BMP2
BMP4
```



Input Gene List (155 / 165)

```
NKX2-5   <==>   NKX2-5   (1482)
MEF2A    <==>   MEF2A    (4205)
GATA4    <==>   GATA4    (2626)
HAND1    <==>   HAND1    (9421)
HAND2    <==>   HAND2    (9464)
TBX5     <==>   TBX5     (6910)
SRF      <==>   SRF      (6722)
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)
```

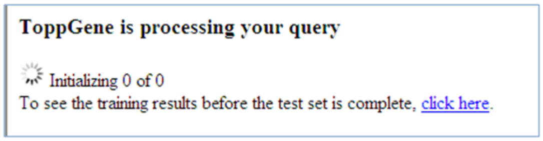
Ignored genes

```
CMYA     <==>   Not found
GATA4    <==>   Duplicated
GJA7     <==>   Not found
HAND1    <==>   Duplicated
HAND2    <==>   Duplicated
NKX2-5   <==>   Duplicated
```

Enrichment analysis parameter

Multiple correction method: Bonferroni

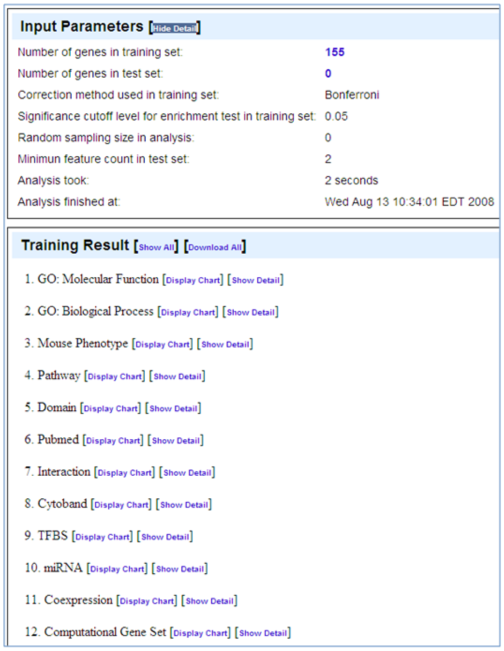
Enrichment significance cutoff level: 0.05



ToppGene is processing your query

Initializing 0 of 0

To see the training results before the test set is complete, click here.

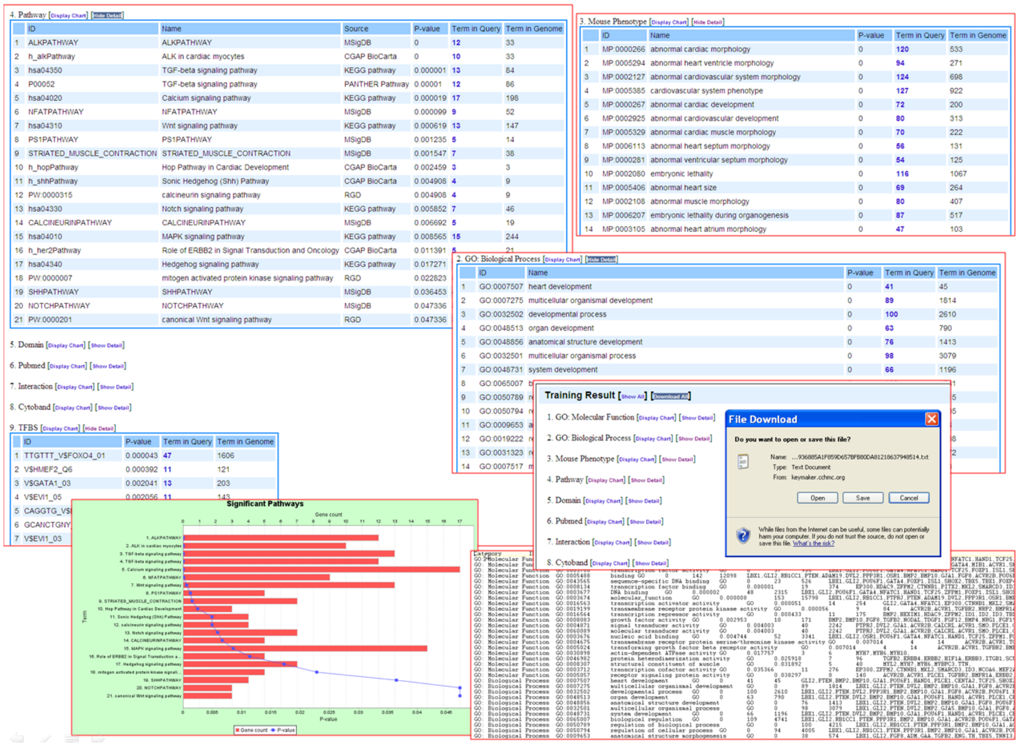


Input Parameters [Hide Detail]

Number of genes in training set:	155
Number of genes in test set:	0
Correction method used in training set:	Bonferroni
Significance cutoff level for enrichment test in training set:	0.05
Random sampling size in analysis:	0
Minimun feature count in test set:	2
Analysis took:	2 seconds
Analysis finished at:	Wed Aug 13 10:34:01 EDT 2008

Training Result [Show All] [Download All]

1. GO: Molecular Function [Display Chart] [Show Detail]
2. GO: Biological Process [Display Chart] [Show Detail]
3. Mouse Phenotype [Display Chart] [Show Detail]
4. Pathway [Display Chart] [Show Detail]
5. Domain [Display Chart] [Show Detail]
6. Pubmed [Display Chart] [Show Detail]
7. Interaction [Display Chart] [Show Detail]
8. Cytoband [Display Chart] [Show Detail]
9. TFBS [Display Chart] [Show Detail]
10. miRNA [Display Chart] [Show Detail]
11. Coexpression [Display Chart] [Show Detail]
12. Computational Gene Set [Display Chart] [Show Detail]



4. Pathway [Display Chart] [Hide Detail]

	ID	Name	Source	P-value	Term in Query	Term in Genome
1	ALKPATHWAY	ALKPATHWAY	MSigDB	0	12	33
2	h_alkPathway	ALK in cardiac myocytes	CGAP BioCarta	0	10	33
3	hsa04350	TGF-beta signaling pathway	KEGG pathway	0.000001	13	84
4	P00052	TGF-beta signaling pathway	PANTHER Pathway	0.00001	12	86
5	hsa04020	Calcium signaling pathway	KEGG pathway	0.000019	17	198
6	NFATPATHWAY	NFATPATHWAY	MSigDB	0.000099	9	52
7	hsa04310	Wnt signaling pathway	KEGG pathway	0.000619	13	147
8	PS1PATHWAY	PS1PATHWAY	MSigDB	0.001235	5	14
9	STRIATED_MUSCLE_CONTRACTION	STRIATED_MUSCLE_CONTRACTION	MSigDB	0.001547	7	38
10	h_hopPathway	Hop Pathway in Cardiac Development	CGAP BioCarta	0.002459	3	3
11	h_shhPathway	Sonic Hedgehog (Shh) Pathway	CGAP BioCarta	0.004908	4	9
12	PW:0000315	calcineurin signaling pathway	RGD	0.004908	4	9
13	hsa04330	Notch signaling pathway	KEGG pathway	0.005852	7	46
14	CALCINEURINPATHWAY	CALCINEURINPATHWAY	MSigDB	0.006692	5	19
15	hsa04010	MAPK signaling pathway	KEGG pathway	0.008565	15	244
16	h_her2Pathway	Role of ERBB2 in Signal Transduction and Oncology	CGAP BioCarta	0.011391	5	21
17	hsa04340	Hedgehog signaling pathway	KEGG pathway	0.017271		
18	PW:0000007	mitogen activated protein kinase signaling pathway	RGD	0.022823		
19	SHHPATHWAY	SHHPATHWAY	MSigDB	0.036453		
20	NOTCHPATHWAY	NOTCHPATHWAY	MSigDB	0.047336		
21	PW:0000201	canonical Wnt signaling pathway	RGD	0.047336		

5. Domain [Display Chart] [Show Detail]

6. Pubmed [Display Chart] [Show Detail]

7. Interaction [Display Chart] [Show Detail]

8. Cytoband [Display Chart] [Show Detail]

9. TFBS [Display Chart] [Hide Detail]

	ID	P-value	Term in Query	Term in Genome
1	TTGTTT_V$FOXO4_01	0.000043	47	1606
2	V$HMEF2_Q6	0.000392	11	121
3	V$GATA1_03	0.002041	13	203
4	V$EVI1_05	0.002056	11	143
5	CAGGTG_V$			
6	GCANCTGNY			
7	V$EVI1_03			

3. Mouse Phenotype [Display Chart] [Hide Detail]

	ID	Name	P-value	Term in Query	Term in Genome
1	MP:0000266	abnormal cardiac morphology	0	120	533
2	MP:0005294	abnormal heart ventricle morphology	0	94	271
3	MP:0002127	abnormal cardiovascular system morphology	0	124	698
4	MP:0005385	cardiovascular system phenotype	0	127	922
5	MP:0000267	abnormal cardiac development	0	72	200
6	MP:0002925	abnormal cardiovascular development	0	80	313
7	MP:0005329	abnormal cardiac muscle morphology	0	70	222
8	MP:0006113	abnormal heart septum morphology	0	56	131
9	MP:0000281	abnormal ventricular septum morphology	0	54	125
10	MP:0002080	embryonic lethality	0	116	1067
11	MP:0005406	abnormal heart size	0	69	264
12	MP:0002108	abnormal muscle morphology	0	80	407
13	MP:0006207	embryonic lethality during organogenesis	0	87	517
14	MP:0003105	abnormal heart atrium morphology	0	47	103

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

	ID	Name	P-value	Term in Query	Term in Genome
1	GO:0007507	heart development	0	41	45
2	GO:0007275	multicellular organismal development	0	89	1814
3	GO:0032502	developmental process	0	100	2610
4	GO:0048513	organ development	0	63	790
5	GO:0048856	anatomical structure development	0	76	1413
6	GO:0032501	multicellular organismal process	0	98	3079
7	GO:0048731	system development	0	66	1196
8	GO:0065007				
9	GO:0050789				
10	GO:0050794				
11	GO:0009653				
12	GO:0019222				
13	GO:0031323				
14	GO:0007517				

Training Result [Show All] [Download All]

1. GO: Molecular Function [Display Chart] [Show Detail]
2. GO: Biological Process [Display Chart] [Show Detail]
3. Mouse Phenotype [Display Chart] [Show Detail]
4. Pathway [Display Chart] [Show Detail]
5. Domain [Display Chart] [Show Detail]
6. Pubmed [Display Chart] [Show Detail]
7. Interaction [Display Chart] [Show Detail]
8. Cytoband [Display Chart] [Show Detail]

File Download

Do you want to open or save this file?

