Comparative Genomics
Gene Regulatory Networks (GRNs)

Anil Jegga
Biomedical Informatics

Contact Information:
Anil Jegga
Biomedical Informatics
Room # 232, S Building 10th Floor
CCHMC
Homepage: http://anil.cchmc.org
Tel: 513-636-0261
E-mail: anil.jegga@cchmc.org

Session 2:
February 24, 2012

Additional exercise available at:
http://anil.cchmc.org/grn.html

Session 1: Overview of GRNs (Feb 23)

a. Computational Approaches
b. Cis-Element Identification
c. Comparative Genomics
d. Regulatory region variations
e. p53 case study

Session 2: Database Session (Feb 24)

a. Genome Browsers
b. Promoter Analysis, TFBS Search
c. Co-regulated gene analysis

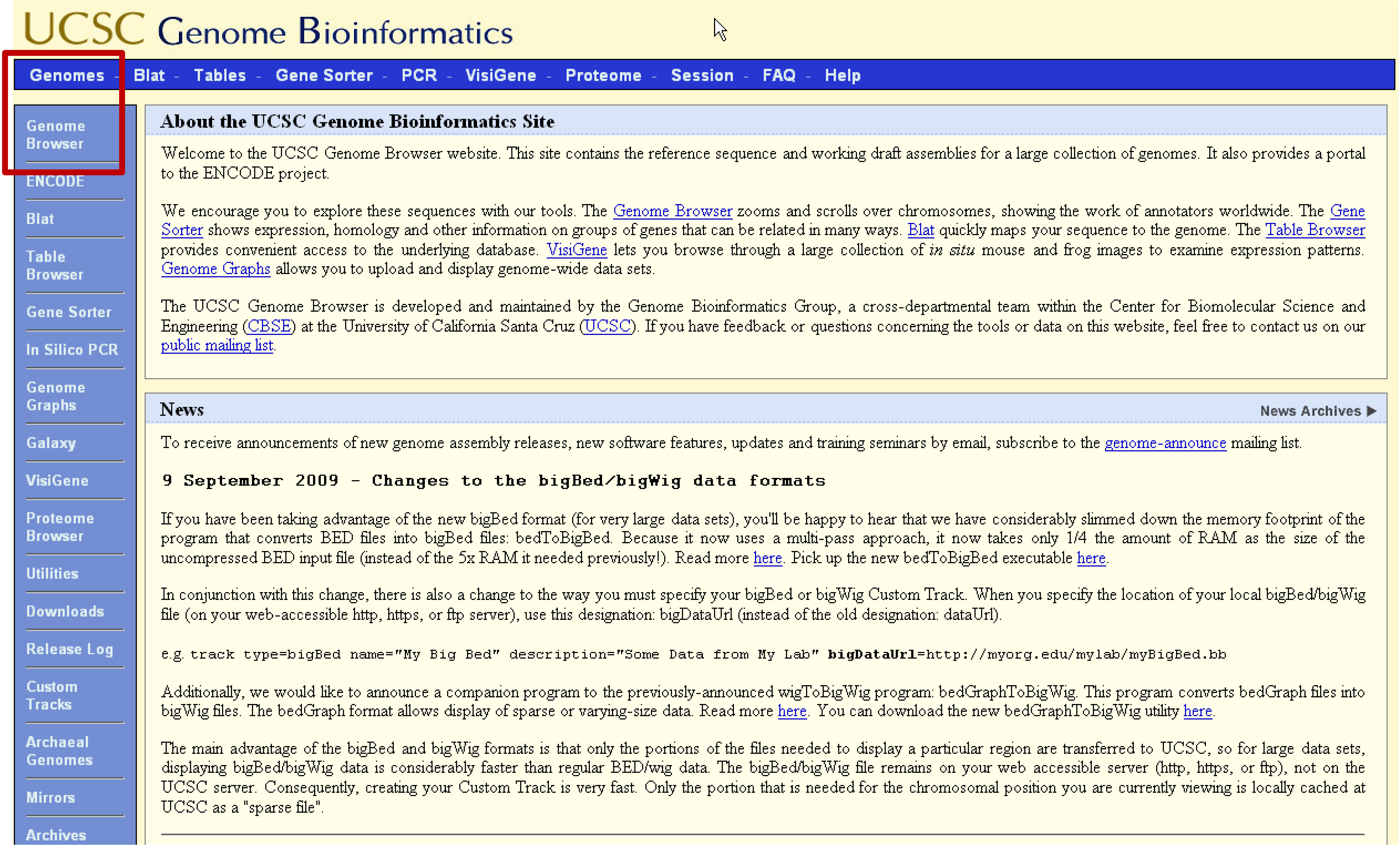
Session 2

(Databases/Servers)

Feb 24, 2012

a. Genome Browsers
b. Promoter Analysis, TFBS Search
c. Co-regulated gene analysis

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



About the UCSC Genome Bioinformatics Site

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

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

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

News

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

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

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

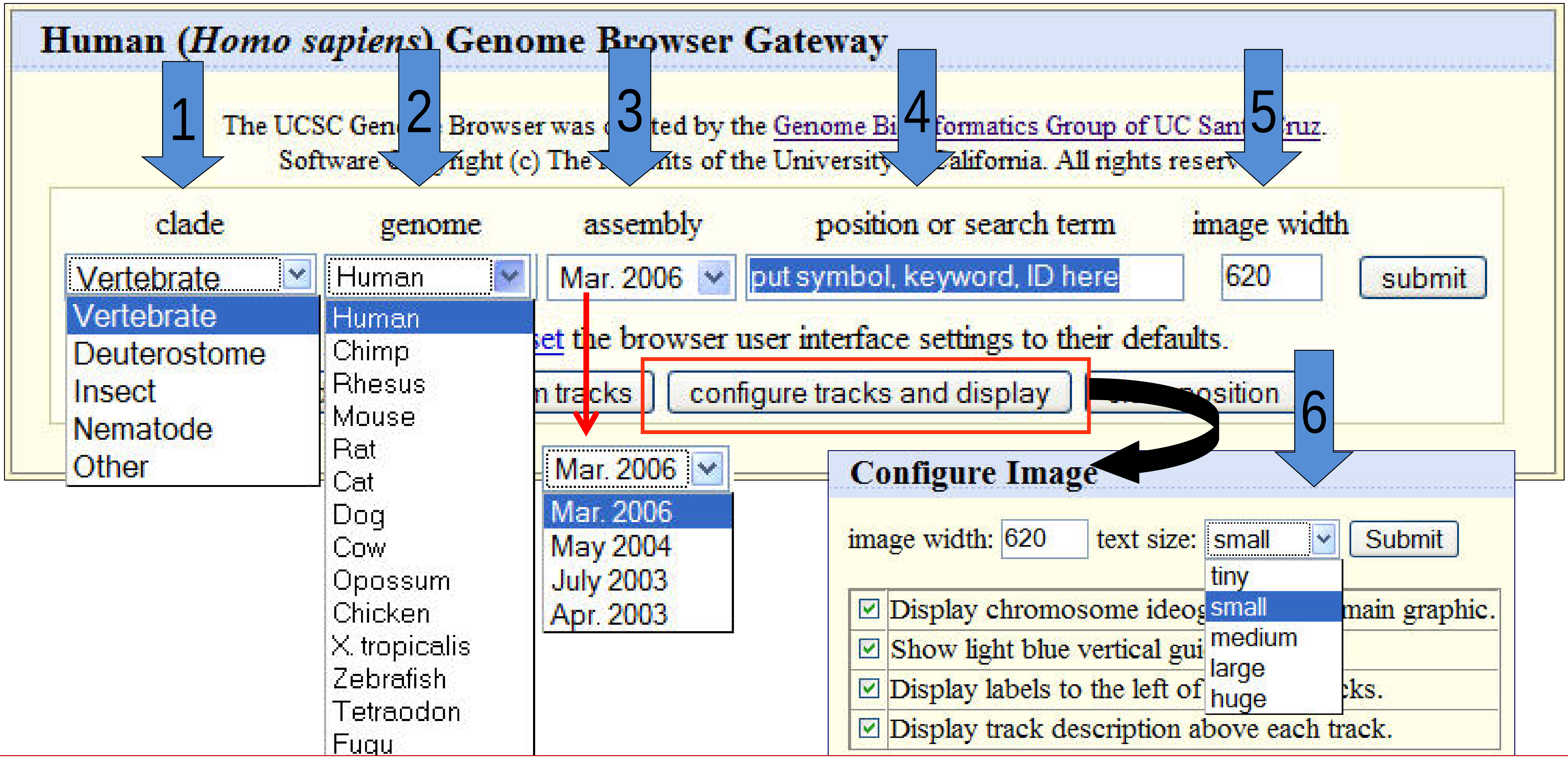
In conjunction with this change, there is also a change to the way you must specify your bigBed or bigWig Custom Track. When you specify the location of your local bigBed/bigWig file (on your web-accessible http, https, or ftp server), use this designation: bigDataUrl (instead of the old designation: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

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

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

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

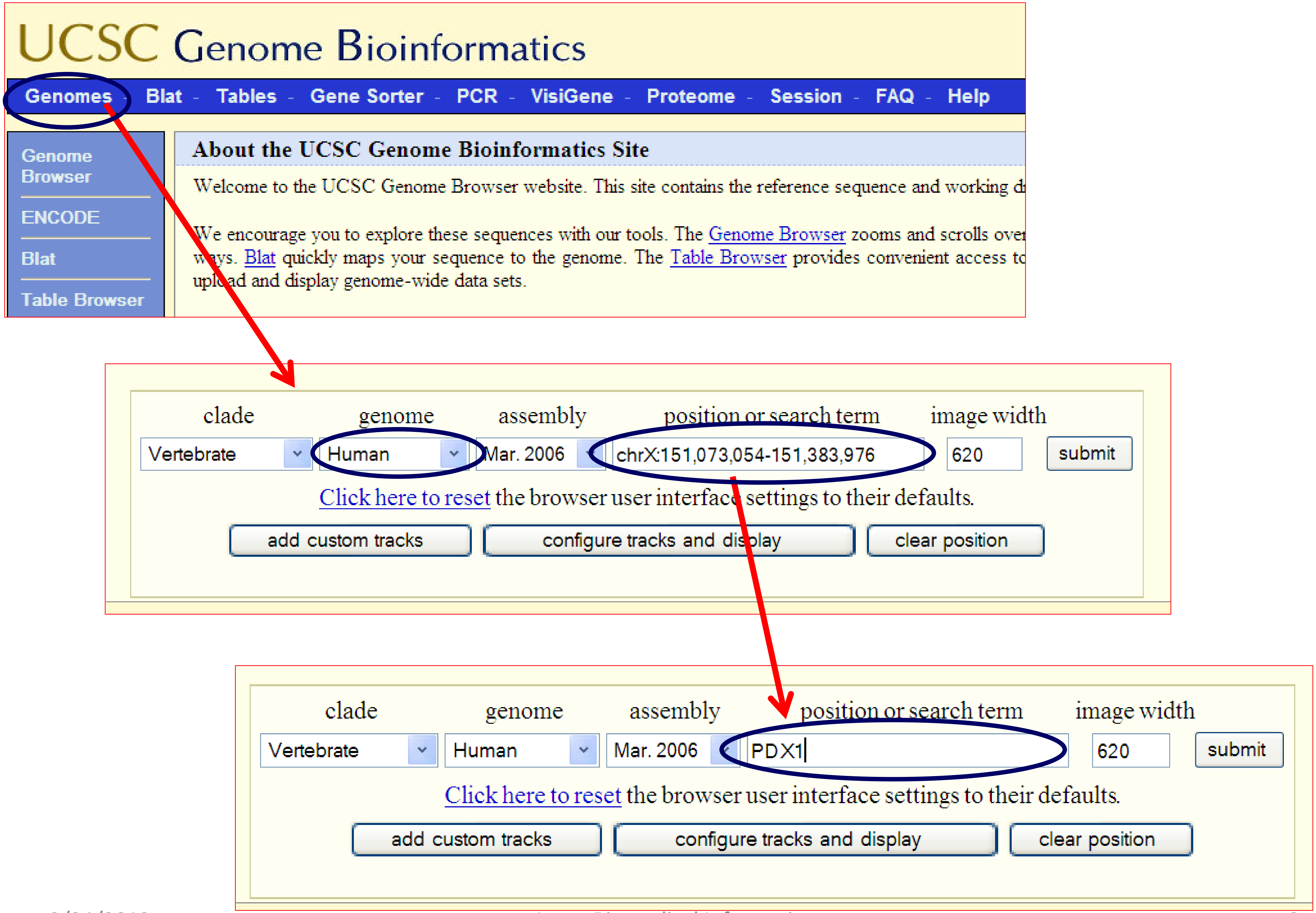
Genome Browser (http://genome.ucsc.edu)



Genome Browser Gateway choices:

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

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



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.

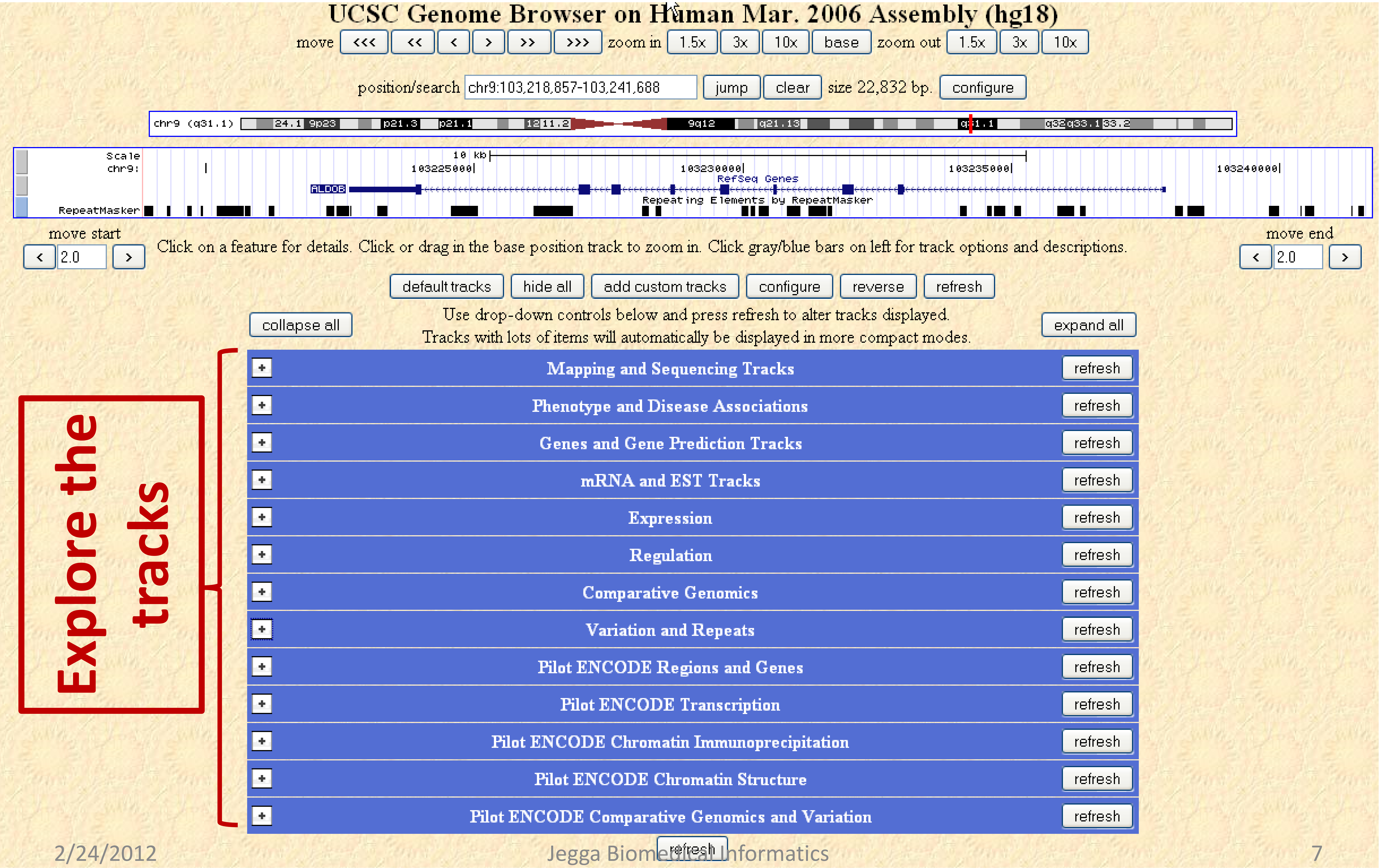
clade	genome	assembly	position or search term	image width
Vertebrate	Human	Mar. 2006	chrX:151,073,054-151,383,976	620

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

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

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

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



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

position/search chr9:103,218,857-103,241,688

size 22,832 bp.

move start 2.0

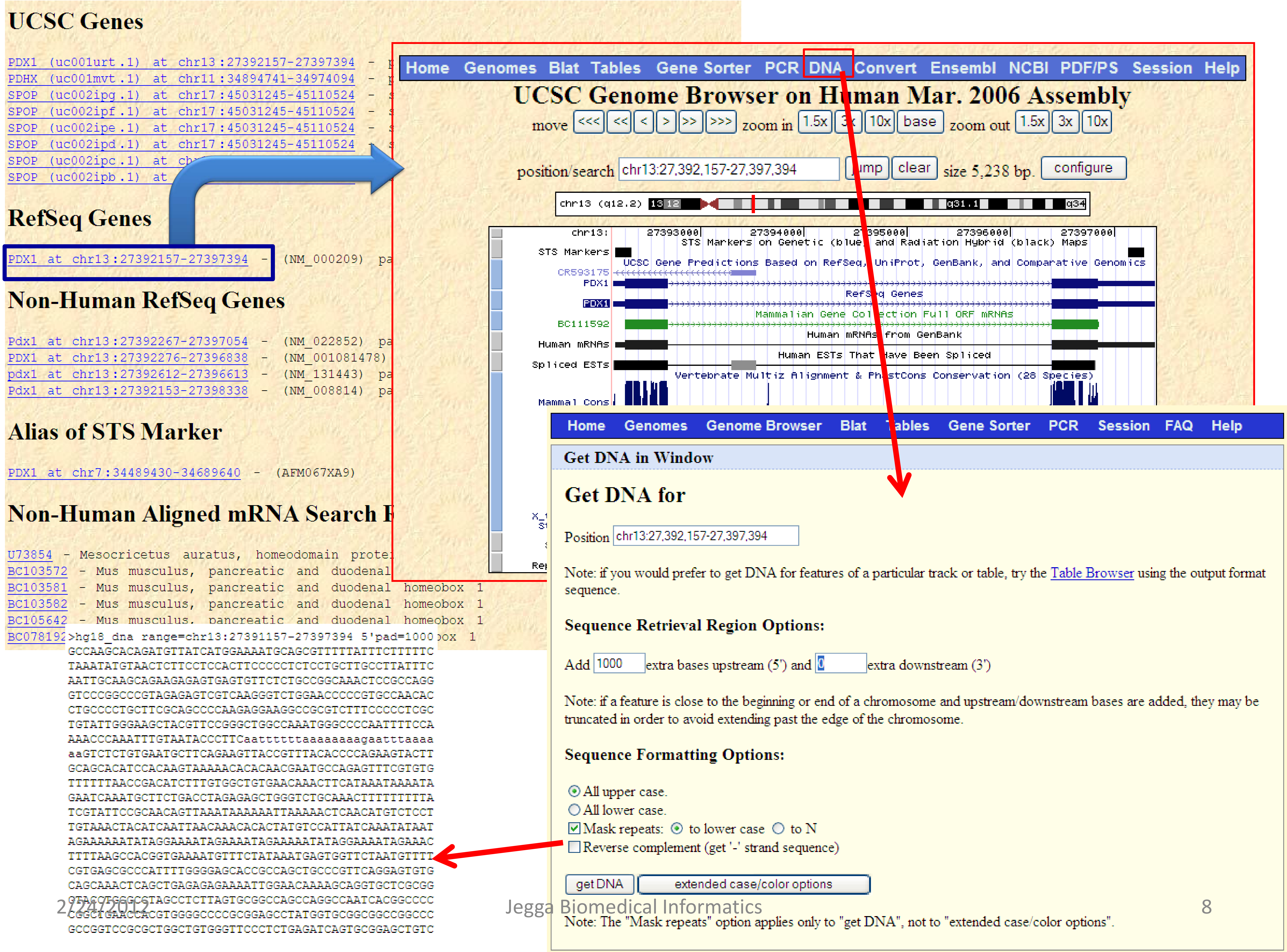
move end 2.0

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

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

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

Explore the tracks



UCSC Genes

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

RefSeq Genes

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

Non-Human RefSeq Genes

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

Alias of STS Marker

PDX1 at chr7:34489430-34689640 - (AFM067XA9)

Non-Human Aligned mRNA Search 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
BC078192 box 1

UCSC Genome Browser on Human Mar. 2006 Assembly

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

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:

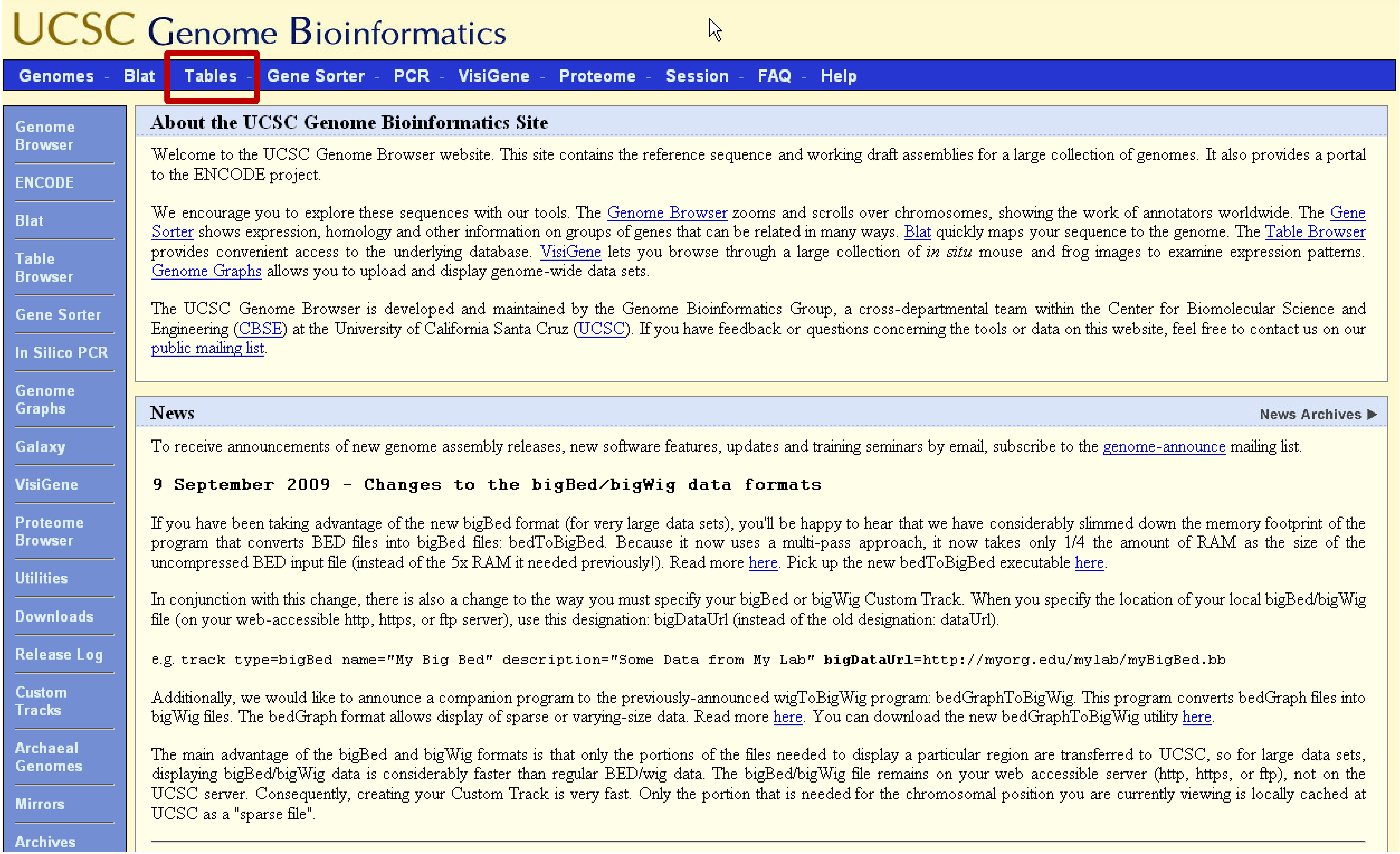
● All upper case.
○ All lower case.
☑ Mask repeats: ● to lower case ○ to N
☐ Reverse complement (get '-' strand sequence)

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

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

What if I want to download promoter sequences for several genes at a time?

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



About the UCSC Genome Bioinformatics Site

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

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

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

News

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

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

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

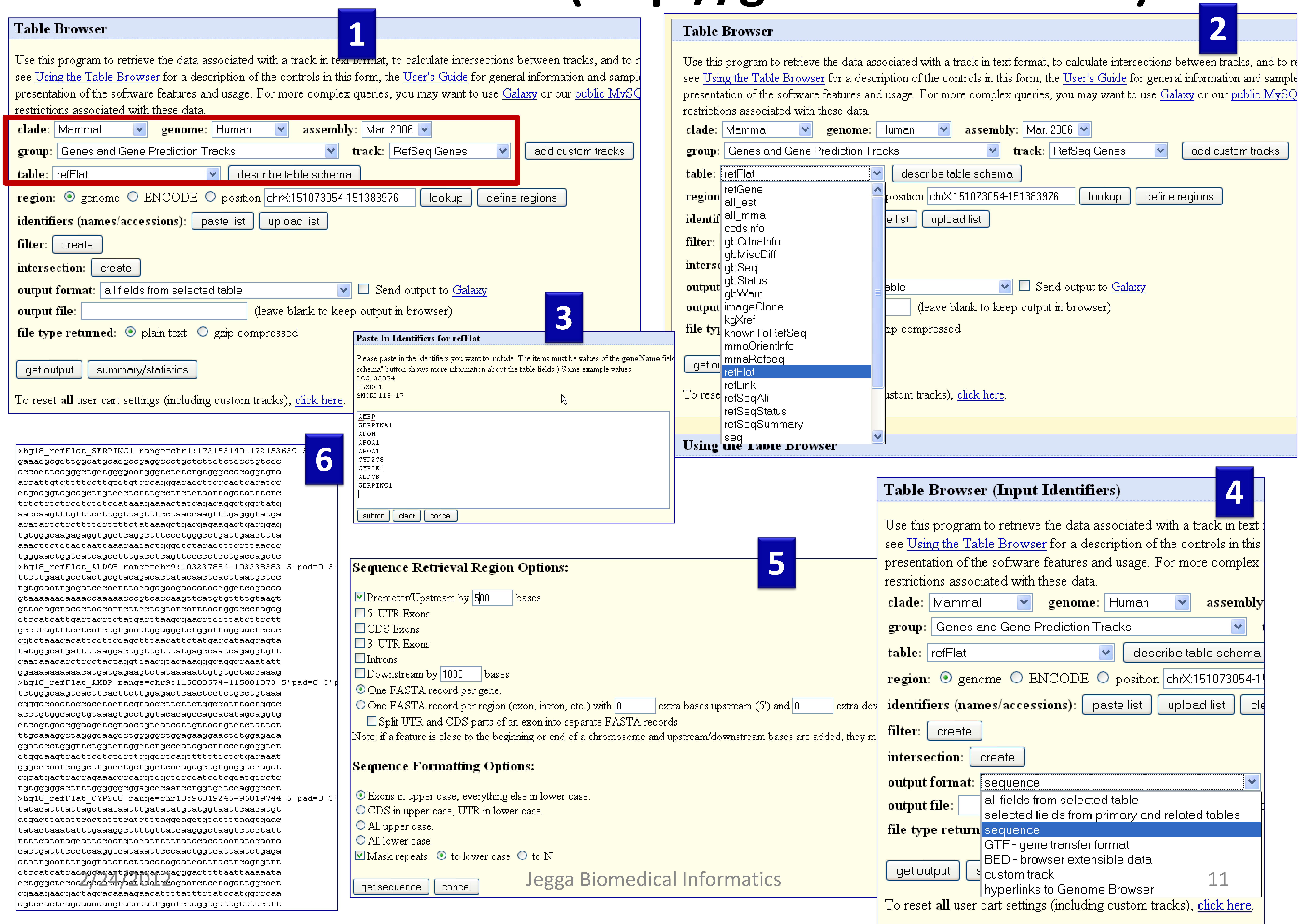
In conjunction with this change, there is also a change to the way you must specify your bigBed or bigWig Custom Track. When you specify the location of your local bigBed/bigWig file (on your web-accessible http, https, or ftp server), use this designation: bigDataUrl (instead of the old designation: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

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

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

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

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



1

Table Browser

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

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

group: Genes and Gene Prediction Tracks **track:** RefSeq Genes

table: refFlat

region: ● genome ○ ENCODE ○ position chrX:151073054-151383976

identifiers (names/accessions):

filter:

intersection:

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

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

2

Table Browser

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

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

group: Genes and Gene Prediction Tracks **track:** RefSeq Genes

table: refFlat

region position chrX:151073054-151383976

identif

filter:

interse

output able ☐ Send output to Galaxy

output (leave blank to keep output in browser)

file ty zip compressed

To rese ustom tracks), click here.

Using the Table Browser

3

Paste In Identifiers for refFlat

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

```
LOC133874
PLXDC1
SNORD115-17
```

```
AMBP
SERPINA1
APOH
APOA1
APOA1
CYP2C8
CYP2E1
ALDOB
SERPINC1
```

4

Table Browser (Input Identifiers)

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

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

group: Genes and Gene Prediction Tracks

table: refFlat

region: ● genome ○ ENCODE ○ position chrX:151073054-15

identifiers (names/accessions):

filter:

intersection:

output format: sequence

output file:

file type return

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

5

Sequence Retrieval Region Options:

☑ Promoter/Upstream by 500 bases

☐ 5' UTR Exons

☐ CDS Exons

☐ 3' UTR Exons

☐ Introns

☐ Downstream by 1000 bases

● One FASTA record per gene.

○ One FASTA record per region (exon, intron, etc.) with 0 extra bases upstream (5') and 0 extra dov

☐ Split UTR and CDS parts of an exon into separate FASTA records

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

Sequence Formatting Options:

● Exons in upper case, everything else in lower case.

○ CDS in upper case, UTR in lower case.

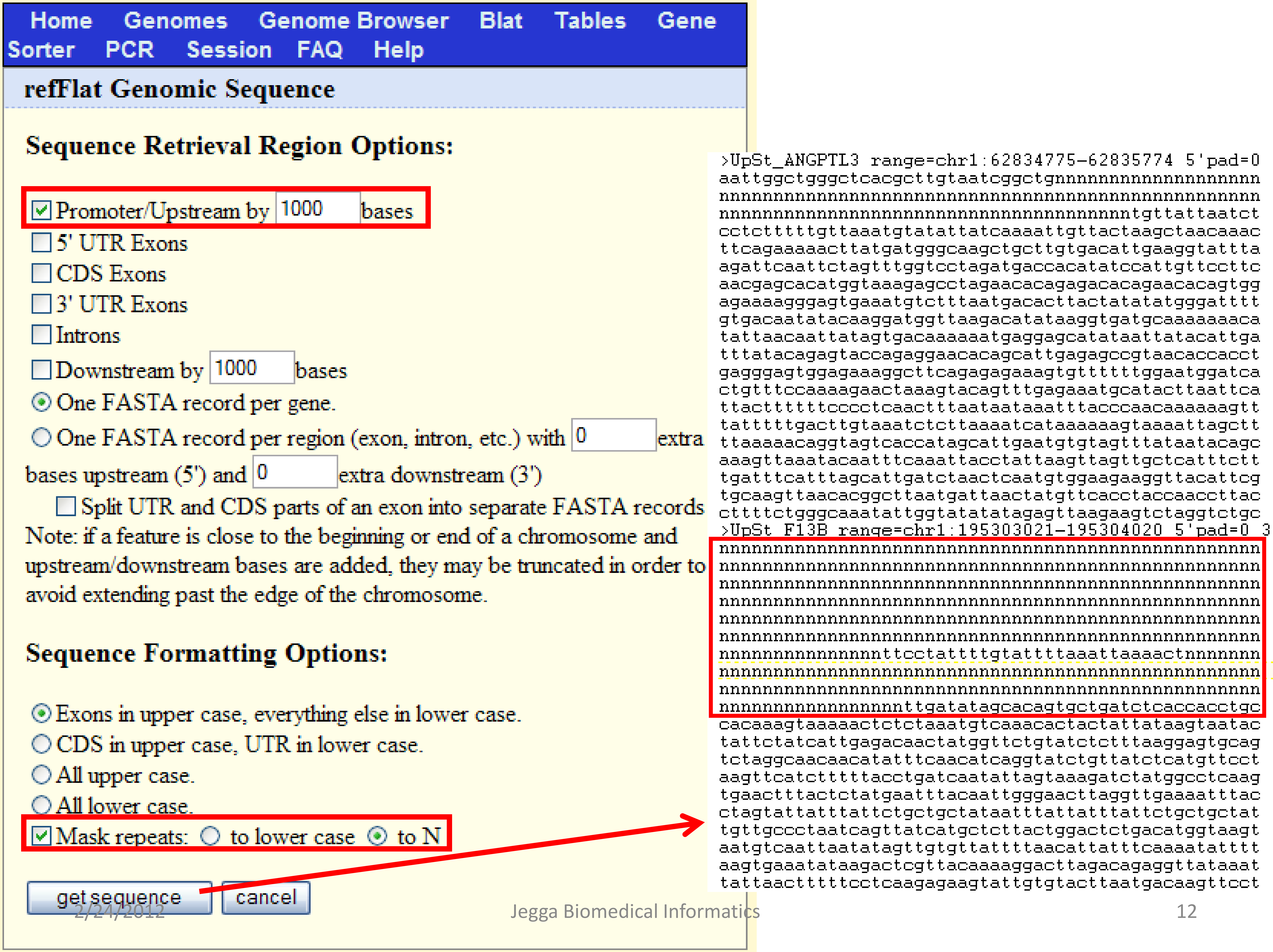
○ All upper case.

○ All lower case.

☑ Mask repeats: ● to lower case ○ to N

6

```
>hg18_refFlat_SERPINC1 range=chr1:172153140-172153639 5
gaaacgcgcttggcatgcaccccgaggccctgctcttctctccctgtccc
accacttcagggctgctggggaatgggtctctctgtgggccacaggtgta
accattgtgttttccttgtctgtgccagggacaccttggcactcagatgc
ctgaaggtagcagcttgtccctctttgccttctctaattagatatttctc
tctctctctccctctctccataaagaaaactatgagagagggtgggtatg
aaccaagtttgtttccttggttagtttcctaaccaagtttgagggtatga
acatactctccttttccttttctataaagctgaggagaagagtgagggag
tgtgggcaagagaggtggctcaggctttccctgggcctgattgaacttta
aaacttctctactaattaaacaacactgggctctacactttgcttaaccc
tgggaactggtcatcagcctttgacctcagttccccctcctgaccagctc
>hg18_refFlat_ALDOB range=chr9:103237884-103238383 5'pad=0 3'
ttcttgaatgcctactgcgtacagacactatacaactcacttaatgctcc
tgtgaaattgagatcccactttacagagaagaaaataacggctcagacaa
gtaaaaaacaaaaccaaaaacccgtcaccaagttcatgtgttttgtaagt
gttacagctacactaacattcttcctagtatcatttaatggaccctagag
ctccatcattgactagctgtatgacttaagggaacctccttatcttcctt
gccttagtttcctcatctgtgaaatggagggtctggattaggaactccac
ggtctaaagacattccttgcagctttaacattctatgagcataaggagta
tatgggcatgatttaaggactggttgtttatgagccaatcagaggtgtt
gaataaacacctccctactaggtcaaggtagaaaggggagggcaaatatt
ggaaaaaaaaaacatgatgagaagtctataaaaattgtgtgctaccaaag
>hg18_refFlat_AMBP range=chr9:115880574-115881073 5'pad=0 3'p
tctgggcaagtcacttcacttcttggagactcaactcctctgcctgtaaa
ggggacaaatagcacctacttcgtaagcttgttgtggggatttactggac
acctgtggcacgtgtaaagtgcctggtacacagccagcacatagcaggtg
ctcagtgaacggaagctcgtaacagtcatcattgttaatgtctctattat
ttgcaaaggctagggcaagcctgggggctggagaaggaactctggagaca
ggatacctgggttctggtcttggctctgcccatagacttccctgaggtct
ctggcaagtcacttcctctccttgggcctcagtttttcctgtgagaaat
gggcccaatcaggcttgacctgctggctcacagagctgtgaggtccagat
ggcatgactcagcagaaaggccaggtcgctccccatcctcgcatgccctc
tgtgggggacttttgggggcggagcccaatcctggtgctccagggccct
>hg18_refFlat_CYP2C8 range=chr10:96819245-96819744 5'pad=0 3'
tatacatttattagctaataatttgatatatgtatggtaattcaacatgt
atgagttatattcactatttcatgtttaggcagctgtattttaagtgaac
tatactaaatatttgaaaggcttttgttatcaagggctaagtctcctatt
ttttgatatagcattacaatgtacatttttatacacaaaatatagaata
cactgatttccctcaaggtcataaattcccaactggtcattaatctgaga
atattgaattttgagtatattctaacatagaatcatttacttcagtgttt
ctccatcatcacagcacattggaacaaccagggactttaattaaaaata
cctgggctccaa[illegible]agaatctcctagattggcact
ggaaagaaggagtaggacaaaagaacattttatttctatccatgggccaa
agtccactcagaaaaaaagtataaattggatctaggtgattgtttacttt
```



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

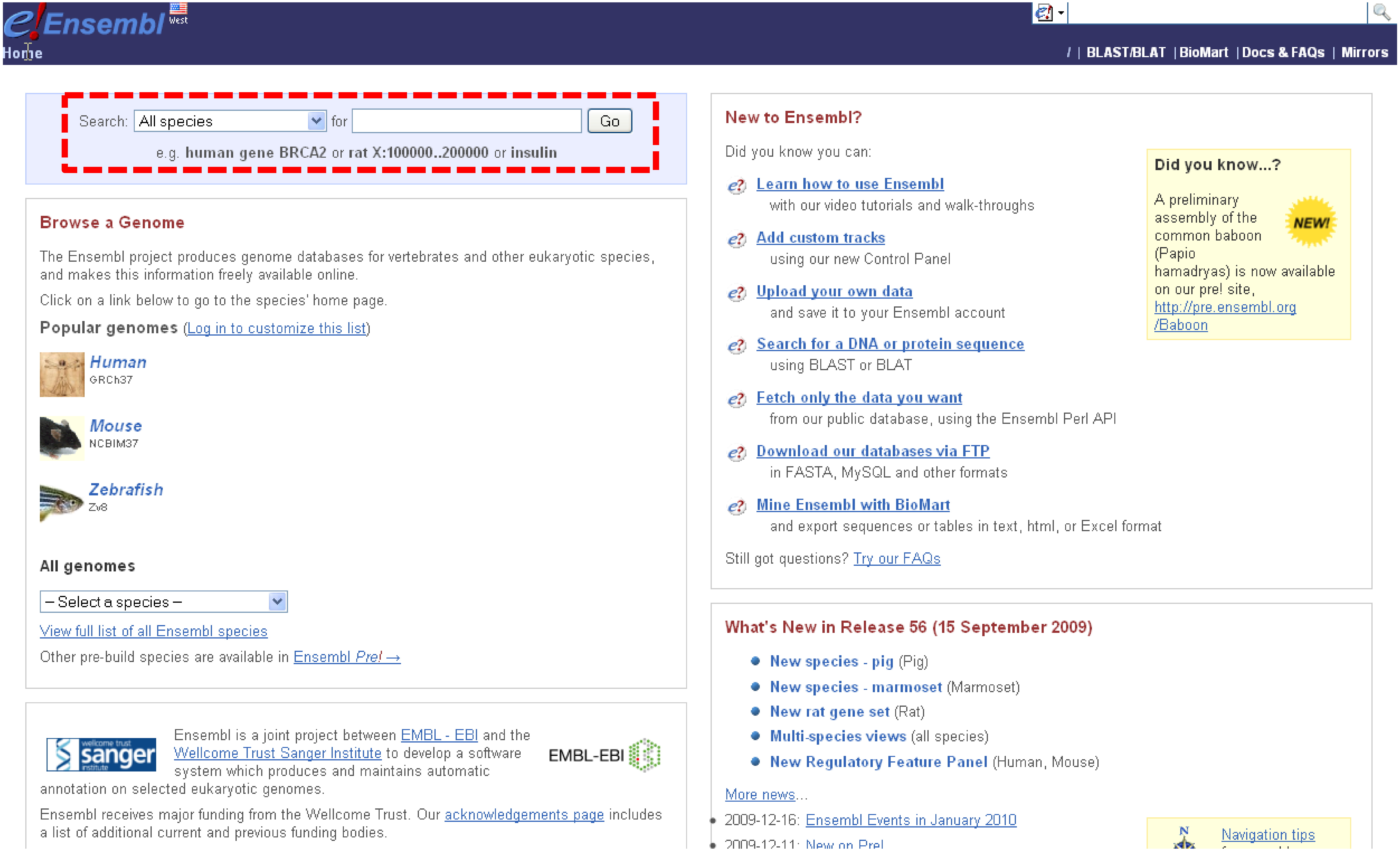
Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

