Detection and visualization of conserved *cis*-element structures and their annotation using NLP-based extraction and GFF-formatting of biomedical literature-derived SO features

Anil Jegga & Jing Chen
Biomedical Informatics
Cincinnati Children's Hospital Research Foundation

Objectives

- Extracting sequence regulatory features from GenBank and MedLine
- Representation of the known regulatory regions in the context of computationally identified putative regulatory regions (GenomeTrafac).
- Export of GenomeTrafac tracks in GFF - Display in UCSC genome browser.

Regulatory Region	A DNA sequence that controls the expression of a gene.
→ Attenuator	A sequence segment located between the promoter and a structural gene that causes partial termination of transcription.
→ Enhancer	A cis-acting sequence that increases the utilization of (some) eukaryotic promoters, and can function in either orientation and in any location (upstream or downstream) relative to the promoter. (Enhancer: ≥1 TFBS?) (Intragenic OR Intergenic)
→ PolyA signal	The recognition sequence necessary for endonuclease cleavage of an RNA transcript that is followed by polyadenylation; consensus=AATAAA.
→ Promoter	The region on a DNA molecule involved in RNA polymerase binding to initiate transcription.
→ Silencer	Combination of short DNA sequence elements which suppress the transcription of an adjacent gene or genes. (Silencer = One or more TFBS?)
→ Splice enhancer	Region of a transcript that regulates splicing.
→ Terminator	The sequence of DNA located either at the end of the transcript that causes RNA polymerase to terminate transcription.
→ TF binding site	A region of a molecule that binds to a transcription factor.

Missing Terms in SOFA?
MeSH Definitions

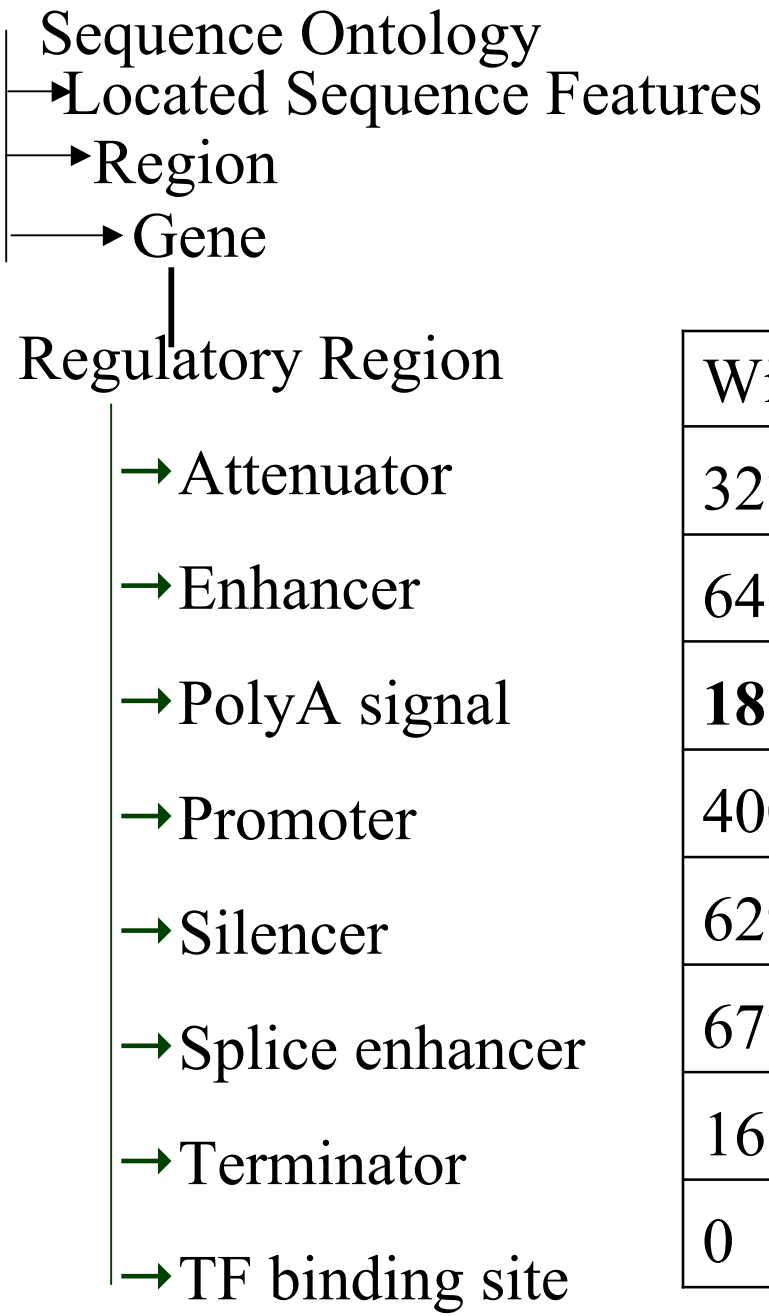
- Locus Control Region: “A regulatory region first identified in the human beta-globin locus but subsequently found in other loci. The region is believed to regulate TRANSCRIPTION by opening and remodeling CHROMATIN structure. It may also have ENHANCER activity”.
- Insulator: “Nucleic acid regulatory sequences that limit or oppose the action of ENHANCER ELEMENTS and define the boundary between differentially regulated gene loci”.
- Operator: “The regulatory elements of an OPERON to which activators or repressors bind thereby effecting the transcription of GENES in the operon”.