Name: ...936885A1F859D657BFB80DA81218637948514.txt

Type: Text Document

From: keymaker.cchmc.org

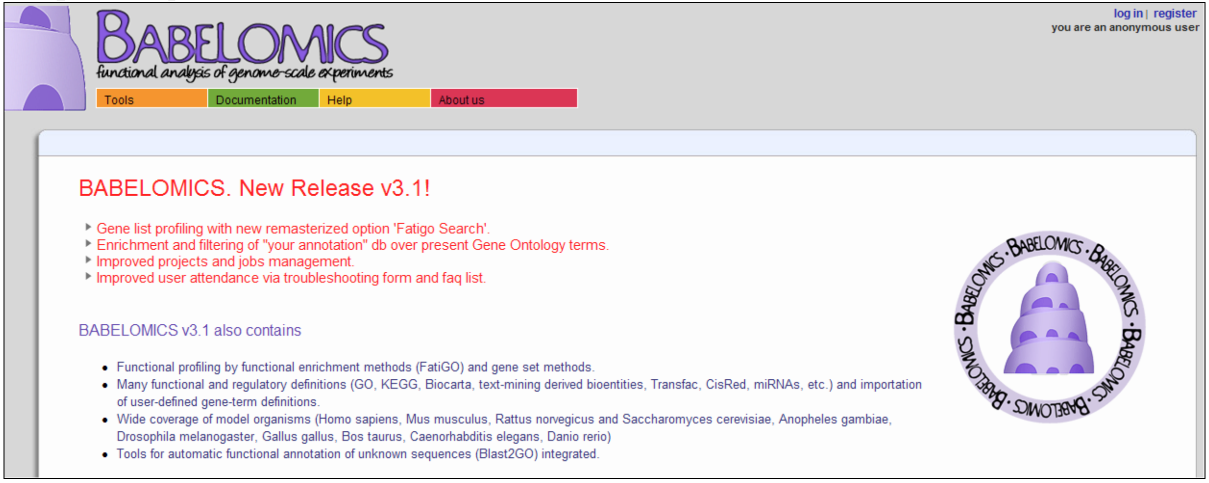
While files from the Internet can be useful, some files can potentially harm your computer. If you do not trust the source, do not open or save this file. What's the risk?

Open
Save
Cancel
Significant Pathways
Gene count
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Term
1. ALKPATHWAY
2. ALK in cardiac myocytes
3. TGF-beta signaling pathway
4. TGF-beta signaling pathway
5. Calcium signaling pathway
6. NFATPATHWAY
7. Wnt signaling pathway
8. PS1PATHWAY
9. STRIATED_MUSCLE_CONTRACTION
10. Hop Pathway in Cardiac Development
11. Sonic Hedgehog (Shh) Pathway
12. calcineurin signaling pathway
13. Notch signaling pathway
14. CALCINEURINPATHWAY
15. MAPK signaling pathway
16. Role of ERBB2 in Signal Transduction a...
17. Hedgehog signaling pathway
18. mitogen activated protein kinase signali...
19. SHHPATHWAY
20. NOTCHPATHWAY
21. canonical Wnt signaling pathway
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
P-value
Gene count
P-value

Babelomics (FatioGO)

Description: In addition to GO terms it can test simultaneously for KEGG pathways, Interpro motifs, SwissProt keywords, TFBSs and CisRed motifs. The distribution of any combination (or all) of the terms between two groups of genes can be simultaneously tested by means of a Fisher exact test. All the P-values are adjusted by FDR.

URL: http://www.fatigo.org
Access: Free web-based
Input: Gene symbols
Output: Graphical and downloadable as text files

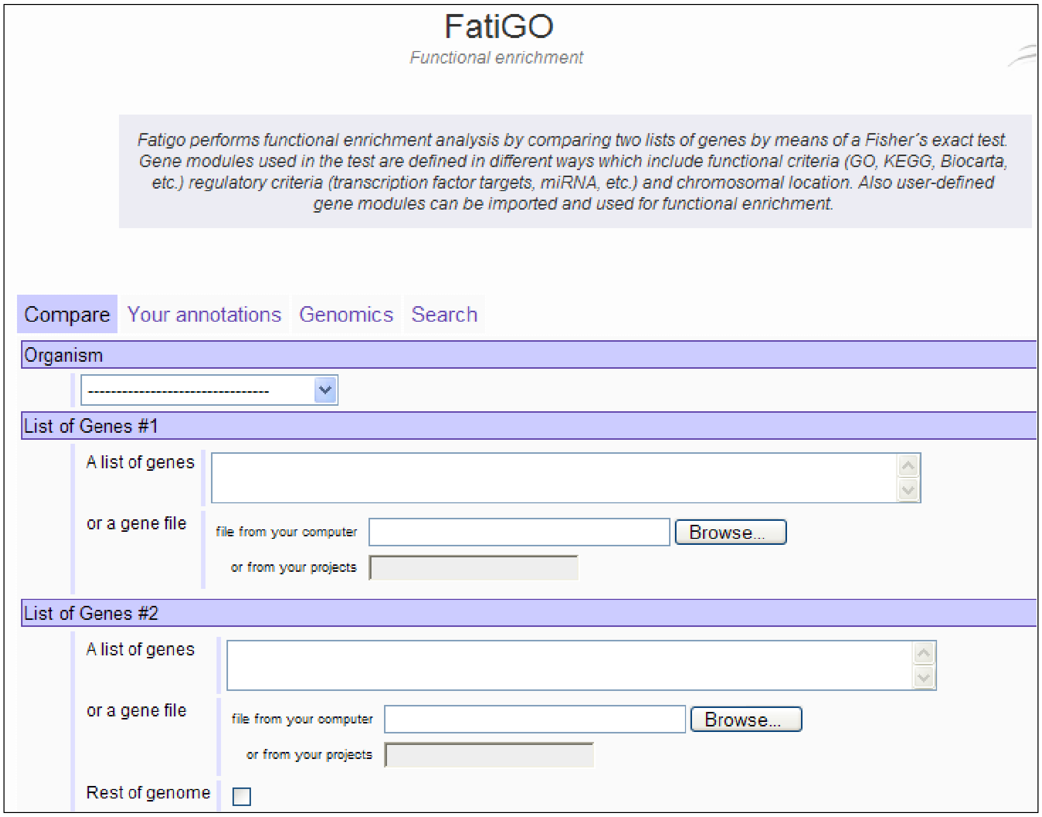

BABELOMICS
functional analysis of genome-scale experiments
Tools
Documentation
Help
About us
log in | register
you are an anonymous user

BABELOMICS. New Release v3.1!

- Gene list profiling with new remasterized option 'Fatigo Search'.
- Enrichment and filtering of "your annotation" db over present Gene Ontology terms.
- Improved projects and jobs management.
- Improved user attendance via troubleshooting form and faq list.

BABELOMICS v3.1 also contains

- Functional profiling by functional enrichment methods (FatiGO) and gene set methods.
- Many functional and regulatory definitions (GO, KEGG, Biocarta, text-mining derived bioentities, Transfac, CisRed, miRNAs, etc.) and importation of user-defined gene-term definitions.
- Wide coverage of model organisms (Homo sapiens, Mus musculus, Rattus norvegicus and Saccharomyces cerevisiae, Anopheles gambiae, Drosophila melanogaster, Gallus gallus, Bos taurus, Caenorhabditis elegans, Danio rerio)
- Tools for automatic functional annotation of unknown sequences (Blast2GO) integrated.



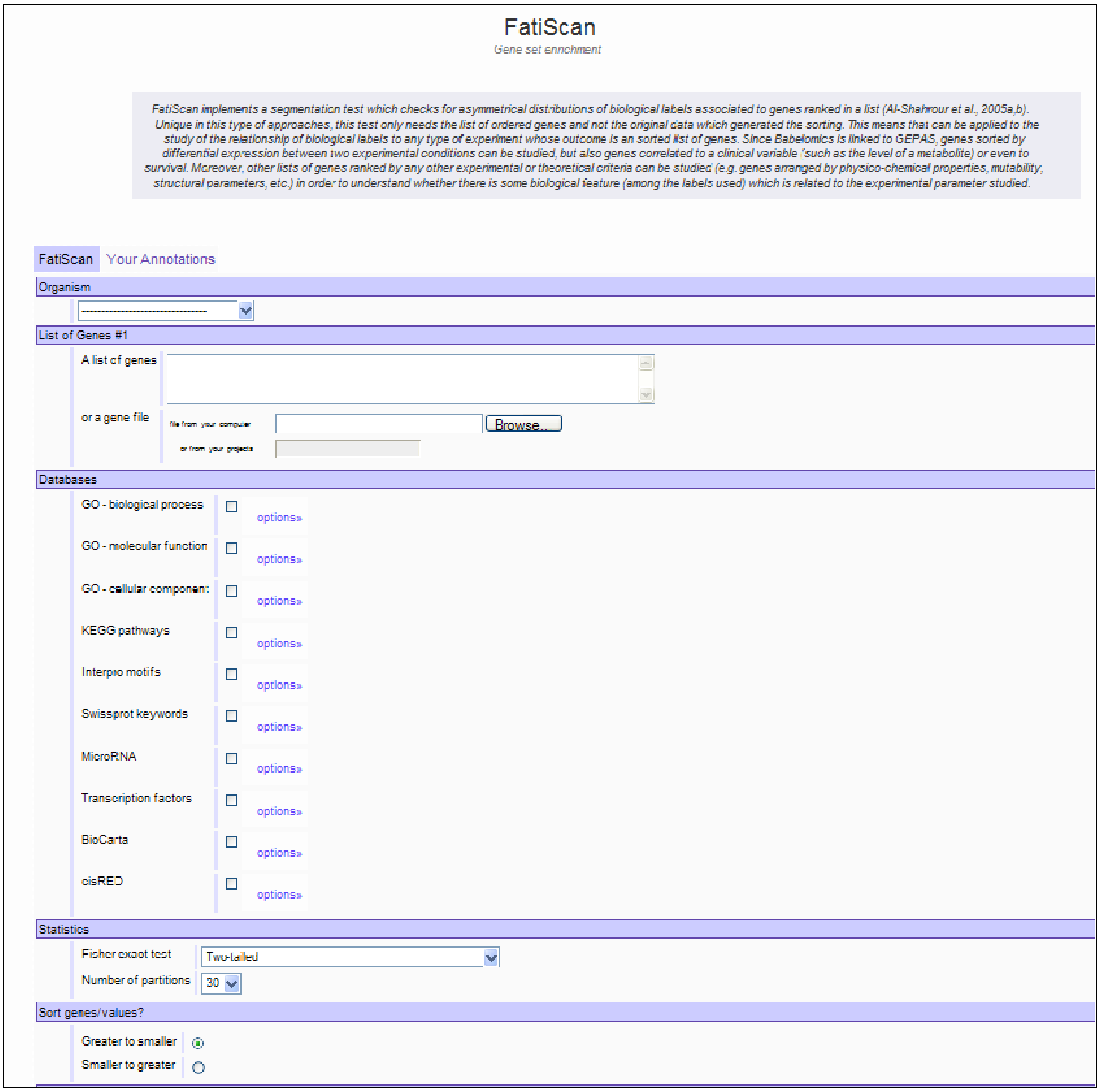
FatiGO
Functional enrichment

Fatigo performs functional enrichment analysis by comparing two lists of genes by means of a Fisher´s exact test. Gene modules used in the test are defined in different ways which include functional criteria (GO, KEGG, Biocarta, etc.) regulatory criteria (transcription factor targets, miRNA, etc.) and chromosomal location. Also user-defined gene modules can be imported and used for functional enrichment.