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
cctcttttgttaaatgtatattatcaaattgttactaagctaacaaac
ttcagaaaacttatgatgggcaagctgcttgtgacattgaaggtattta
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
nnnnnnnnnnnnnnnttcctatttgtatttaaattaaactnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnttgatatagcacagtgctgatctcaccacctgc
cacaaagtaaaactctctaaatgtcaaacactactattataagtaatac
tattctatcattgagacaactatggttctgtatctctttaaggagtgcag
tctaggcaacaacatatttcaacatcaggtatctgttatctcatgttcct
aagttcatcttttttacctgatcaatattagtaaagatctatggcctcaag
tgaactttactctatgaatttacaattgggaacttaggttgaaaatttac
ctagtattatttattctgctgctataatttattatttattctgctgctat
tgttgccctaatcagttatcatgctcttactggactctgacatggtaagt
aatgtcaattaatatagttgtgttattttaacattatttcaaaatatttt
aagtgaaatataagactcgttacaaaggacttagacagaggttataaat
tattaacttttttcctcaagagaagtattgtgtacttaatgacaagttcct
```

Other Genome Browsers: ENSEMBL
http://www.ensembl.org



Search: All species for Go

e.g. **human gene BRCA2** or **rat X:100000..200000** or **insulin**

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

- *Human* GRCh37
- *Mouse* NCBIM37
- *Zebrafish* Zv8

All genomes

– Select a species –

View full list of all Ensembl species

Other pre-build species are available in Ensembl *Pre!* →

Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies.

New to Ensembl?

Did you know you can:

- Learn how to use Ensembl with our video tutorials and walk-throughs
- Add custom tracks using our new Control Panel
- Upload your own data and save it to your Ensembl account
- Search for a DNA or protein sequence using BLAST or BLAT
- Fetch only the data you want from our public database, using the Ensembl Perl API
- Download our databases via FTP in FASTA, MySQL and other formats
- Mine Ensembl with BioMart and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs

Did you know...?

A preliminary assembly of the common baboon (Papio hamadryas) is now available on our pre! site, http://pre.ensembl.org/Baboon

What's New in Release 56 (15 September 2009)

- **New species - pig** (Pig)
- **New species - marmoset** (Marmoset)
- **New rat gene set** (Rat)
- **Multi-species views** (all species)
- **New Regulatory Feature Panel** (Human, Mouse)

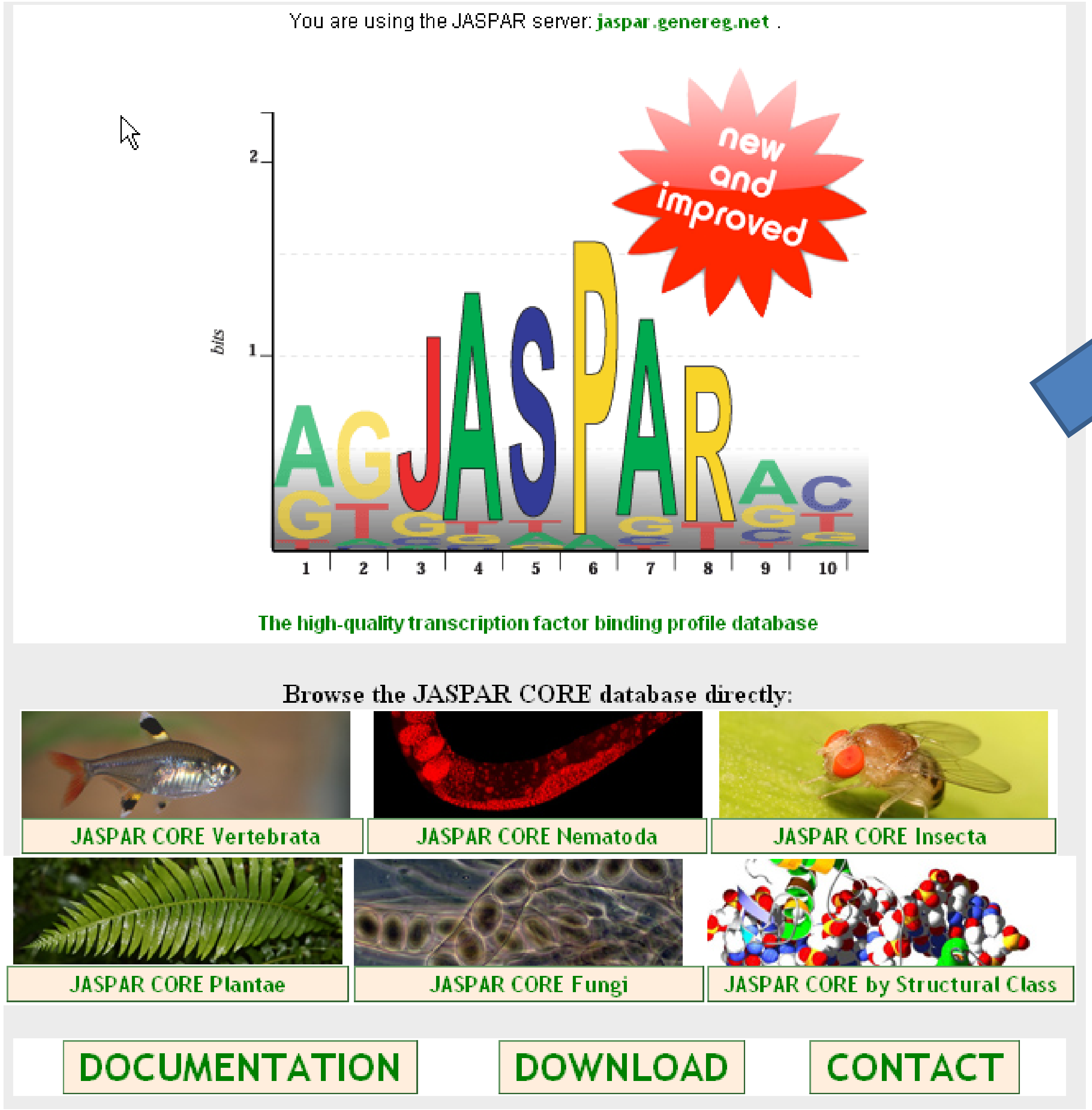
More news...

- 2009-12-16: Ensembl Events in January 2010
- 2009-12-11: New on Pre!

Navigation tips

I have a promoter sequence and how do I scan it for known TFBSs?

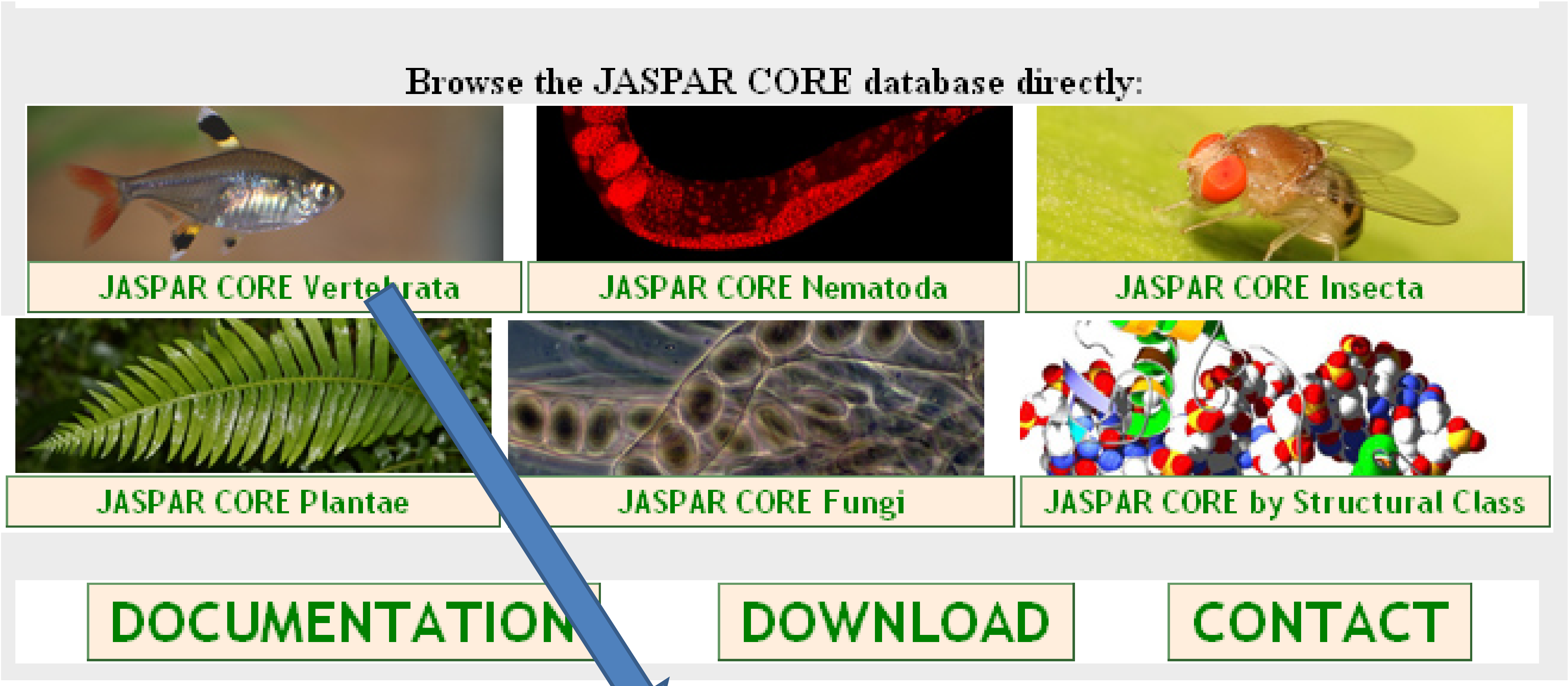
JASPAR: http://jaspar.genereg.net



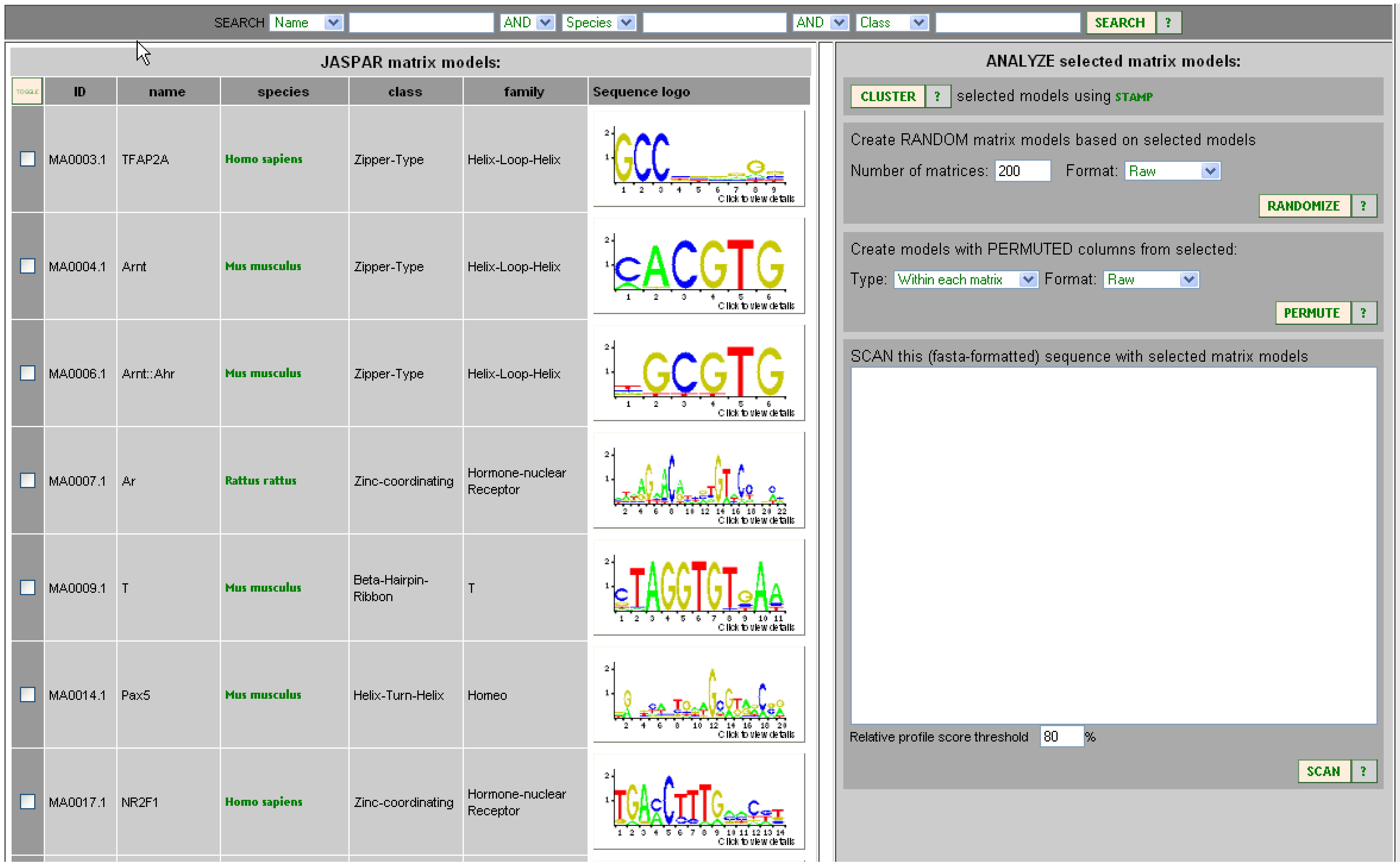
You are using the JASPAR server: jaspar.genereg.net .

The high-quality transcription factor binding profile database

Browse the JASPAR CORE database directly:



Browse the JASPAR CORE database directly:



JASPAR matrix models:

TOGGLE	ID	name	species	class	family	Sequence logo
☐	MA0003.1	TFAP2A	**Homo sapiens**	Zipper-Type	Helix-Loop-Helix	Click to view details
☐	MA0004.1	Arnt	**Mus musculus**	Zipper-Type	Helix-Loop-Helix	Click to view details
☐	MA0006.1	Arnt::Ahr	**Mus musculus**	Zipper-Type	Helix-Loop-Helix	Click to view details
☐	MA0007.1	Ar	**Rattus rattus**	Zinc-coordinating	Hormone-nuclear Receptor	Click to view details
☐	MA0009.1	T	**Mus musculus**	Beta-Hairpin-Ribbon	T	Click to view details
☐	MA0014.1	Pax5	**Mus musculus**	Helix-Turn-Helix	Homeo	Click to view details
☐	MA0017.1	NR2F1	**Homo sapiens**	Zinc-coordinating	Hormone-nuclear Receptor	Click to view details

ANALYZE selected matrix models:

CLUSTER ? selected models using STAMP

Create RANDOM matrix models based on selected models

Number of matrices: 200 Format: Raw

RANDOMIZE ?

Create models with PERMUTED columns from selected:

Type: Within each matrix Format: Raw

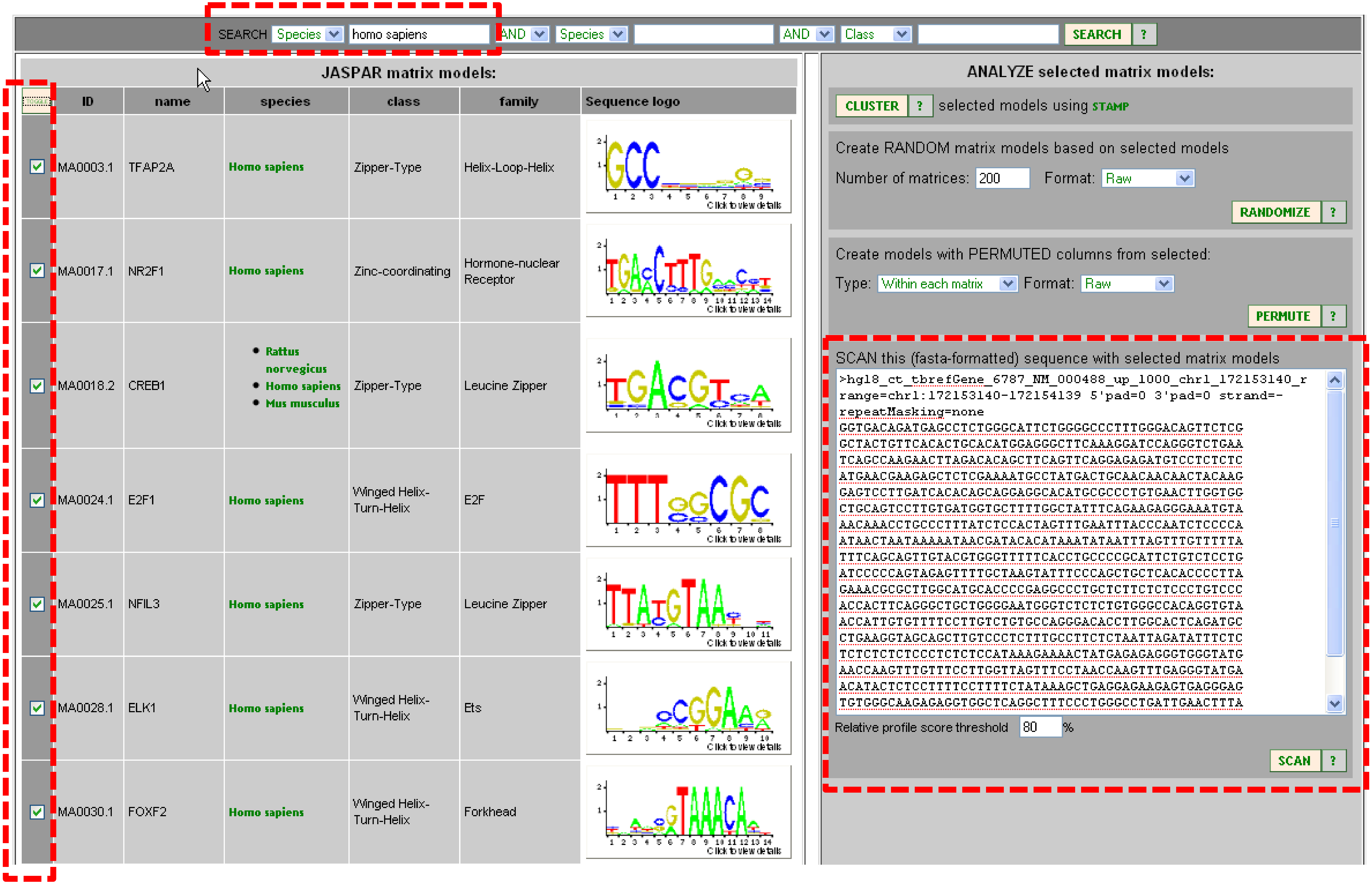
PERMUTE ?

SCAN this (fasta-formatted) sequence with selected matrix models

Relative profile score threshold 80 %

SCAN ?

JASPAR: http://jaspar.genereg.net



SEARCH Species homo sapiens AND Species AND Class

JASPAR matrix models:

	ID	name	species	class	family	Sequence logo
☑	MA0003.1	TFAP2A	**Homo sapiens**	Zipper-Type	Helix-Loop-Helix	Click to view details
☑	MA0017.1	NR2F1	**Homo sapiens**	Zinc-coordinating	Hormone-nuclear Receptor	Click to view details
☑	MA0018.2	CREB1	• **Rattus norvegicus** • **Homo sapiens** • **Mus musculus**	Zipper-Type	Leucine Zipper	Click to view details
☑	MA0024.1	E2F1	**Homo sapiens**	Winged Helix-Turn-Helix	E2F	Click to view details
☑	MA0025.1	NFIL3	**Homo sapiens**	Zipper-Type	Leucine Zipper	Click to view details
☑	MA0028.1	ELK1	**Homo sapiens**	Winged Helix-Turn-Helix	Ets	Click to view details
☑	MA0030.1	FOXF2	**Homo sapiens**	Winged Helix-Turn-Helix	Forkhead	Click to view details

ANALYZE selected matrix models:

CLUSTER selected models using STAMP

Create RANDOM matrix models based on selected models

Number of matrices: 200 Format: Raw

Create models with PERMUTED columns from selected:

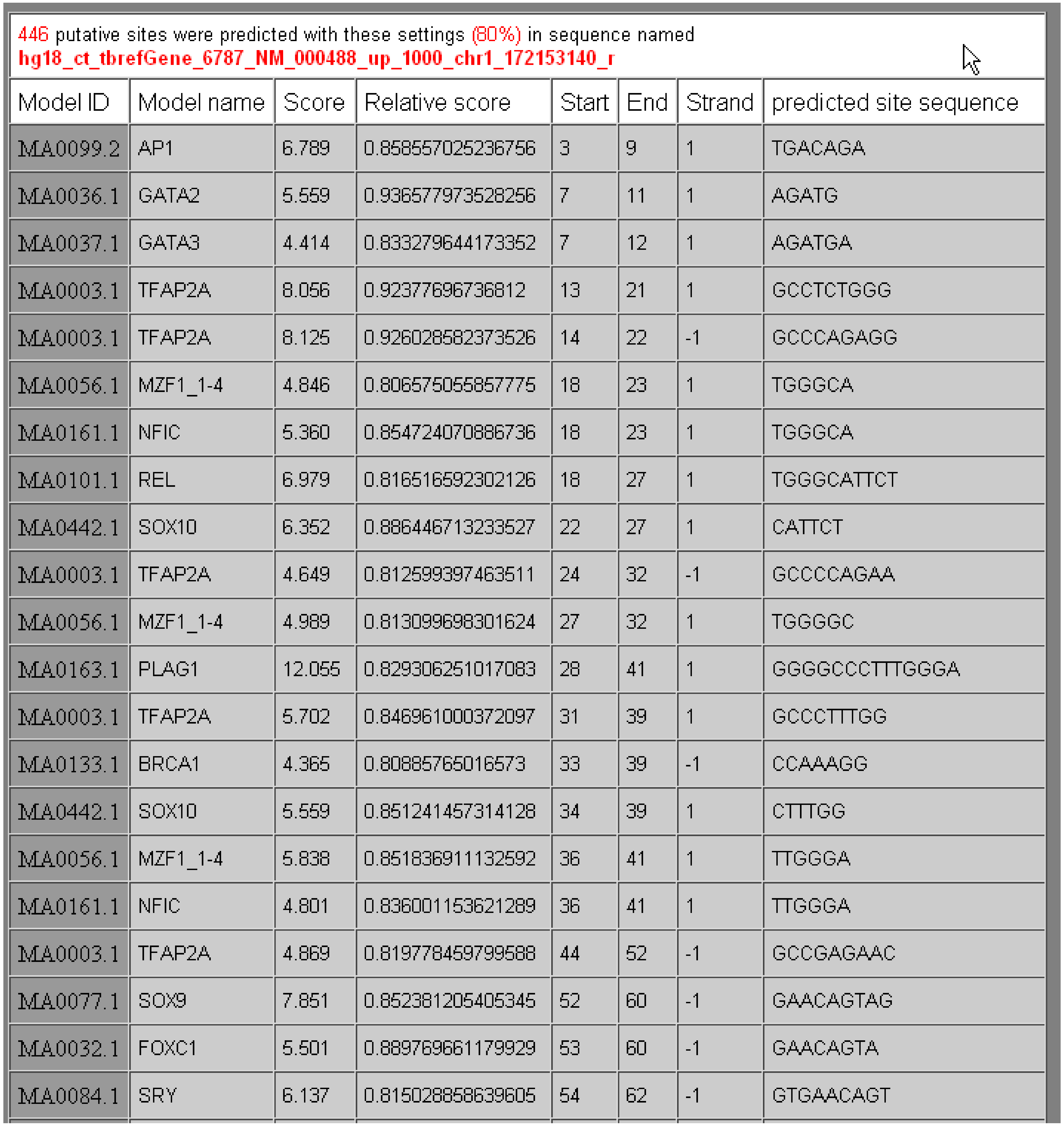
Type: Within each matrix Format: Raw

SCAN this (fasta-formatted) sequence with selected matrix models

```
>hg18_ct_tbrefGene_6787_NM_000488_up_1000_chr1_172153140_r
range=chr1:172153140-172154139 5'pad=0 3'pad=0 strand=-
repeatMasking=none
GGTGACAGATGAGCCTCTGGGCATTCTGGGGCCCTTTGGGACAGTTCTCG
GCTACTGTTCACACTGCACATGGAGGGCTTCAAAGGATCCAGGGTCTGAA
TCAGCCAAGAACTTAGACACAGCTTCAGTTCAGGAGAGATGTCCTCTCTC
ATGAACGAAGAGCTCTCGAAAATGCCTATGACTGCAACAACAACTACAAG
GAGTCCTTGATCACACAGCAGGAGGCACATGCGCCCTGTGAACTTGGTGG
CTGCAGTCCTTGTGATGGTGCTTTTGGCTATTTCAGAAGAGGGAAATGTA
AACAAACCTGCCCTTTATCTCCACTAGTTTGAATTTACCCAATCTCCCCA
ATAACTAATAAAAATAACGATACACATAAATATAATTTAGTTTGTTTTTA
TTTCAGCAGTTGTACGTGGGTTTTTCACCTGCCCCGCATTCTGTCTCCTG
ATCCCCCAGTAGAGTTTTGCTAAGTATTTCCCAGCTGCTCACACCCCTTA
GAAACGCGCTTGGCATGCACCCCGAGGCCCTGCTCTTCTCTCCCTGTCCC
ACCACTTCAGGGCTGCTGGGGAATGGGTCTCTCTGTGGGCCACAGGTGTA
ACCATTGTGTTTTCCTTGTCTGTGCCAGGGACACCTTGGCACTCAGATGC
CTGAAGGTAGCAGCTTGTCCCTCTTTGCCTTCTCTAATTAGATATTTCTC
TCTCTCTCTCCCTCTCTCCATAAAGAAAACTATGAGAGAGGGTGGGTATG
AACCAAGTTTGTTTCCTTGGTTAGTTTCCTAACCAAGTTTGAGGGTATGA
ACATACTCTCCTTTTCCTTTTCTATAAAGCTGAGGAGAAGAGTGAGGGAG
TGTGGGCAAGAGAGGTGGCTCAGGCTTTCCCTGGGCCTGATTGAACTTTA
```

Relative profile score threshold 80 %

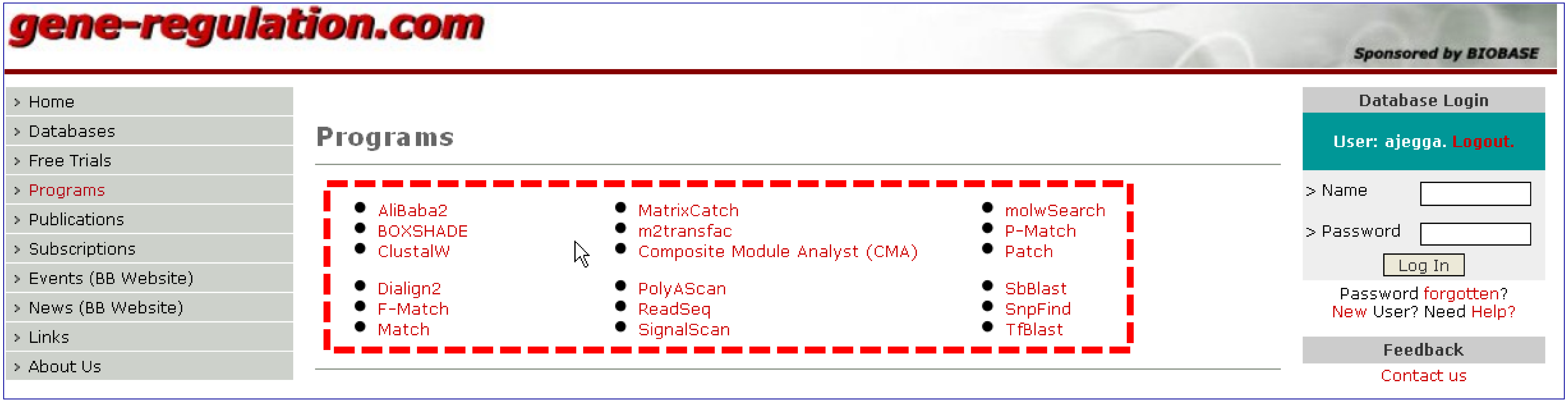
JASPAR: http://jaspar.genereg.net



446 putative sites were predicted with these settings (80%) in sequence named
hg18_ct_tbrefGene_6787_NM_000488_up_1000_chr1_172153140_r

Model ID	Model name	Score	Relative score	Start	End	Strand	predicted site sequence
MA0099.2	AP1	6.789	0.858557025236756	3	9	1	TGACAGA
MA0036.1	GATA2	5.559	0.936577973528256	7	11	1	AGATG
MA0037.1	GATA3	4.414	0.833279644173352	7	12	1	AGATGA
MA0003.1	TFAP2A	8.056	0.92377696736812	13	21	1	GCCTCTGGG
MA0003.1	TFAP2A	8.125	0.926028582373526	14	22	-1	GCCCAGAGG
MA0056.1	MZF1_1-4	4.846	0.806575055857775	18	23	1	TGGGCA
MA0161.1	NFIC	5.360	0.854724070886736	18	23	1	TGGGCA
MA0101.1	REL	6.979	0.816516592302126	18	27	1	TGGGCATTCT
MA0442.1	SOX10	6.352	0.886446713233527	22	27	1	CATTCT
MA0003.1	TFAP2A	4.649	0.812599397463511	24	32	-1	GCCCCAGAA
MA0056.1	MZF1_1-4	4.989	0.813099698301624	27	32	1	TGGGGC
MA0163.1	PLAG1	12.055	0.829306251017083	28	41	1	GGGGCCCTTTGGGA
MA0003.1	TFAP2A	5.702	0.846961000372097	31	39	1	GCCCTTTGG
MA0133.1	BRCA1	4.365	0.80885765016573	33	39	-1	CCAAAGG
MA0442.1	SOX10	5.559	0.851241457314128	34	39	1	CTTTGG
MA0056.1	MZF1_1-4	5.838	0.851836911132592	36	41	1	TTGGGA
MA0161.1	NFIC	4.801	0.836001153621289	36	41	1	TTGGGA
MA0003.1	TFAP2A	4.869	0.819778459799588	44	52	-1	GCCGAGAAC
MA0077.1	SOX9	7.851	0.852381205405345	52	60	-1	GAACAGTAG
MA0032.1	FOXC1	5.501	0.889769661179929	53	60	-1	GAACAGTA
MA0084.1	SRY	6.137	0.815028858639605	54	62	-1	GTGAACAGT

Gene-Regulation: http://www.gene-regulation.com



gene-regulation.com

Sponsored by BIOBASE

- Home
- Databases
- Free Trials
- Programs
- Publications
- Subscriptions
- Events (BB Website)
- News (BB Website)
- Links
- About Us

Programs

- AliBaba2
- BOXSHADE
- ClustalW
- Dialign2
- F-Match
- Match
- MatrixCatch
- m2transfac
- Composite Module Analyst (CMA)
- PolyAScan
- ReadSeq
- SignalScan
- molwSearch
- P-Match
- Patch
- SbBlast
- SnpFind
- TfBlast

Database Login

User: ajegga. Logout.

> Name

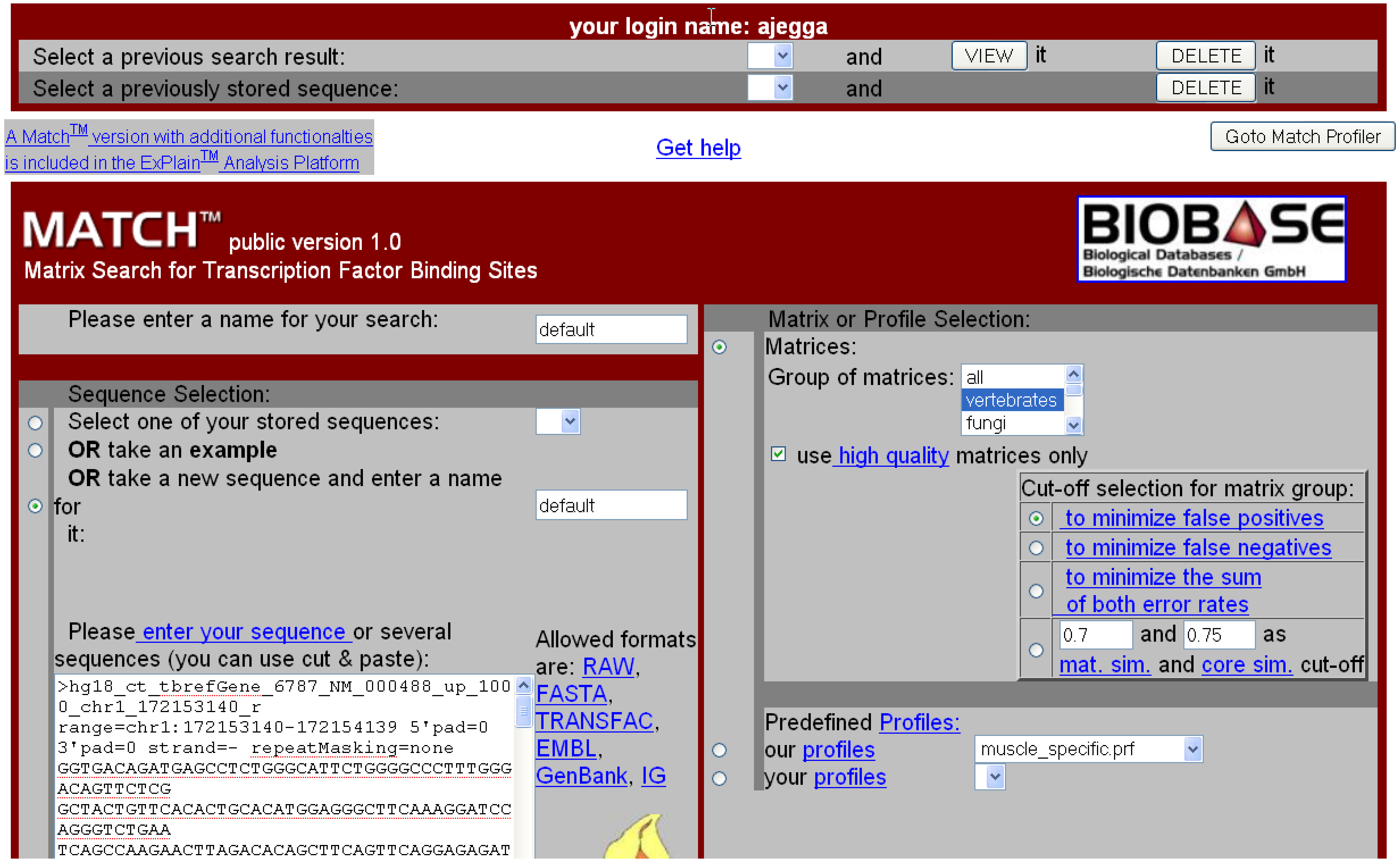
> Password

Log In

Password forgotten?
New User? Need Help?

Feedback

Contact us



your login name: ajegga

Select a previous search result: and VIEW it DELETE it

Select a previously stored sequence: and DELETE it

A Match^{TM} version with additional functionalities is included in the ExPlain^{TM} Analysis Platform

Get help

Goto Match Profiler

MATCH™ public version 1.0
Matrix Search for Transcription Factor Binding Sites

BIOBASE Biological Databases / Biologische Datenbanken GmbH

Please enter a name for your search: default

Sequence Selection:

○ Select one of your stored sequences:
○ OR take an example
● OR take a new sequence and enter a name for it: default

Please enter your sequence or several sequences (you can use cut & paste):