Enhancer: 64636 entries (GenBank Nucleotide Sequences)!

Enhancer [Limited to “Feature Key”]: 1045

GenBank Search

```
LOCUS       AF187727                 315 bp    DNA     linear   PRI 23-AU
DEFINITION  Homo sapiens APOB intestinal enhancer, complete sequence.
ACCESSION   AF187727
VERSION     AF187727.1  GI:9885269
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 315)
  AUTHORS   Antes,T.J., Goodart,S.A., Huynh,C., Sullivan,M., Young,S.G. and
            Levy-Wilson,B.
  TITLE     Identification and characterization of a 315-base pair enhancer,
            located more than 55 kilobases 5' of the apolipoprotein B gene,
            that confers expression in the intestine
  JOURNAL   J. Biol. Chem. 275 (34), 26637-26648 (2000)
  MEDLINE   20409023
   PUBMED   10859308
FEATURES             Location/Qualifiers
     source          1..315
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="2"
     enhancer          1..315
                     /note="APOB intestinal enhancer"
ORIGIN
        1 aattcaattc tcctgccctt gcta[illegible]agat agacatgcac
```



GenBank Search

Without Feature key	With Feature key
327	108
64636	1045
18187	**54430**
400552	11334
629	0
677	0
161930	4107
0	0

E.g. Enhancer AND “Genes”[MeSH];
“Enhancer” AND “Genes”[MeSH]

MedLine Search

Term	plain	quoted	plain + Genes	quoted + Genes
enhancer	23861	21519	7387	6744
silencer	1153	1153	425	425
enhancer by bound factor	1092	1092	272	272
exonic splice enhancer	75	75	13	13
intronic splice enhancer	46	46	5	5
splice enhancer	349	16	98	2

Enhancer: 21519 Abstracts

Enhancer [MeSH]: 10429

"Enhancer Elements (Genetics)"[MeSH]: 10429

"Genes"[MeSH] AND "Enhancer Elements (Genetics)"[MeSH]: 4042

MedLine Search

2163: Brickner AG, Gossage DL, Dusing MR, Wiginton DA.

Identification of a murine homolog of the human adenosine deaminase thymic enhancer.Gene. 1995 Dec 29;167(1-2):261-6.

No. of nucleotide sequences related to first 500 abstracts: 5776

(= a total of ~46208 sequences)!

1: U72392 Mus musculus adenosine deaminase (Ada) gene, thymic enhancer region, partial first intron gi|1613865|gb|U72392.1|MMU72392[1613865]

2: NM_007398 Mus musculus adenosine deaminase (Ada), mRNA gi|31982514|ref|NM_007398.2|[31982514]

1: J Orthop Res. 2004 Jul;22(4):751-8.

A silencer element in the cartilage oligomeric matrix protein gene regulates chondrocyte-specific expression.

Issack PS, Liu CJ, Prazak L, Di Cesare PE. Department of Orthopaedic Surgery, Musculoskeletal Research Center, Hospital for Joint Diseases, NYU, 301 East 17th Street, New York, NY, USA.