Organism

List of Genes #1
A list of genes
or a gene file
file from your computer Browse...
or from your projects

List of Genes #2
A list of genes
or a gene file
file from your computer Browse...
or from your projects
Rest of genome □



FatiScan

Gene set enrichment

FatiScan implements a segmentation test which checks for asymmetrical distributions of biological labels associated to genes ranked in a list (Al-Shahrour et al., 2005a,b). Unique in this type of approaches, this test only needs the list of ordered genes and not the original data which generated the sorting. This means that can be applied to the study of the relationship of biological labels to any type of experiment whose outcome is an sorted list of genes. Since Babelomics is linked to GEPAS, genes sorted by differential expression between two experimental conditions can be studied, but also genes correlated to a clinical variable (such as the level of a metabolite) or even to survival. Moreover, other lists of genes ranked by any other experimental or theoretical criteria can be studied (e.g. genes arranged by physico-chemical properties, mutability, structural parameters, etc.) in order to understand whether there is some biological feature (among the labels used) which is related to the experimental parameter studied.

Organism

List of Genes #1

A list of genes

or a gene file — file from your computer — or from your projects

Databases

- GO - biological process □ options»
- GO - molecular function □ options»
- GO - cellular component □ options»
- KEGG pathways □ options»
- Interpro motifs □ options»
- Swissprot keywords □ options»
- MicroRNA □ options»
- Transcription factors □ options»
- BioCarta □ options»
- cisRED □ options»

Statistics

Fisher exact test: Two-tailed

Number of partitions: 30

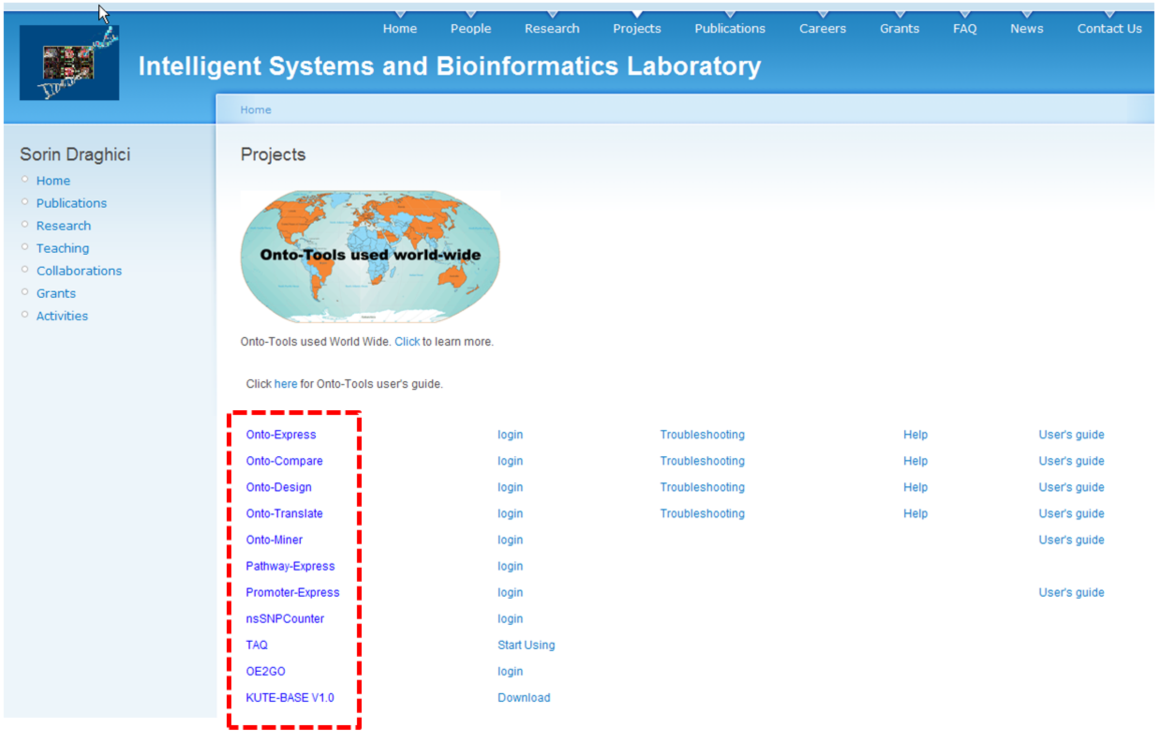
Sort genes/values?

- Greater to smaller ●
- Smaller to greater ○

OntoExpress

Description: OntoExpress constructs functional profiles (using Gene Ontology terms) for the following categories: biochemical function, biological process, cellular role, cellular component, molecular function and chromosome location. Statistical significance values are calculated for each category (Draghici et.al, Genomics, 81(2), 2003).
URL: http://vortex.cs.wayne.edu/projects.htm

http://vortex.cs.wayne.edu/projects.htm



Intelligent Systems and Bioinformatics Laboratory

Home

Sorin Draghici

- Home
- Publications
- Research
- Teaching
- Collaborations
- Grants
- Activities

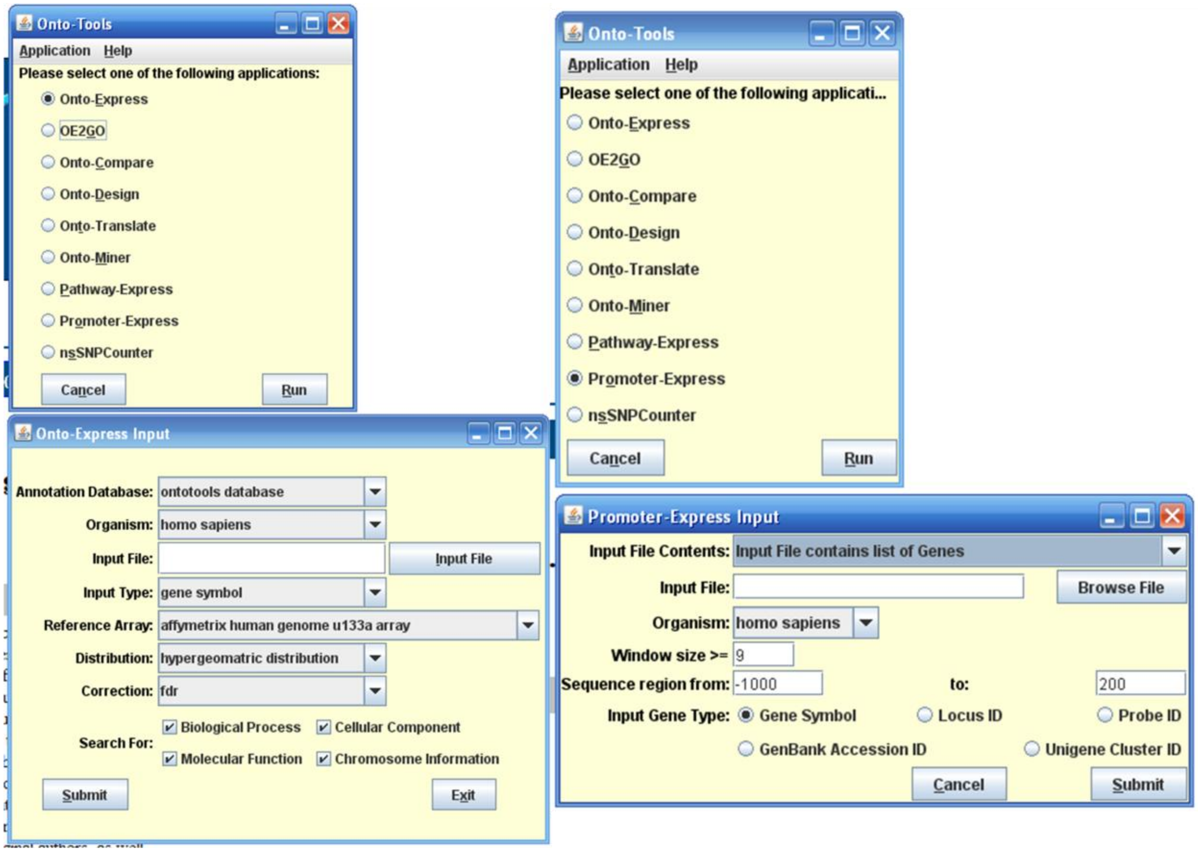
Projects

Onto-Tools used world-wide

Onto-Tools used World Wide. Click to learn more.

Click here for Onto-Tools user's guide.