```
>hg18_ct_tbrefGene_6787_NM_000488_up_100
0_chr1_172153140_r
range=chr1:172153140-172154139 5'pad=0
3'pad=0 strand=- repeatMasking=none
GGTGACAGATGAGCCTCTGGGCATTCTGGGGCCCTTTGGG
ACAGTTCTCG
GCTACTGTTCACACTGCACATGGAGGGCTTCAAAGGATCC
AGGGTCTGAA
TCAGCCAAGAACTTAGACACAGCTTCAGTTCAGGAGAGAT
```

Allowed formats are: RAW, FASTA, TRANSFAC, EMBL, GenBank, IG

Matrix or Profile Selection:

● Matrices:

Group of matrices: all / vertebrates / fungi

☑ use high quality matrices only

Cut-off selection for matrix group:

● to minimize false positives
○ to minimize false negatives
○ to minimize the sum of both error rates
○ 0.7 and 0.75 as mat. sim. and core sim. cut-off

Predefined Profiles:

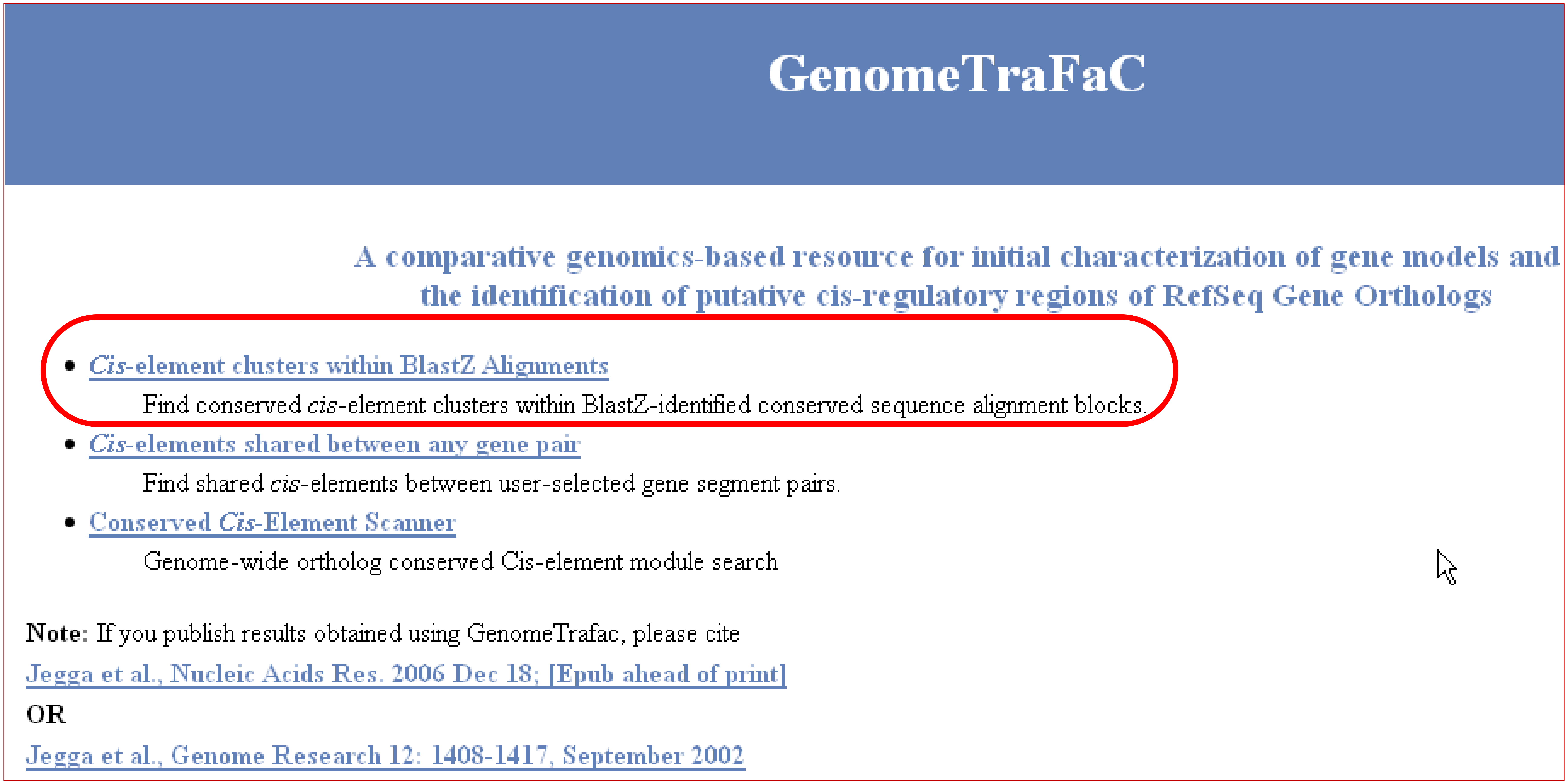
○ our profiles: muscle_specific.prf
○ your profiles

Need to have an account (free for academic use)

How can I identify putative regulatory regions for a gene or microRNA?

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

GenomeTrafac: http://genometrafac.cchmc.org



GenomeTraFaC

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

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

Note: If you publish results obtained using GenomeTrafac, please cite

Jegga et al., Nucleic Acids Res. 2006 Dec 18; [Epub ahead of print]

OR

Jegga et al., Genome Research 12: 1408-1417, September 2002

GenomeTrafac: http://genometrafac.cchmc.org

Basic Search

Description

mir-122a

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

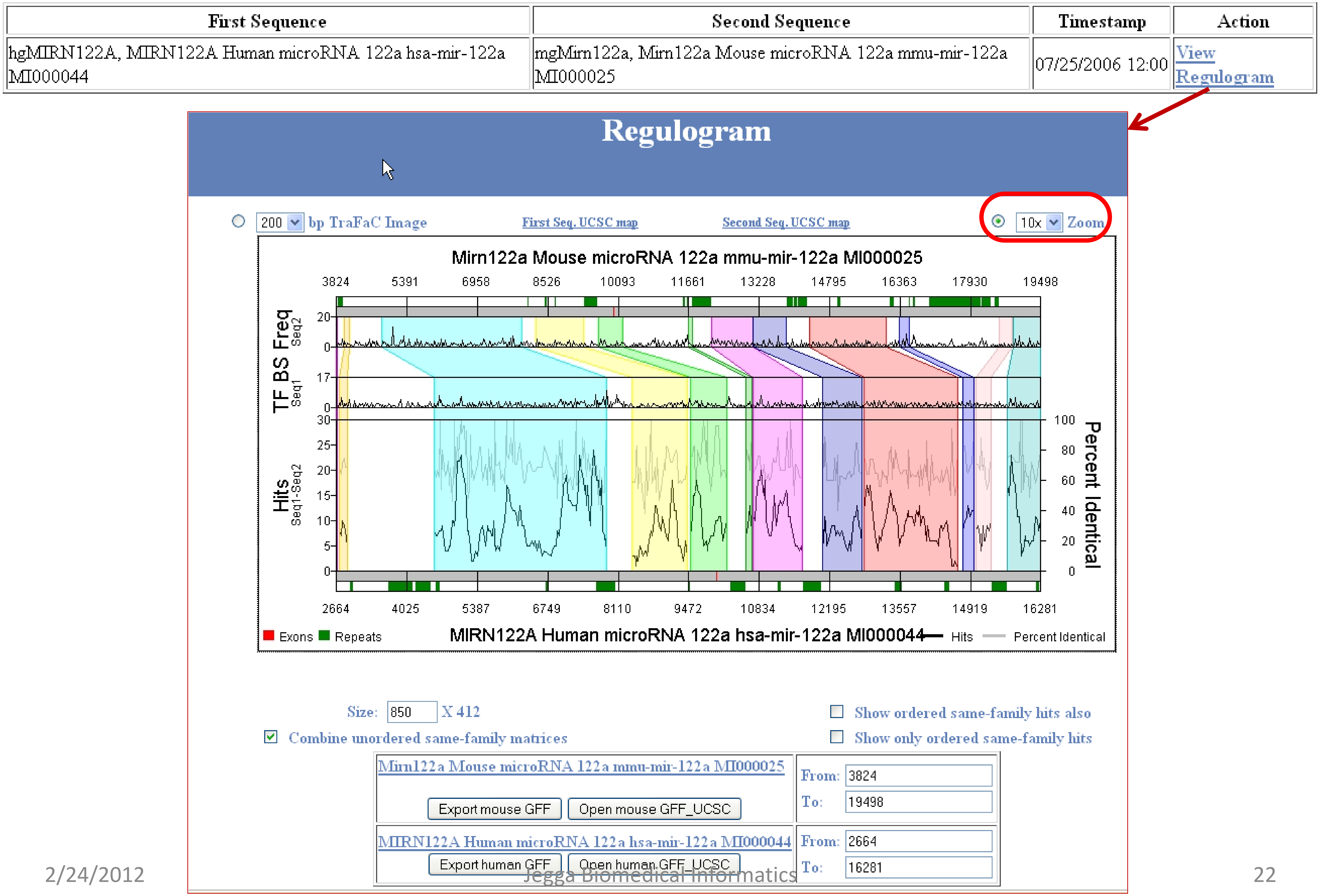
Select	Query
●	Disease *(Always use* ***Disease Selector****)*
○	Pathway *(Always use* ***Pathway Selector****)*
○	Gene ontology *(Always use* ***Ontology Selector****)*
○	Mammalian phenotype *(Always use* ***Phenotype Selector****)*
○	Select gene family from the list
○	Select custom group from the list

Query took 1.514 s

(2 genes meet the search criteria)

✓	Query Term	Accession Number	Name
☑	MIR-122A	hgMIRN122A	MIRN122A Human microRNA 122a hsa-mir-122a MI000044
☑	MIR-122A	mgMirn122a	Mirn122a Mouse microRNA 122a mmu-mir-122a MI000025

GenomeTrafac: http://genometrafac.cchmc.org



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

Regulogram

○ 200 bp TraFaC Image

First Seq. UCSC map

Second Seq. UCSC map

● 10x Zoom

Size: 850 X 412

☑ Combine unordered same-family matrices

☐ Show ordered same-family hits also

☐ Show only ordered same-family hits

Mirn122a Mouse microRNA 122a mmu-mir-122a MI000025

Export mouse GFF | Open mouse GFF_UCSC

From: 3824

To: 19498

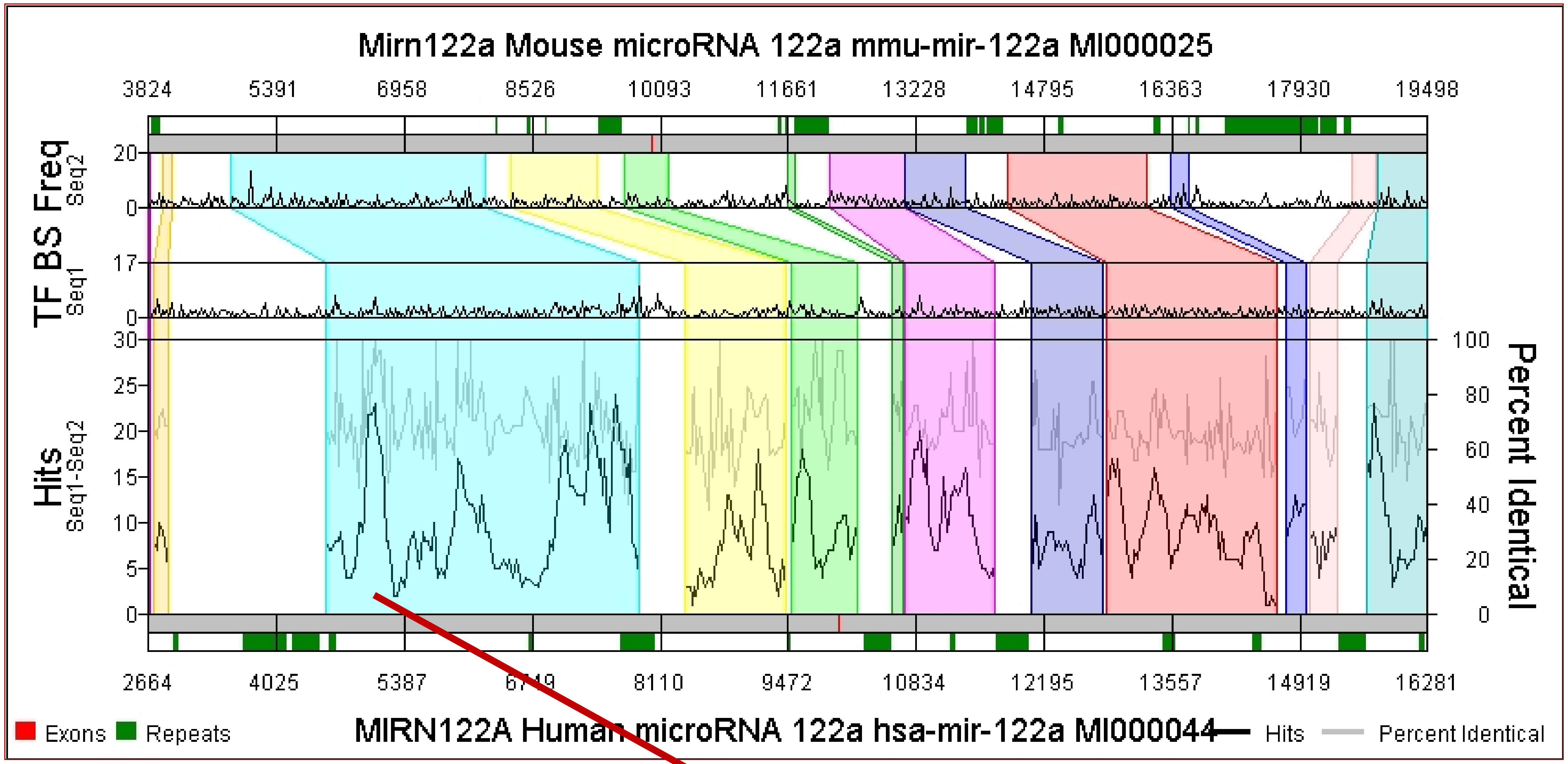
MIRN122A Human microRNA 122a hsa-mir-122a MI000044

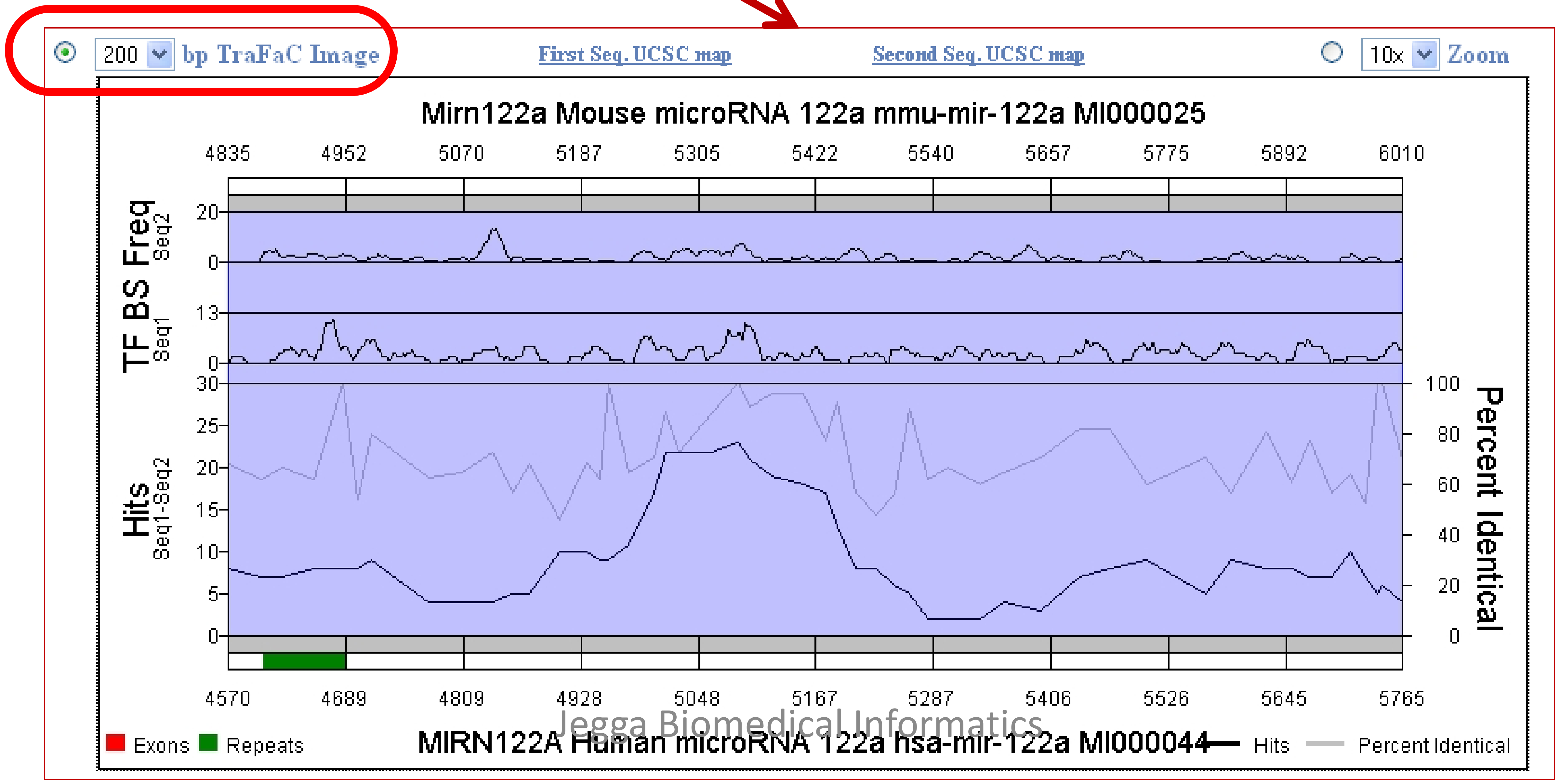
Export human GFF | Open human GFF_UCSC

From: 2664

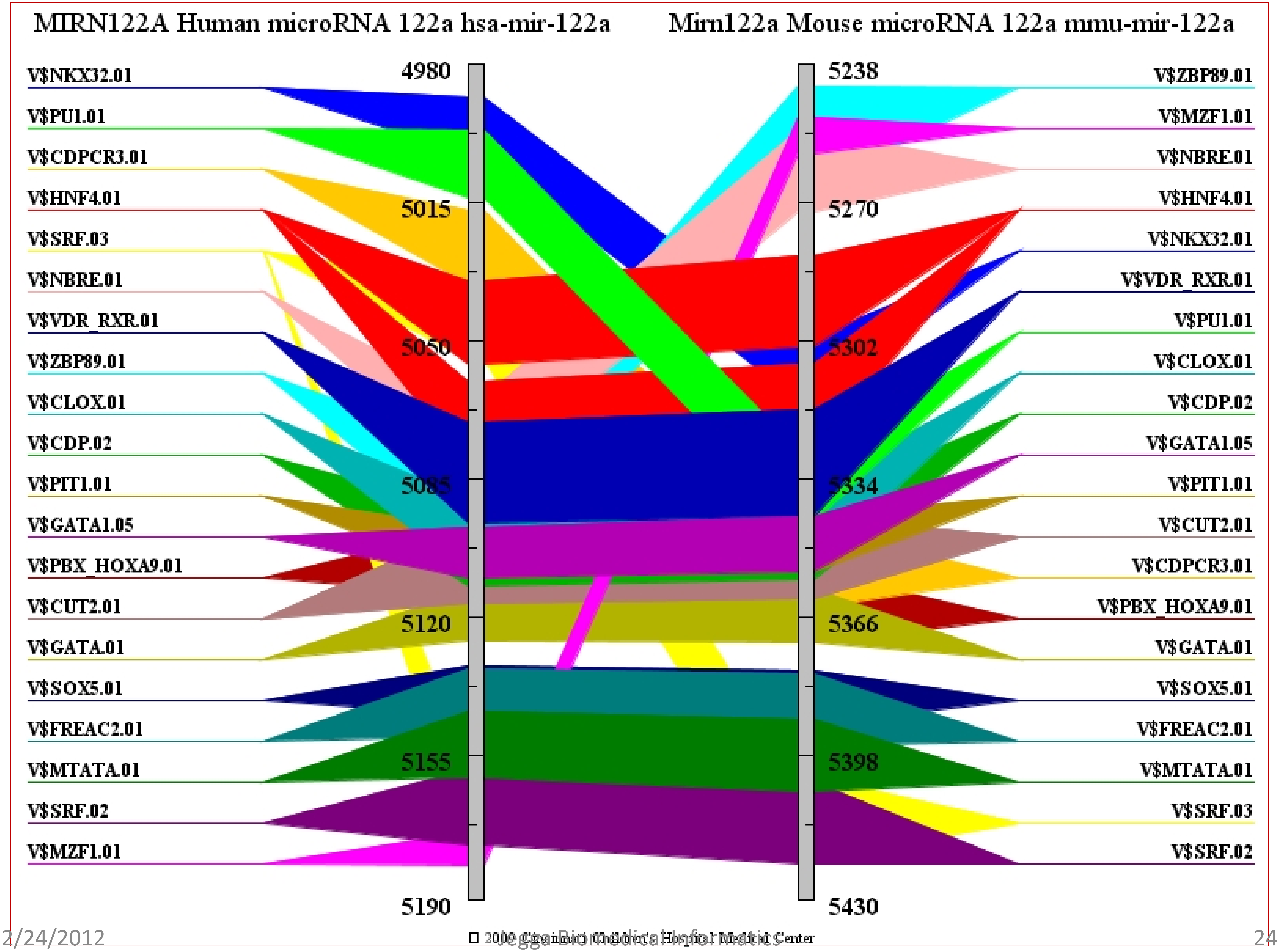
To: 16281

GenomeTrafac: http://genometrafac.cchmc.org

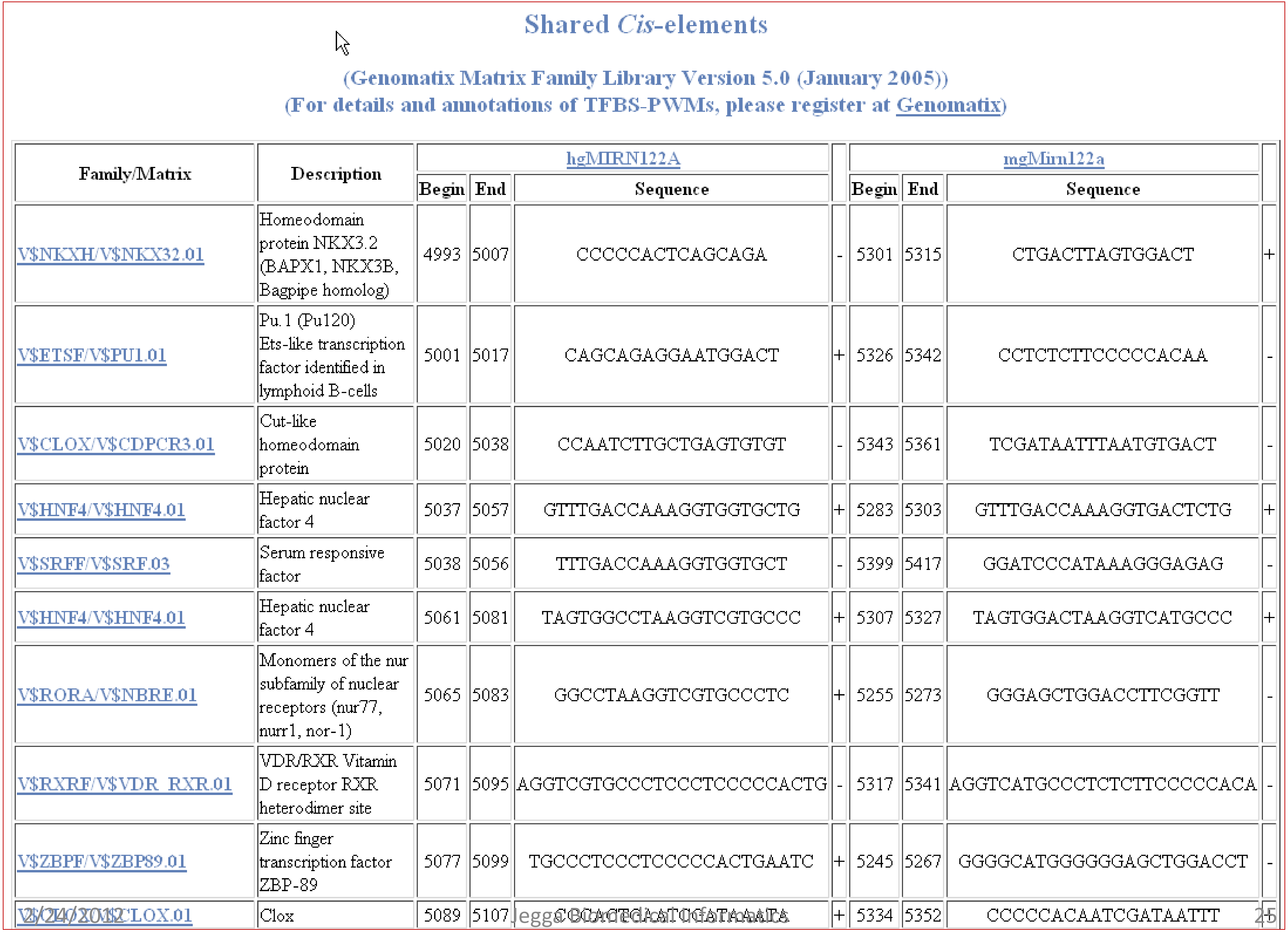




GenomeTrafac: http://genometrafac.cchmc.org



GenomeTrafac: http://genometrafac.cchmc.org

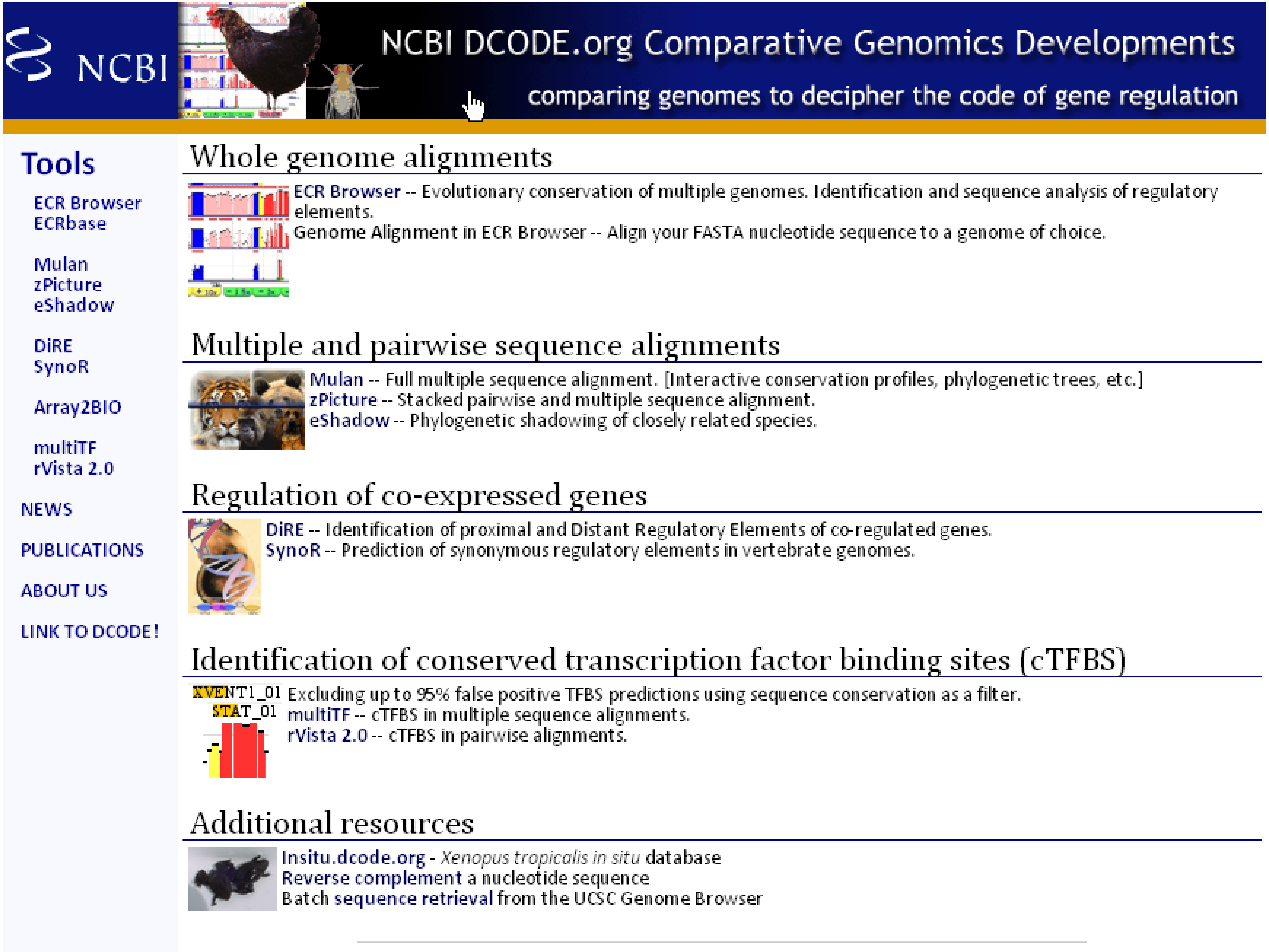


Shared *Cis*-elements

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

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

DCODE: http://www.dcode.org/



NCBI

NCBI DCODE.org Comparative Genomics Developments

comparing genomes to decipher the code of gene regulation

Tools

ECR Browser
ECRbase

Mulan
zPicture
eShadow

DiRE
SynoR

Array2BIO

multiTF
rVista 2.0

NEWS

PUBLICATIONS

ABOUT US

LINK TO DCODE!

Whole genome alignments

ECR Browser -- Evolutionary conservation of multiple genomes. Identification and sequence analysis of regulatory elements.
Genome Alignment in ECR Browser -- Align your FASTA nucleotide sequence to a genome of choice.

Multiple and pairwise sequence alignments

Mulan -- Full multiple sequence alignment. [Interactive conservation profiles, phylogenetic trees, etc.]
zPicture -- Stacked pairwise and multiple sequence alignment.
eShadow -- Phylogenetic shadowing of closely related species.

Regulation of co-expressed genes

DiRE -- Identification of proximal and Distant Regulatory Elements of co-regulated genes.
SynoR -- Prediction of synonymous regulatory elements in vertebrate genomes.

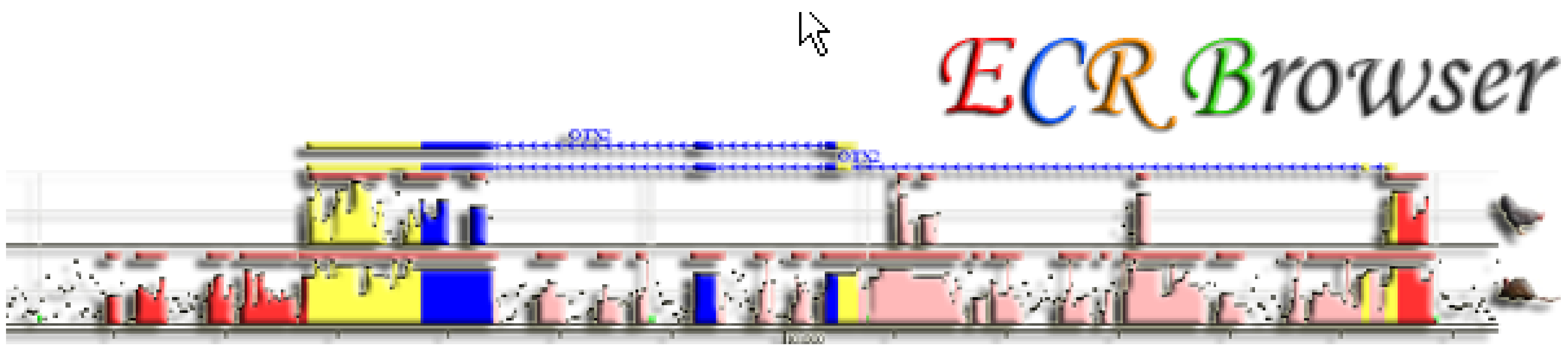
Identification of conserved transcription factor binding sites (cTFBS)

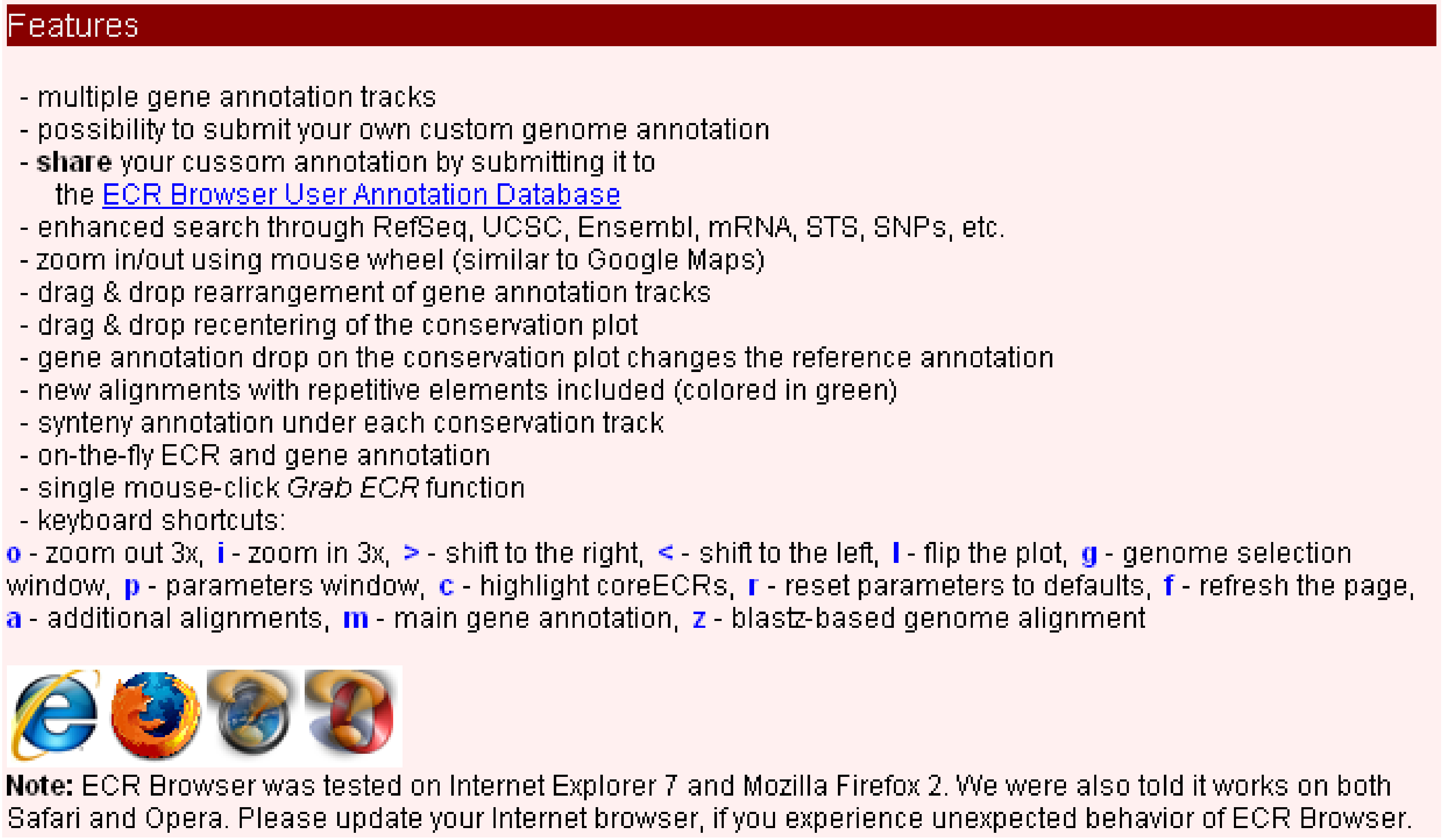
Excluding up to 95% false positive TFBS predictions using sequence conservation as a filter.
multiTF -- cTFBS in multiple sequence alignments.
rVista 2.0 -- cTFBS in pairwise alignments.

Additional resources

Insitu.dcode.org - *Xenopus tropicalis in situ* database
Reverse complement a nucleotide sequence
Batch sequence retrieval from the UCSC Genome Browser

ECR Browser: http://ecrbrowser.dcode.org/



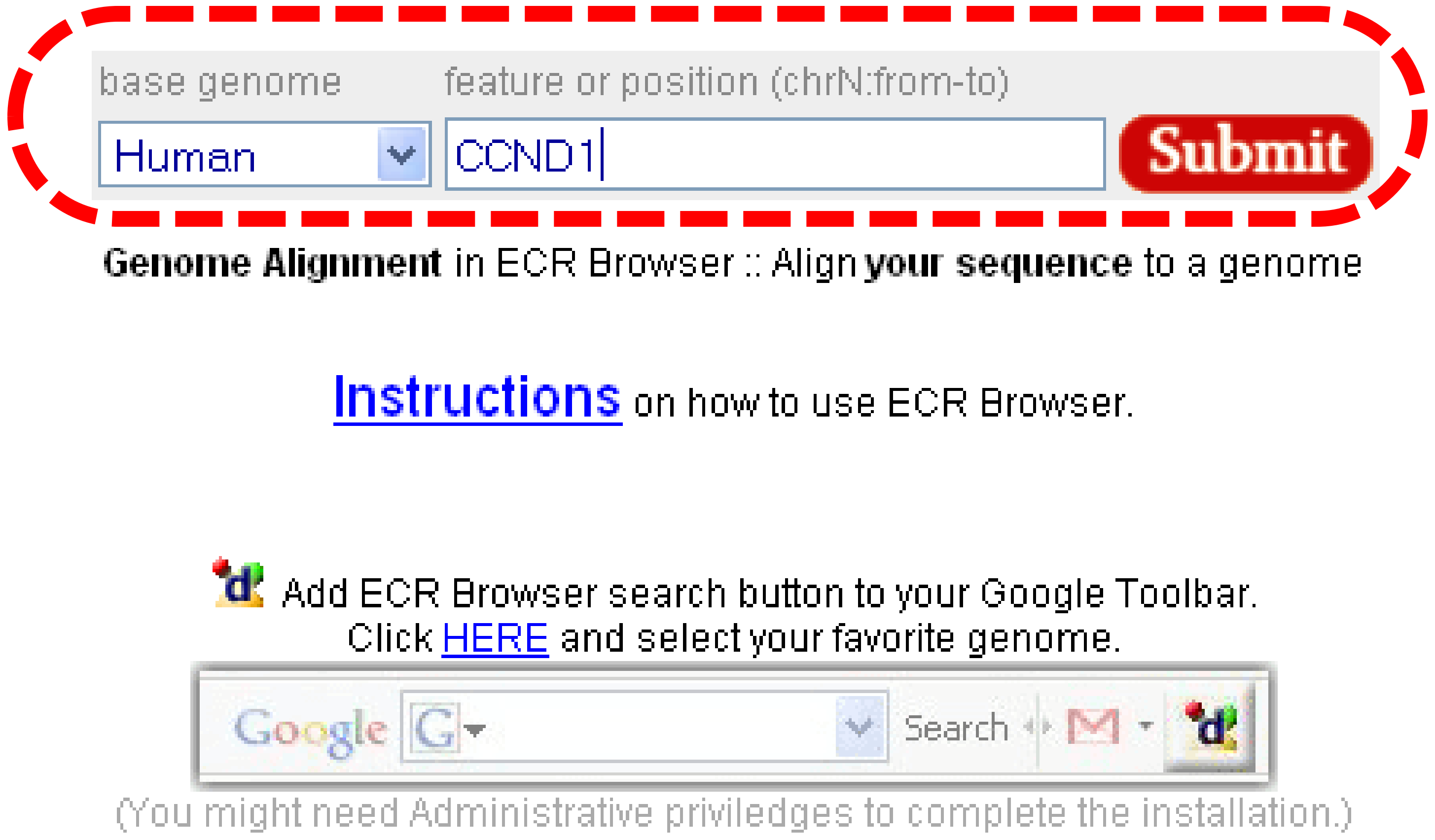


Features

- multiple gene annotation tracks
- possibility to submit your own custom genome annotation
- **share** your cussom annotation by submitting it to the ECR Browser User Annotation Database
- enhanced search through RefSeq, UCSC, Ensembl, mRNA, STS, SNPs, etc.
- zoom in/out using mouse wheel (similar to Google Maps)
- drag & drop rearrangement of gene annotation tracks
- drag & drop recentering of the conservation plot
- gene annotation drop on the conservation plot changes the reference annotation
- new alignments with repetitive elements included (colored in green)
- synteny annotation under each conservation track
- on-the-fly ECR and gene annotation
- single mouse-click *Grab ECR* function
- keyboard shortcuts:

o - zoom out 3x, **i** - zoom in 3x, **>** - shift to the right, **<** - shift to the left, **l** - flip the plot, **g** - genome selection window, **p** - parameters window, **c** - highlight coreECRs, **r** - reset parameters to defaults, **f** - refresh the page, **a** - additional alignments, **m** - main gene annotation, **z** - blastz-based genome alignment

Note: ECR Browser was tested on Internet Explorer 7 and Mozilla Firefox 2. We were also told it works on both Safari and Opera. Please update your Internet browser, if you experience unexpected behavior of ECR Browser.



base genome: Human

feature or position (chrN:from-to): CCND1

Genome Alignment in ECR Browser :: Align **your sequence** to a genome

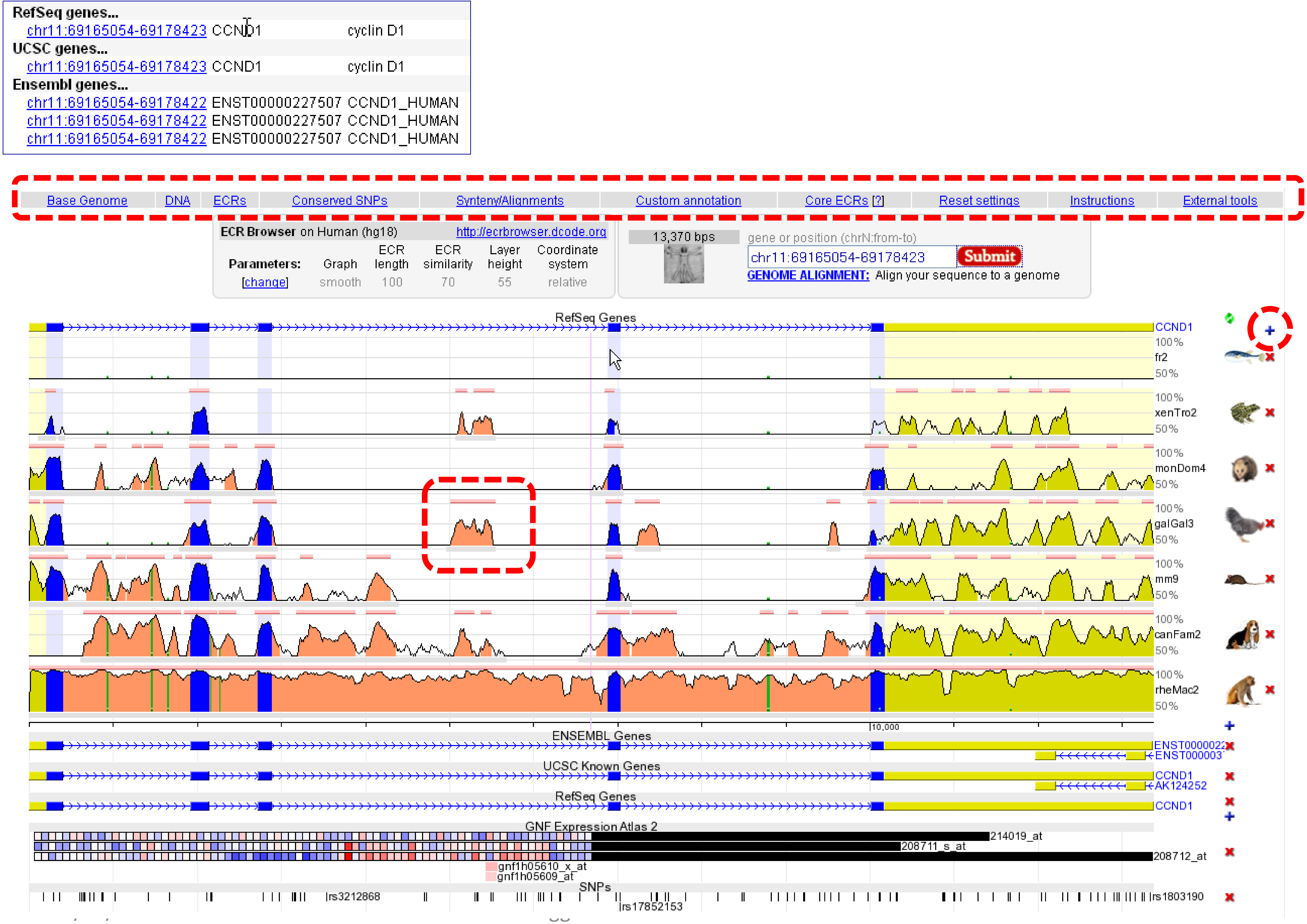
Instructions on how to use ECR Browser.

Add ECR Browser search button to your Google Toolbar. Click HERE and select your favorite genome.

(You might need Administrative priviledges to complete the installation.)

Multispecies (not limited to pairwise comparisons)

ECR Browser: http://ecrbrowser.dcode.org/



RefSeq genes...		
chr11:69165054-69178423	CCND1	cyclin D1
UCSC genes...		
chr11:69165054-69178423	CCND1	cyclin D1
Ensembl genes...		
chr11:69165054-69178422	ENST00000227507	CCND1_HUMAN
chr11:69165054-69178422	ENST00000227507	CCND1_HUMAN
chr11:69165054-69178422	ENST00000227507	CCND1_HUMAN

ECR Browser on Human (hg18) http://ecrbrowser.dcode.org

Parameters:	Graph	ECR length	ECR similarity	Layer height	Coordinate system
[change]	smooth	100	70	55	relative

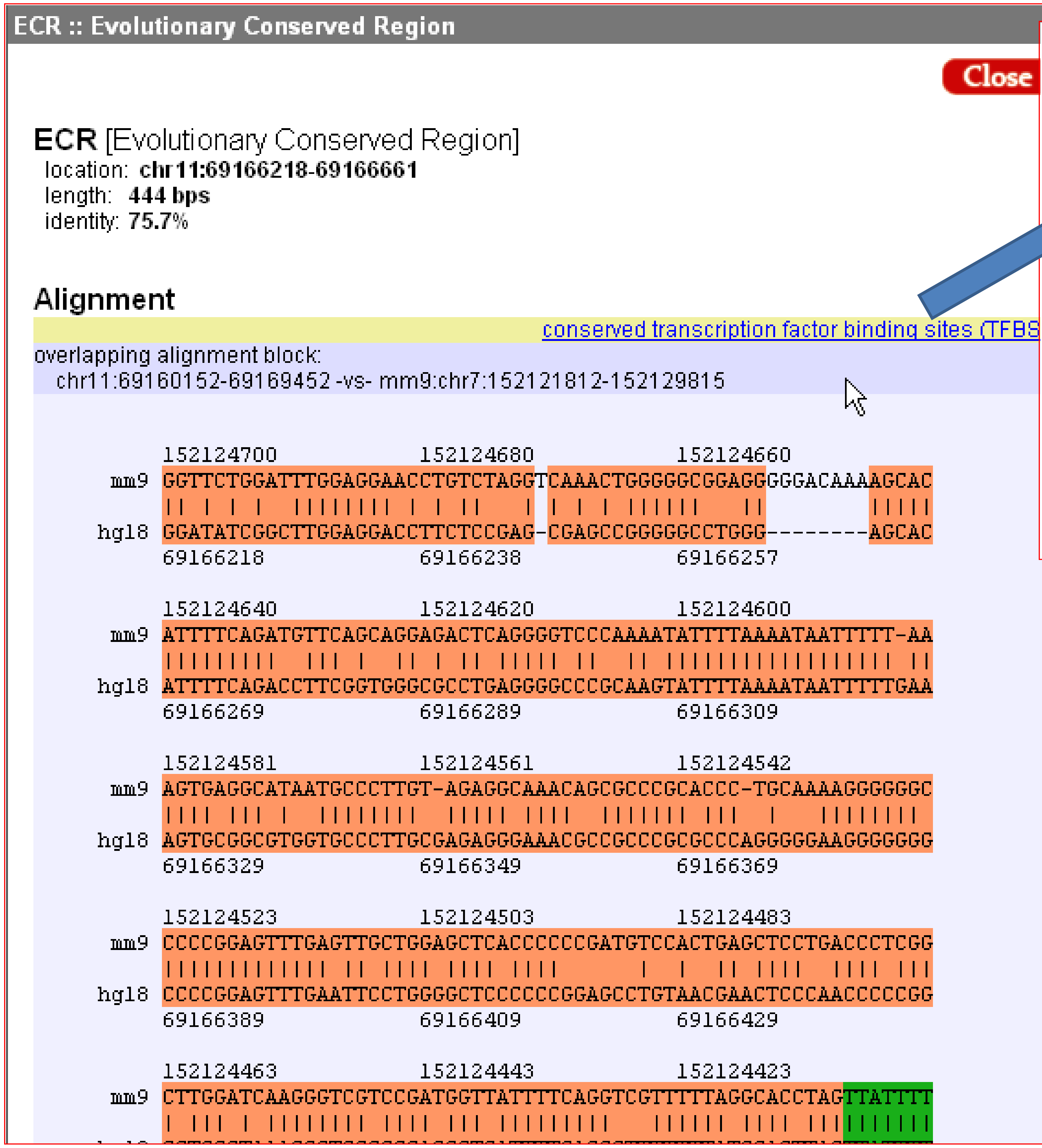
13,370 bps

gene or position (chrN:from-to)

chr11:69165054-69178423 Submit

GENOME ALIGNMENT: Align your sequence to a genome

ECR Browser: http://ecrbrowser.dcode.org/



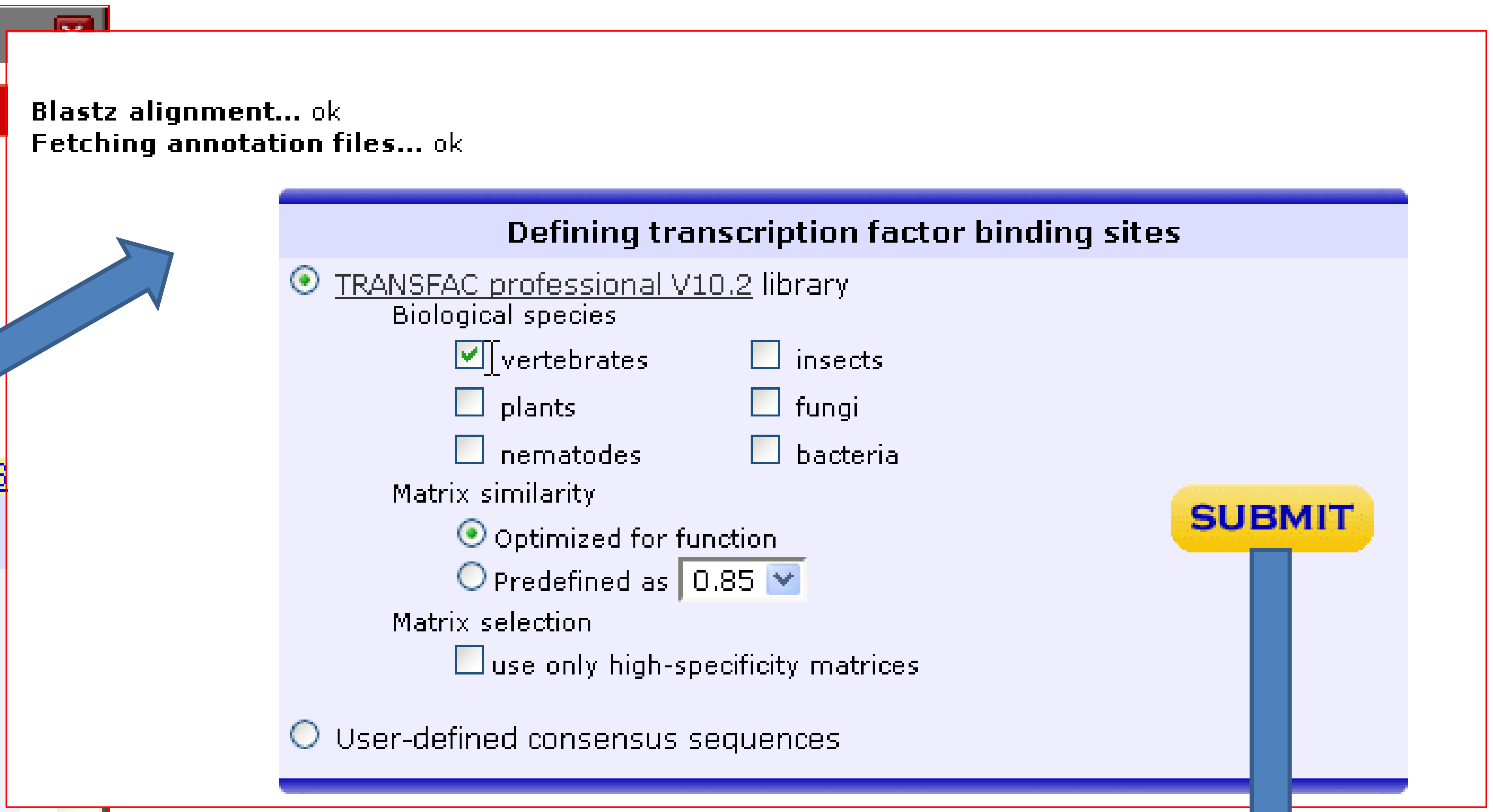
ECR [Evolutionary Conserved Region]
location: **chr11:69166218-69166661**
length: **444 bps**
identity: **75.7%**

Alignment

conserved transcription factor binding sites (TFBS

overlapping alignment block:
chr11:69160152-69169452 -vs- mm9:chr7:152121812-152129815

```
        152124700           152124680           152124660
 mm9 GGTTCTGGATTTGGAGGAACCTGTCTAGGTCAAACTGGGGGCGGAGGGGGACAAAAGCAC
     || | | |  |||||||| | | ||   | | | | ||||||   ||        |||||
hg18 GGATATCGGCTTGGAGGACCTTCTCCGAG-CGAGCCGGGGGCCTGGG--------AGCAC
     69166218            69166238            69166257

        152124640           152124620           152124600
 mm9 ATTTTCAGATGTTCAGCAGGAGACTCAGGGGTCCCAAAATATTTTAAAATAATTTTT-AA
     |||||||||  ||| |  || | || ||||| ||  || |||||||||||||||||| ||
hg18 ATTTTCAGACCTTCGGTGGGCGCCTGAGGGGCCCGCAAGTATTTTAAAATAATTTTTGAA
     69166269            69166289            69166309

        152124581           152124561           152124542
 mm9 AGTGAGGCATAATGCCCTTGT-AGAGGCAAACAGCGCCCGCACCC-TGCAAAAGGGGGGC
     |||| ||| |  ||||||||  ||||| ||||  ||||||| |||  |   |||||||| 
hg18 AGTGCGGCGTGGTGCCCTTGCGAGAGGGAAACGCCGCCCGCGCCCAGGGGGAAGGGGGGG
     69166329            69166349            69166369

        152124523           152124503           152124483
 mm9 CCCCGGAGTTTGAGTTGCTGGAGCTCACCCCCCGATGTCCACTGAGCTCCTGACCCTCGG
     ||||||||||||| || |||| |||| ||||      |  |  || ||||  |||| |||
hg18 CCCCGGAGTTTGAATTCCTGGGGCTCCCCCCGGAGCCTGTAACGAACTCCCAACCCCCGG
     69166389            69166409            69166429