Onto-Express	login	Troubleshooting	Help	User's guide
Onto-Compare	login	Troubleshooting	Help	User's guide
Onto-Design	login	Troubleshooting	Help	User's guide
Onto-Translate	login	Troubleshooting	Help	User's guide
Onto-Miner	login			User's guide
Pathway-Express	login			
Promoter-Express	login			User's guide
nsSNPCounter	login			
TAQ	Start Using			
OE2GO	login			
KUTE-BASE V1.0	Download			



Please select one of the following applications:

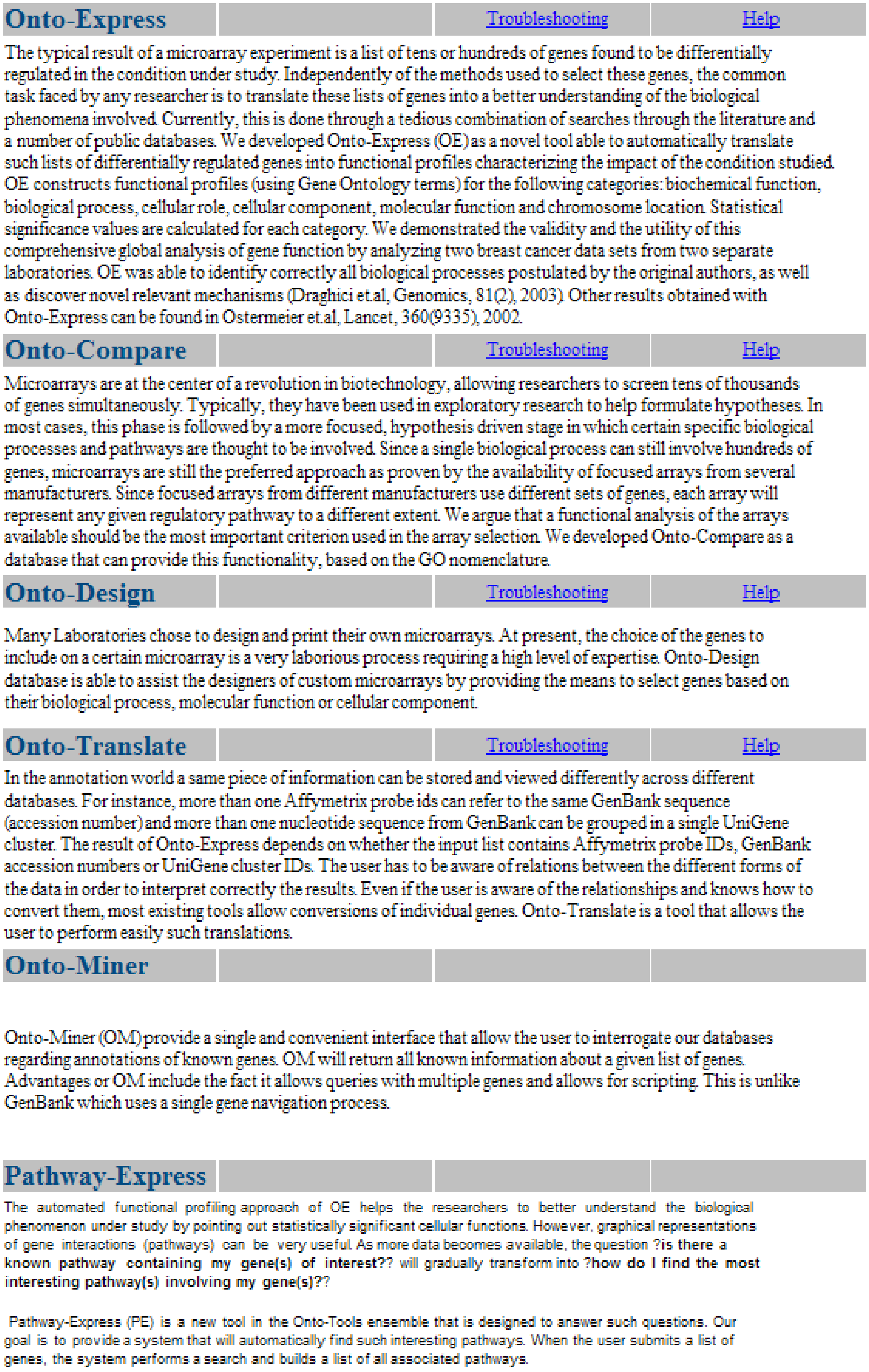
- ● Onto-Express
- ○ OE2GO
- ○ Onto-Compare
- ○ Onto-Design
- ○ Onto-Translate
- ○ Onto-Miner
- ○ Pathway-Express
- ○ Promoter-Express
- ○ nsSNPCounter

Annotation Database:	ontotools database
Organism:	homo sapiens
Input File:	
Input Type:	gene symbol
Reference Array:	affymetrix human genome u133a array
Distribution:	hypergeomatric distribution
Correction:	fdr
Search For:	☑ Biological Process ☑ Cellular Component ☑ Molecular Function ☑ Chromosome Information

Please select one of the following applicati...

- ○ Onto-Express
- ○ OE2GO
- ○ Onto-Compare
- ○ Onto-Design
- ○ Onto-Translate
- ○ Onto-Miner
- ○ Pathway-Express
- ● Promoter-Express
- ○ nsSNPCounter

Input File Contents:	Input File contains list of Genes
Input File:	
Organism:	homo sapiens
Window size >=	9
Sequence region from:	-1000 to: 200
Input Gene Type:	● Gene Symbol ○ Locus ID ○ Probe ID ○ GenBank Accession ID ○ Unigene Cluster ID



Onto-Express

Troubleshooting Help

The typical result of a microarray experiment is a list of tens or hundreds of genes found to be differentially regulated in the condition under study. Independently of the methods used to select these genes, the common task faced by any researcher is to translate these lists of genes into a better understanding of the biological phenomena involved. Currently, this is done through a tedious combination of searches through the literature and a number of public databases. We developed Onto-Express (OE) as a novel tool able to automatically translate such lists of differentially regulated genes into functional profiles characterizing the impact of the condition studied. OE constructs functional profiles (using Gene Ontology terms) for the following categories: biochemical function, biological process, cellular role, cellular component, molecular function and chromosome location. Statistical significance values are calculated for each category. We demonstrated the validity and the utility of this comprehensive global analysis of gene function by analyzing two breast cancer data sets from two separate laboratories. OE was able to identify correctly all biological processes postulated by the original authors, as well as discover novel relevant mechanisms (Draghici et.al, Genomics, 81(2), 2003) Other results obtained with Onto-Express can be found in Ostermeier et.al, Lancet, 360(9335), 2002.

Onto-Compare

Troubleshooting Help

Microarrays are at the center of a revolution in biotechnology, allowing researchers to screen tens of thousands of genes simultaneously. Typically, they have been used in exploratory research to help formulate hypotheses. In most cases, this phase is followed by a more focused, hypothesis driven stage in which certain specific biological processes and pathways are thought to be involved. Since a single biological process can still involve hundreds of genes, microarrays are still the preferred approach as proven by the availability of focused arrays from several manufacturers. Since focused arrays from different manufacturers use different sets of genes, each array will represent any given regulatory pathway to a different extent. We argue that a functional analysis of the arrays available should be the most important criterion used in the array selection. We developed Onto-Compare as a database that can provide this functionality, based on the GO nomenclature.

Onto-Design

Troubleshooting Help

Many Laboratories chose to design and print their own microarrays. At present, the choice of the genes to include on a certain microarray is a very laborious process requiring a high level of expertise. Onto-Design database is able to assist the designers of custom microarrays by providing the means to select genes based on their biological process, molecular function or cellular component.

Onto-Translate

Troubleshooting Help

In the annotation world a same piece of information can be stored and viewed differently across different databases. For instance, more than one Affymetrix probe ids can refer to the same GenBank sequence (accession number) and more than one nucleotide sequence from GenBank can be grouped in a single UniGene cluster. The result of Onto-Express depends on whether the input list contains Affymetrix probe IDs, GenBank accession numbers or UniGene cluster IDs. The user has to be aware of relations between the different forms of the data in order to interpret correctly the results. Even if the user is aware of the relationships and knows how to convert them, most existing tools allow conversions of individual genes. Onto-Translate is a tool that allows the user to perform easily such translations.

Onto-Miner

Onto-Miner (OM) provide a single and convenient interface that allow the user to interrogate our databases regarding annotations of known genes. OM will return all known information about a given list of genes. Advantages or OM include the fact it allows queries with multiple genes and allows for scripting. This is unlike GenBank which uses a single gene navigation process.

Pathway-Express

The automated functional profiling approach of OE helps the researchers to better understand the biological phenomenon under study by pointing out statistically significant cellular functions. However, graphical representations of gene interactions (pathways) can be very useful. As more data becomes available, the question ?**is there a known pathway containing my gene(s) of interest?**? will gradually transform into ?**how do I find the most interesting pathway(s) involving my gene(s)?**?

Pathway-Express (PE) is a new tool in the Onto-Tools ensemble that is designed to answer such questions. Our goal is to provide a system that will automatically find such interesting pathways. When the user submits a list of genes, the system performs a search and builds a list of all associated pathways.

Chapter 4: Identification of Regulatory Regions: Using Trafac and Other Related Tools

*Before going any further, please check the **GenomeTrafac** database (**http://genometrafac.cchmc.org**) which has more than 12,000 human-mouse gene pairs and about 200 microRNAs already analysed for potential regulatory regions. The genes or microRNAs you are interested in might be already there. This will save you the trouble of uploading them through trafac. You DON'T need any account to access the database.*

Introduction

Trafac (http://trafac.cchmc.org) is a web-accessible system, developed by us at Biomedical Informatics, to detect and visualize constitutionally similar clusters of transcription factor binding sites between a pair of genes.

Method

We first identify regions of genomic sequence conservation between two related but yet divergent species. Potential transcription factor binding sites, based on the TRANSFAC Professional Library, are independently predicted for both the genomic sequences. A JAVA servlet is then used to parse the results of this sequence conservation and transcription factor binding sites (*cis*-elements) data into an Oracle database. The database is mined for the detection of clusters of *cis*-elements in common between the two genes.

Output

1. **Trafacgram**: A high-resolution graphical image depicting the relative structural arrangement of the shared transcription factor (TF) binding sites within each sequence.
2. **Regulogram**: To better understand the occurrence of conserved *cis*-element clusters as a function of entire genomic regions, rather than discrete homologous blocks, we developed a *cis*-element hit-density graph (Regulogram) that depicts the density of shared *cis*-elements occurring within a moving window through conserved regions.

Utility

1. To find conserved TF binding sites between two orthologous genes in the context of sequence similarity.
2. It can be a valuable filtration tool for identifying potential novel regulatory regions, hitherto unknown.
3. It helps in comparison of heterologous genes (for e.g. two genes with similar expression).
4. It helps in identifying shared TF binding sites within genes that exhibit coordinate expression.
5. Finally, it also helps in understanding the constitution of regulatory regions of tissue specific genes.

What do you need?

1. Genomic Sequences: Exons, Introns, upstream and downstream regions.
2. Exon coordinates or positions in the genomic sequence.
3. Masking the repeats.
4. List of transcription factor binding sites.
5. Sequence alignment data where applicable.

How to Use Trafac?

From the home page of Trafac you can approach to the analysis by taking either of the two routes.

1. Cis-element Clusters within BlastZ Alignments: To Find conserved cis-clusters within BLASTZ-identified conserved sequence alignment blocks. Using this link you would be able to visualize the alignment between an orthologous pair of genes (mostly human and mouse sequences). Most importantly, you can view the common putative TF binding sites shared by the human and mouse genes in the context of the conserved regions. This utility is limited to a pairwise comparison of only those sequences for which the alignment data is present in the database. But, then if you are interested in a particular gene(s) you can upload the requisite input data to visualize the results. Alternatively, you can send us the sequences and we will do the rest for you.
2. Cis-elements Shared Between any Gene Pair: To Find cis-element clusters between user-selected gene segment pairs. This link would take you to explore the genes for regulatory elements irrespective of the sequence similarity. The main advantage of this route is you can compare any gene with any other gene or known promoters/enhancers in the Trafac database. If you are analyzing a group of co-expressed or coordinately regulated genes, this approach is recommended, especially when you know the transcription start site.

You can search the trafac database either by the

- Accession number: A GenBank or Celera accession number can be used.
- Name/Symbol of the gene: Trafac supports the gene nomenclature approved by the HUGO.
- Description/any term: Enter any term, for example, human, mouse, repair, etc. Please make sure to enter only one term.