        152124463           152124443           152124423
 mm9 CTTGGATCAAGGGTCGTCCGATGGTTATTTTCAGGTCGTTTTTAGGCACCTAGTTATTTT
     | ||| | |||||||| |||| ||| |||||||||   |||||| |||| ||||||||||
```

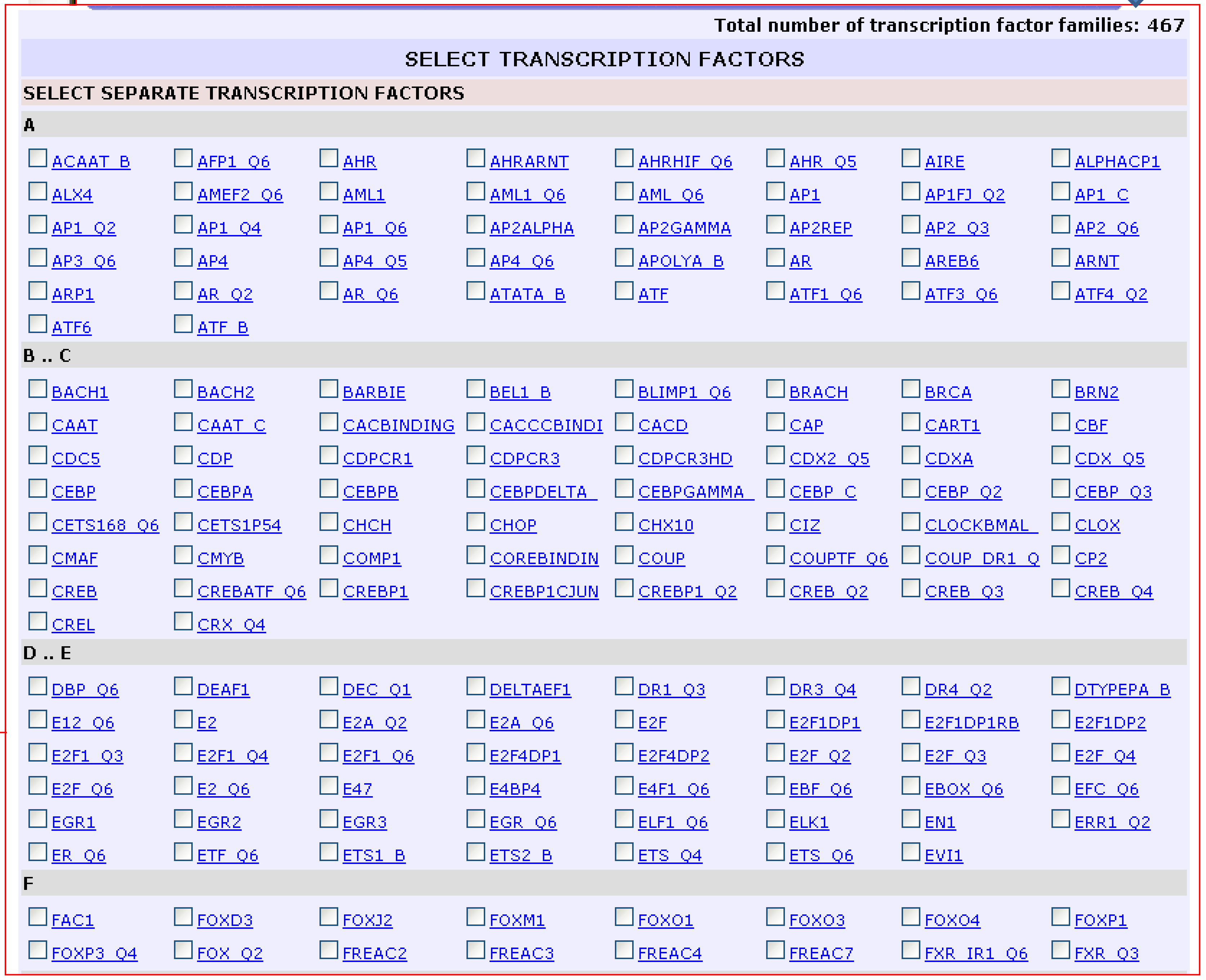


Blastz alignment... ok
Fetching annotation files... ok

Defining transcription factor binding sites

● TRANSFAC professional V10.2 library
Biological species
☑ vertebrates ☐ insects
☐ plants ☐ fungi
☐ nematodes ☐ bacteria
Matrix similarity
● Optimized for function
○ Predefined as 0.85
Matrix selection
☐ use only high-specificity matrices

○ User-defined consensus sequences



Total number of transcription factor families: 467

SELECT TRANSCRIPTION FACTORS

SELECT SEPARATE TRANSCRIPTION FACTORS

A

☐ ACAAT_B ☐ AFP1_Q6 ☐ AHR ☐ AHRARNT ☐ AHRHIF_Q6 ☐ AHR_Q5 ☐ AIRE ☐ ALPHACP1
☐ ALX4 ☐ AMEF2_Q6 ☐ AML1 ☐ AML1_Q6 ☐ AML_Q6 ☐ AP1 ☐ AP1FJ_Q2 ☐ AP1_C
☐ AP1_Q2 ☐ AP1_Q4 ☐ AP1_Q6 ☐ AP2ALPHA ☐ AP2GAMMA ☐ AP2REP ☐ AP2_Q3 ☐ AP2_Q6
☐ AP3_Q6 ☐ AP4 ☐ AP4_Q5 ☐ AP4_Q6 ☐ APOLYA_B ☐ AR ☐ AREB6 ☐ ARNT
☐ ARP1 ☐ AR_Q2 ☐ AR_Q6 ☐ ATATA_B ☐ ATF ☐ ATF1_Q6 ☐ ATF3_Q6 ☐ ATF4_Q2
☐ ATF6 ☐ ATF_B

B .. C

☐ BACH1 ☐ BACH2 ☐ BARBIE ☐ BEL1_B ☐ BLIMP1_Q6 ☐ BRACH ☐ BRCA ☐ BRN2
☐ CAAT ☐ CAAT_C ☐ CACBINDING ☐ CACCCBINDI ☐ CACD ☐ CAP ☐ CART1 ☐ CBF
☐ CDC5 ☐ CDP ☐ CDPCR1 ☐ CDPCR3 ☐ CDPCR3HD ☐ CDX2_Q5 ☐ CDXA ☐ CDX_Q5
☐ CEBP ☐ CEBPA ☐ CEBPB ☐ CEBPDELTA_ ☐ CEBPGAMMA_ ☐ CEBP_C ☐ CEBP_Q2 ☐ CEBP_Q3
☐ CETS168_Q6 ☐ CETS1P54 ☐ CHCH ☐ CHOP ☐ CHX10 ☐ CIZ ☐ CLOCKBMAL_ ☐ CLOX
☐ CMAF ☐ CMYB ☐ COMP1 ☐ COREBINDIN ☐ COUP ☐ COUPTF_Q6 ☐ COUP_DR1_Q ☐ CP2
☐ CREB ☐ CREBATF_Q6 ☐ CREBP1 ☐ CREBP1CJUN ☐ CREBP1_Q2 ☐ CREB_Q2 ☐ CREB_Q3 ☐ CREB_Q4
☐ CREL ☐ CRX_Q4

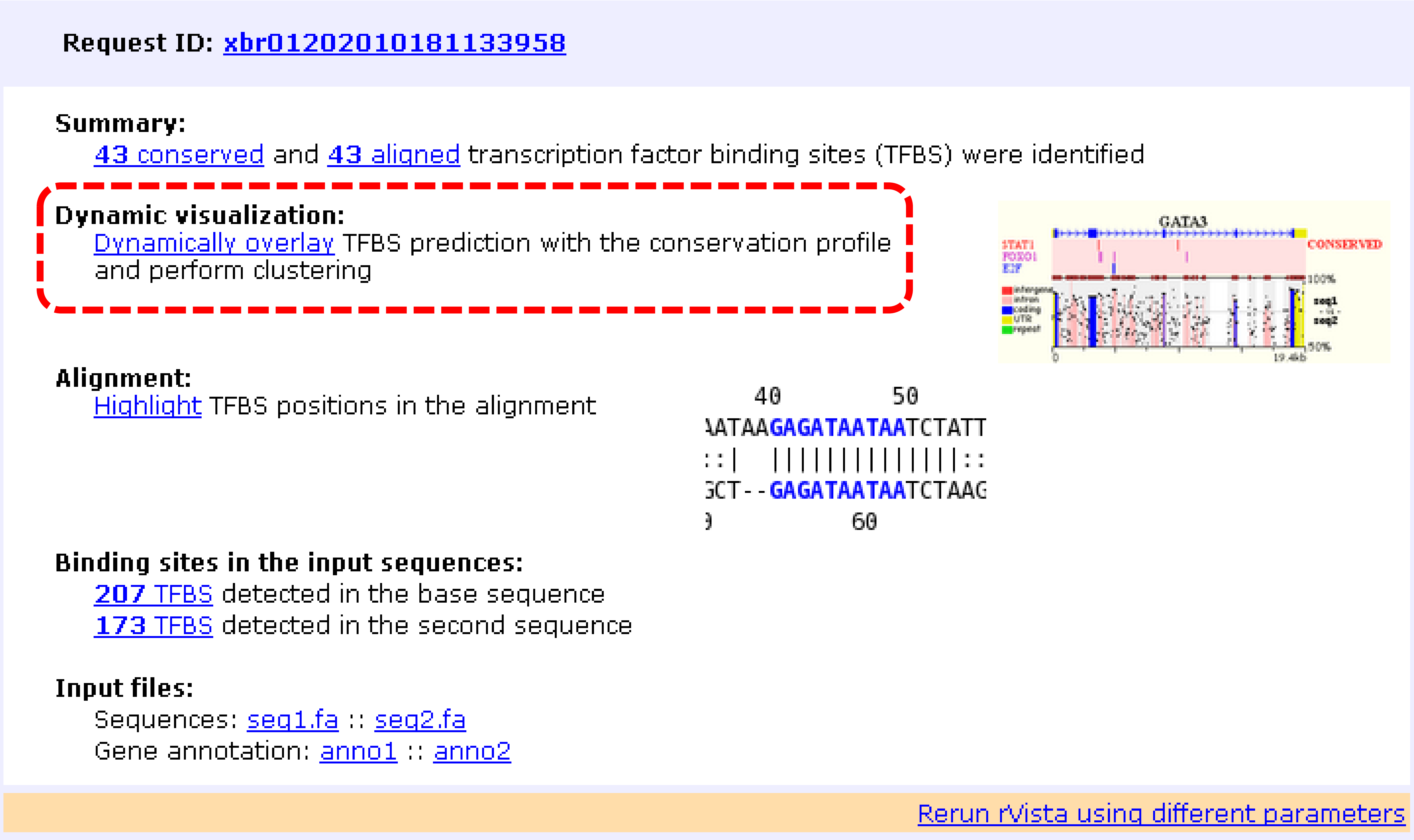
D .. E

☐ DBP_Q6 ☐ DEAF1 ☐ DEC_Q1 ☐ DELTAEF1 ☐ DR1_Q3 ☐ DR3_Q4 ☐ DR4_Q2 ☐ DTYPEPA_B
☐ E12_Q6 ☐ E2 ☐ E2A_Q2 ☐ E2A_Q6 ☐ E2F ☐ E2F1DP1 ☐ E2F1DP1RB ☐ E2F1DP2
☐ E2F1_Q3 ☐ E2F1_Q4 ☐ E2F1_Q6 ☐ E2F4DP1 ☐ E2F4DP2 ☐ E2F_Q2 ☐ E2F_Q3 ☐ E2F_Q4
☐ E2F_Q6 ☐ E2_Q6 ☐ E47 ☐ E4BP4 ☐ E4F1_Q6 ☐ EBF_Q6 ☐ EBOX_Q6 ☐ EFC_Q6
☐ EGR1 ☐ EGR2 ☐ EGR3 ☐ EGR_Q6 ☐ ELF1_Q6 ☐ ELK1 ☐ EN1 ☐ ERR1_Q2
☐ ER_Q6 ☐ ETF_Q6 ☐ ETS1_B ☐ ETS2_B ☐ ETS_Q4 ☐ ETS_Q6 ☐ EVI1

F

☐ FAC1 ☐ FOXD3 ☐ FOXJ2 ☐ FOXM1 ☐ FOXO1 ☐ FOXO3 ☐ FOXO4 ☐ FOXP1
☐ FOXP3_Q4 ☐ FOX_Q2 ☐ FREAC2 ☐ FREAC3 ☐ FREAC4 ☐ FREAC7 ☐ FXR_IR1_Q6 ☐ FXR_Q3

ECR Browser: http://ecrbrowser.dcode.org/



Request ID: xbr01202010181133958

Summary:
43 conserved and 43 aligned transcription factor binding sites (TFBS) were identified

Dynamic visualization:
Dynamically overlay TFBS prediction with the conservation profile and perform clustering

Alignment:
Highlight TFBS positions in the alignment

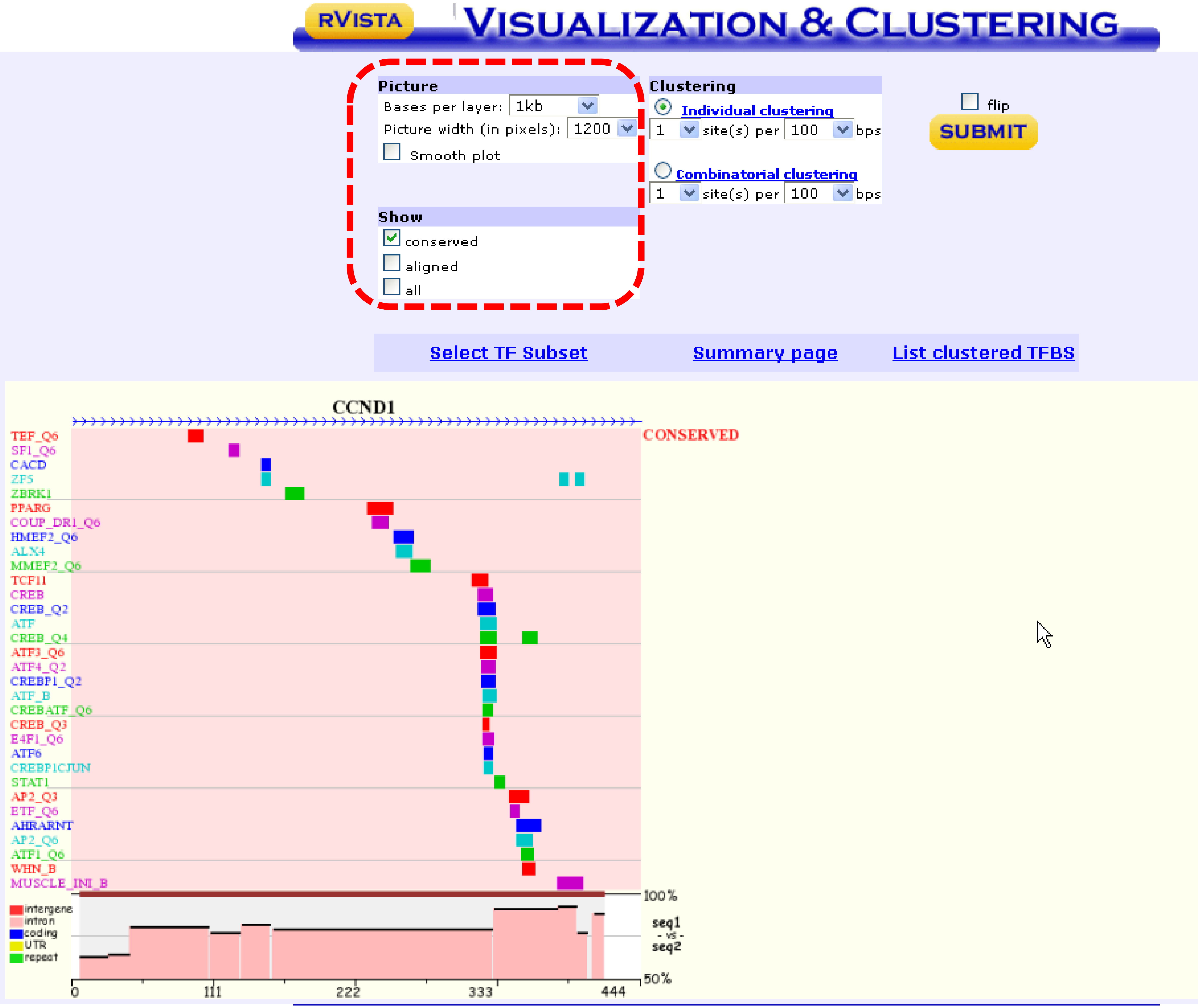
```
     40          50
ATAAGAGATAATAATCTATT
::|  |||||||||||||::
CT--GAGATAATAATCTAAG
           60
```

Binding sites in the input sequences:
207 TFBS detected in the base sequence
173 TFBS detected in the second sequence

Input files:
Sequences: seq1.fa :: seq2.fa
Gene annotation: anno1 :: anno2

Rerun rVista using different parameters

ECR Browser: http://ecrbrowser.dcode.org/



Picture

Bases per layer: 1kb

Picture width (in pixels): 1200

☐ Smooth plot

Show

☑ conserved

☐ aligned

☐ all

Clustering

● Individual clustering
1 site(s) per 100 bps

○ Combinatorial clustering
1 site(s) per 100 bps

☐ flip

RESOURCES - URLs: Summary

Application/Resource	URL
Genome Browser	http://genome.ucsc.edu
JASPAR	http://jaspar.genereg.net/
Gene Regulation	http://www.gene-regulation.com
GenomeTrafac	http://genometrafac.cchmc.org
DCODE	http://www.dcode.org/

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

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

- Known transcription factor binding sites (TFBS)
  - Conserved
    - oPOSSUM
    - DiRE
  - Non-conserved
    - Pscan
    - MatInspector (*Licensed)
- Unknown TFBS or Novel motifs
  - Conserved
    - oPOSSUM
    - Weeder-H
  - Non-conserved
    - MEME
    - Weeder

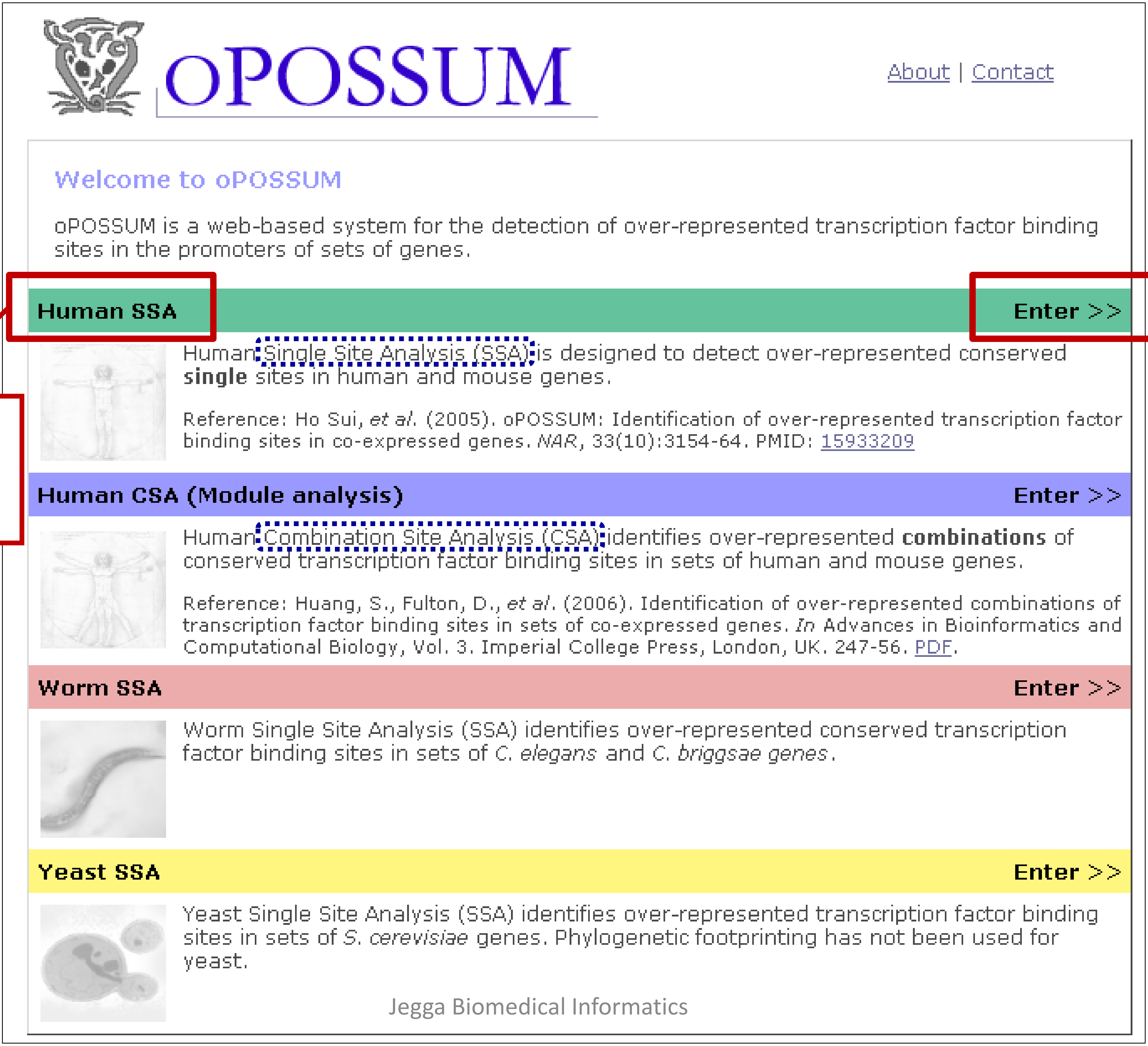
1. Each of these applications support different forms of input. Very few support probeset IDs.
2. **Red Font**: Input sequence required; Do not support gene symbols, gene IDs, or accession numbers. The advantage is you can use them for scanning sequences from any species.
3. *Licensed software: We have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

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

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

- ❑ Known transcription factor binding sites (TFBS)
  - ❖ Conserved
    - oPOSSUM
    - DiRE
  - ❖ Non-conserved
    - Pscan
    - MatInspector (*Licensed)
- ❑ Unknown TFBS or Novel motifs
  - ❖ Conserved
    - oPOSSUM
    - Weeder-H
  - ❖ Non-conserved
    - MEME
    - Weeder

oPOSSUM (http://burgundy.cmmt.ubc.ca/oPOSSUM/)



oPOSSUM

About | Contact

Welcome to oPOSSUM

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

Human SSA **Enter >>**

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

Reference: Ho Sui, *et al.* (2005). oPOSSUM: Identification of over-represented transcription factor binding sites in co-expressed genes. *NAR*, 33(10):3154-64. PMID: 15933209

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

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

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

Worm SSA **Enter >>**

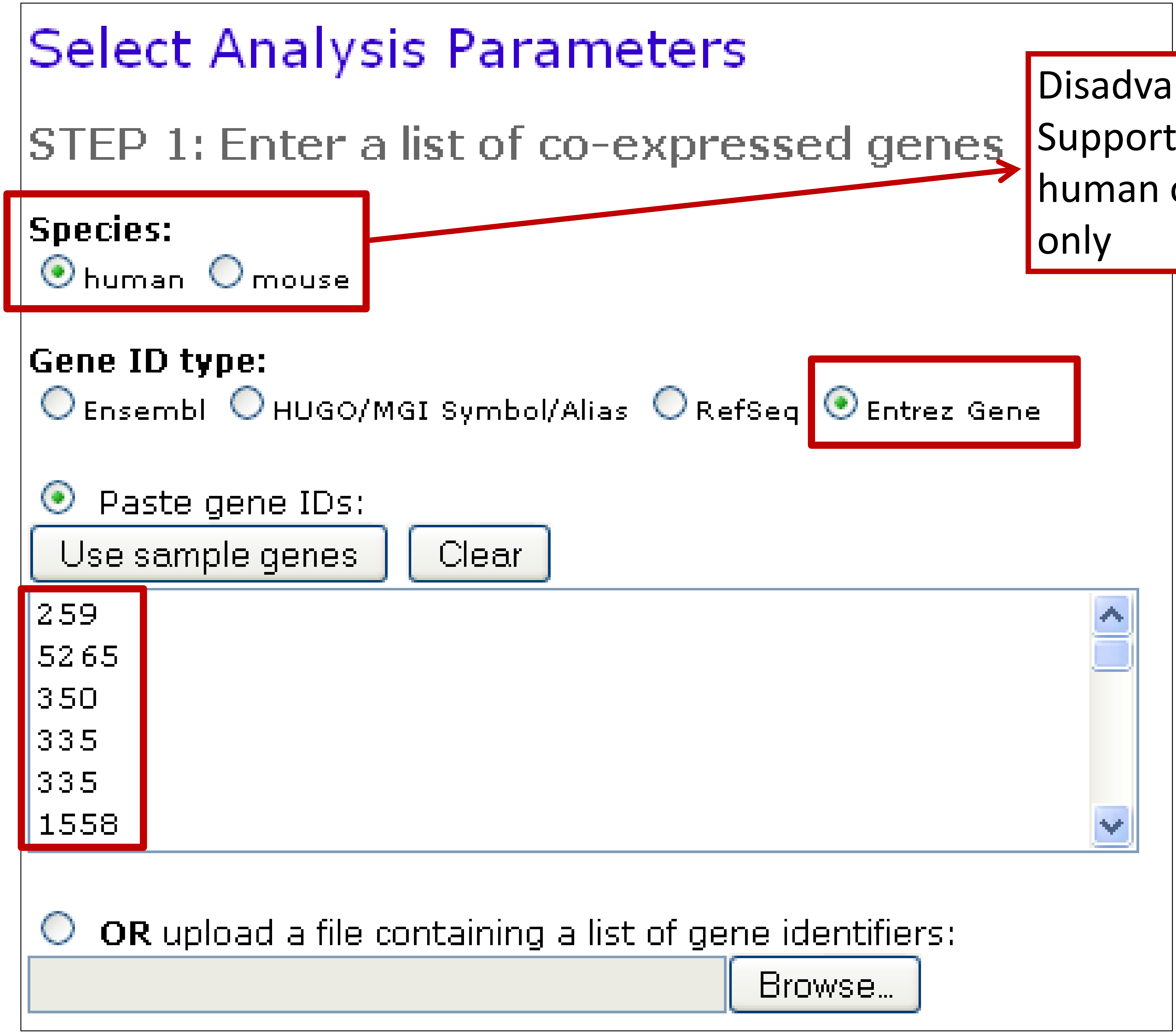
Worm Single Site Analysis (SSA) identifies over-represented conserved transcription factor binding sites in sets of *C. elegans* and *C. briggsae genes*.

Yeast SSA **Enter >>**

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

Supports human and mouse

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



Select Analysis Parameters

STEP 1: Enter a list of co-expressed genes

Species:
● human ○ mouse

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

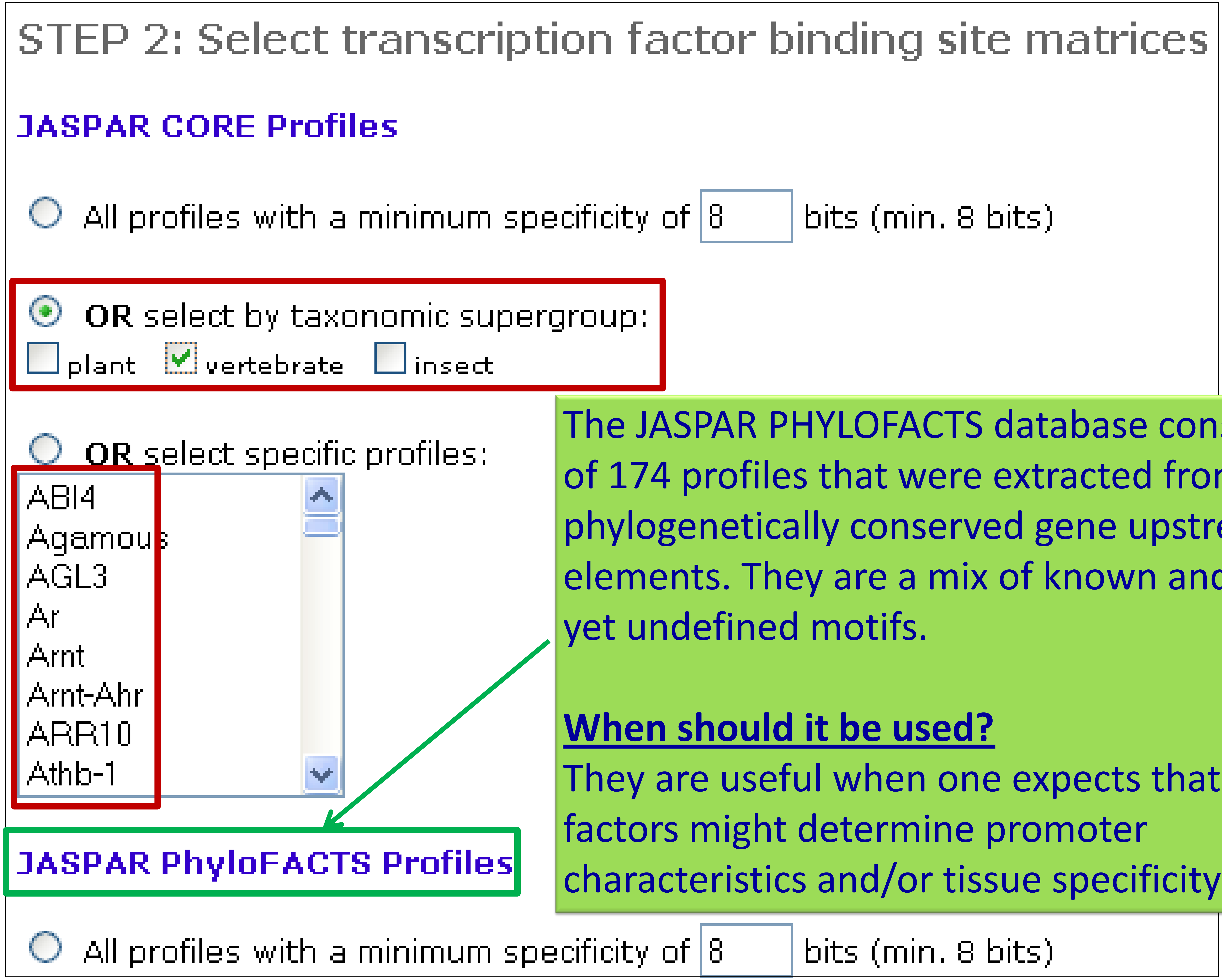
● Paste gene IDs:

259
5265
350
335
335
1558

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

Disadvantage: Supports either human or mouse only

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



STEP 2: Select transcription factor binding site matrices

JASPAR CORE Profiles

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

● **OR** select by taxonomic supergroup:
☐ plant ☑ vertebrate ☐ insect

○ **OR** select specific profiles:

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

JASPAR PhyloFACTS Profiles

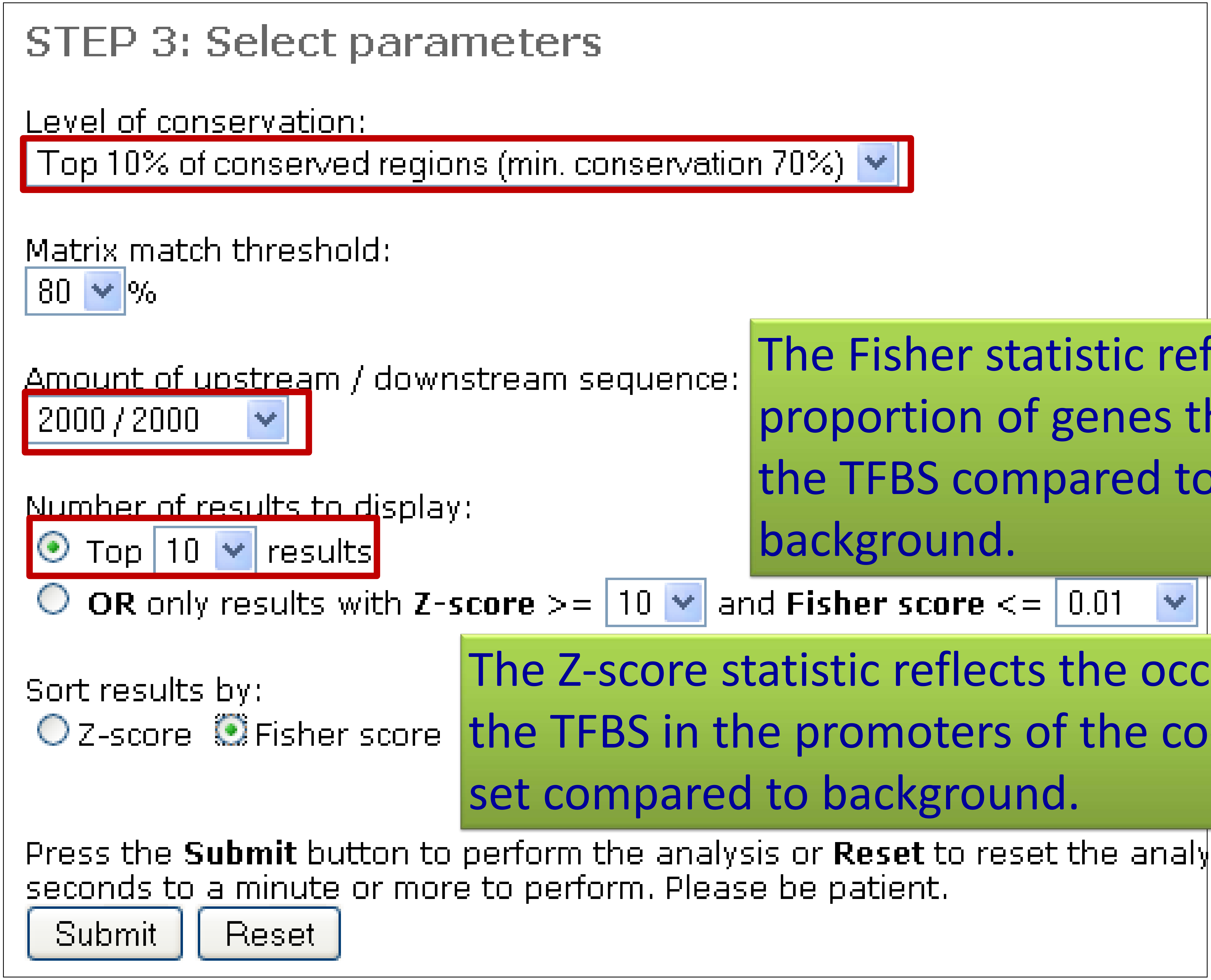
○ All profiles with a minimum specificity of 8 bits (min. 8 bits)

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

When should it be used?

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

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



STEP 3: Select parameters

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

Matrix match threshold:
80 %

Amount of upstream / downstream sequence:
2000 / 2000

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

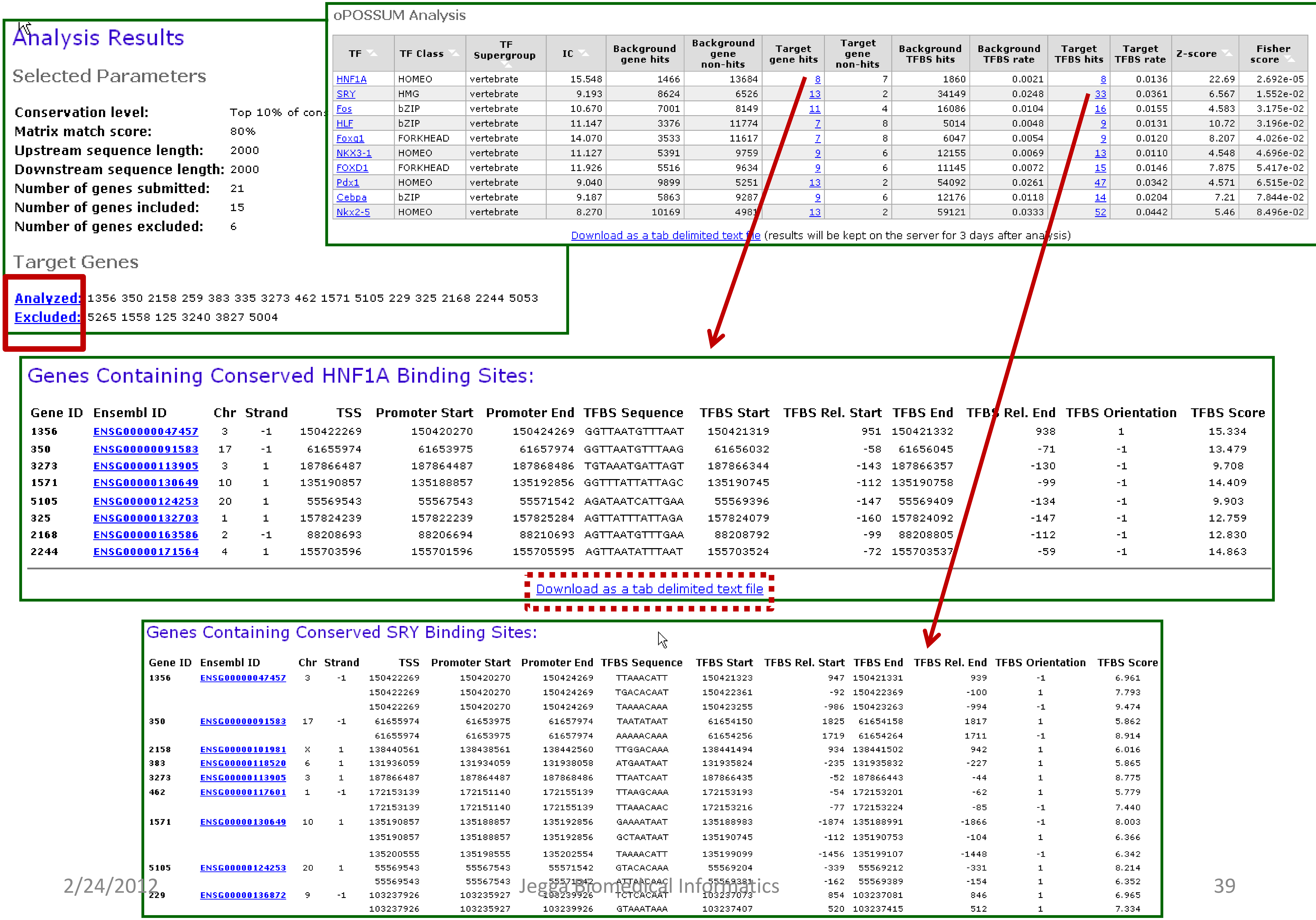
Sort results by:
○ Z-score ● Fisher score

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

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

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

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



Analysis Results

Selected Parameters

Conservation level:	Top 10% of cons
Matrix match score:	80%
Upstream sequence length:	2000
Downstream sequence length:	2000
Number of genes submitted:	21
Number of genes included:	15
Number of genes excluded:	6

Target Genes

Analyzed: 1356 350 2158 259 383 335 3273 462 1571 5105 229 325 2168 2244 5053

Excluded: 5265 1558 125 3240 3827 5004

oPOSSUM Analysis

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

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

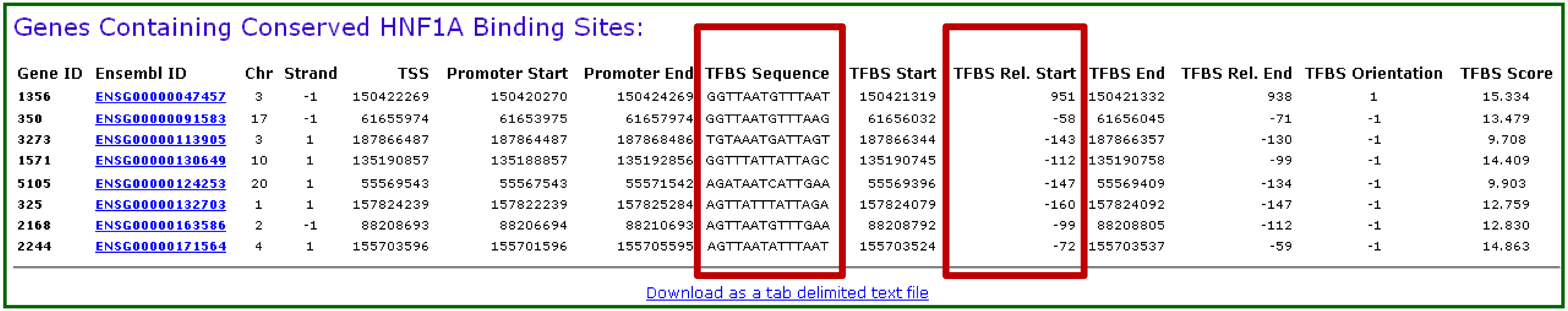
Genes Containing Conserved HNF1A Binding Sites:

Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1356	ENSG00000047457	3	-1	150422269	150420270	150424269	GGTTAATGTTTAAT	150421319	951	150421332	938	1	15.334
350	ENSG00000091583	17	-1	61655974	61653975	61657974	GGTTAATGTTTAAG	61656032	-58	61656045	-71	-1	13.479
3273	ENSG00000113905	3	1	187866487	187864487	187868486	TGTAAATGATTAGT	187866344	-143	187866357	-130	-1	9.708
1571	ENSG00000130649	10	1	135190857	135188857	135192856	GGTTTATTATTAGC	135190745	-112	135190758	-99	-1	14.409
5105	ENSG00000124253	20	1	55569543	55567543	55571542	AGATAATCATTGAA	55569396	-147	55569409	-134	-1	9.903
325	ENSG00000132703	1	1	157824239	157822239	157825284	AGTTATTTATTAGA	157824079	-160	157824092	-147	-1	12.759
2168	ENSG00000163586	2	-1	88208693	88206694	88210693	AGTTAATGTTTGAA	88208792	-99	88208805	-112	-1	12.830
2244	ENSG00000171564	4	1	155703596	155701596	155705595	AGTTAATATTTAAT	155703524	-72	155703537	-59	-1	14.863

Download as a tab delimited text file

Genes Containing Conserved SRY Binding Sites:

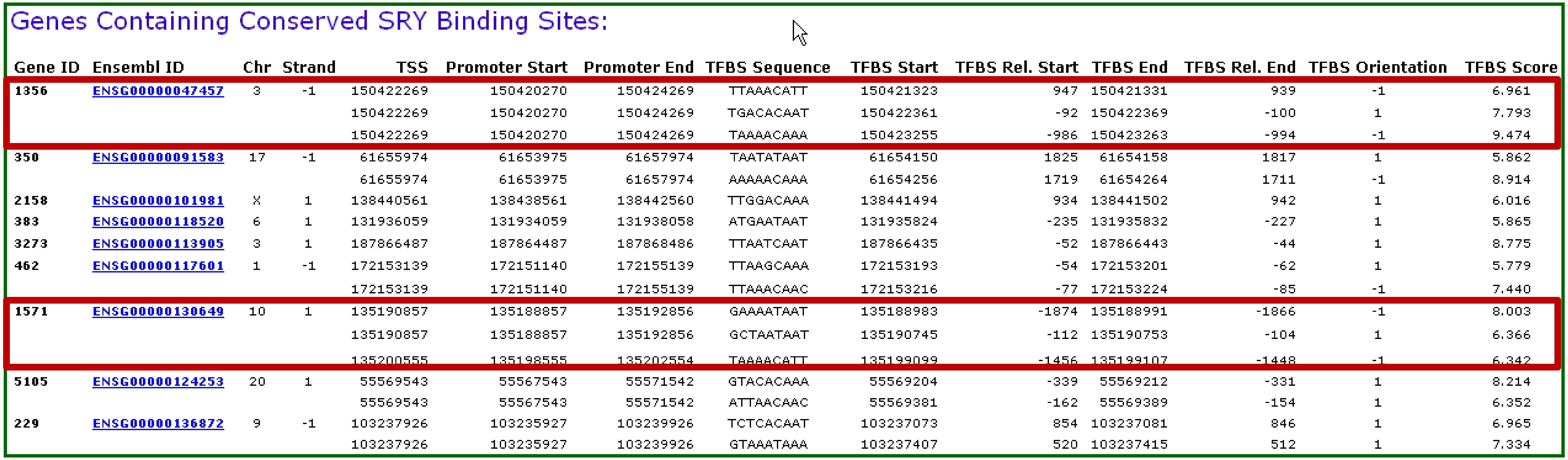
Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1356	ENSG00000047457	3	-1	150422269	150420270	150424269	TTAAACATT	150421323	947	150421331	939	-1	6.961
				150422269	150420270	150424269	TGACACAAT	150422361	-92	150422369	-100	1	7.793
				150422269	150420270	150424269	TAAAACAAA	150423255	-986	150423263	-994	-1	9.474
350	ENSG00000091583	17	-1	61655974	61653975	61657974	TAATATAAT	61654150	1825	61654158	1817	1	5.862
				61655974	61653975	61657974	AAAAACAAA	61654256	1719	61654264	1711	-1	8.914
2158	ENSG00000101981	X	1	138440561	138438561	138442560	TTGGACAAA	138441494	934	138441502	942	1	6.016
383	ENSG00000118520	6	1	131936059	131934059	131938058	ATGAATAAT	131935824	-235	131935832	-227	1	5.865
3273	ENSG00000113905	3	1	187866487	187864487	187868486	TTAATCAAT	187866435	-52	187866443	-44	1	8.775
462	ENSG00000117601	1	-1	172153139	172151140	172155139	TTAAGCAAA	172153193	-54	172153201	-62	1	5.779
				172153139	172151140	172155139	TTAAACAAC	172153216	-77	172153224	-85	-1	7.440
1571	ENSG00000130649	10	1	135190857	135188857	135192856	GAAAATAAT	135188983	-1874	135188991	-1866	-1	8.003
				135190857	135188857	135192856	GCTAATAAT	135190745	-112	135190753	-104	1	6.366
				135200555	135198555	135202554	TAAAACATT	135199099	-1456	135199107	-1448	-1	6.342
5105	ENSG00000124253	20	1	55569543	55567543	55571542	GTACACAAA	55569204	-339	55569212	-331	1	8.214
				55569543	55567543	55571542	ATTAACAAC	55569381	-162	55569389	-154	1	6.352
229	ENSG00000136872	9	-1	103237926	103235927	103239926	TCTCACAAT	103237073	854	103237081	846	1	6.965
				103237926	103235927	103239926	GTAAATAAA	103237407	520	103237415	512	1	7.334



Genes Containing Conserved HNF1A Binding Sites:

Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1356	ENSG00000047457	3	-1	150422269	150420270	150424269	GGTTAATGTTTAAT	150421319	951	150421332	938	1	15.334
350	ENSG00000091583	17	-1	61655974	61653975	61657974	GGTTAATGTTTAAG	61656032	-58	61656045	-71	-1	13.479
3273	ENSG00000113905	3	1	187866487	187864487	187868486	TGTAAATGATTAGT	187866344	-143	187866357	-130	-1	9.708
1571	ENSG00000130649	10	1	135190857	135188857	135192856	GGTTTATTATTAGC	135190745	-112	135190758	-99	-1	14.409
5105	ENSG00000124253	20	1	55569543	55567543	55571542	AGATAATCATTGAA	55569396	-147	55569409	-134	-1	9.903
325	ENSG00000132703	1	1	157824239	157822239	157825284	AGTTATTTATTAGA	157824079	-160	157824092	-147	-1	12.759
2168	ENSG00000163586	2	-1	88208693	88206694	88210693	AGTTAATGTTTGAA	88208792	-99	88208805	-112	-1	12.830
2244	ENSG00000171564	4	1	155703596	155701596	155705595	AGTTAATATTTAAT	155703524	-72	155703537	-59	-1	14.863

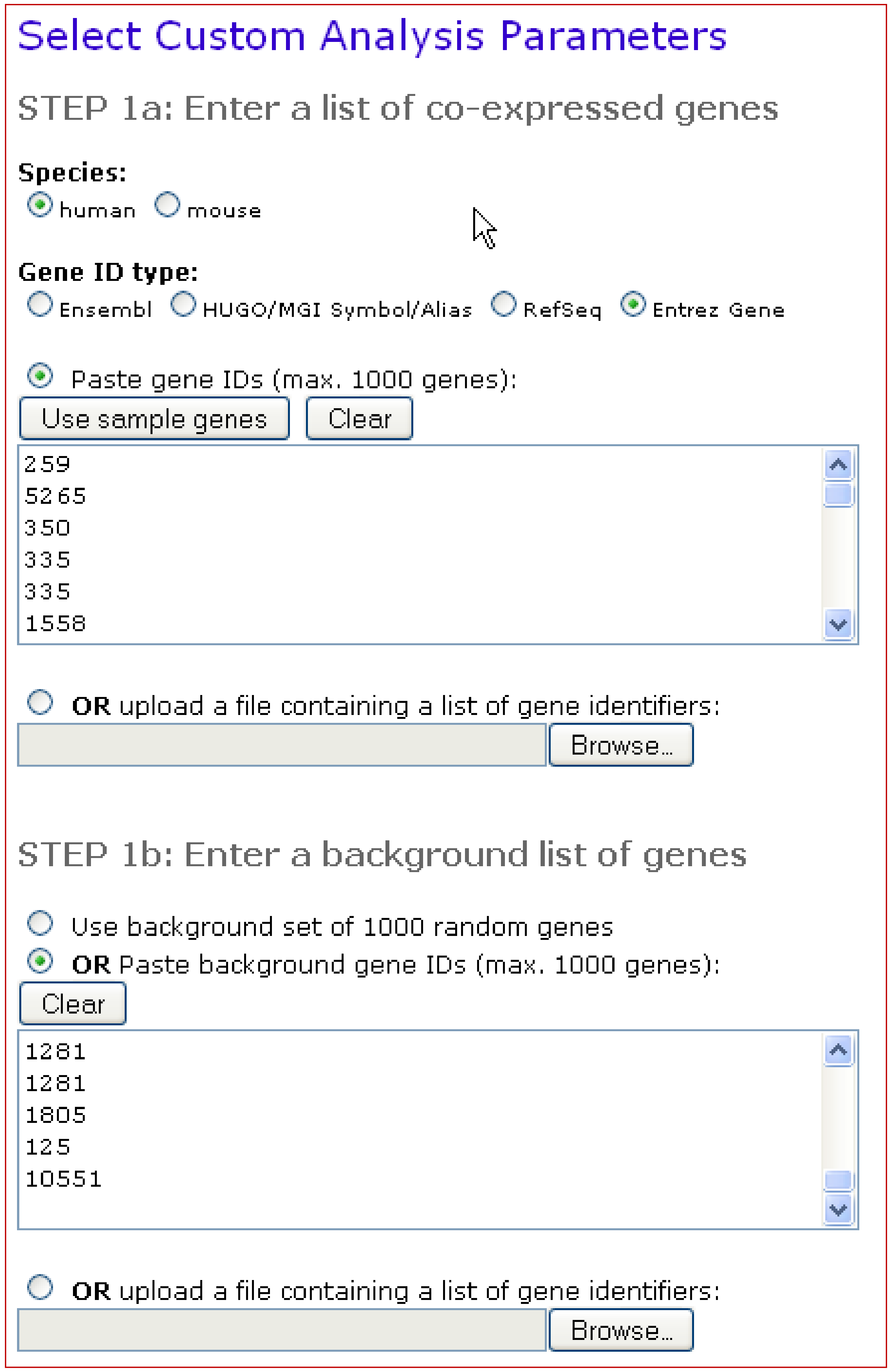
Download as a tab delimited text file



Genes Containing Conserved SRY Binding Sites:

Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1356	ENSG00000047457	3	-1	150422269	150420270	150424269	TTAAACATT	150421323	947	150421331	939	-1	6.961
				150422269	150420270	150424269	TGACACAAT	150422361	-92	150422369	-100	1	7.793
				150422269	150420270	150424269	TAAAACAAA	150423255	-986	150423263	-994	-1	9.474
350	ENSG00000091583	17	-1	61655974	61653975	61657974	TAATATAAT	61654150	1825	61654158	1817	1	5.862
				61655974	61653975	61657974	AAAAACAAA	61654256	1719	61654264	1711	-1	8.914
2158	ENSG00000101981	X	1	138440561	138438561	138442560	TTGGACAAA	138441494	934	138441502	942	1	6.016
383	ENSG00000118520	6	1	131936059	131934059	131938058	ATGAATAAT	131935824	-235	131935832	-227	1	5.865
3273	ENSG00000113905	3	1	187866487	187864487	187868486	TTAATCAAT	187866435	-52	187866443	-44	1	8.775
462	ENSG00000117601	1	-1	172153139	172151140	172155139	TTAAGCAAA	172153193	-54	172153201	-62	1	5.779
				172153139	172151140	172155139	TTAAACAAC	172153216	-77	172153224	-85	-1	7.440
1571	ENSG00000130649	10	1	135190857	135188857	135192856	GAAAATAAT	135188983	-1874	135188991	-1866	-1	8.003
				135190857	135188857	135192856	GCTAATAAT	135190745	-112	135190753	-104	1	6.366
				135200555	135198555	135202554	TAAAACATT	135199099	-1456	135199107	-1448	-1	6.342
5105	ENSG00000124253	20	1	55569543	55567543	55571542	GTACACAAA	55569204	-339	55569212	-331	1	8.214
				55569543	55567543	55571542	ATTAACAAC	55569381	-162	55569389	-154	1	6.352
229	ENSG00000136872	9	-1	103237926	103235927	103239926	TCTCACAAT	103237073	854	103237081	846	1	6.965
				103237926	103235927	103239926	GTAAATAAA	103237407	520	103237415	512	1	7.334

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



Select Custom Analysis Parameters

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

Species:
● human ○ mouse

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

● Paste gene IDs (max. 1000 genes):

259
5265
350
335
335
1558

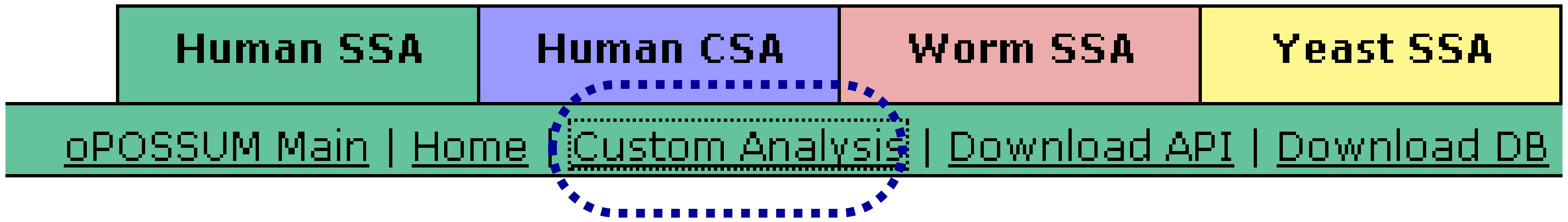
○ **OR** upload a file containing a list of gene identifiers:

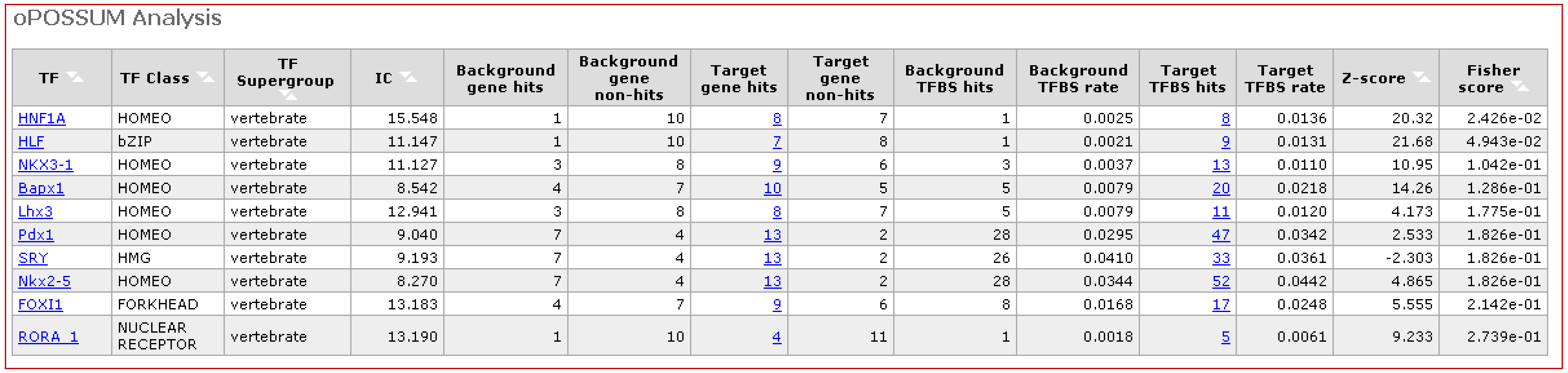
STEP 1b: Enter a background list of genes

○ Use background set of 1000 random genes
● **OR** Paste background gene IDs (max. 1000 genes):

1281
1281
1805
125
10551

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

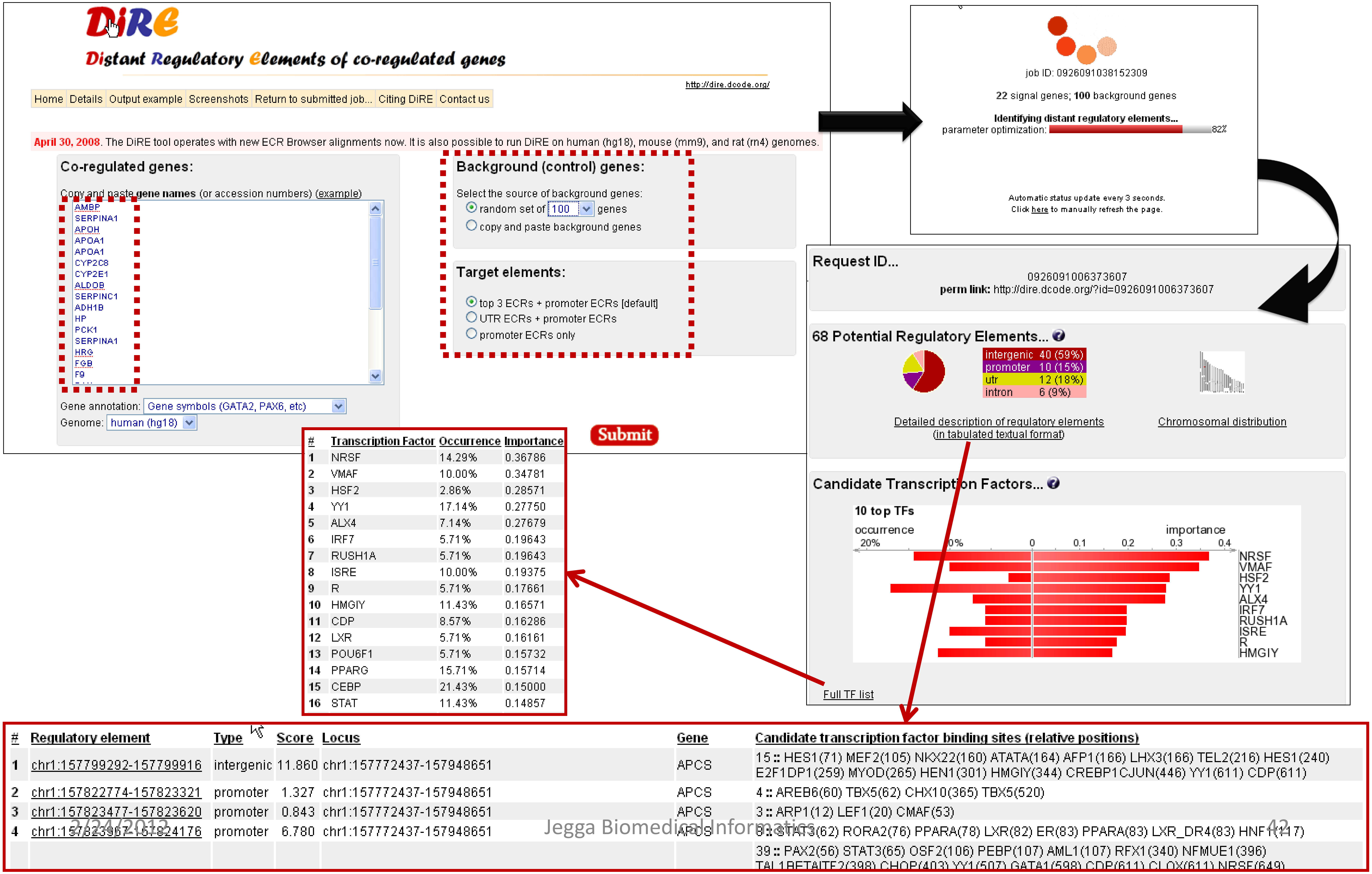




oPOSSUM Analysis

TF	TF Class	TF Supergroup	IC	Background gene hits	Background gene non-hits	Target gene hits	Target gene non-hits	Background TFBS hits	Background TFBS rate	Target TFBS hits	Target TFBS rate	Z-score	Fisher score
HNF1A	HOMEO	vertebrate	15.548	1	10	8	7	1	0.0025	8	0.0136	20.32	2.426e-02
HLF	bZIP	vertebrate	11.147	1	10	7	8	1	0.0021	9	0.0131	21.68	4.943e-02
NKX3-1	HOMEO	vertebrate	11.127	3	8	9	6	3	0.0037	13	0.0110	10.95	1.042e-01
Bapx1	HOMEO	vertebrate	8.542	4	7	10	5	5	0.0079	20	0.0218	14.26	1.286e-01
Lhx3	HOMEO	vertebrate	12.941	3	8	8	7	5	0.0079	11	0.0120	4.173	1.775e-01
Pdx1	HOMEO	vertebrate	9.040	7	4	13	2	28	0.0295	47	0.0342	2.533	1.826e-01
SRY	HMG	vertebrate	9.193	7	4	13	2	26	0.0410	33	0.0361	-2.303	1.826e-01
Nkx2-5	HOMEO	vertebrate	8.270	7	4	13	2	28	0.0344	52	0.0442	4.865	1.826e-01
FOXI1	FORKHEAD	vertebrate	13.183	4	7	9	6	8	0.0168	17	0.0248	5.555	2.142e-01
RORA_1	NUCLEAR RECEPTOR	vertebrate	13.190	1	10	4	11	1	0.0018	5	0.0061	9.233	2.739e-01

DiRE (http://dire.dcode.org/)



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

Co-regulated genes:

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

AMBP
SERPINA1
APOH
APOA4
APOA1
CYP2C8
CYP2E1
ALDOB
SERPINC1
ADH1B
HP
PCK1
SERPINA1
HRG
FGB
F9

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

Genome: human (hg18)

Background (control) genes:

Select the source of background genes:

● random set of 100 genes

○ copy and paste background genes

Target elements:

● top 3 ECRs + promoter ECRs [default]

○ UTR ECRs + promoter ECRs

○ promoter ECRs only

job ID: 0926091038152309

22 signal genes; **100** background genes

Identifying distant regulatory elements...

parameter optimization: 82%

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

Request ID...

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

68 Potential Regulatory Elements...

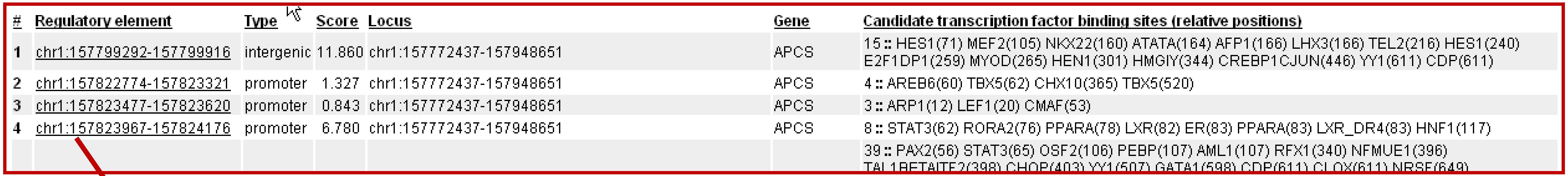
intergenic	40	(59%)
promoter	10	(15%)
utr	12	(18%)
intron	6	(9%)

Candidate Transcription Factors...

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

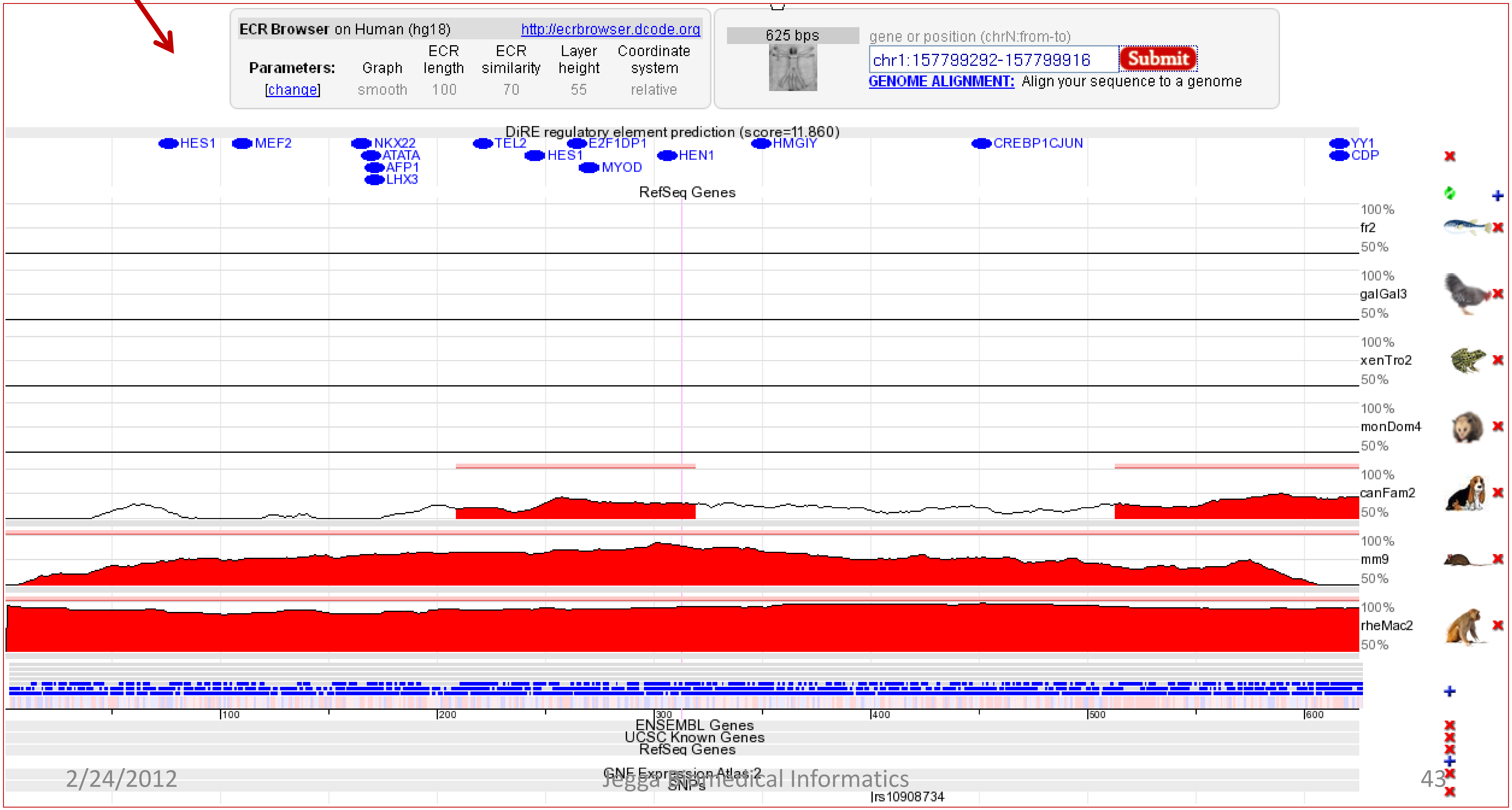
#	Regulatory element	Type	Score	Locus	Gene	Candidate transcription factor binding sites (relative positions)
1	chr1:157799292-157799916	intergenic	11.860	chr1:157772437-157948651	APCS	15 :: HES1(71) MEF2(105) NKX22(160) ATATA(164) AFP1(166) LHX3(166) TEL2(216) HES1(240) E2F1DP1(259) MYOD(265) HEN1(301) HMGIY(344) CREBP1CJUN(446) YY1(611) CDP(611)
2	chr1:157822774-157823321	promoter	1.327	chr1:157772437-157948651	APCS	4 :: AREB6(60) TBX5(62) CHX10(365) TBX5(520)
3	chr1:157823477-157823620	promoter	0.843	chr1:157772437-157948651	APCS	3 :: ARP1(12) LEF1(20) CMAF(53)
4	chr1:15[illegible]-157824176	promoter	6.780	chr1:157772437-157948651	APCS	8 :: STAT3(62) RORA2(76) PPARA(78) LXR(82) ER(83) PPARA(83) LXR_DR4(83) HNF1(117)
						39 :: PAX2(56) STAT3(65) OSF2(106) PEBP(107) AML1(107) RFX1(340) NFMUE1(396) TAL1BETAITF2(398) CHOP(403) YY1(507) GATA1(598) CDP(611) CLOX(611) NRSF(649)

DiRE (http://dire.dcode.org/)



#	Regulatory element	Type	Score	Locus	Gene	Candidate transcription factor binding sites (relative positions)
1	chr1:157799292-157799916	intergenic	11.860	chr1:157772437-157948651	APCS	15 :: HES1(71) MEF2(105) NKX22(160) ATATA(164) AFP1(166) LHX3(166) TEL2(216) HES1(240) E2F1DP1(259) MYOD(265) HEN1(301) HMGIY(344) CREBP1CJUN(446) YY1(611) CDP(611)
2	chr1:157822774-157823321	promoter	1.327	chr1:157772437-157948651	APCS	4 :: AREB6(60) TBX5(62) CHX10(365) TBX5(520)
3	chr1:157823477-157823620	promoter	0.843	chr1:157772437-157948651	APCS	3 :: ARP1(12) LEF1(20) CMAF(53)
4	chr1:157823967-157824176	promoter	6.780	chr1:157772437-157948651	APCS	8 :: STAT3(62) RORA2(76) PPARA(78) LXR(82) ER(83) PPARA(83) LXR_DR4(83) HNF1(117)
						39 :: PAX2(56) STAT3(65) OSF2(106) PEBP(107) AML1(107) RFX1(340) NFMUE1(396) TAL1BETAITF2(398) CHOP(403) YY1(507) GATA1(598) CDP(611) CLOX(611) NRSF(649)

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



ECR Browser on Human (hg18) http://ecrbrowser.dcode.org

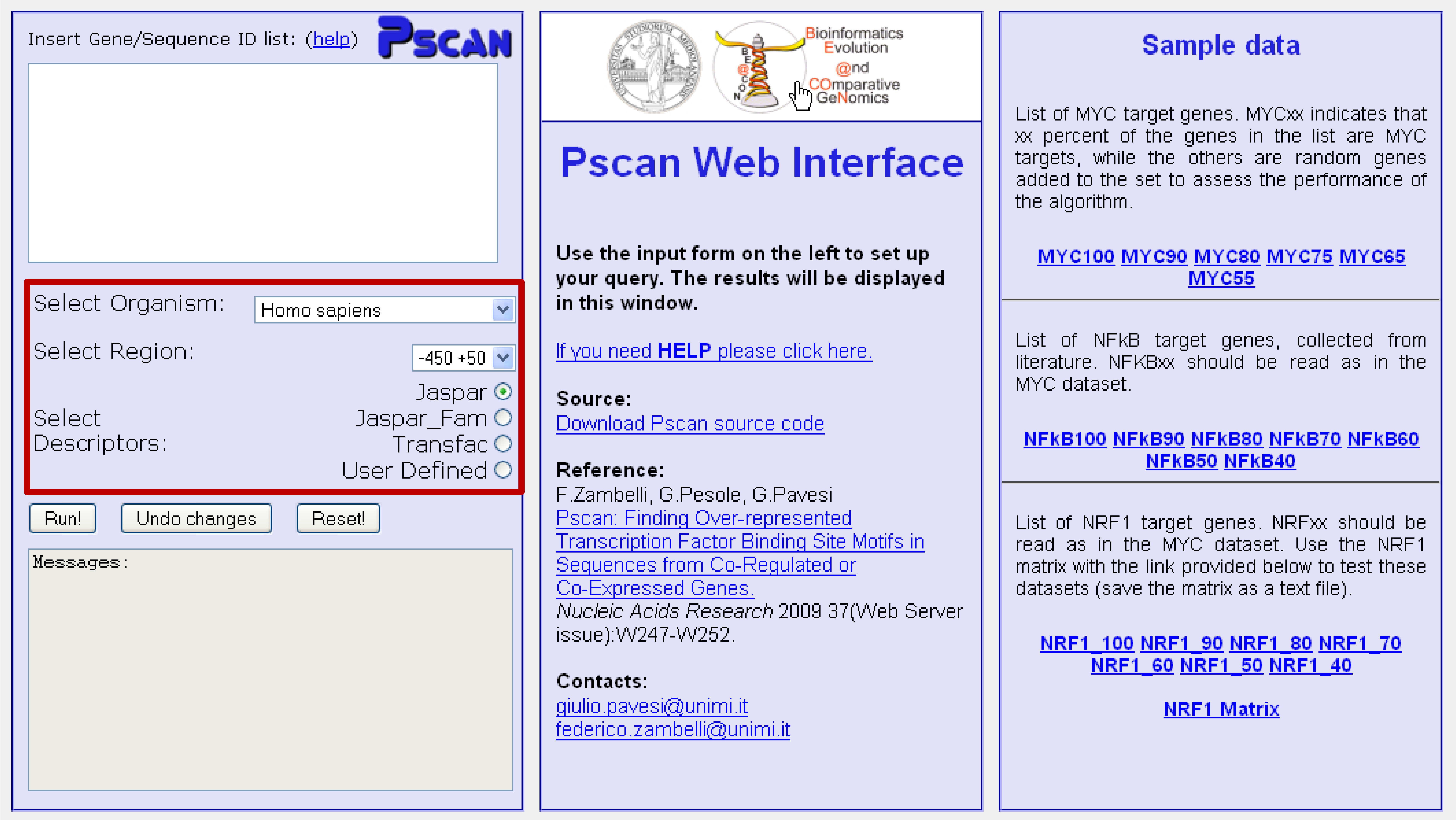
Parameters:	Graph	ECR length	ECR similarity	Layer height	Coordinate system
[change]	smooth	100	70	55	relative

625 bps

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

GENOME ALIGNMENT: Align your sequence to a genome

Pscan (http://159.149.109.9/pscan)



Insert Gene/Sequence ID list: (help)

Select Organism: Homo sapiens

Select Region: -450 +50

Select Descriptors:
Jaspar ●
Jaspar_Fam ○
Transfac ○
User Defined ○

Pscan Web Interface

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

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

Source:
Download Pscan source code

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

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

Sample data

List of MYC target genes. MYCxx indicates that xx percent of the genes in the list are MYC targets, while the others are random genes added to the set to assess the performance of the algorithm.

MYC100 MYC90 MYC80 MYC75 MYC65 MYC55

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

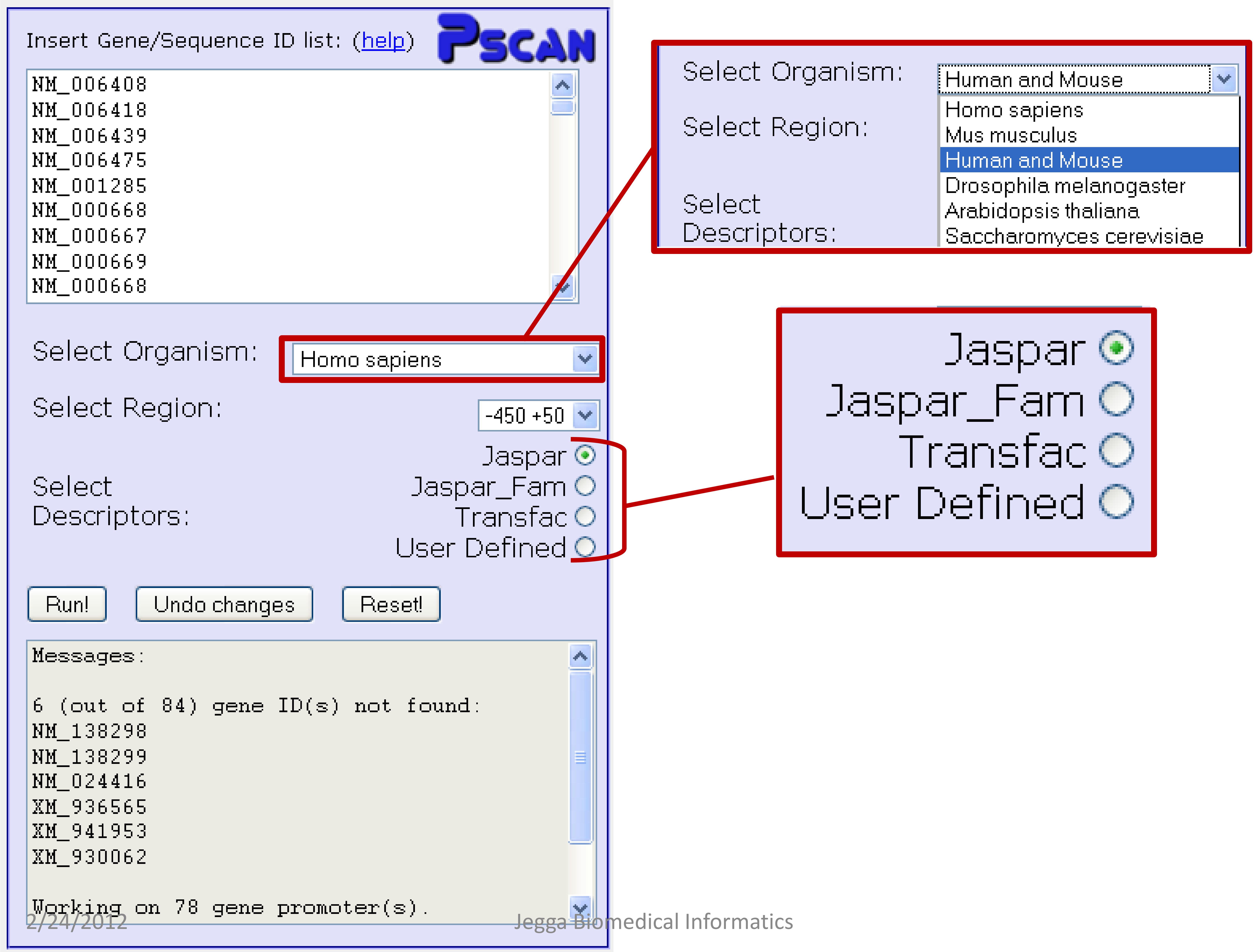
NFkB100 NFkB90 NFkB80 NFkB70 NFkB60 NFkB50 NFkB40

List of NRF1 target genes. NRFxx should be read as in the MYC dataset. Use the NRF1 matrix with the link provided below to test these datasets (save the matrix as a text file).

NRF1_100 NRF1_90 NRF1_80 NRF1_70 NRF1_60 NRF1_50 NRF1_40

NRF1 Matrix

Pscan (http://159.149.109.9/pscan)



Insert Gene/Sequence ID list: (help)

```
NM_006408
NM_006418
NM_006439
NM_006475
NM_001285
NM_000668
NM_000667
NM_000669
NM_000668
```

Select Organism: Homo sapiens

Select Region: -450 +50

Select Descriptors:
● Jaspar
○ Jaspar_Fam
○ Transfac
○ User Defined

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

Select Region:

Select Descriptors:

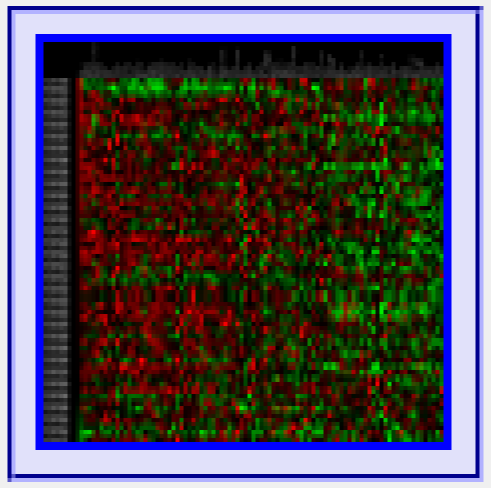
```
Messages:

6 (out of 84) gene ID(s) not found:
NM_138298
NM_138299
NM_024416
XM_936565
XM_941953
XM_930062

Working on 78 gene promoter(s).
```

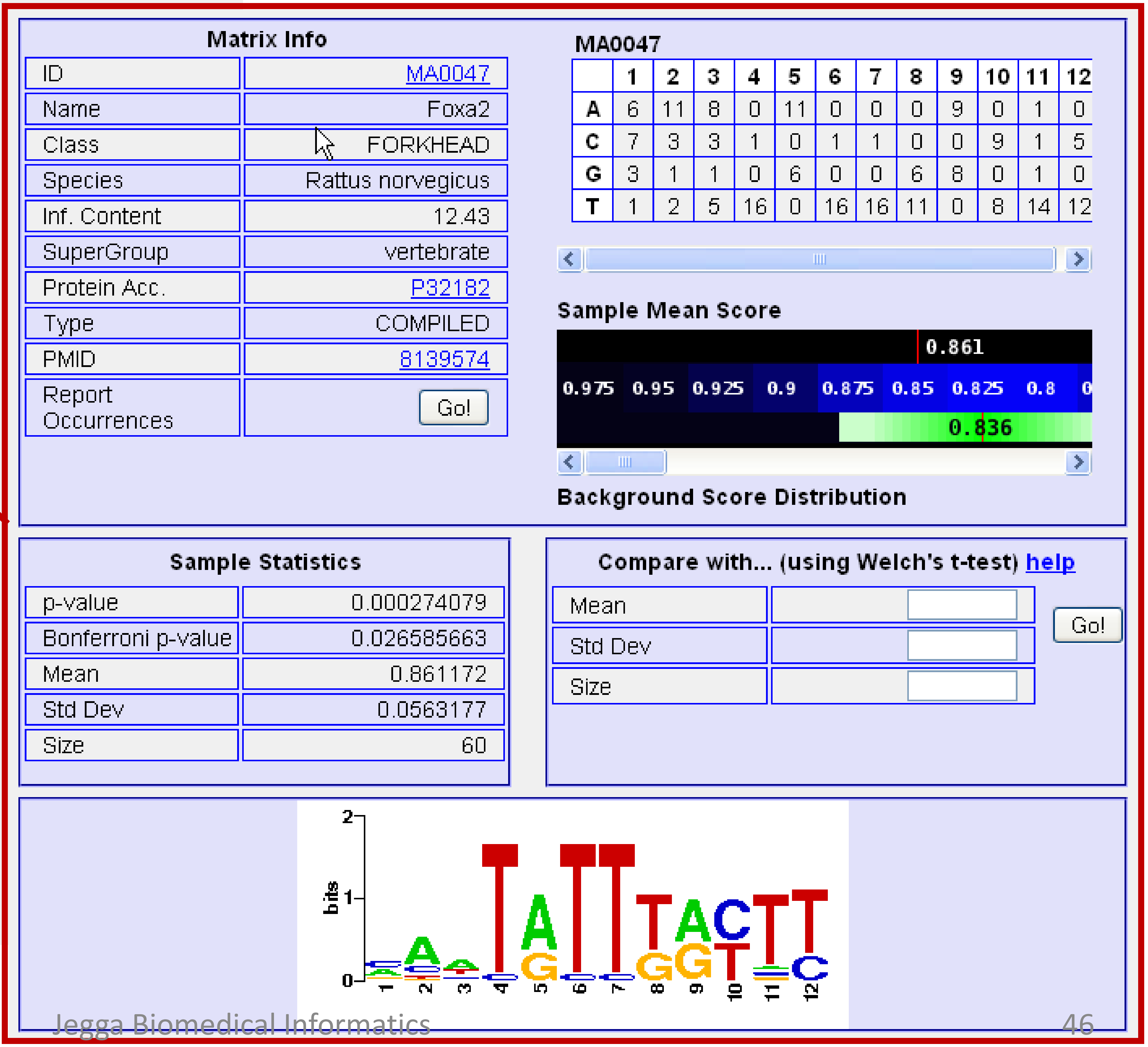
Pscan (http://159.149.109.9/pscan)

View Text Results

97 TF profiles used

Matrix Name	P-value
TBP	1.59074e-08
Foxa2	0.000274079
FOXL1	0.000657034
MEF2A	0.000657227
Hand1-Tcfe2a	0.000697277
Nobox	0.000790445
FOXI1	0.000804377
PBX1	0.00124224
SRF	0.00124647
Evi1	0.00128699
TEAD1	0.00212538
Lhx3	0.00303459
Foxq1	0.00355502
Prrx2	0.00486451
Lhx3	0.00527407
NKX3-1	0.00590862
NFIL3	0.00642618
REL	0.00685234
Pax6	0.00765503
Foxd3	0.00776631
HNF1A	0.00783389
Cebpa	0.00920516
Nkx2-5	0.00940039





Matrix Info

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

MA0047

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

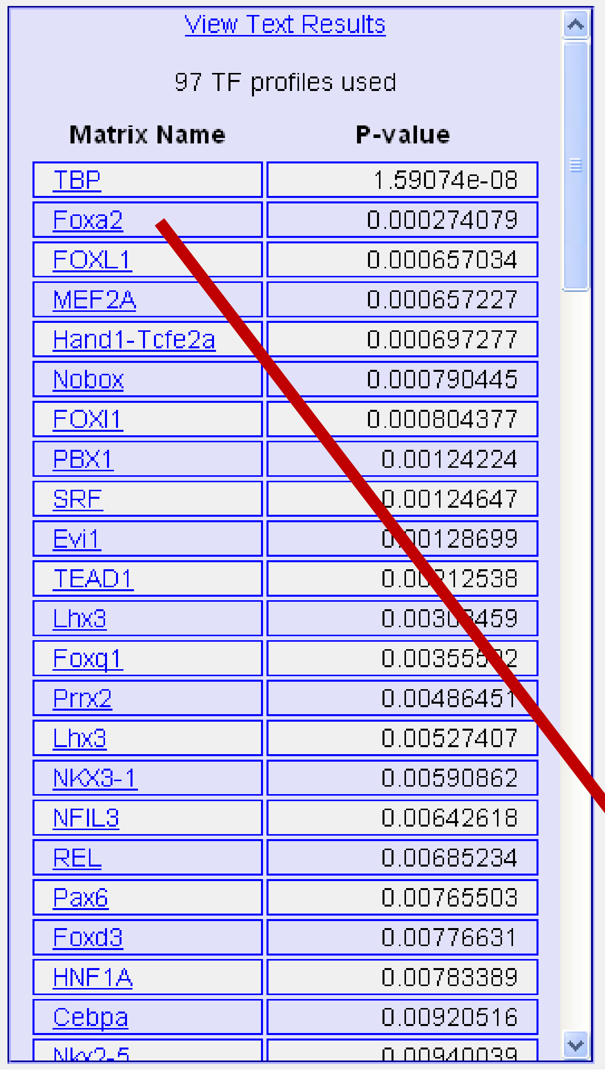
Sample Statistics

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

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

Mean	
Std Dev	
Size	

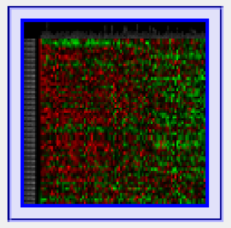
Pscan (http://159.149.109.9/pscan)

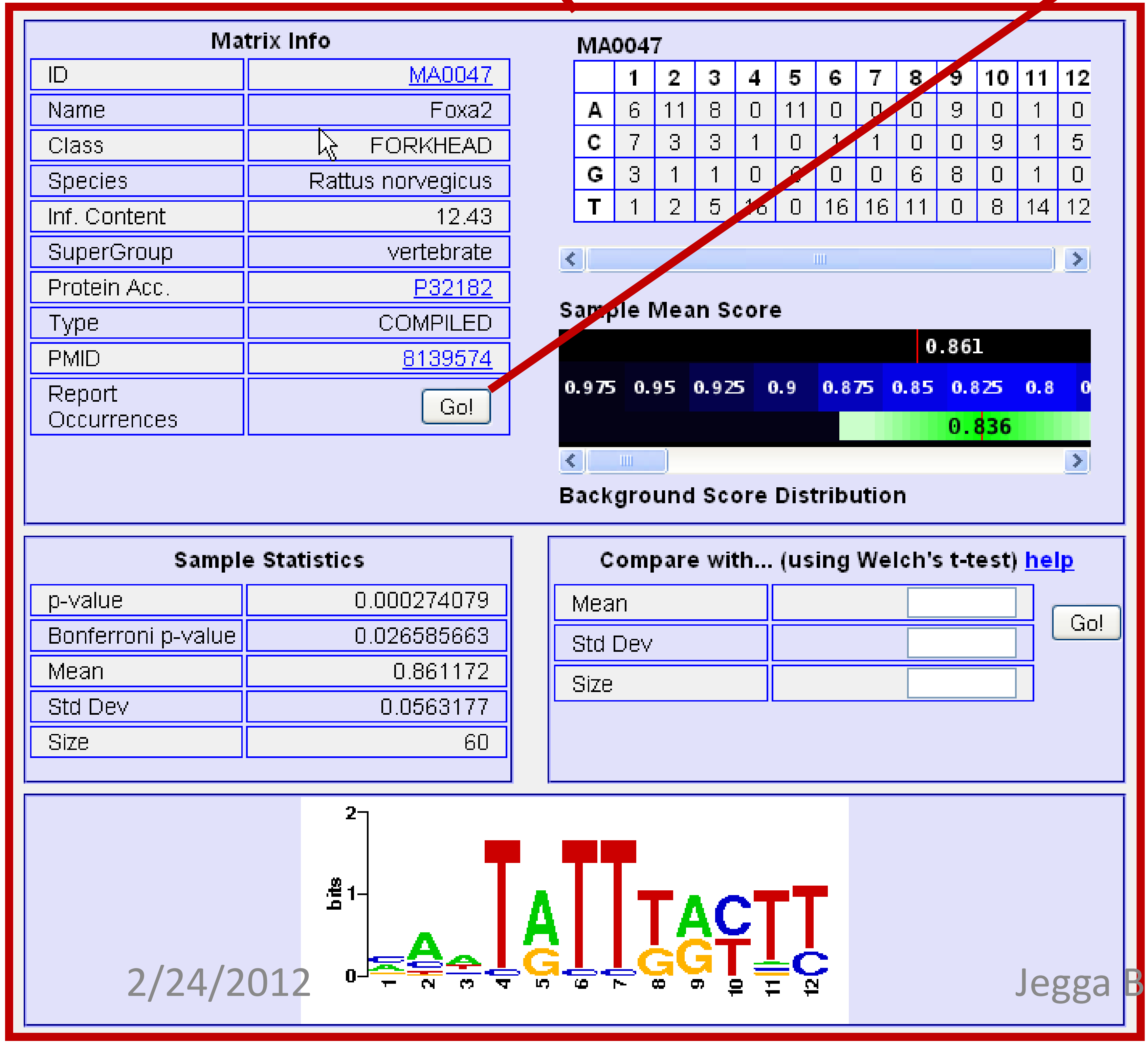


View Text Results

97 TF profiles used

Matrix Name	P-value
TBP	1.59074e-08
Foxa2	0.000274079
FOXL1	0.000657034
MEF2A	0.000657227
Hand1-Tcfe2a	0.000697277
Nobox	0.000790445
FOXI1	0.000804377
PBX1	0.00124224
SRF	0.00124647
Evi1	0.00128699
TEAD1	0.00[illegible]12538
Lhx3	0.003[illegible]459
Foxq1	0.00355[illegible]2
Prrx2	0.00486451
Lhx3	0.00527407
NKX3-1	0.00590862
NFIL3	0.00642618
REL	0.00685234
Pax6	0.00765503
Foxd3	0.00776631
HNF1A	0.00783389
Cebpa	0.00920516
Nkx2-5	0.00940039





Matrix Info

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

MA0047

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

Sample Mean Score

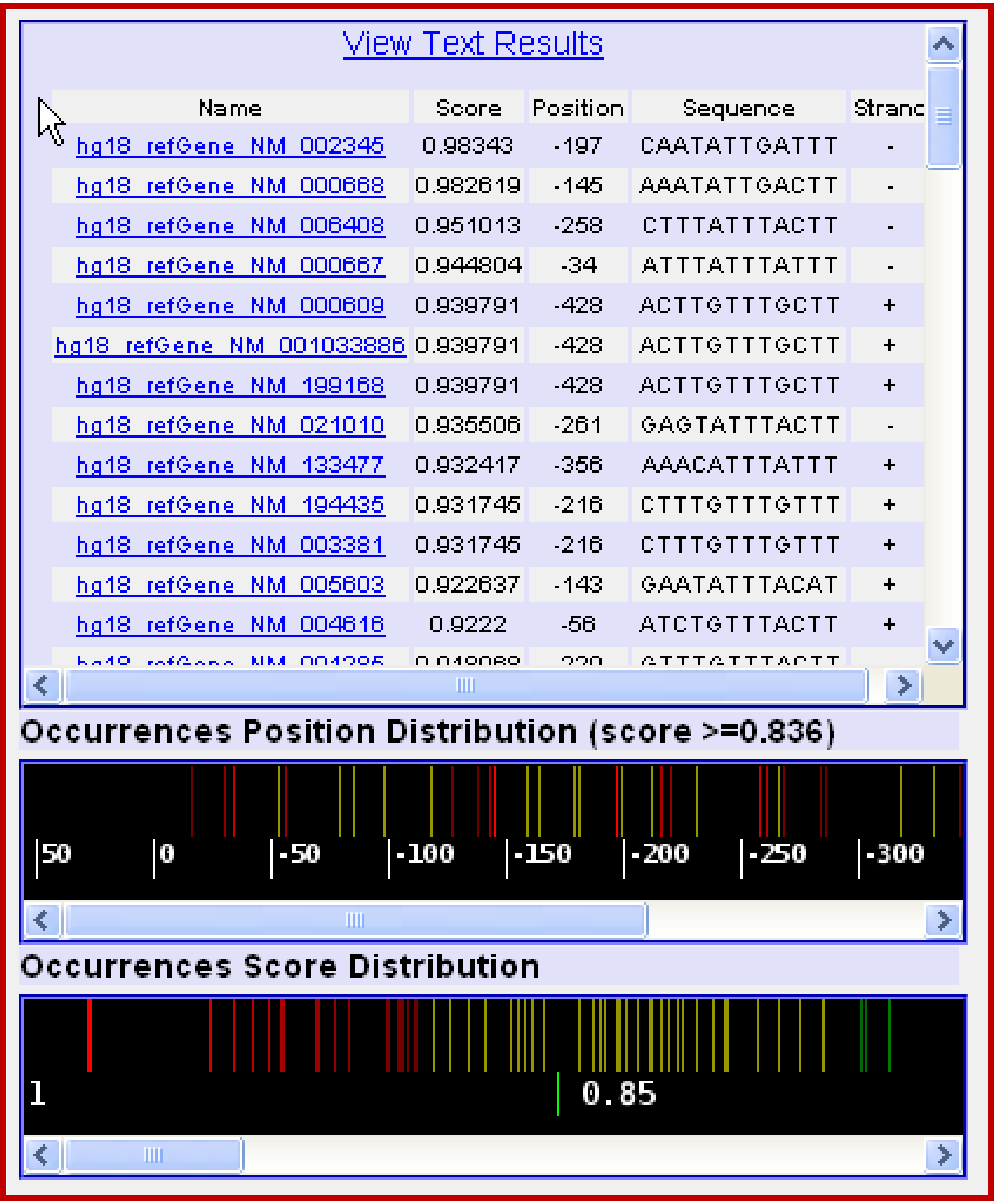
Background Score Distribution

Sample Statistics

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

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

Mean	
Std Dev	
Size	



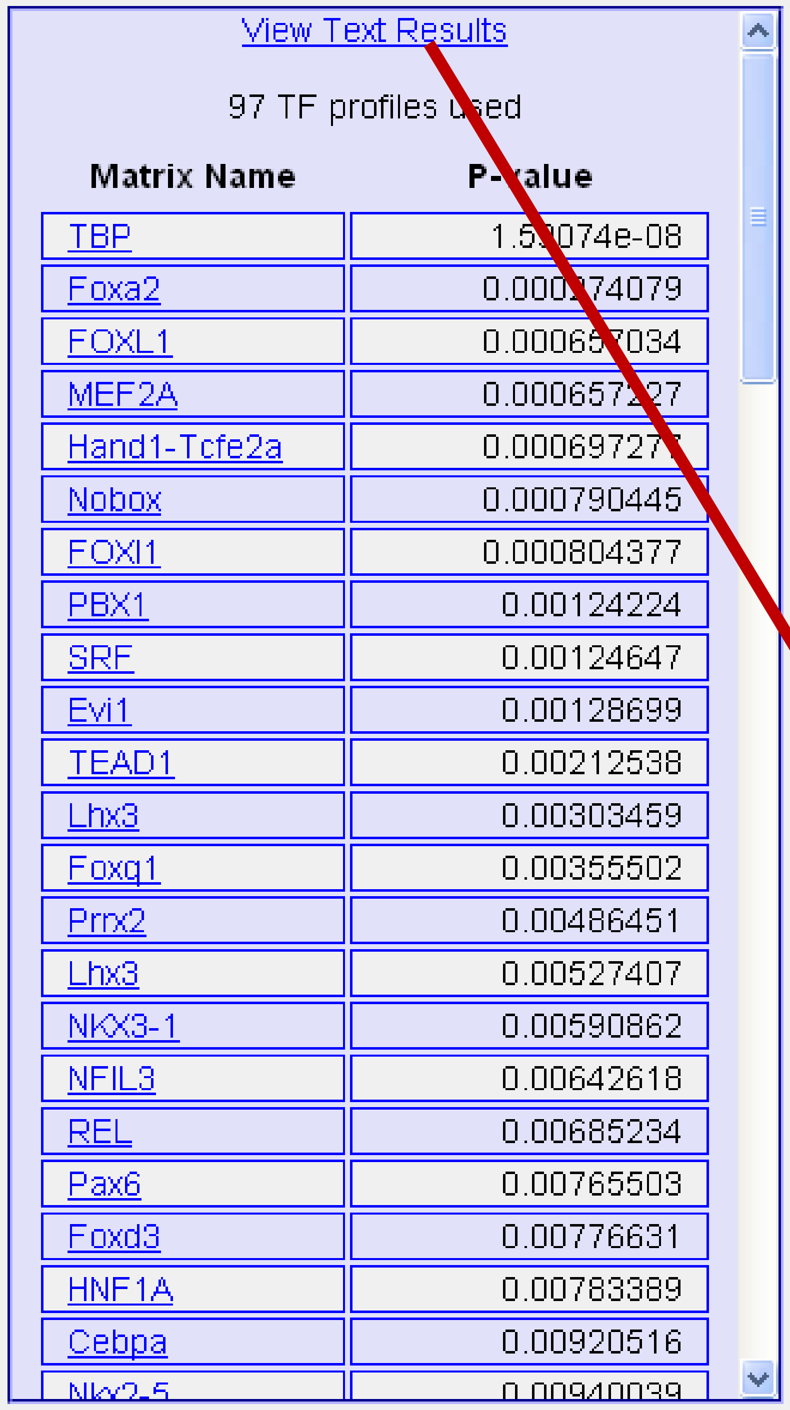
View Text Results

Name	Score	Position	Sequence	Stranc
hg18_refGene_NM_002345	0.98343	-197	CAATATTGATTT	-
hg18_refGene_NM_000668	0.982619	-145	AAATATTGACTT	-
hg18_refGene_NM_006408	0.951013	-258	CTTTATTTACTT	-
hg18_refGene_NM_000667	0.944804	-34	ATTTATTTATTT	-
hg18_refGene_NM_000609	0.939791	-428	ACTTGTTTGCTT	+
hg18_refGene_NM_001033886	0.939791	-428	ACTTGTTTGCTT	+
hg18_refGene_NM_199168	0.939791	-428	ACTTGTTTGCTT	+
hg18_refGene_NM_021010	0.935506	-261	GAGTATTTACTT	-
hg18_refGene_NM_133477	0.932417	-356	AAACATTTATTT	+
hg18_refGene_NM_194435	0.931745	-216	CTTTGTTTGTTT	+
hg18_refGene_NM_003381	0.931745	-216	CTTTGTTTGTTT	+
hg18_refGene_NM_005603	0.922637	-143	GAATATTTACAT	+
hg18_refGene_NM_004616	0.9222	-56	ATCTGTTTACTT	+

Occurrences Position Distribution (score >=0.836)

Occurrences Score Distribution

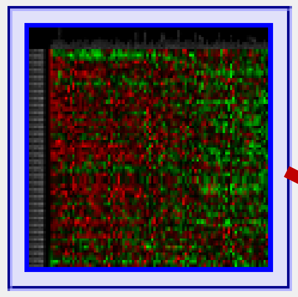
Pscan (http://159.149.109.9/pscan)

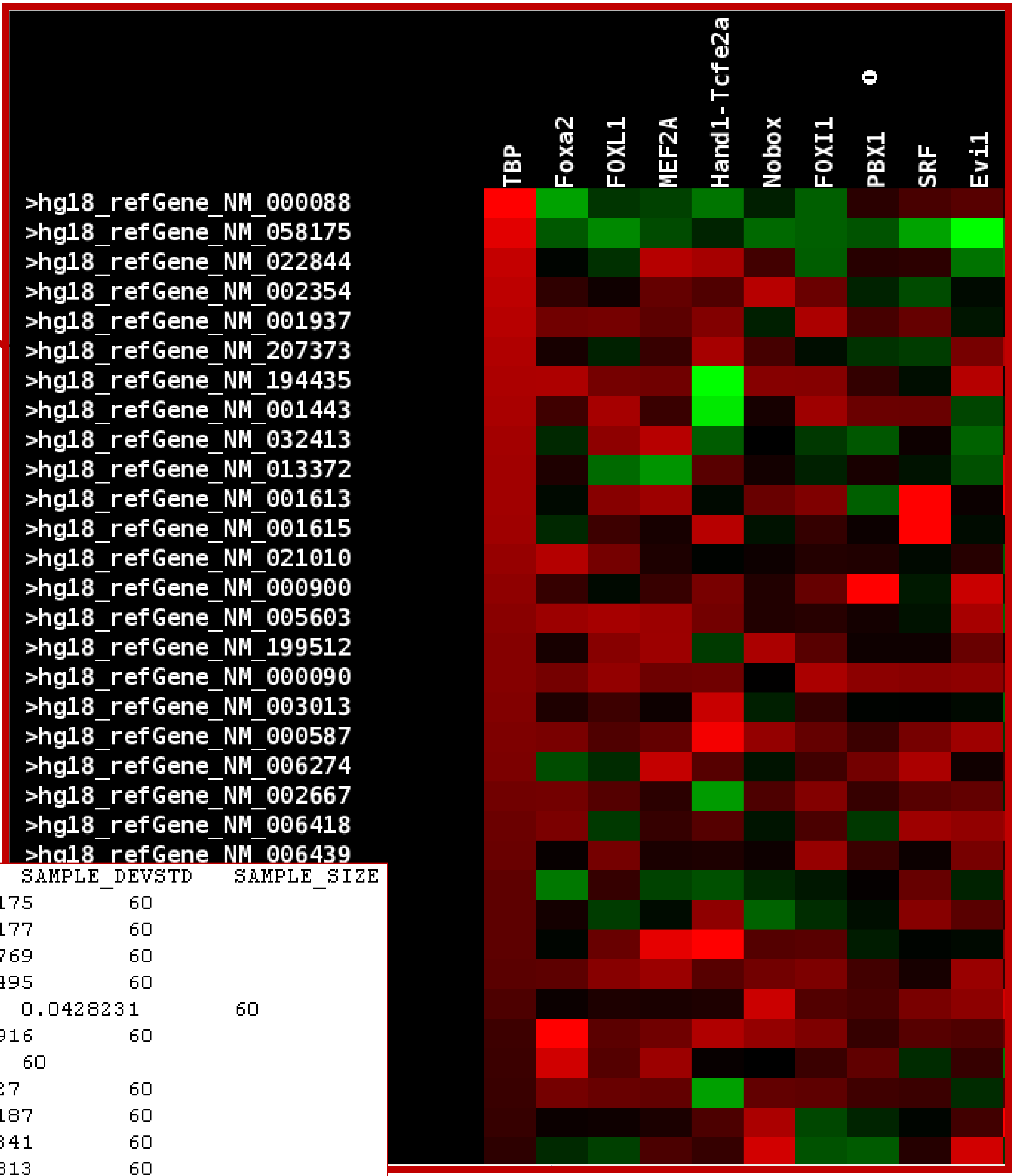


View Text Results

97 TF profiles used

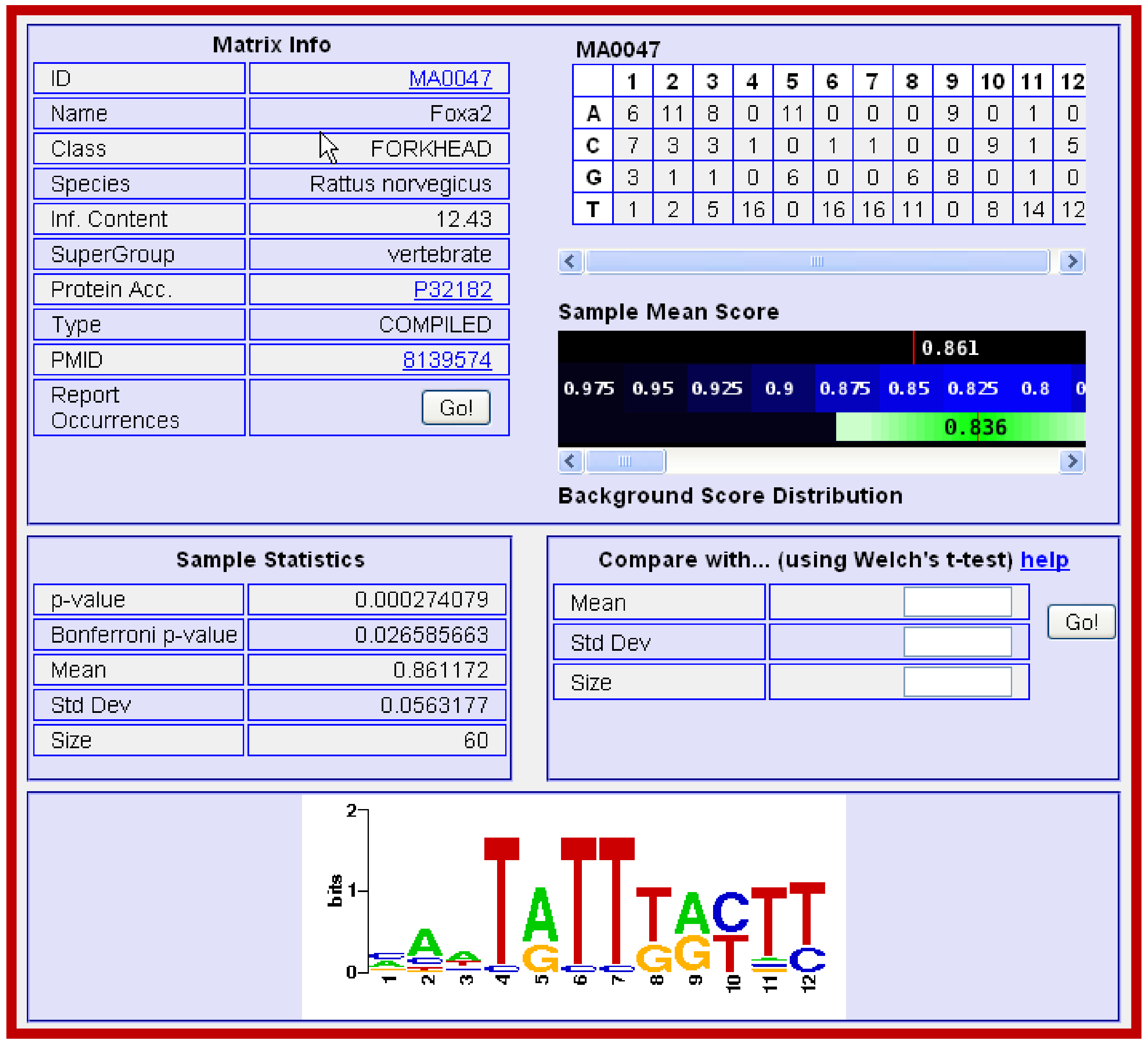
Matrix Name	P-value
TBP	1.59074e-08
Foxa2	0.000274079
FOXL1	0.000657034
MEF2A	0.000657227
Hand1-Tcfe2a	0.000697277
Nobox	0.000790445
FOXI1	0.000804377
PBX1	0.00124224
SRF	0.00124647
Evi1	0.00128699
TEAD1	0.00212538
Lhx3	0.00303459
Foxq1	0.00355502
Prrx2	0.00486451
Lhx3	0.00527407
NKX3-1	0.00590862
NFIL3	0.00642618
REL	0.00685234
Pax6	0.00765503
Foxd3	0.00776631
HNF1A	0.00783389
Cebpa	0.00920516
Nkx2-5	0.00940039