Sequence Group: You can also select the genes based upon the gene group. For example, selecting "DNA repair" from the list of the available groups would display all genes belonging to DNA repair group and which have the BlastZ alignment data entered to Trafac. *Please note that the list of genes under each of the groups at present is not exhaustive*. We are in the process of building up the database and adding more genes and more groups. If you are interested in any particular gene or group please let us know so that we can add them to the Trafac database.

Sequence Selector: Sequence selector page has two parts. The top one is the query part wherein you can enter one or more than one search criteria. The second part or the lower half displays your search results once you click on the search.

Using Sequence Selector for identifying potential conserved regulatory regions:

1. Enter one or more than one search criteria and click the Search button. Alternatively you can choose one from the existing groups of genes.
2. The results will be displayed in the lower half.
3. For example, using the term "Human" for description displays a list of entries in the lower half.
4. The results table has check boxes in the first column against each entry. You can check one or more to view the sequences.
5. Check your selected entries and click on Select. This would take you to the BlastZ alignments page. The first two columns of the table show the sequence information for the human and mouse sequences followed by a date of entry column. The last column shows three options. The **view** option would take you to the Local alignments page. This is nothing but a summary view of the sequence alignment information. The second option "**PIP**" leads to a graphic display of the alignment image. This PIP is generated using the PipMaker software. This is a pdf file so you need to have an Adobe Acrobat Reader installed to view this. The third option is the "**Regulogram**", a cis-element hit density graph in the context of sequence similarity.

Using Sequence Selector for identifying constitutionally similar regulatory sites:

1. Enter one or more than one search criteria and click the Search button. Alternatively you can choose one from the existing groups of genes
2. The results will be displayed in the lower half.
3. For example, using the term "Human" for description displays a list of entries in the lower half.
4. The results table has check boxes in the first column against each entry. You can check the first sequence and click. You will be prompted to select the second sequence. After selecting both the sequences, you will be led to the TraFaC query Page.
5. The **TraFaC Query Page** allows the user to alter the various parameters like sequence extent, matrices, image size, comparison parameters, whether based on matrix families or individual matrices, etc.

Uploading Sequences:

If you wish to upload the genes of your interest and those which are not already present in TraFac database, you need to have an account. You can obtain one by sending a mail (anil.jegga@chmcc.org or bhuvana.sakthivel@cchmc.org). If you have only one or two genes to be analyzed, you can mail us the sequences or GenBank accession number(s) or gene symbols and we can upload the genes for you.

Multi-Upload Page
From the Multi-Upload page, you can parse all the input files to the TraFac server and see the results. There are certain rules you need to follow when uploading the input files.

- All the input files should be strictly in the recommended format (in text format for all except the MatInspector files and the PIP, which are in html format and pdf respectively).
- An accession number should always be entered whenever you are uploading any of the files except MatInspector files.
- The exon file, repeat mask and the actual sequence file upload is optional. However, we advise you to upload all of these so that you can comprehend your results more clearly.

The requisite input files can be uploaded from your local system using the browse button. Click the upload button when you are through. Uploading may take some time depending upon your sequence size. You can upload any of these data in parts but do remember to associate with a unique ID, which is the accession number. An "upload successful" message indicates that you are ready to view your results.

Input Files

a. Sequence Files
b. Exon Files
c. RepeatMasker Output Files
d. PipMaker Output Files
e. MatInspector/Match Output Files

1. **Sequence Files**: The nucleotide sequence files need to be in the fasta format.
2. **RepeatMasker**: The RepeatMasker web application is used to mask the repeats before aligning them. Given below are the brief instructions for using the RepeatMasker. You can find a detailed set on the RepeatMasker page. Use the ‘browse’ button to select the fasta file created above, or copy the fasta sequence into the text box. Select the "html" return format, the appropriate DNA Source and press Submit Sequence. Save the Matches as a text file to your computer by selecting File->Save As… from the main menu. You can use copy and paste to save the Repeat Mask output. The mask output alone is required. Do not save the masked sequence.
3. **Exon Files**: An optional text file, providing the positions of transcriptional units in the first sequence. The directionality of a gene (< or >), its start and end positions, and name should be on one line, followed by separate lines specifying the start-positions and end-positions of each exon. An optional line beginning with a "+" character can indicate the first and last nucleotides of the translated region (including the initiation codon, Met, and the stop codon). Blank lines are ignored. Exons must be specified in order of increasing address even if the gene is on the reverse strand (<). For example, the Exons file might begin as follows:
   ```
   > 100 800 Gene 1
   100 200
   300 400
   600 800
   ```
4. **PipMaker Output Files**: PipMaker computes alignments of similar regions in two DNA sequences. The resulting alignments are summarized with a Percent Identity Plot (PIP). It generates graphical output as a PDF document by default. For TraFaC,

you need the **Blastz alignment file**, the **summary file** and an optional **PIP** file. These are referred as "**text**", "**concise**" and "**pip**" files respectively in the PipMaker output data, which you receive by an E-mail. We give here the brief instructions for using the Advanced PipMaker. For detailed instructions refer the advanced PipMaker instructions page. Follow the Advanced PipMaker instructions (the files you have created above can be used). Enter your E-mail address. Press the Submit button. When the alignment is finished, you will receive an E-mail containing multiple attachments. Of the different output files **Trafac** needs only three of them *viz*., ***Concise Alignment file***: This file has a summary of the sequence alignment information. ***Text file***: This is a verbose file and is actually a BLASTZ alignment file. ***PIP***: a .pdf file of the percent identity plot. Store the first two files as text files.

5. **MatInspector *Professional*/Match Output Files**: You can use either of these programs for detection of transcription factor binding sites. MatInspector professional or Match is a tool that utilizes a library of matrix descriptions for transcription factor binding sites to locate matches in sequences of unlimited length. To access both these programs you need to have registered with them and it is free. You can find the detailed instructions on these pages. We however, would like to highlight some of the points like:Be certain to choose the appropriate matrix family e.g. vertebrates. And follow the default options for the rest of the parameters. When you have the output, save the MatInspector files as html files. But in case of Match files, save them as text files.

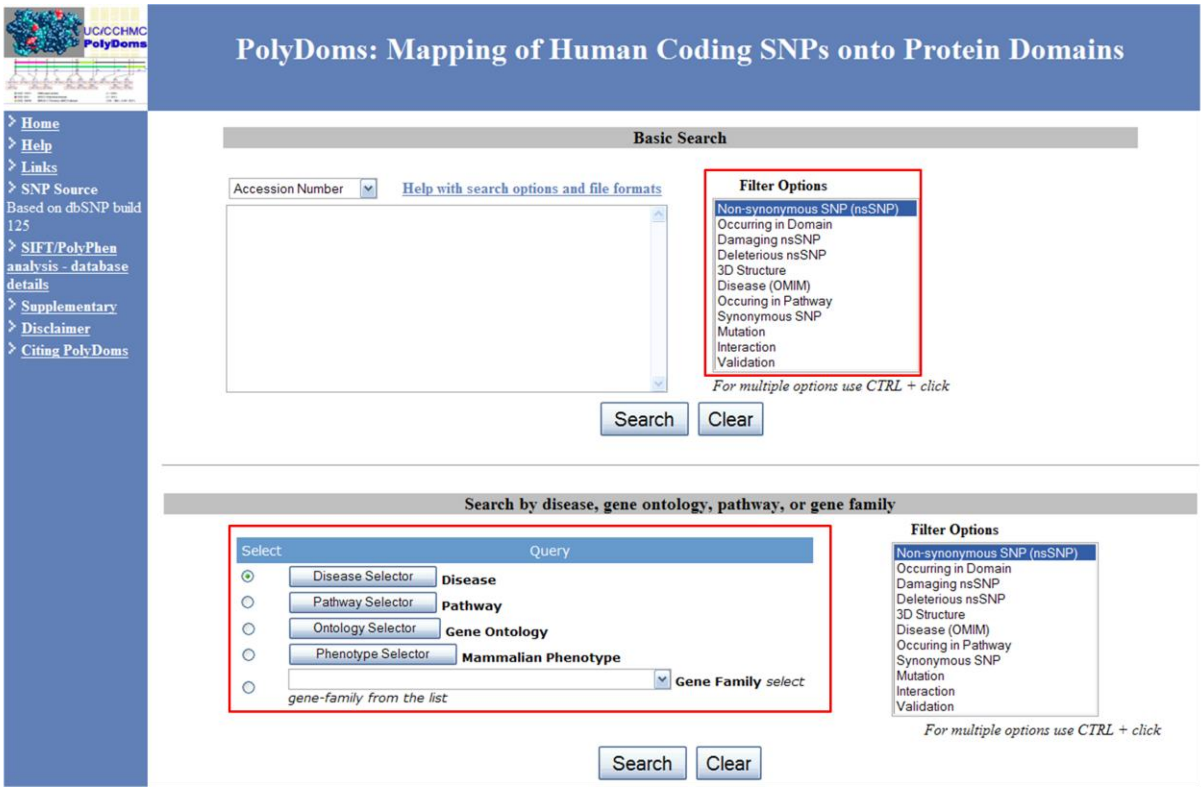
Once you have all of the requisite files, log on to Trafac. From Advanced Tools, select Upload/Parse Sequence Data.