```
TF_NAME MATRIX_ID        Z_SCORE P_VALUE SAMPLE_AVERAGE   BACKGROUND_AVERAGE      SAMPLE_DEVSTD   SAMPLE_SIZE
TBP     MA0108  5.52625 1.59074e-08     0.859494        0.816811        0.0493175       60
Foxa2   MA0047  3.45141 0.000274079     0.861172        0.836541        0.0563177       60
FOXL1   MA0033  3.21019 0.000657034     0.918922        0.891892        0.0523769       60
MEF2A   MA0052  3.20941 0.000657227     0.810217        0.777574        0.0710495       60
Hand1-Tcfe2a    MA0092  3.19081 0.000697277     0.895392        0.879315        0.0428231       60
Nobox   MA0125  3.15645 0.000790445     0.880691        0.854103        0.0553916       60
FOXI1   MA0042  3.15094 0.000804377     0.85506 0.825768        0.0653533       60
PBX1    MA0070  3.02199 0.00124224      0.802727        0.781876        0.048527        60
SRF     MA0083  3.02072 0.00124647      0.759912        0.740766        0.0462187       60
Evi1    MA0029  3.01105 0.00128699      0.769179        0.746618        0.0546341       60
TEAD1   MA0090  2.85352 0.00212538      0.827634        0.810446        0.0565813       60
Lhx3    MA0134  2.7417  0.00303459      0.834887        0.804852        0.0683236       60
Foxq1   MA0040  2.68955 0.00355502      0.825721        0.802203        0.0522346       60
Prrx2   MA0075  2.58254 0.00486451      0.92434 0.89143 0.0897038       60
Lhx3    MA0135  2.55439 0.00527407      0.798641        0.773954        0.0704547       60
NKX3-1  MA0124  2.51437 0.00590862      0.853125        0.829214        0.0728452       60
NFIL3   MA0025  2.48463 0.00642618      0.80459 0.783263        0.0663844       60
REL     MA0101  2.46099 0.00685234      0.8843  0.869993        0.0500624       60
Pax6    MA0069  2.42112 0.00765503      0.778947        0.766632        0.0436411       60
Foxd3   MA0041  2.4171  0.00776631      0.857889        0.837233        0.0604468       60
HNF1A   MA0046  2.41362 0.00783389      0.790719        0.770942        0.0615964       60
```

Pscan (http://159.149.109.9/pscan)



Matrix Info

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

MA0047

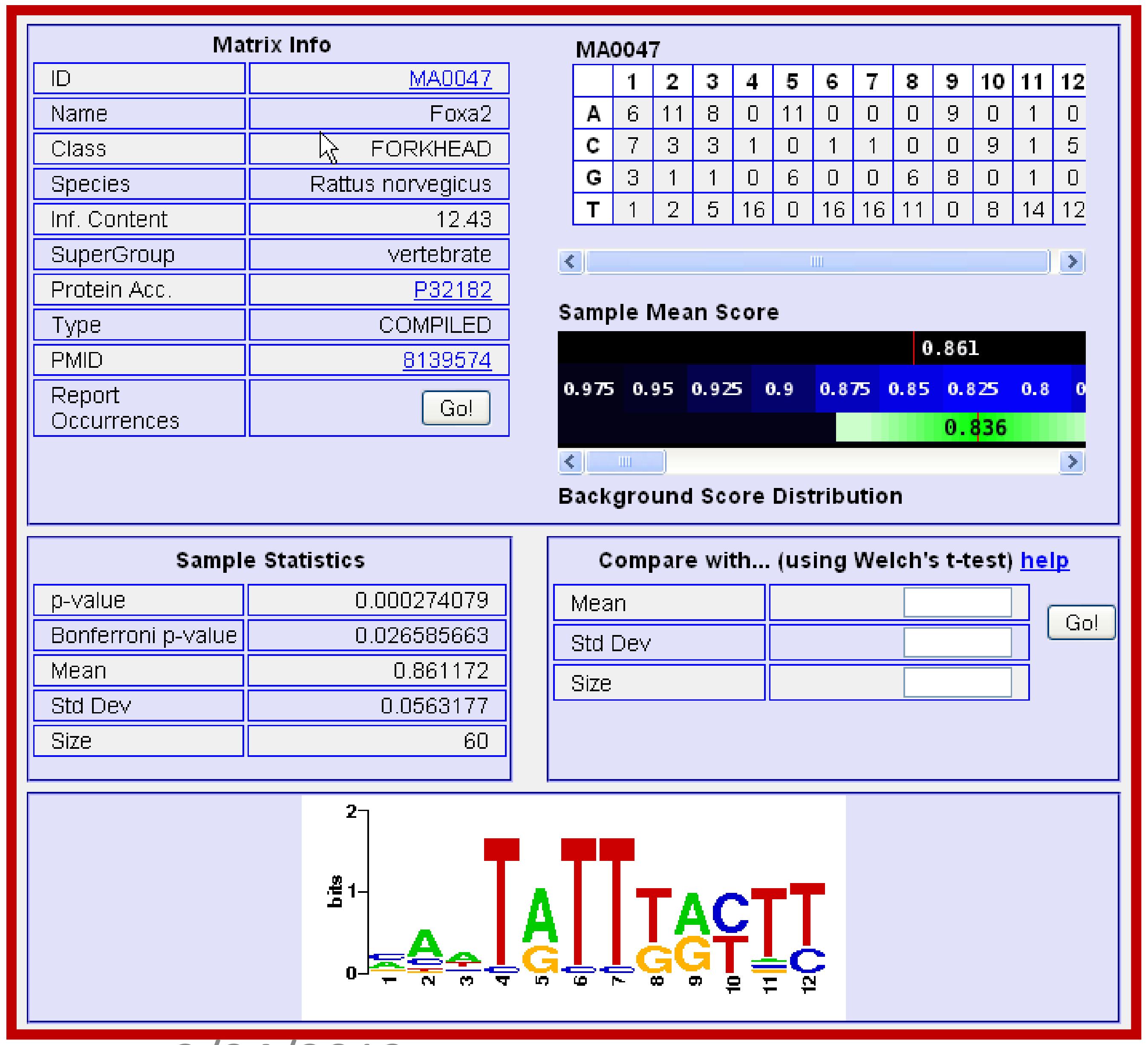
	1	2	3	4	5	6	7	8	9	10	11	12
A	6	11	8	0	11	0	0	0	9	0	1	0
C	7	3	3	1	0	1	1	0	0	9	1	5
G	3	1	1	0	6	0	0	6	8	0	1	0
T	1	2	5	16	0	16	16	11	0	8	14	12

Sample Statistics

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

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

Mean	
Std Dev	
Size	



Matrix Info

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

MA0047

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

Sample Statistics

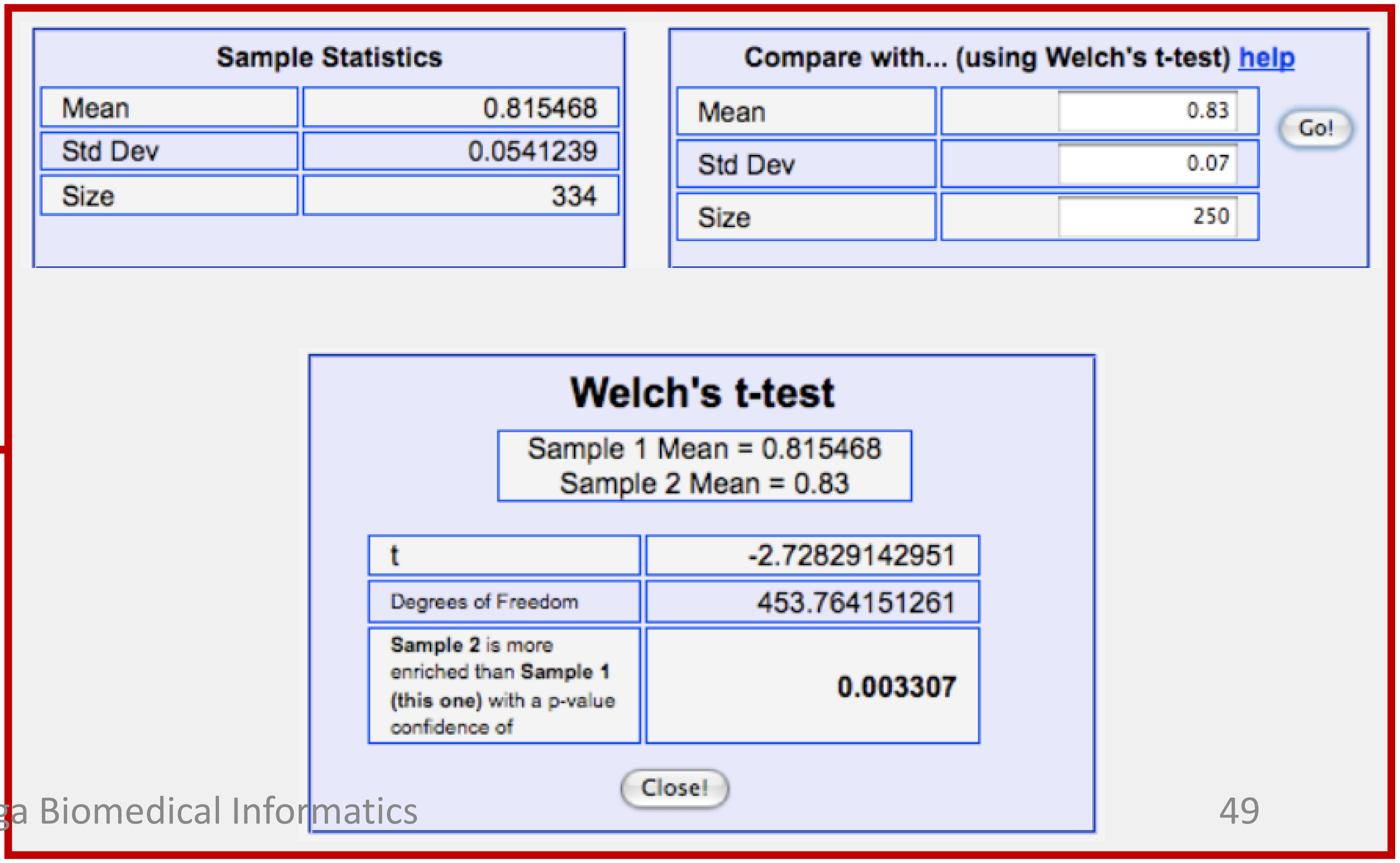
p-value	0.000274079
Bonferroni p-value	0.026585663
Mean	0.861172
Std Dev	0.0563177
Size	60

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

Mean	
Std Dev	
Size	

Comparing different input gene sets:

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



Sample Statistics

Mean	0.815468
Std Dev	0.0541239
Size	334

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

Mean	0.83
Std Dev	0.07
Size	250

Welch's t-test

Sample 1 Mean = 0.815468
Sample 2 Mean = 0.83

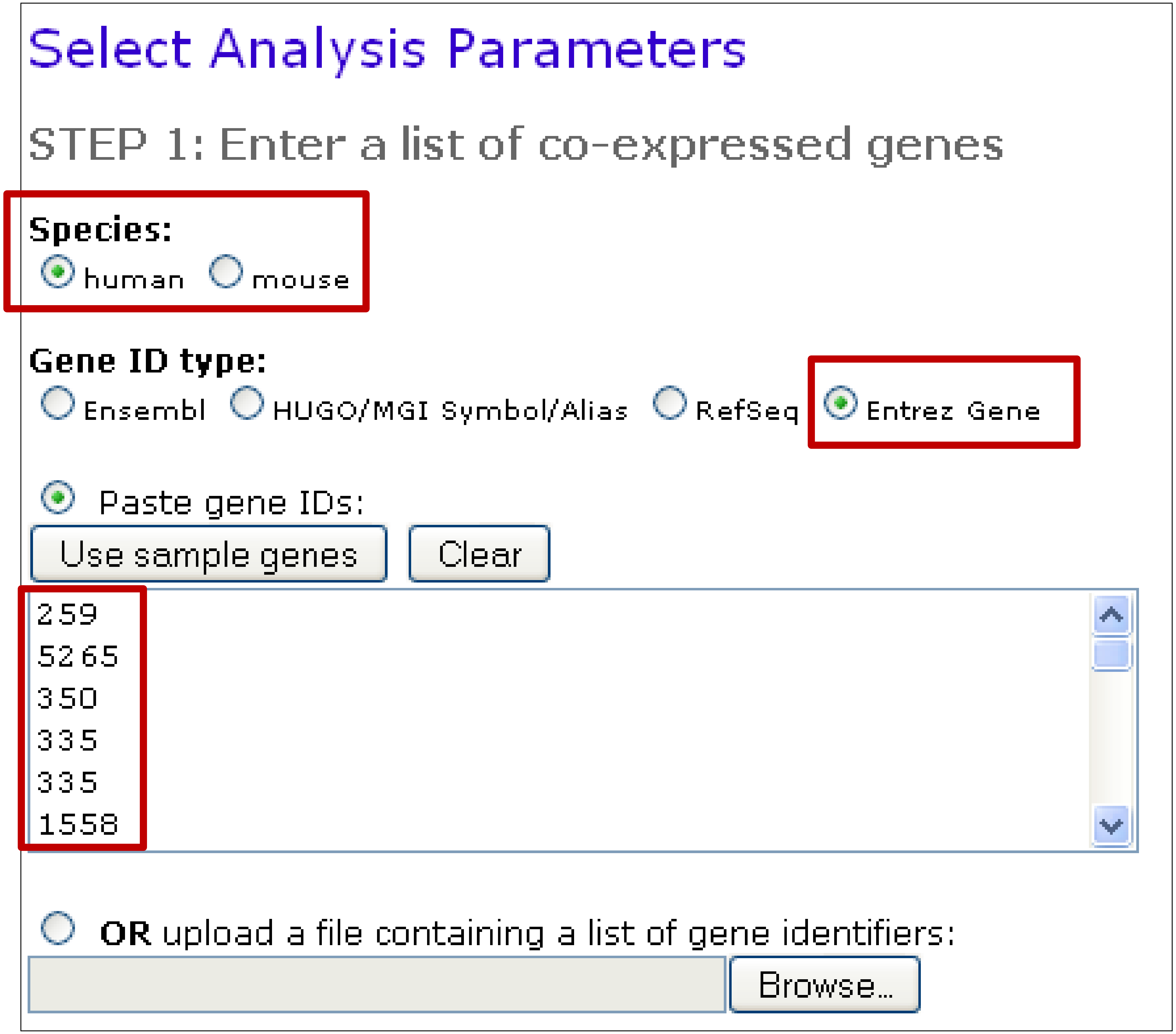
t	-2.72829142951
Degrees of Freedom	453.764151261
Sample 2 is more enriched than **Sample 1** **(this one)** with a p-value confidence of	**0.003307**

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

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

- Known transcription factor binding sites (TFBS)
  - Conserved
    - oPOSSUM
    - DiRE
  - Non-conserved
    - Pscan
    - MatInspector (*Licensed)
- Unknown TFBS or Novel motifs
  - Conserved
    - oPOSSUM
    - Weeder-H
  - Non-conserved
    - MEME
    - Weeder

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



Select Analysis Parameters

STEP 1: Enter a list of co-expressed genes

Species:
● human ○ mouse

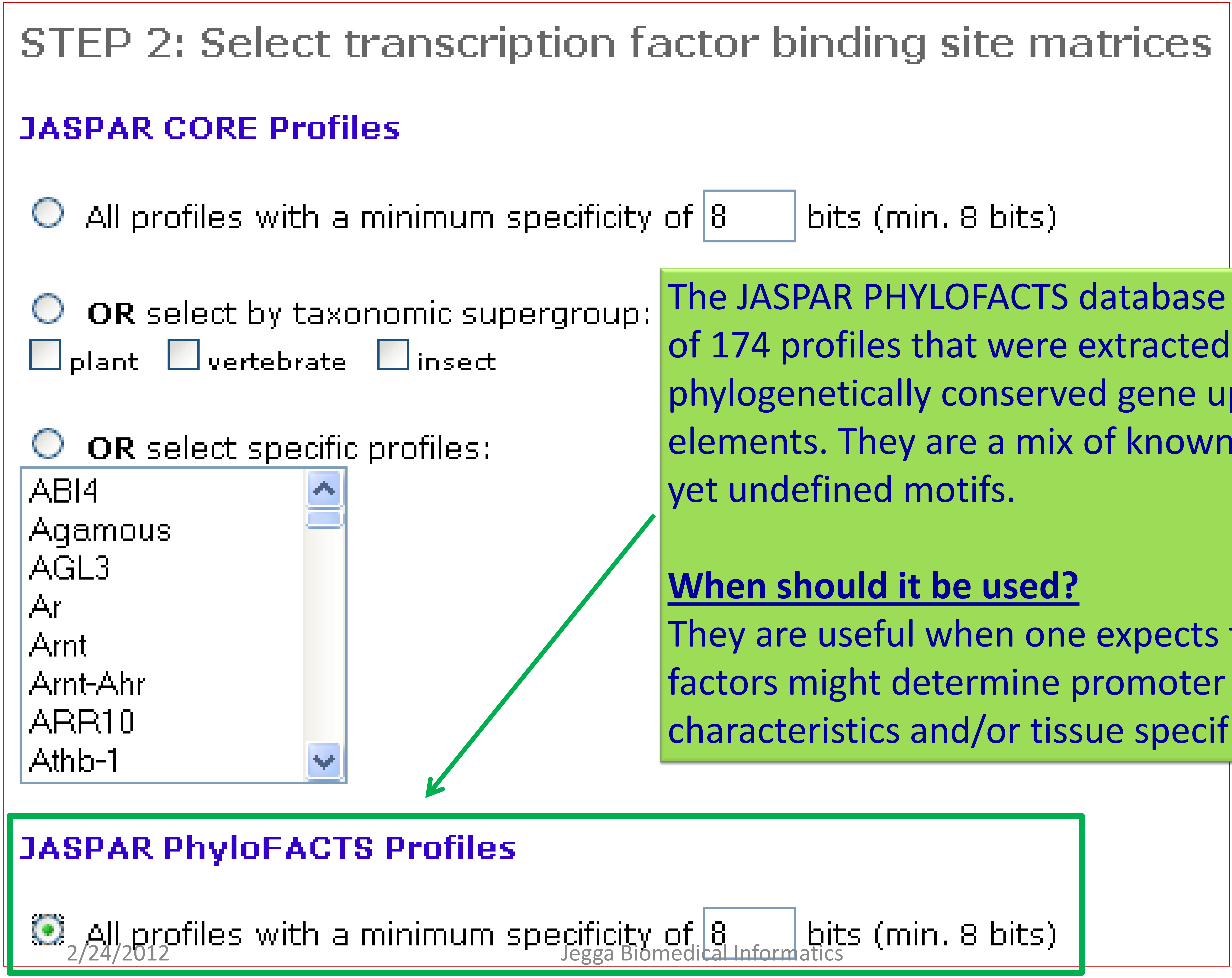
Gene ID type:
○ Ensembl ○ HUGO/MGI Symbol/Alias ○ RefSeq ● Entrez Gene

● Paste gene IDs:

259
5265
350
335
335
1558

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

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



STEP 2: Select transcription factor binding site matrices

JASPAR CORE Profiles

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

○ **OR** select by taxonomic supergroup:
☐ plant ☐ vertebrate ☐ insect

○ **OR** select specific profiles:

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

JASPAR PhyloFACTS Profiles

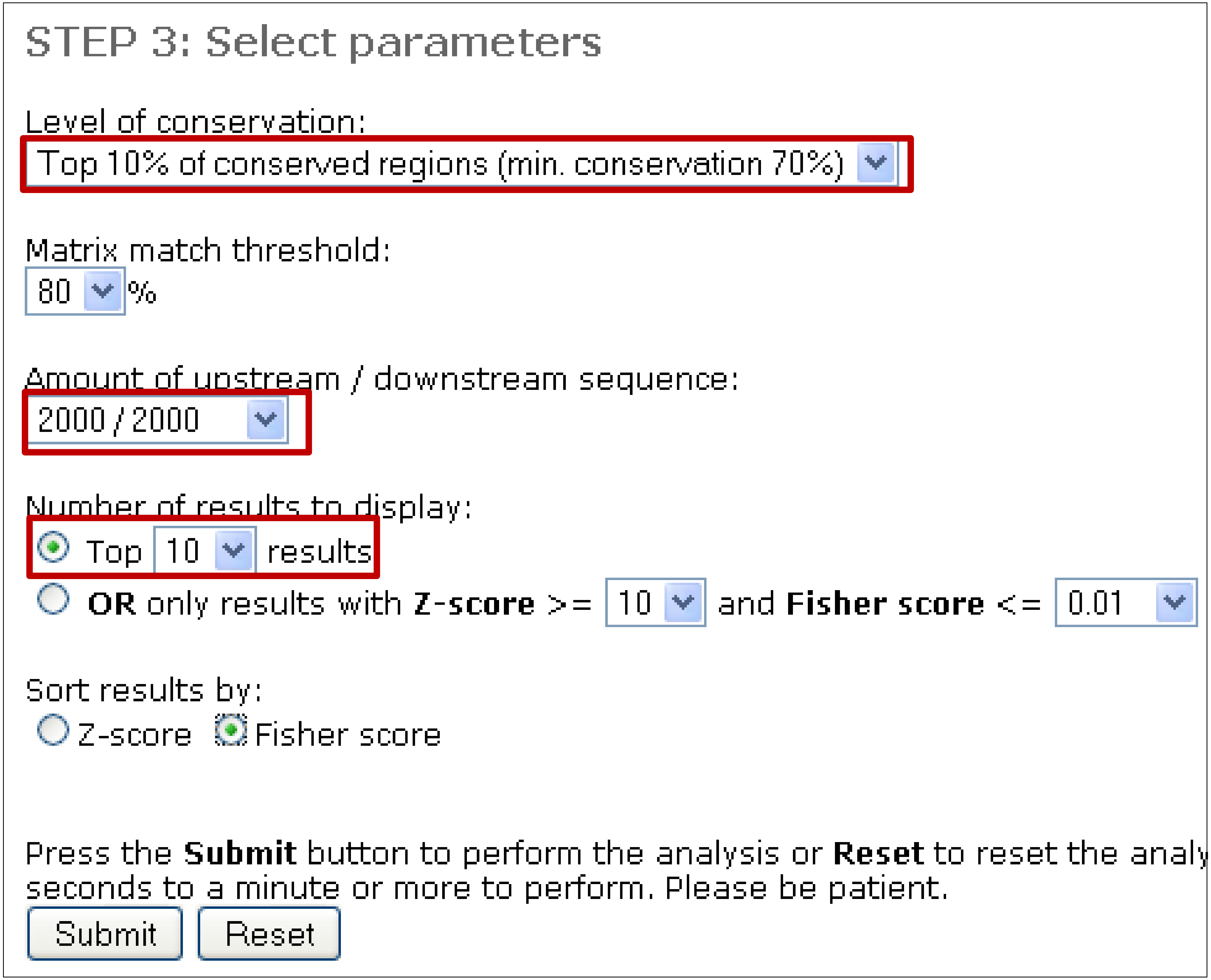
● All profiles with a minimum specificity of 8 bits (min. 8 bits)

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

When should it be used?

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

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



STEP 3: Select parameters

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

Matrix match threshold:
80 %

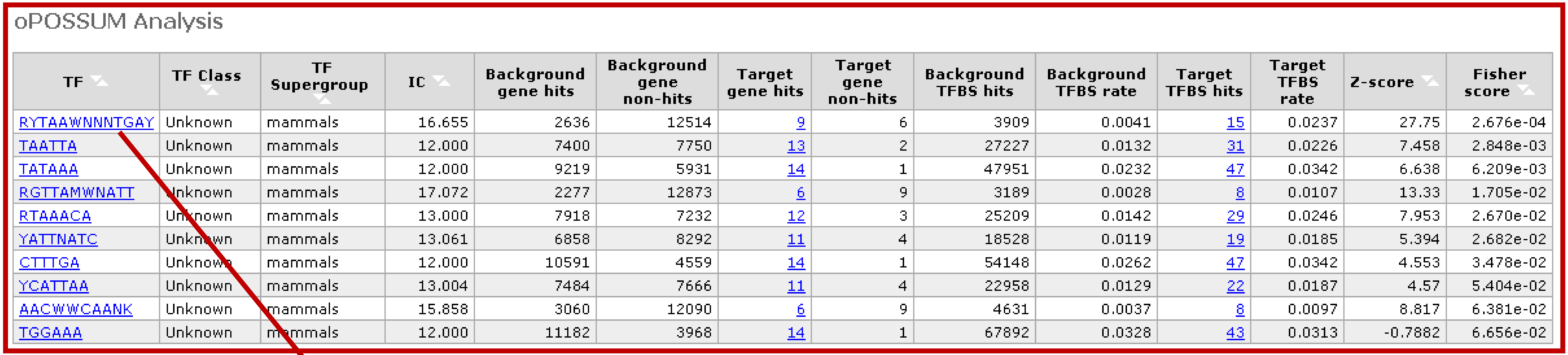
Amount of upstream / downstream sequence:
2000 / 2000

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

Sort results by:
○ Z-score ● Fisher score

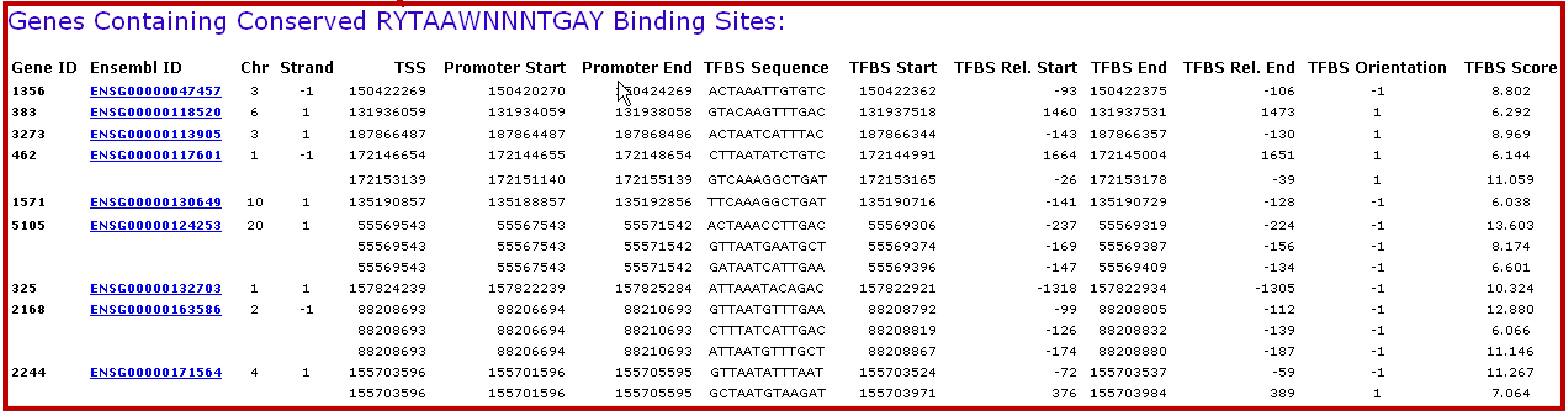
Press the **Submit** button to perform the analysis or **Reset** to reset the analy
seconds to a minute or more to perform. Please be patient.

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



oPOSSUM Analysis

TF	TF Class	TF Supergroup	IC	Background gene hits	Background gene non-hits	Target gene hits	Target gene non-hits	Background TFBS hits	Background TFBS rate	Target TFBS hits	Target TFBS rate	Z-score	Fisher score
RYTAAWNNNTGAY	Unknown	mammals	16.655	2636	12514	9	6	3909	0.0041	15	0.0237	27.75	2.676e-04
TAATTA	Unknown	mammals	12.000	7400	7750	13	2	27227	0.0132	31	0.0226	7.458	2.848e-03
TATAAA	Unknown	mammals	12.000	9219	5931	14	1	47951	0.0232	47	0.0342	6.638	6.209e-03
RGTTAMWNATT	Unknown	mammals	17.072	2277	12873	6	9	3189	0.0028	8	0.0107	13.33	1.705e-02
RTAAACA	Unknown	mammals	13.000	7918	7232	12	3	25209	0.0142	29	0.0246	7.953	2.670e-02
YATTNATC	Unknown	mammals	13.061	6858	8292	11	4	18528	0.0119	19	0.0185	5.394	2.682e-02
CTTTGA	Unknown	mammals	12.000	10591	4559	14	1	54148	0.0262	47	0.0342	4.553	3.478e-02
YCATTAA	Unknown	mammals	13.004	7484	7666	11	4	22958	0.0129	22	0.0187	4.57	5.404e-02
AACWWCAANK	Unknown	mammals	15.858	3060	12090	6	9	4631	0.0037	8	0.0097	8.817	6.381e-02
TGGAAA	Unknown	mammals	12.000	11182	3968	14	1	67892	0.0328	43	0.0313	-0.7882	6.656e-02



Genes Containing Conserved RYTAAWNNNTGAY Binding Sites:

Gene ID	Ensembl ID	Chr	Strand	TSS	Promoter Start	Promoter End	TFBS Sequence	TFBS Start	TFBS Rel. Start	TFBS End	TFBS Rel. End	TFBS Orientation	TFBS Score
1356	ENSG00000047457	3	-1	150422269	150420270	150424269	ACTAAATTGTGTC	150422362	-93	150422375	-106	-1	8.802
383	ENSG00000118520	6	1	131936059	131934059	131938058	GTACAAGTTTGAC	131937518	1460	131937531	1473	1	6.292
3273	ENSG00000113905	3	1	187866487	187864487	187868486	ACTAATCATTTAC	187866344	-143	187866357	-130	1	8.969
462	ENSG00000117601	1	-1	172146654	172144655	172148654	CTTAATATCTGTC	172144991	1664	172145004	1651	1	6.144
				172153139	172151140	172155139	GTCAAAGGCTGAT	172153165	-26	172153178	-39	1	11.059
1571	ENSG00000130649	10	1	135190857	135188857	135192856	TTCAAAGGCTGAT	135190716	-141	135190729	-128	-1	6.038
5105	ENSG00000124253	20	1	55569543	55567543	55571542	ACTAAACCTTGAC	55569306	-237	55569319	-224	-1	13.603
				55569543	55567543	55571542	GTTAATGAATGCT	55569374	-169	55569387	-156	-1	8.174
				55569543	55567543	55571542	GATAATCATTGAA	55569396	-147	55569409	-134	-1	6.601
325	ENSG00000132703	1	1	157824239	157822239	157825284	ATTAAATACAGAC	157822921	-1318	157822934	-1305	-1	10.324
2168	ENSG00000163586	2	-1	88208693	88206694	88210693	GTTAATGTTTGAA	88208792	-99	88208805	-112	-1	12.880
				88208693	88206694	88210693	CTTTATCATTGAC	88208819	-126	88208832	-139	-1	6.066
				88208693	88206694	88210693	ATTAATGTTTGCT	88208867	-174	88208880	-187	-1	11.146
2244	ENSG00000171564	4	1	155703596	155701596	155705595	GTTAATATTTAAT	155703524	-72	155703537	-59	-1	11.267
				155703596	155701596	155705595	GCTAATGTAAGAT	155703971	376	155703984	389	1	7.064

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

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

- ❑ Known transcription factor binding sites (TFBS)
  - ❖ Conserved
    - oPOSSUM
    - DiRE
  - ❖ Non-conserved
    - Pscan
    - MatInspector (*Licensed)
- ❑ Unknown TFBS or Novel motifs
  - ❖ Conserved
    - oPOSSUM
    - Weeder-H
  - ❖ Non-conserved
    - MEME
    - Weeder

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

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

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

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

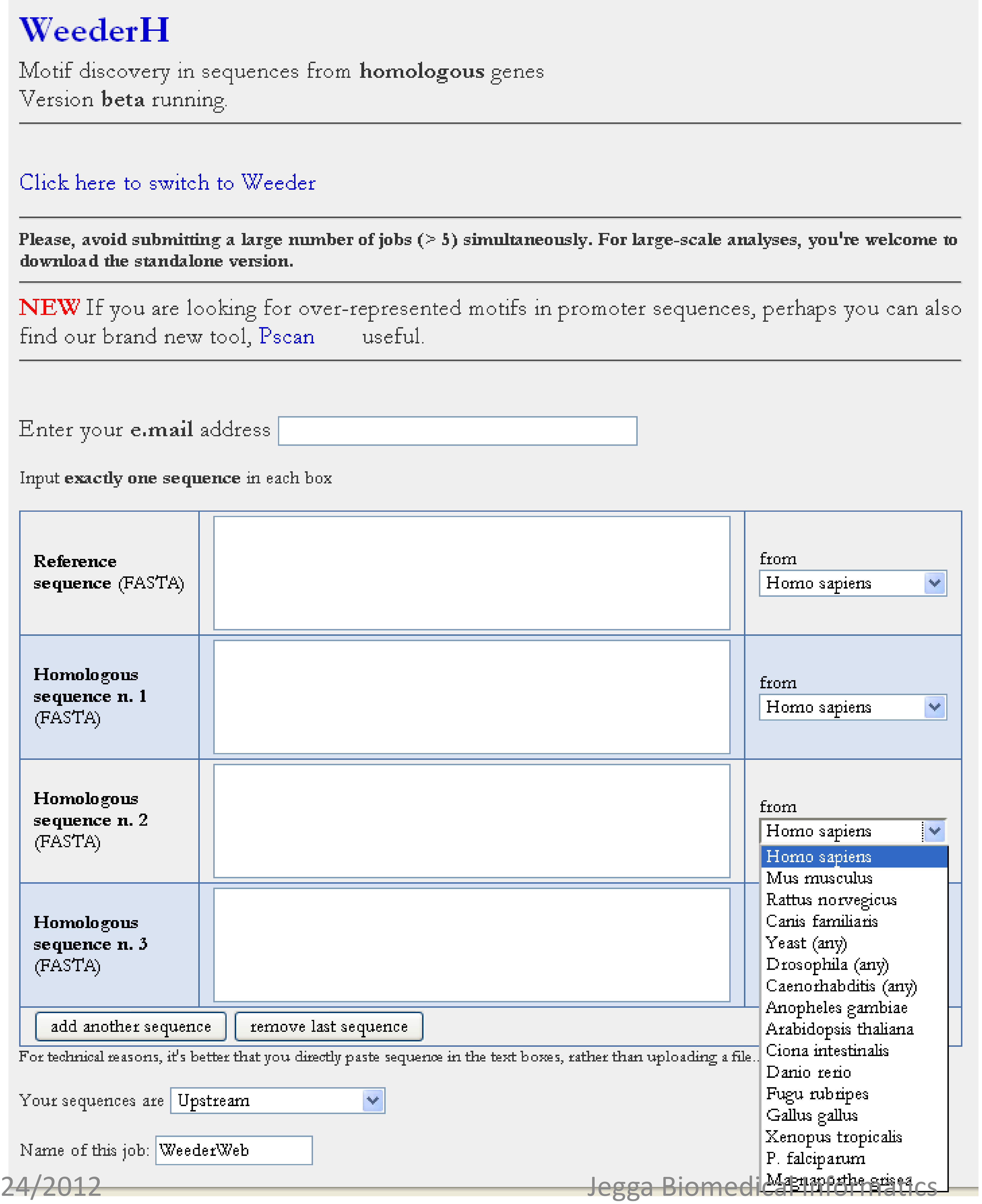
- Known transcription factor binding sites (TFBS)
  - Conserved
    - oPOSSUM
    - DiRE
  - Non-conserved
    - Pscan
    - MatInspector (*Licensed)

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

- Unknown TFBS or Novel motifs
  - Conserved
    - oPOSSUM
    - Weeder-H
  - Non-conserved
    - MEME
    - Weeder

Use the fetched promoter/upstream sequences for the following analyses

WeederH (http://159.149.109.9/pscan)



WeederH

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

Click here to switch to Weeder

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

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

Enter your **e.mail** address

Input **exactly one sequence** in each box

Reference sequence (FASTA)		from Homo sapiens
Homologous sequence n. 1 (FASTA)		from Homo sapiens
Homologous sequence n. 2 (FASTA)		from Homo sapiens
Homologous sequence n. 3 (FASTA)		

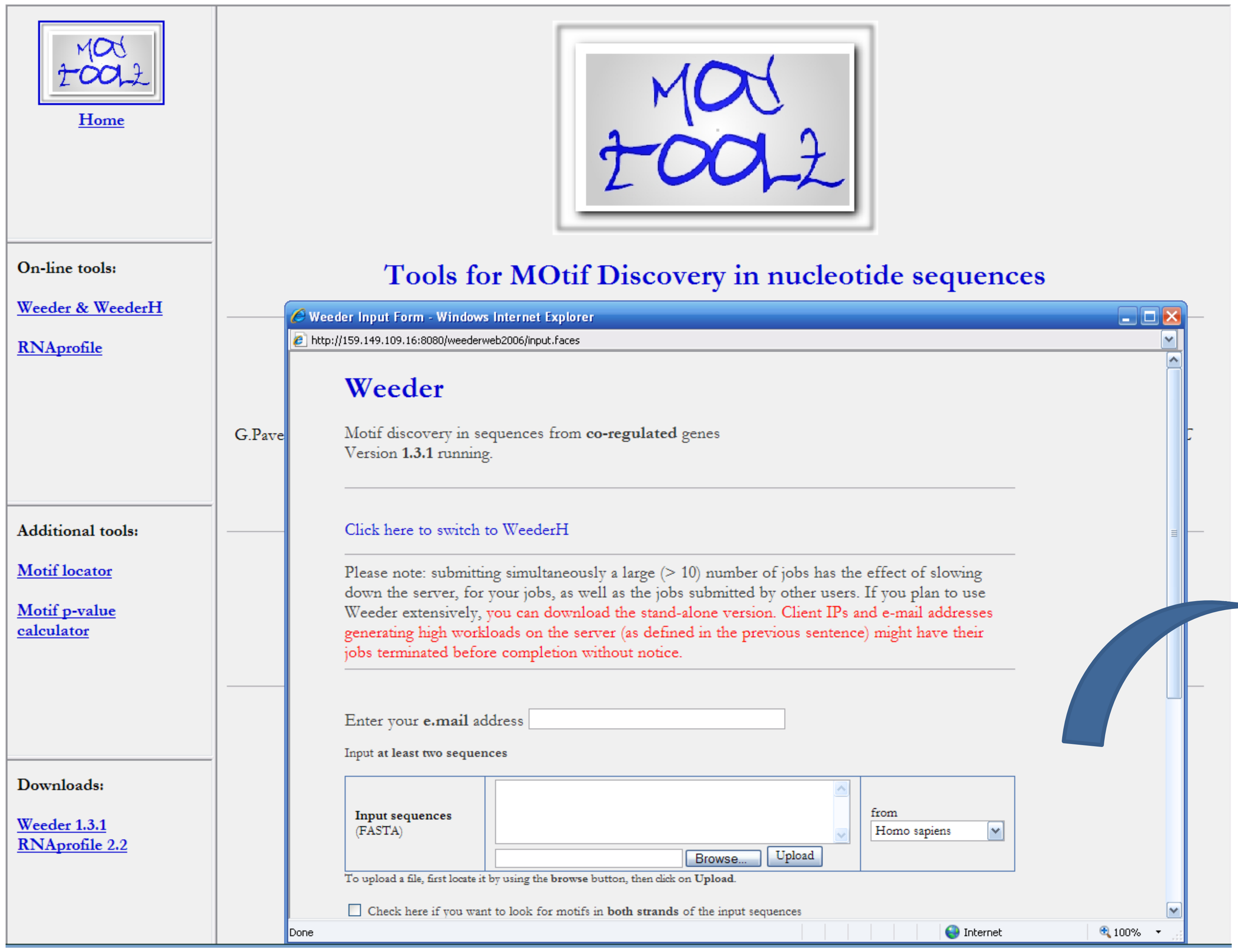
For technical reasons, it's better that you directly paste sequence in the text boxes, rather than uploading a file.

Your sequences are Upstream

Name of this job: WeederWeb

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

Weeder (http://159.149.109.9/modtools/)



On-line tools:

Weeder & WeederH

RNAprofile

Additional tools:

Motif locator

Motif p-value calculator

Downloads:

Weeder 1.3.1
RNAprofile 2.2

Tools for MOtif Discovery in nucleotide sequences

G.Pave

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

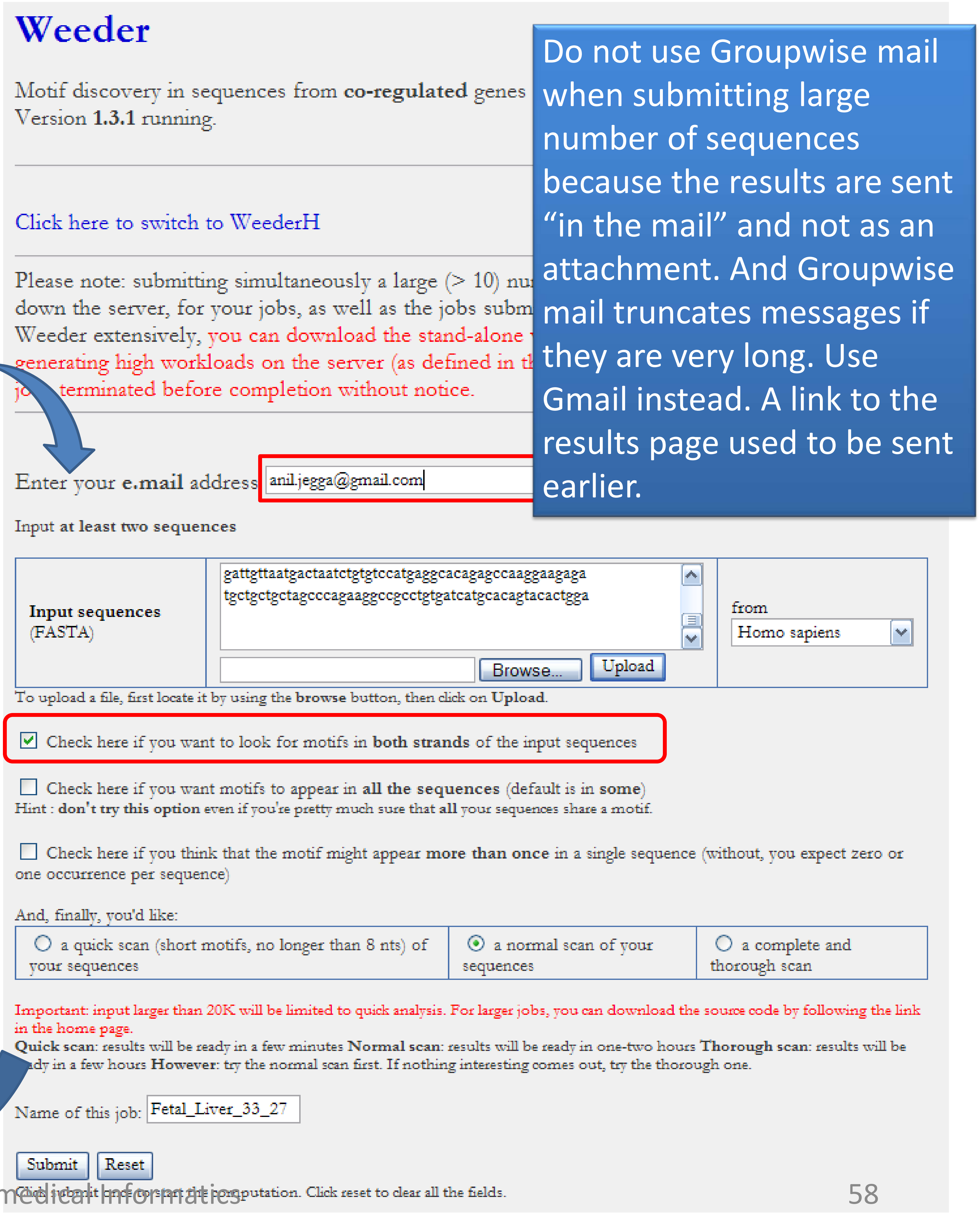
Input **at least two** sequences

Input sequences (FASTA)

from Homo sapiens

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



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) nu
down the server, for your jobs, as well as the jobs subm
Weeder extensively, you can download the stand-alone
generating high workloads on the server (as defined in t
jo 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

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:

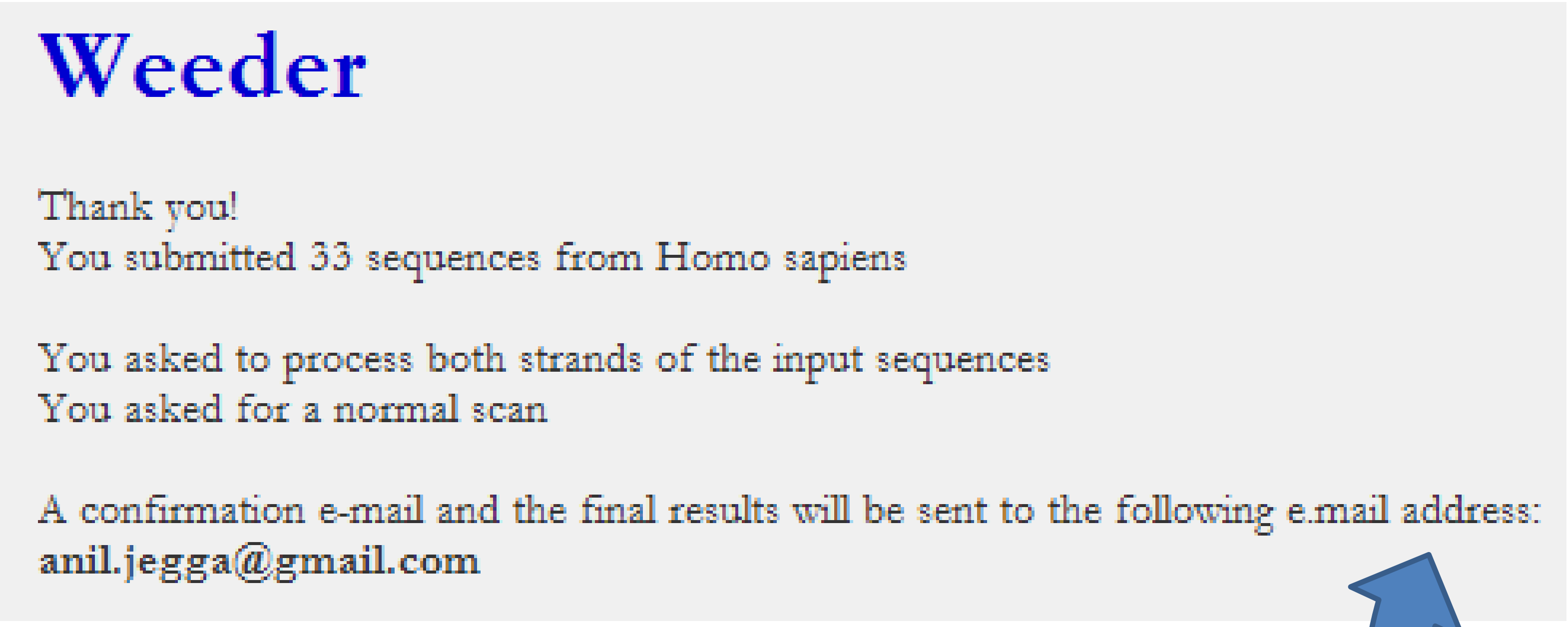
○ a quick scan (short motifs, no longer than 8 nts) of your sequences
● a normal scan of your sequences
○ a complete and thorough scan

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 ady 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

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

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



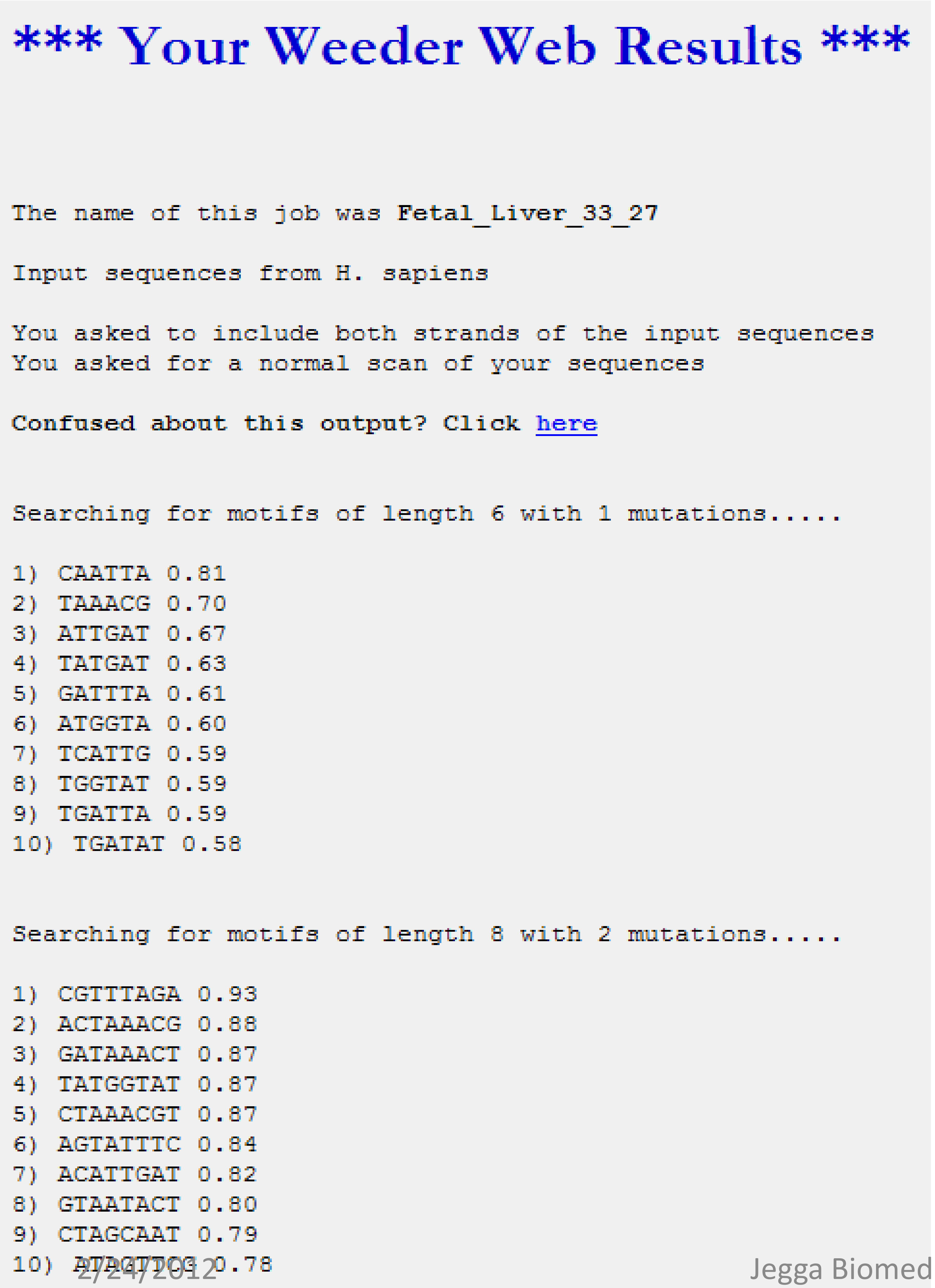
Weeder

Thank you!
You submitted 33 sequences from Homo sapiens

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

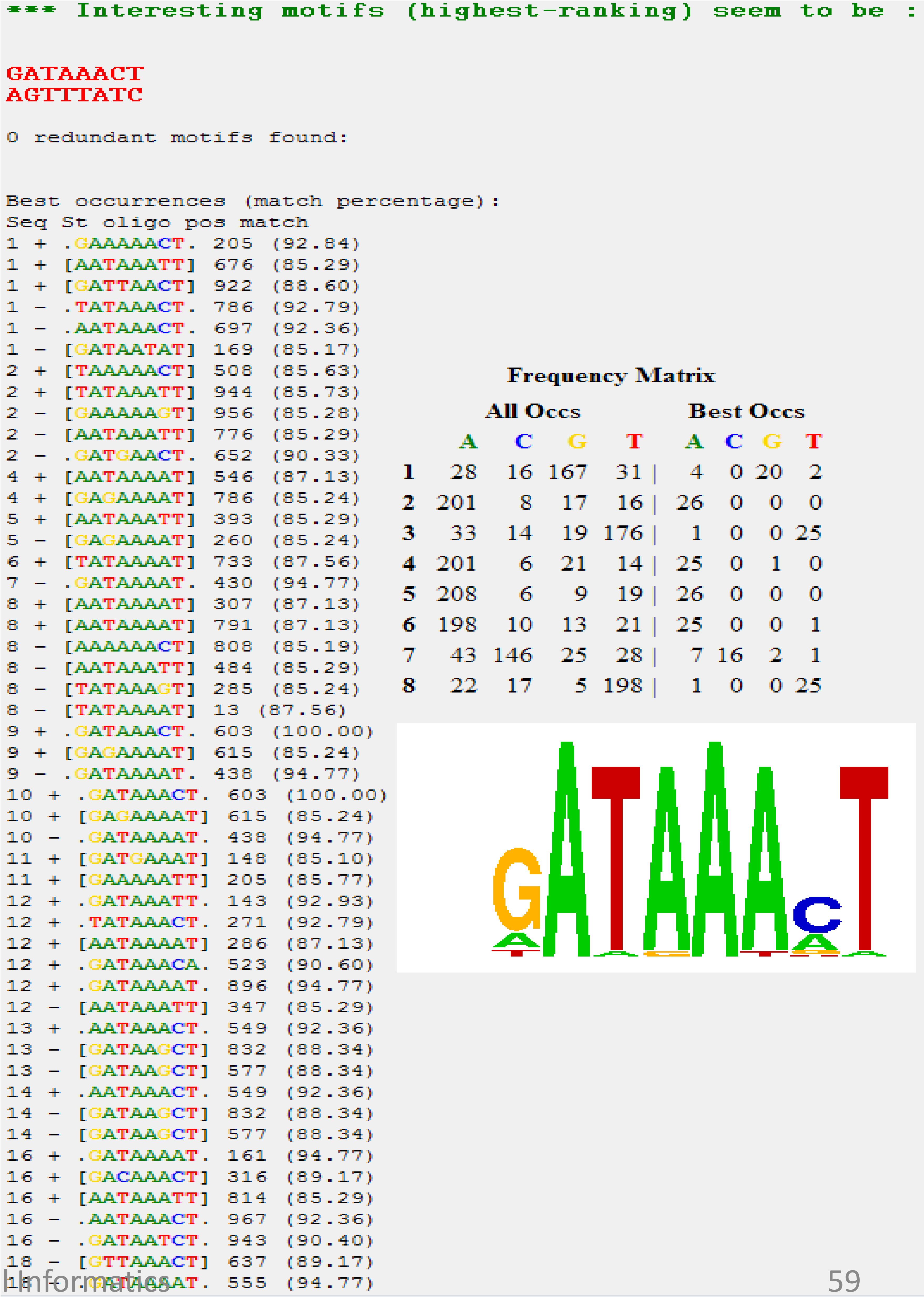
A confirmation e-mail and the final results will be sent to the following e.mail address:
anil.jegga@gmail.com

Weeder (http://159.149.109.9/modtools/)



*** Your Weeder Web Results ***

```
The name of this job was Fetal_Liver_33_27

Input sequences from H. sapiens

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

Confused about this output? Click here


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

1) CAATTA 0.81
2) TAAACG 0.70
3) ATTGAT 0.67
4) TATGAT 0.63
5) GATTTA 0.61
6) ATGGTA 0.60
7) TCATTG 0.59
8) TGGTAT 0.59
9) TGATTA 0.59
10) TGATAT 0.58


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

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



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

GATAAACT
AGTTTATC

0 redundant motifs found:


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

Frequency Matrix

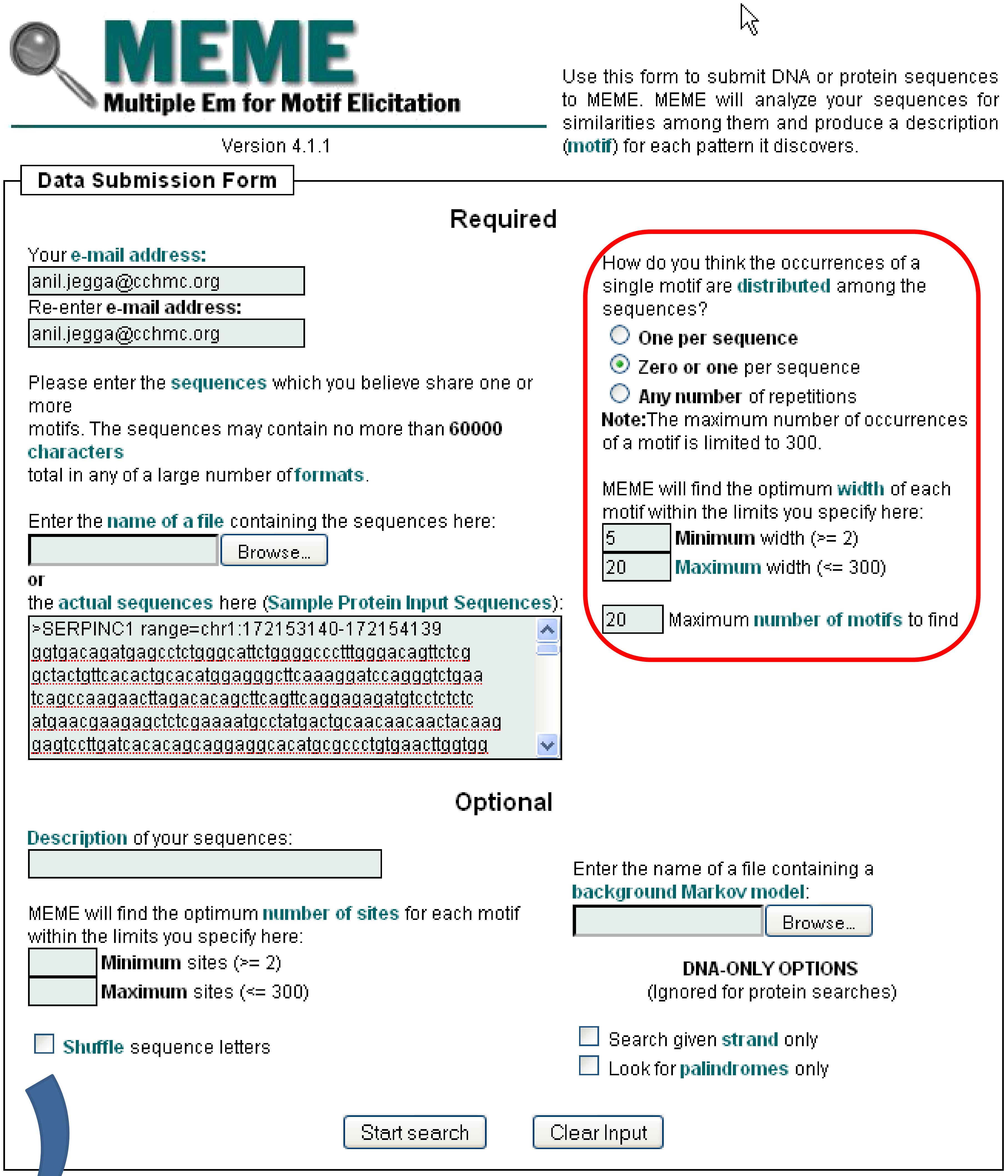
	All Occs				**Best Occs**			
	A	**C**	**G**	**T**	**A**	**C**	**G**	**T**
1	28	16	167	31	4	0	20	2
2	201	8	17	16	26	0	0	0
3	33	14	19	176	1	0	0	25
4	201	6	21	14	25	0	1	0
5	208	6	9	19	26	0	0	0
6	198	10	13	21	25	0	0	1
7	43	146	25	28	7	16	2	1
8	22	17	5	198	1	0	0	25

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

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

Your MEME results consist of:

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



Version 4.1.1

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.

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 **60000 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 Protein Input Sequences**):

```
>SERPINC1 range=chr1:172153140-172154139
ggtgacagatgagcctctgggcattctggggcccttgggacagttctcg
gctactgttcacactgcacatggagggcttcaaaggatccagggtctgaa
tcagccaagaacttagacacagcttcagttcaggagagatgtcctctctc
atgaacgaagagctctcgaaaatgcctatgactgcaacaacaactacaag
gagtccttgatcacacagcaggaggcacatgcgccctgtgaacttggtgg
```

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
Note: The maximum number of occurrences of a motif is limited to 300.

MEME will find the optimum **width** of each motif within the limits you specify here:
5 **Minimum** width (>= 2)
20 **Maximum** width (<= 300)

20 Maximum **number of motifs** to find

Optional

Description of your sequences:

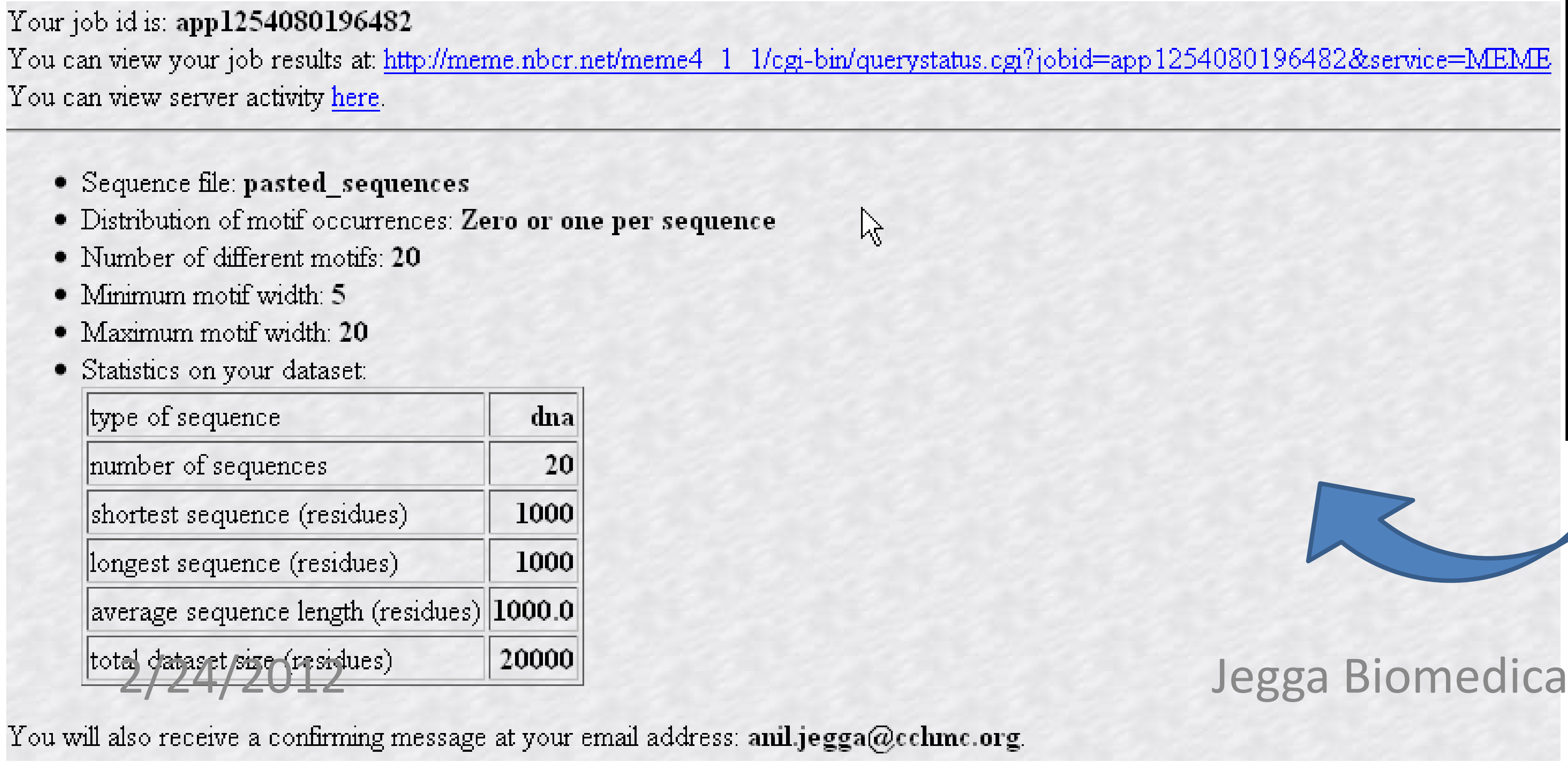
MEME will find the optimum **number of sites** for each motif within the limits you specify here:
Minimum sites (>= 2)
Maximum sites (<= 300)

☐ **Shuffle** sequence letters

Enter the name of a file containing a **background Markov model**:

DNA-ONLY OPTIONS
(Ignored for protein searches)

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



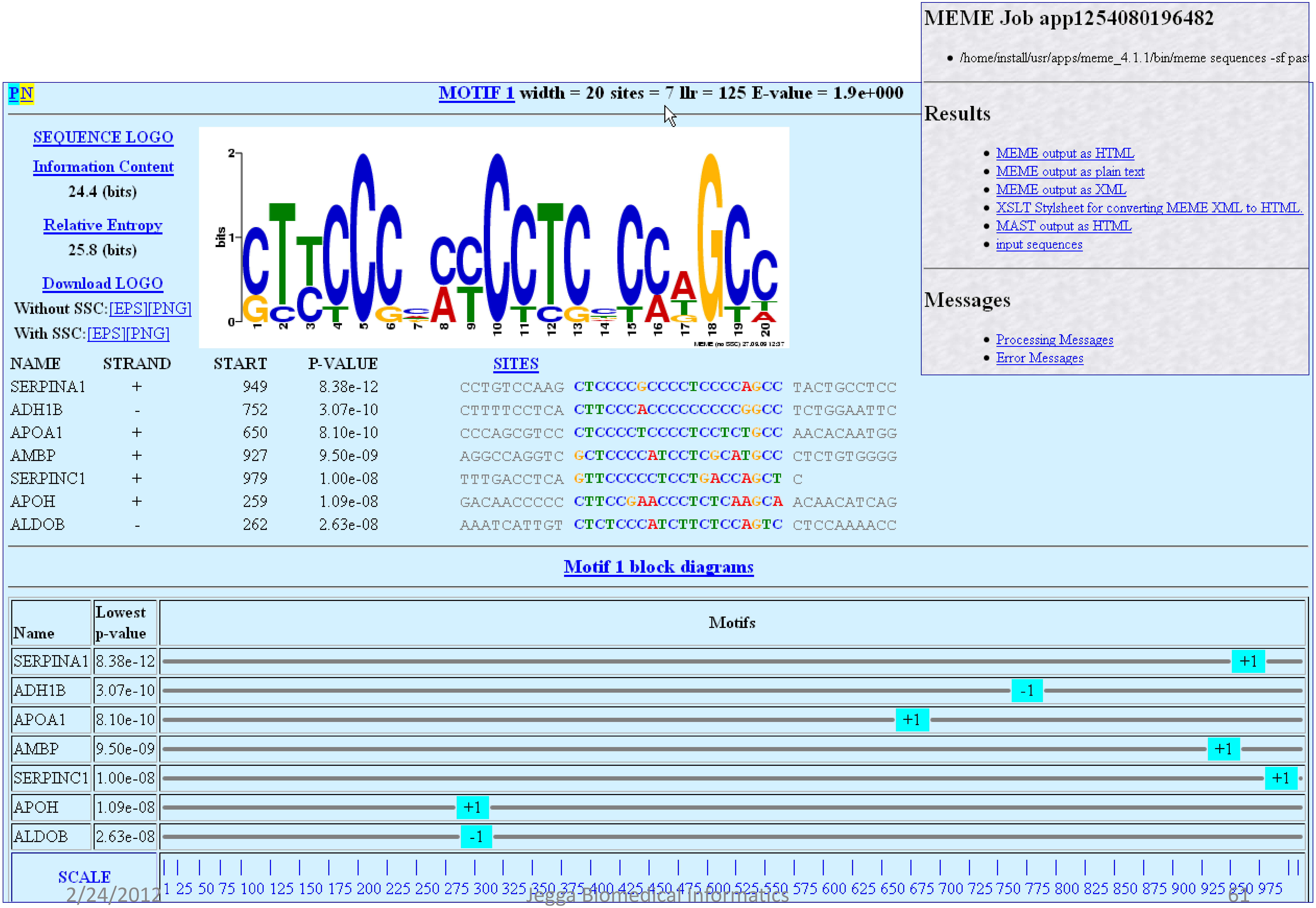
Your job id is: **app1254080196482**
You can view your job results at: http://meme.nbcr.net/meme4_1_1/cgi-bin/querystatus.cgi?jobid=app1254080196482&service=MEME
You can view server activity here.

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

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

You will also receive a confirming message at your email address: **anil.jegga@cchmc.org**.

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



MEME Job app1254080196482

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

Results

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

Messages

- Processing Messages
- Error Messages

PN

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

SEQUENCE LOGO

Information Content

24.4 (bits)

Relative Entropy

25.8 (bits)

Download LOGO

Without SSC:[EPS][PNG]

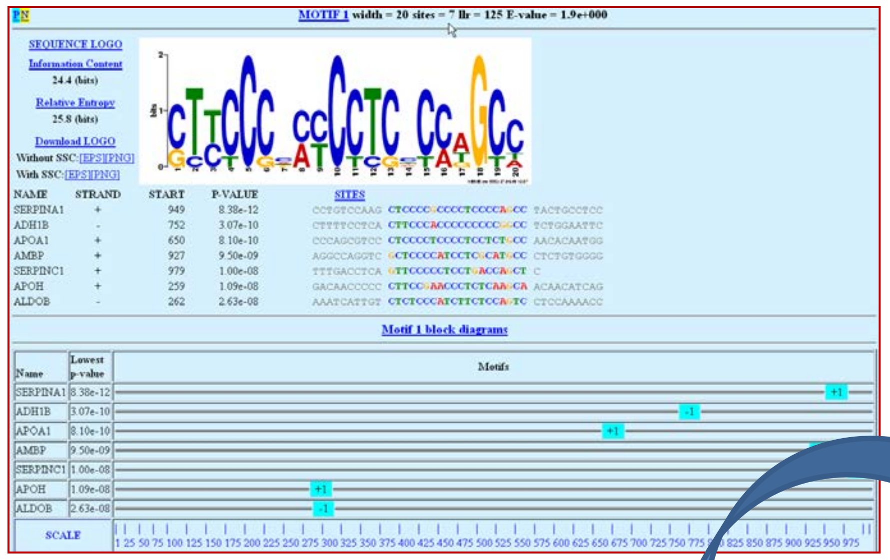
With SSC:[EPS][PNG]

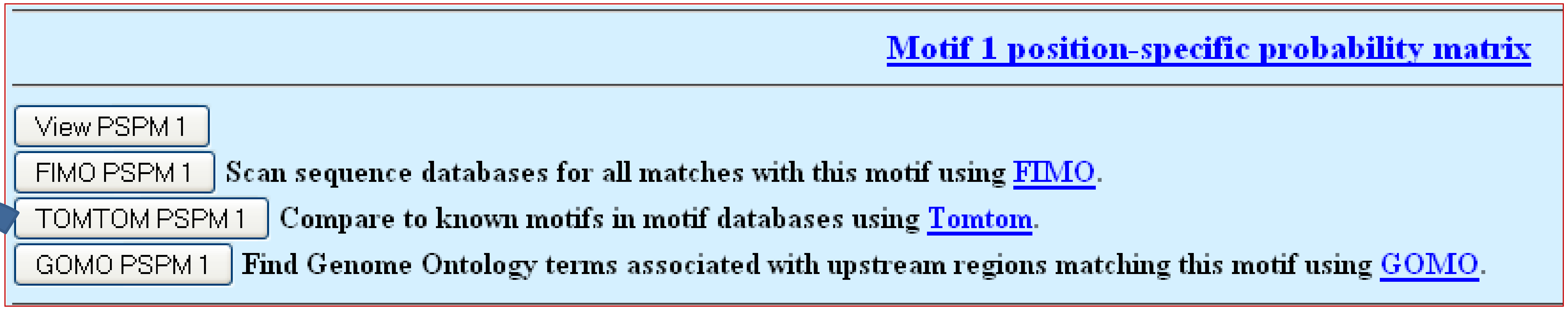
NAME	STRAND	START	P-VALUE	SITES		
SERPINA1	+	949	8.38e-12	CCTGTCCAAG	CTCCCCGCCCCTCCCCAGCC	TACTGCCTCC
ADH1B	-	752	3.07e-10	CTTTTCCTCA	CTTCCCACCCCCCCCCGGCC	TCTGGAATTC
APOA1	+	650	8.10e-10	CCCAGCGTCC	CTCCCCTCCCCTCCTCTGCC	AACACAATGG
AMBP	+	927	9.50e-09	AGGCCAGGTC	GCTCCCCATCCTCGCATGCC	CTCTGTGGGG
SERPINC1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

Motif 1 block diagrams

Name	Lowest p-value	Motifs
SERPINA1	8.38e-12	+1
ADH1B	3.07e-10	-1
APOA1	8.10e-10	+1
AMBP	9.50e-09	+1
SERPINC1	1.00e-08	+1
APOH	1.09e-08	+1
ALDOB	2.63e-08	-1
SCALE		1 25 50 75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525 550 575 600 625 650 675 700 725 750 775 800 825 850 875 900 925 950 975

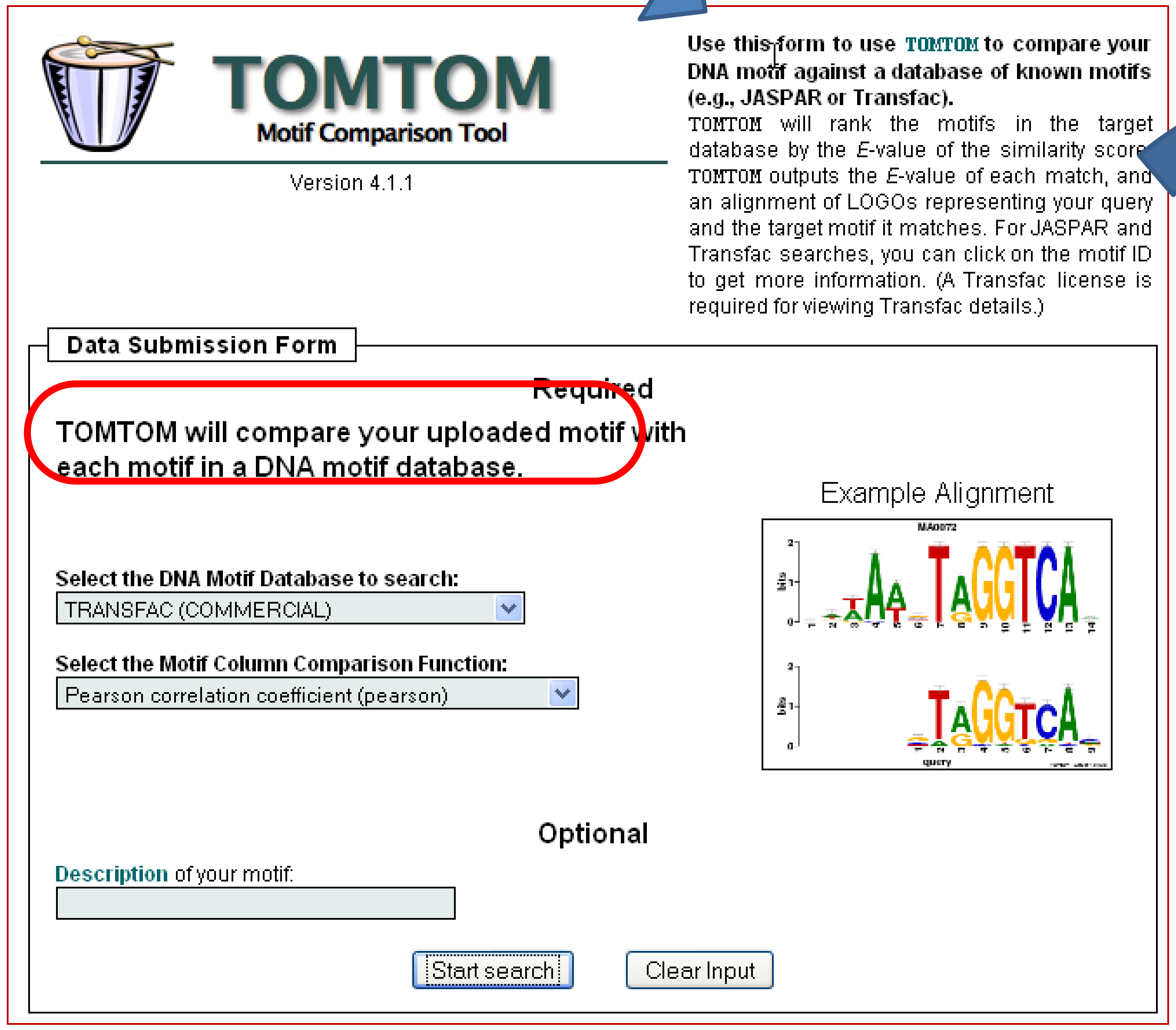
MEME (http://meme.sdsc.edu)





Motif 1 position-specific probability matrix

View PSPM 1	
FIMO PSPM 1	Scan sequence databases for all matches with this motif using FIMO.
TOMTOM PSPM 1	Compare to known motifs in motif databases using Tomtom.
GOMO PSPM 1	Find Genome Ontology terms associated with upstream regions matching this motif using GOMO.



Version 4.1.1

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

Data Submission Form

Required

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

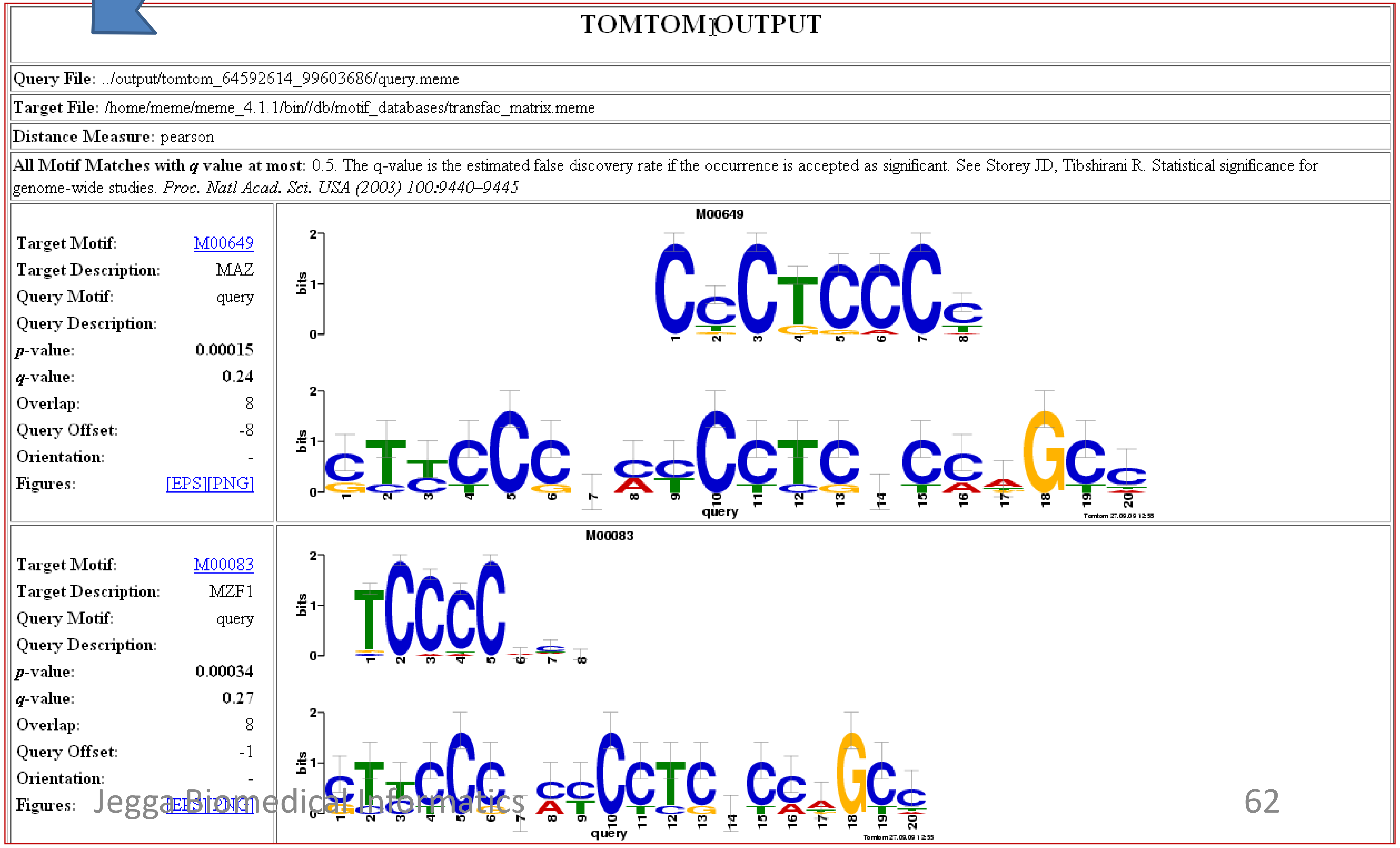
Select the DNA Motif Database to search: TRANSFAC (COMMERCIAL)

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

Optional

Description of your motif:

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



TOMTOM OUTPUT

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

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

Distance Measure: pearson

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

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

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

Summary

Cis-Element Finding Matrix

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

RESOURCES - URLs: Summary

Application/Resource	URL
oPOSSUM	http://burgundy.cmmt.ubc.ca/oPOSSUM/
DiRE	http://dire.dcode.org/
Weeder-H	http://159.149.109.9/modtools/
Weeder	http://159.149.109.9/modtools/
Pscan	http://159.149.109.9/pscan
MEME	http://meme.sdsc.edu/
MatInspector	http://www.genomatix.de/
Genome Browser	http://genome.ucsc.edu
ECR Browser	http://ecrbrowser.dcode.org

Additional exercise available at:
http://anil.cchmc.org/grn.html