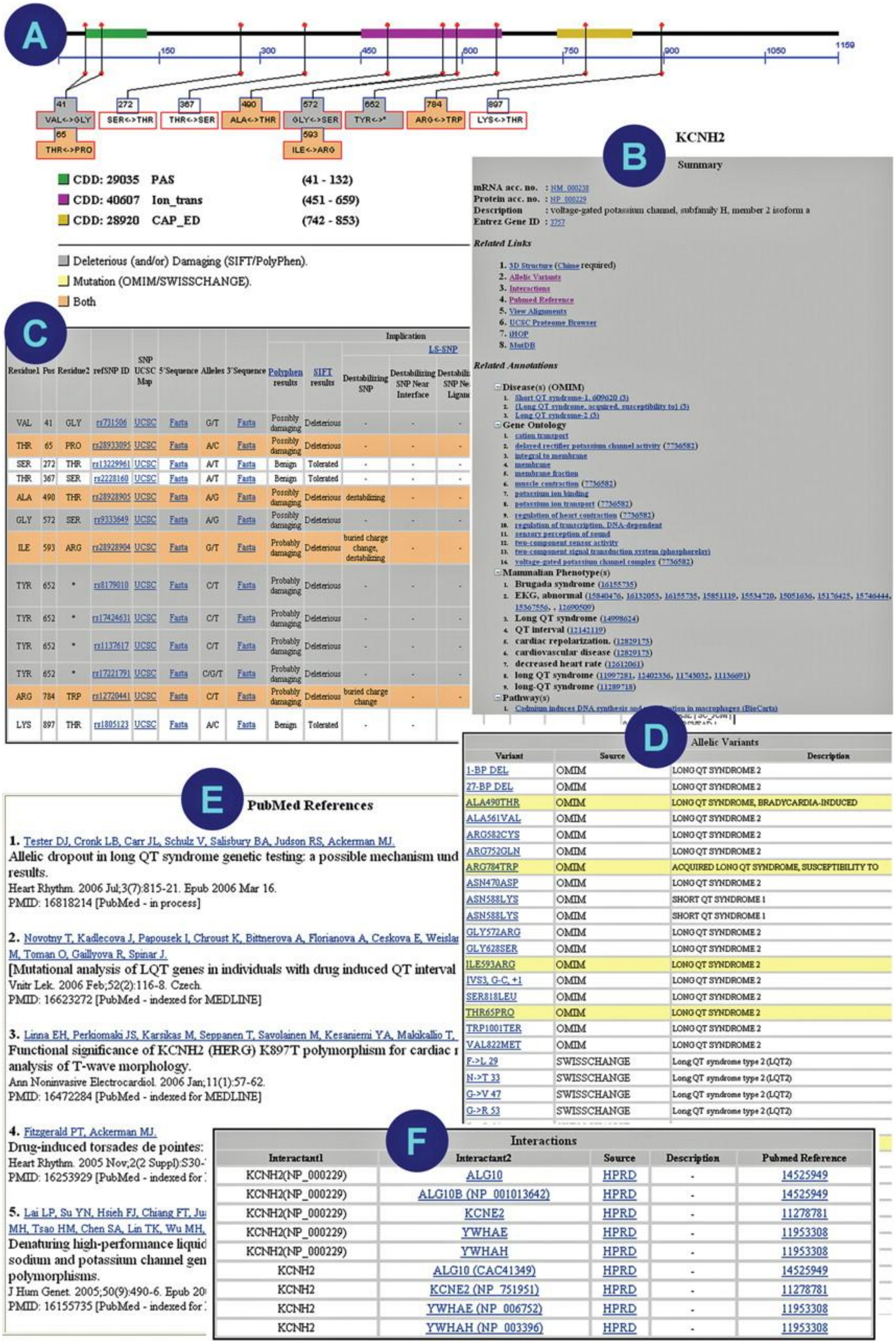
Related Tools

Concise Scanner (Conserved Cis-Element Scanner; http://concise-scanner.cchmc.org)**:** Concise Scanner was developed primarily as an answer to the question as to what are the potential additional target genes for an identified regulatory module(s). Tools based on the phlyogenetic footprinting like TraFaC (and GenomeTraFaC, the human-mouse gene regulatory region repository created using TraFaC server), and others while helping in identification of potential regulatory regions provide little or no information as to through what genes the transcription factors (TFs) exert their function in the living system. Providing a complement to the above listed phylogenetic approaches, we developed Concise (Conserved Cis Element) Scanner that undertakes a more targeted search, finding phylogenetically conserved regulatory targets of defined transcription factors whose DNA binding site specificity is known. It identifies potential targets of one or more clusters of transcription factors with a defined cis-regulatory target specificity, using human and mouse genomes. It enables you to select one or more transcription binding sites and search all genes in the GenomeTrafac database for clusters containing the selected site(s). Within each cluster, you can view the exact position of each binding site.

Chapter 5: Annotation of Coding Single Nucleotide Polymorphisms: Using PolyDoms

Description: There is a wide gap between the growing number of reported nsSNPs in human population and the functional consequences of these nsSNPs, and the major challenge lies in distinguishing the functionally significant and potentially disease-related ones from the rest. PolyDoms is based on the hypothesis that if an nsSNP alters the ability of gene products to function normally within the biological pathways or processes, the consequences might either alter disease susceptibility or resistance or result in disease itself or affect the therapeutic regimen. We mapped the coding SNPs (synonymous and non-synonymous) of all the proteins onto their know protein 3D structure and functional domains. We used the coding SNP data from the dbSNP. The database supplemented with a variety of functional implication prediction algorithms like SIFT, PolyPhen, LS-SNP, etc. The web interface supports a variety of queries (GO terms, disease terms, gene families, etc.). Please refer to the "Help" section (http://info.chmcc.org/help/polydoms/index.html) on PolyDoms home page (http://polydoms.cchmc.org). For any problems or questions or analysis, send a mail to anil.jegga@cchmc.org
Access: Free web-based
Input: Gene symbols, accession numbers, rsSNP IDs, disease terms, GO Terms, etc.
Output: Graphical and results are downloadable as a spreadsheet.





🟩 CDD: 29035 PAS (41 - 132)
🟪 CDD: 40607 Ion_trans (451 - 659)
🟨 CDD: 28920 CAP_ED (742 - 853)

▨ Deleterious (and/or) Damaging (SIFT/PolyPhen).
🟨 Mutation (OMIM/SWISSCHANGE).
🟧 Both

KCNH2

Summary

mRNA acc. no. : NM_000238
Protein acc. no. : NP_000229
Description : voltage-gated potassium channel, subfamily H, member 2 isoform a
Entrez Gene ID : 3757

Related Links

1. 3D Structure (Chime required)
2. Allelic Variants
3. Interactions
4. Pubmed Reference
5. View Alignments
6. UCSC Proteome Browser
7. iHOP
8. MutDB

Related Annotations

⊟ Disease(s) (OMIM)
1. Short QT syndrome-1, 609620 (3)
2. {Long QT syndrome, acquired, susceptibility to} (3)
3. Long QT syndrome-2 (3)

⊟ Gene Ontology
1. cation transport
2. delayed rectifier potassium channel activity (7736582)
3. integral to membrane
4. membrane
5. membrane fraction
6. muscle contraction (7736582)
7. potassium ion binding
8. potassium ion transport (7736582)
9. regulation of heart contraction (7736582)
10. regulation of transcription, DNA-dependent
11. sensory perception of sound
12. two-component sensor activity
13. two-component signal transduction system (phosphorelay)
14. voltage-gated potassium channel complex (7736582)

⊟ Mammalian Phenotype(s)
1. **Brugada syndrome** (16155735)
2. **EKG, abnormal** (15840476, 16132053, 16155735, 15851119, 15534720, 15051636, 15176425, 15746444, 15367556, , 12690509)
3. **Long QT syndrome** (14998624)
4. **QT interval** (12142119)
5. **cardiac repolarization.** (12829173)
6. **cardiovascular disease** (12829173)
7. **decreased heart rate** (12612061)
8. **long QT syndrome** (11997281, 12402336, 11743032, 11136691)
9. **long-QT syndrome** (11289718)

⊟ Pathway(s)
1. Cadmium induces DNA synthesis and [illegible] tion in macrophages (BioCarta)

								Implication				
										LS-SNP		
Residue1	Pos	Residue2	refSNP ID	SNP UCSC Map	5'Sequence	Alleles	3'Sequence	Polyphen results	SIFT results	Destabilizing SNP	Destabilizing SNP Near Interface	Destabili SNP Ne Ligan
VAL	41	GLY	rs731506	UCSC	Fasta	G/T	Fasta	Possibly damaging	Deleterious	-	-	-
THR	65	PRO	rs28933095	UCSC	Fasta	A/C	Fasta	Possibly damaging	Deleterious	-	-	-
SER	272	THR	rs13229961	UCSC	Fasta	A/T	Fasta	Benign	Tolerated	-	-	-
THR	367	SER	rs2228160	UCSC	Fasta	A/T	Fasta	Benign	Tolerated	-	-	-
ALA	490	THR	rs28928905	UCSC	Fasta	A/G	Fasta	Possibly damaging	Deleterious	destabilizing	-	-
GLY	572	SER	rs9333649	UCSC	Fasta	A/G	Fasta	Possibly damaging	Deleterious	-	-	-
ILE	593	ARG	rs28928904	UCSC	Fasta	G/T	Fasta	Probably damaging	Deleterious	buried charge change, destabilizing	-	-
TYR	652	*	rs8179010	UCSC	Fasta	C/T	Fasta	Probably damaging	Deleterious	-	-	-
TYR	652	*	rs17424631	UCSC	Fasta	C/T	Fasta	Probably damaging	Deleterious	-	-	-
TYR	652	*	rs1137617	UCSC	Fasta	C/T	Fasta	Probably damaging	Deleterious	-	-	-
TYR	652	*	rs17221791	UCSC	Fasta	C/G/T	Fasta	Probably damaging	Deleterious	-	-	-
ARG	784	TRP	rs12720441	UCSC	Fasta	C/T	Fasta	Probably damaging	Deleterious	buried charge change	-	-
LYS	897	THR	rs1805123	UCSC	Fasta	A/C	Fasta	Benign	Tolerated	-	-	

Allelic Variants

Variant	Source	Description
1-BP DEL	OMIM	LONG QT SYNDROME 2
27-BP DEL	OMIM	LONG QT SYNDROME 2
ALA490THR	OMIM	LONG QT SYNDROME, BRADYCARDIA-INDUCED
ALA561VAL	OMIM	LONG QT SYNDROME 2
ARG582CYS	OMIM	LONG QT SYNDROME 2
ARG752GLN	OMIM	LONG QT SYNDROME 2
ARG784TRP	OMIM	ACQUIRED LONG QT SYNDROME, SUSCEPTIBILITY TO
ASN470ASP	OMIM	LONG QT SYNDROME 2
ASN588LYS	OMIM	SHORT QT SYNDROME 1
ASN588LYS	OMIM	SHORT QT SYNDROME 1
GLY572ARG	OMIM	LONG QT SYNDROME 2
GLY628SER	OMIM	LONG QT SYNDROME 2
ILE593ARG	OMIM	LONG QT SYNDROME 2
IVS3, G-C, +1	OMIM	LONG QT SYNDROME 2
SER818LEU	OMIM	LONG QT SYNDROME 2
THR65PRO	OMIM	LONG QT SYNDROME 2
TRP1001TER	OMIM	LONG QT SYNDROME 2
VAL822MET	OMIM	LONG QT SYNDROME 2
F->L 29	SWISSCHANGE	Long QT syndrome type 2 (LQT2)
N->T 33	SWISSCHANGE	Long QT syndrome type 2 (LQT2)
G->V 47	SWISSCHANGE	Long QT syndrome type 2 (LQT2)
G->R 53	SWISSCHANGE	Long QT syndrome type 2 (LQT2)

PubMed References

1. Tester DJ, Cronk LB, Carr JL, Schulz V, Salisbury BA, Judson RS, Ackerman MJ.
Allelic dropout in long QT syndrome genetic testing: a possible mechanism und
results.
Heart Rhythm. 2006 Jul;3(7):815-21. Epub 2006 Mar 16.
PMID: 16818214 [PubMed - in process]

2. Novotny T, Kadlecova J, Papousek I, Chroust K, Bittnerova A, Florianova A, Ceskova E, Weislar
M, Toman O, Gaillyova R, Spinar J.
[Mutational analysis of LQT genes in individuals with drug induced QT interval
Vnitr Lek. 2006 Feb;52(2):116-8. Czech.
PMID: 16623272 [PubMed - indexed for MEDLINE]

3. Linna EH, Perkiomaki JS, Karsikas M, Seppanen T, Savolainen M, Kesaniemi YA, Makikallio T,
Functional significance of KCNH2 (HERG) K897T polymorphism for cardiac r
analysis of T-wave morphology.
Ann Noninvasive Electrocardiol. 2006 Jan;11(1):57-62.
PMID: 16472284 [PubMed - indexed for MEDLINE]

4. Fitzgerald PT, Ackerman MJ.
Drug-induced torsades de pointes:
Heart Rhythm. 2005 Nov;2(2 Suppl):S30-
PMID: 16253929 [PubMed - indexed for

5. Lai LP, Su YN, Hsieh FJ, Chiang FT, Ju
MH, Tsao HM, Chen SA, Lin TK, Wu MH,
Denaturing high-performance liquid
sodium and potassium channel gen
polymorphisms.
J Hum Genet. 2005;50(9):490-6. Epub 20
PMID: 16155735 [PubMed - indexed for

Interactions

Interactant1	Interactant2	Source	Description	Pubmed Reference
KCNH2(NP_000229)	ALG10	HPRD	-	14525949
KCNH2(NP_000229)	ALG10B (NP_001013642)	HPRD	-	14525949
KCNH2(NP_000229)	KCNE2	HPRD	-	11278781
KCNH2(NP_000229)	YWHAE	HPRD	-	11953308
KCNH2(NP_000229)	YWHAH	HPRD	-	11953308
KCNH2	ALG10 (CAC41349)	HPRD	-	14525949
KCNH2	KCNE2 (NP_751951)	HPRD	-	11278781
KCNH2	YWHAE (NP_006752)	HPRD	-	11953308
KCNH2	YWHAH (NP_003396)	HPRD	-	11953308

EXERCISE: Transcriptome Analysis and Annotation

1. Using the following two gene lists, identify conserved and non-conserved common binding sites within the upstream 500 bp (***Hint***: when downloading the promoter sequences use 500 bp)

   Gene List 1:
   ADORA1
   AGTRAP
   CD37
   COL3A1
   COL4A2
   COL6A1
   COL6A3
   DCN
   E2F4
   FN1
   LOC374395
   LOXL1
   LRP3
   LRP5
   LUM
   MMP9
   NUMA1
   PPARBP
   PPARD
   PPP2R1A
   PSG9
   PTGDS
   PTPRN
   ROM1
   SNN
   SPARC
   THBS2

   Gene List 2:
   APAF1
   BAD
   BAX
   BCL2
   BID
   BIRC2
   BNIP3L
   CASP1
   CASP2
   CHUK
   CYCS

DFFA
DFFB
FADD
FAS
MDM2
MYC
NFKB1
NFKBIA
PRF1
RELA
RIPK1
TNF
TNFRSF10B
TP53
TP73
TRAF1

2. UCSC Browser related:
    a. Find genes that are predominately expressed in the mouse pancreas, determine the expression pattern of the human ortholog of one such gene and obtain the genomic sequence of the human gene.
    b. Obtain a list of SNPs in a single gene (*PLG*) using the UCSC Table Browser and annotate them using NCBI's dbSNP batch processor.
    c. Using one of the gene lists from (1) download all noncoding SNPs occurring within upstream 1 kb region
    d. Using the same gene list download all coding SNPs. How many of these SNPs are predicted to be deleterious (***Hint***: use PolyDoms once you obtain the list of SNPs. ***Alternate approach***: You can directly use the gene symbols and download the annotated cSNPs including putative deleterious ones from PolyDoms)

3. Using the gene lists (based on published microarray data; Appendix 2), and the applications CisMols, GenomeTrafac and ConciseScanner:
    a. Find the potential common transcription factor binding site clusters that could be responsible for the co-expression.
    b. Find additional genome-wide targets for the signature cis-regulatory modules identified by CisMols (***Hint***: use the feature ConciseScanner to find genome-wide additional targets for the shared cis-cluster obtained through CisMols analyzer).

4. Use PolyDoms applications for the following:
    a. Using one of the gene lists from Appendix 2, annotate all the coding SNPs and find out what SNPs you would consider for designing a diagnostic chip or recommending for re-sequencing?
    b. What coding SNPs are potentially deleterious for the apoptosis pathway?
    c. How many proteins are involved in the DNA repair and how many of these have at least one coding nsSNP that occurs in a functional domain and is predicted as deleterious/damaging?

5. Using gene list 2 from (1) above, find out with which other differentially expressed gene lists do they overlap (use MSigDB Annotation or L2L or ToppGene). Are there any gene expression profiling studies where in these genes are down- or up-regulated?
6. Using gene list 1 from (1) above, obtain the mouse orthologs. For the mouse orthologs, download the 3'UTR sequences.

APPENDIX 1: Other Useful bioinformatics resources and tools

1. Bioinformatics Resources: **http://anil.cchmc.org**
2. Sequence Manipulation Suite: **http://anil.chmcc.org/sms/**
3. RepeatMasker: To mask the repeat elements in a genomic sequence. **http://www.repeatmasker.org/**
4. PipMaker: To compute alignments of similar regions in DNA sequences. **http://www.bx.psu.edu/**
5. Exon Mapper: **http://pbil.univ-lyon1.fr/sim4.php**
6. EPD Database: Eukaryotic Promoter Database - **http://www.epd.isb-sib.ch/**
7. Exon Mapper: **http://pbil.univ-lyon1.fr/sim4.php**

APPENDIX 2: Co-expressed gene lists

1: Nitric Oxide. 2002 Nov;7(3):165-86.
A DNA microarray study of nitric oxide-induced genes in mouse hepatocytes: implications for hepatic heme oxygenase-1 expression in ischemia/reperfusion.
Zamora R, Vodovotz Y, Aulak KS, Kim PK, Kane JM 3rd, Alarcon L, Stuehr DJ, Billiar TR.

#NAME inos_10_dn
#DESCRIPTION Ten most-downregulated genes following iNOS induction in hepatocytes
#GENES: CD151, EIF5A, EEF2, CD81, PKM2, ACT6

#NAME inos_10_up
#DESCRIPTION Ten most-upregulated genes following iNOS induction in hepatocytes
#GENES: EED, CSRP1, PCNA, HMOX1, MCM2, CDK2, MCM6, GNB1, TUBB1

2: J Nutr. 2004 Apr;134(4):762-70.
Gene expression profiling in human preadipocytes and adipocytes by microarray analysis.
Urs S, Smith C, Campbell B, Saxton AM, Taylor J, Zhang B, Snoddy J, Jones Voy B, Moustaid-Moussa N.

#NAME adip_human_dn
#DESCRIPTION Down-regulated in primary human adipocytes, versus preadipocytes
#GENES: PPARD, CEBPA, MMP2, SNN, SPARC, COL5A1, DCN, COL3A1, LRP3, COL6A3, PSG9, ATRAP, CD37, ROM1, COL4A2, LUM, PPP2R1A, LRP5, LOX, PTPRN, OKL38, IL18BP, THBS4, FN1, THBS1, LOXL1, COL6A1, ADORA1, MMP9, PPARBP, E2F4, PTGDS, THBS2

#NAME adip_human_up
#DESCRIPTION Up-regulated in primary human adipocytes, versus preadipocytes
#GENES: PFKFB3, ABCE1, PLCD1, AGTRL1, CROC4, HSD11B2, AGT, FABP4, DGKG, PTPN21, PTPRZ1, SCD, FABP5, RXRA, SMARCB1, COL1A2, CRYAB, DGAT1, ZNF336, LRP8, CTSG, 3-PAP, APM1, DPT, CAP2, IL22R, SCAP1, USP8, LYPLA1, HPCA, STAT5B, CYB5, E2F5, ALDH6A1, MMP7, LBP, GPD1, GLUL, GPX3, INSR, FXYD1, FACL2, ALDH1A2, MGST1, MAP4K3, MASP1, ECM2, PTPRS, CEBPD, KCNH2, ATP2B2, ACOX3, SPTBN4, TNFAIP2, LIPE, VN, FABP7, UCP4, LPL, ADFP, PPAR- , E2F1, IGFBP2, CHST1, GDF8, ADORA2B, ATP8A2, ATIP1, LIPC, REQ, PLEK, APOB, TAP1, AMT, PLIN, TFCP2, RXRB

3: Science. 2000 Mar 31;287(5462):2486-92.
Mitotic misregulation and human aging.
Ly DH, Lockhart DJ, Lerner RA, Schultz PG.

#NAME middleage_dn
#DESCRIPTION Downregulated in fibroblasts from middle-age individuals, compared to young

#GENES: CCNB, PLK, FOXM1, KIF11, PTGS2, KIF2C, CENPA, CDC20, H2AFX, KIF23, HMGN2, UBE2C, CCNF, CCNA, CENPF, MYB

#NAME middleage_up
#DESCRIPTION Upregulated in fibroblasts from middle-age individuals, compared to young
#GENES: COL15A1, TNFRSF11B, SERPINB2, COL6A2, IL8, FMOD, MMP12, DPT, CST6, COMP, THBS2, PTGS1, CRYBB2, MMP10, PRSS11

4: Oncogene. 2001 Jun 21;20(28):3674-82.
Distinctive gene expression profiles associated with Hepatitis B virus x protein.
Wu CG, Salvay DM, Forgues M, Valerie K, Farnsworth J, Markin RS, Wang XW.

#NAME hbx_dn
#DESCRIPTION Downregulated by expression of Hepatitis HBx protein in hepatocytes
#GENES: CD4, GSTA4, GLG1, WT1, TGFB1, MAP3K1, IL6, APR-3, TP53, CDKN1A, GSTM5, APC, GAS6

#NAME hbx_up
#DESCRIPTION Upregulated by expression of Hepatitis HBx protein in hepatocytes
#GENES: CCNI, DAD1, GSTM4, AP4B1, CDK4, TYMS, TNFRSF6, MYC, CCND3, PDCD2, PTK9, AP4S1, IGF1R, BCL2L1, TUBG1, TUBA4, TUBG2, VCL, IFNGR1, SLC5A1, MFNG, IFNAR2, CASP4, AP4E1, CDKN3

5: Nat Rev Cancer. 2002 Jan;2(1):38-47.
Hypoxia--a key regulatory factor in tumour growth.
Harris AL.

#NAME hypoxia_review
#DESCRIPTION Genes known to be induced by hypoxia
#GENES: EDN1, PFKP, MMP13, HSF, AK3, BIK, TGM2, P4HA, TEK, CDKN1B, SLC2A3, TF, CCNG2, CD99, SAT, FTL, PFKL, BNIP3, TH, RP1, STC1, HIF2A, PDGFB, VIM, IL8, LDHA, SPP1, CA9, PTGS2, PRPS1, BHLHB2, HK1, IGFBP2, TGFB1, APEX1, TFRC, ALDOA, CCL2, CA12, IL6, SLC2A1, ANGPT2, ACAT, L1CAM, TAGLN, HMOX1, FLT1, ANXA, TGFB3, PGF, IGF2, VEGF, IGFBP1, DDIT3, FOS, LRP8, ENPEP, HK2, G22P1, NOS, ADRA, HIF1A, TGFA, ENO1, PKM2, FGF3, HDAC, CDKN1A, ITGA, BNIP3L, ADM, XRCC5, EDN2, MIF, NFKB1, SERPINE, TXN, IGFBP3, COL5A1, F3, JUN, GAPD, PLAUR, TFF3, EPO, CP, HGF, PGK1

6: Science. 1999 Aug 27;285(5432):1390-3.
Gene expression profile of aging and its retardation by caloric restriction.
Lee CK, Klopp RG, Weindruch R, Prolla TA.

#NAME aged_mouse_muscle_dn
#DESCRIPTION Downregulated in the gastrocnemius muscle of aged adult mice (30-month) vs. young adult (5-month)

#GENES: IL6ST, CALM3, SIN3A, GFER, USP4, ABCB4, COL1A2, PRKCSH, PTPRR, POLA2, PLA2G7, HNRPD, COL1A1, PPP1R2, PRSS15, CLTB, FDFT1, PMP22, PSMB8, MYH2, TST, BMP8B, ADAM28, SRPR, PSMC3, CDC2L2, PPP3CC, S100A10, RAI2, NR2F1, PHOX2A, WNT4

#NAME aged_mouse_muscle_up
#DESCRIPTION Upregulated in the gastrocnemius muscle of aged adult mice (30-month) vs. young adult (5-month)
#GENES: GDF9, ARF5, MFAP5, HSPA6, HSPB1, ETV4, TFAP2B, ISLR, GADD45A, CKMT2, USP53, ATF3, ACTR1B, PBEF1, RAB1A, DCTN1, STARD7, DDX5, TM4SF3, U2AF2, SOX17, RAB21, AP3S2, CDC42, PLAGL1, AMY2B, PRSS11, ZFP90, POU3F2, HINT1, TGFB1I1, TGIF, ARHGDIB

APPENDIX 3: Querying NCBI - Example

Querying the NCBI's Entrez Gene:

For each of these, you can use the "Limits" also.

Purpose	Query	Explanation
find genes mapped to *Mus musculus* chromosome 16 that have orthologs reported in HomoloGene	Mus musculus[orgn] AND 16[chr] AND gene_homologene[filter]	• [orgn] is used to restrict to mouse (*Mus musculus*) to the organism field. Alternatively, you can use the "Limits" form to select mouse only. • [chr] is used to restrict '16' to the chromosome field • gene homologene[filter] is used to restrict records to those processed by HomoloGene.
Find genes mapped to human chromosome 16 that have orthologs reported in HomoloGene and also have an OMIM record associated	16[chr] AND gene_homologene[filter] AND gene_OMIM[filter] AND "Homo sapiens"[orgn]	• Same as previous but an additional filter (OMIM) is used.
find all genes in the NCBI database that are derived from genomes other than mammals and are classified by Gene Ontology to have some association to DNA repair	"dna repair"[go] NOT mammalia[orgn]	• [go] is used to restrict to the field 'Genome Ontology' • Quotes ("dna repair") are necessary to treat the two words (dna and repair) together • [orgn] is used to restrict (as the boolean NOT) to species not classified as

		mammals.

Querying the NCBI's OMIM:

1. What human genes are related to diabetes? Which of those genes are on chromosome 1?
    - i. enter: diabetes in the search box
    - ii. select Limits
    - iii. check the box for chromosome 1
    - iv. press Go

OR enter the following query in the search box and it will return the same results:

```
diabetes[All Fields] AND 1[chr]
```

Note that some of the records retrieved will mention hypertension as part of the text of an entry, rather than in the title. That is because Entrez searches All Fields of a record by default. If you would like to limit retrieval only to records that contain the term diabetes in the Title Word field, check the "Search in Field" box for "Title" on the Limits page, in addition to checking the box for Chromosome 1. The query then would be

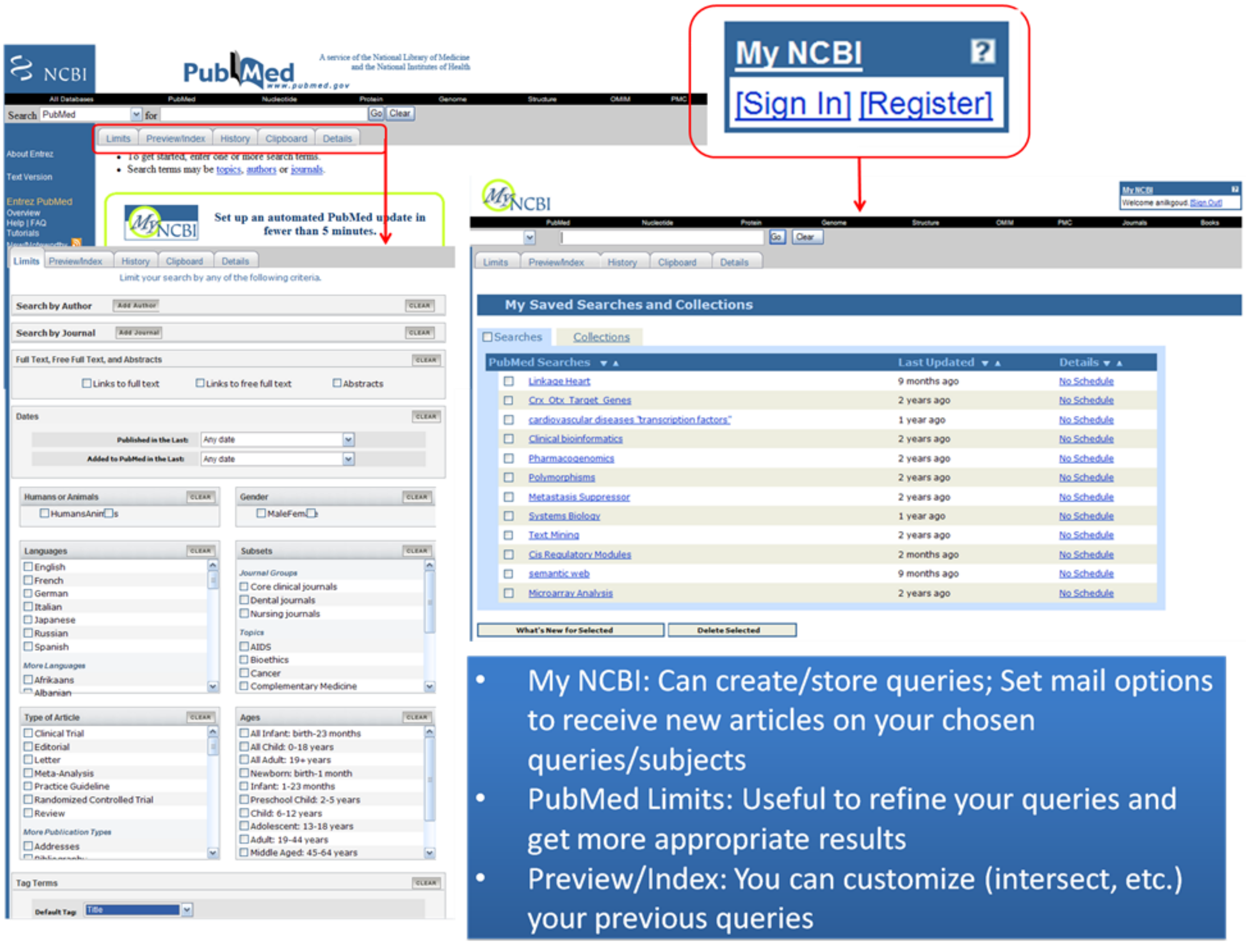
```
diabetes[Title] AND 1[chr]
```

2. List the OMIM entries that describe genes on chromosome 21 and additionally each of which have a clinical synopsis also.

You can do this search in any of these two ways:

- i. Use the Limits page:
    - a. From the OMIM home page, select the Limits option under the search box.
    - b. check the box for chromosome 21.
    - c. Check the box for "clinical synopsis" under the section "Only Records With".
    - d. Press Go.
- ii. Enter the search as a command:
    - a. On the OMIM home page, enter the following in the search box:
    - `b. 21[chr] AND "Clinical Synopsis"[prop]`
    - c. Press Go.

APPENDIX 4: Coping/Keeping up with literature (PubMed) searches



- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.

Set up an automated PubMed update in fewer than 5 minutes.

Limit your search by any of the following criteria.

Search by Author

Search by Journal

Full Text, Free Full Text, and Abstracts

□ Links to full text □ Links to free full text □ Abstracts

Dates

Published in the Last: Any date

Added to PubMed in the Last: Any date

Humans or Animals

□ HumansAnin□s

Gender

□ MaleFem□

Languages

□ English
□ French
□ German
□ Italian
□ Japanese
□ Russian
□ Spanish

More Languages

□ Afrikaans
□ Albanian

Subsets

Journal Groups

□ Core clinical journals
□ Dental journals
□ Nursing journals

Topics

□ AIDS
□ Bioethics
□ Cancer
□ Complementary Medicine

Type of Article

□ Clinical Trial
□ Editorial
□ Letter
□ Meta-Analysis
□ Practice Guideline
□ Randomized Controlled Trial
□ Review

More Publication Types

□ Addresses

Ages

□ All Infant: birth-23 months
□ All Child: 0-18 years
□ All Adult: 19+ years
□ Newborn: birth-1 month
□ Infant: 1-23 months
□ Preschool Child: 2-5 years
□ Child: 6-12 years
□ Adolescent: 13-18 years
□ Adult: 19-44 years
□ Middle Aged: 45-64 years

Tag Terms

Default Tag: Title

Welcome anilkgoud. [Sign Out]

My Saved Searches and Collections

□ Searches Collections

	PubMed Searches ▼ ▲	Last Updated ▼ ▲	Details ▼ ▲
□	Linkage Heart	9 months ago	No Schedule
□	Crx_Otx_Target_Genes	2 years ago	No Schedule
□	cardiovascular diseases "transcription factors"	1 year ago	No Schedule
□	Clinical bioinformatics	2 years ago	No Schedule
□	Pharmacogenomics	2 years ago	No Schedule
□	Polymorphisms	2 years ago	No Schedule
□	Metastasis Suppressor	2 years ago	No Schedule
□	Systems Biology	1 year ago	No Schedule
□	Text Mining	2 years ago	No Schedule
□	Cis Regulatory Modules	2 months ago	No Schedule
□	semantic web	9 months ago	No Schedule
□	Microarray Analysis	2 years ago	No Schedule

- My NCBI: Can create/store queries; Set mail options to receive new articles on your chosen queries/subjects
- PubMed Limits: Useful to refine your queries and get more appropriate results
- Preview/Index: You can customize (intersect, etc.) your previous queries