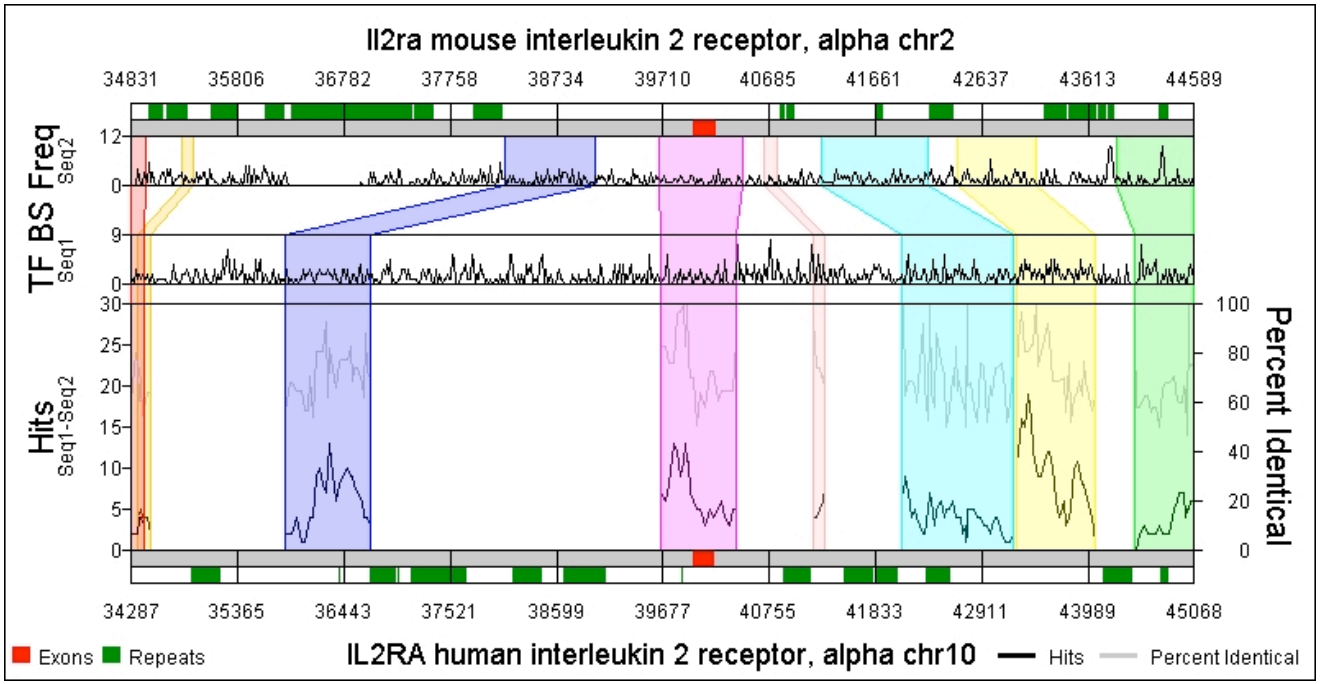
The molecular mechanisms by which mesenchymal cells differentiate into chondrocytes are poorly understood. The cartilage oligomeric matrix protein gene (COMP) encodes a noncollagenous extracellular matrix protein whose expression pattern correlates with chondrocyte differentiation and arthritis. We have used the COMP promoter as a model to identify regulatory sequences necessary for chondrocyte-specific expression and to identify cell type-specific proteins that bind these sequences. We have previously cloned 1.9 kilobases of the 5(') flanking promoter sequence of the murine COMP gene and by deletion analysis have identified two spatially distant chondrocyte-specific regulatory regions. One element is situated proximally **(-125 to -75**), and a second region is located distally **(-1925 to -592**) relative to the transcription start site. In the present study, we performed a finer deletion analysis of the region of the COMP promoter from -1925 to -592 and **identified a silencer region situated between -1775 and -1725**. This **silencer binds sequence-specific protein complexes**; the intensity of these complexes is greater in two different fibroblast cell lines (NIH3T3 and 10T1/2) than in chondrocytic RCS cells. **Competition experiments localized the binding site of these protein complexes from -1775 to -1746; deletion of this 30-bp site results in a selective increase in COMP promoter activity in fibroblasts**. Four tandem repeats of this 30-bp site are sufficient to confer negative transcriptional regulation on a heterologous promoter (SV40) in NIH3T3 fibroblasts. These results suggest that negative regulation of transcription is an important mechanism for chondrocyte-specific expression of the COMP gene. PMID: 15183430 [PubMed - indexed for MEDLINE]

J Neurosci Res. 2002 Sep 15;69(6):784-94.

Nestin enhancer requirements for expression in normal and injured adult CNS. Johansson CB, Lothian C, Molin M, Okano H, Lendahl U. Department of Cell and Molecular Biology, Medical Nobel Institute, Karolinska Institute, Stockholm, Sweden.

The nestin gene is expressed in many CNS stem/progenitor cells, both in the embryo and the adult, and nestin is used commonly as a marker for these cells. In this report we analyze nestin enhancer requirements in the adult CNS, using transgenic mice carrying reporter genes linked to three different nestin enhancer constructs: the genomic rat nestin gene and **5 kb of upstream** nestin sequence (NesPlacZ/3), **636 bp of the rat nestin second intron** (E/nestin:EGFP), and a corresponding **714 bp region from the human second intron** (Nes714tk/lacZ). NesPlacZ/3 and E/nestin:EGFP mice showed reporter gene expression in stem cell-containing regions of brain and spinal cord during normal conditions. NesPlacZ/3 and E/nestin:EGFP mice showed increased expression in spinal cord after injury and NesPlacZ/3 mice displayed elevated expression in the periventricular area of the brain after injury, which was not the case for the E/nestin:EGFP mice. In contrast, no expression in adult CNS in vivo was seen in the Nes714tk/lacZ mice carrying the human enhancer, neither during normal conditions nor after injury. The Nes714 tk/lacZ mice, however, expressed the reporter gene in reactive astrocytes and CNS stem cells cultured ex vivo. **Collectively, this suggests a species difference for the nestin enhancer function in adult CNS and that elements outside the second intron enhancer are required for the full injury response in vivo**.
Copyright 2002 Wiley-Liss, Inc. PMID: 12205672 [PubMed - indexed for MEDLINE]

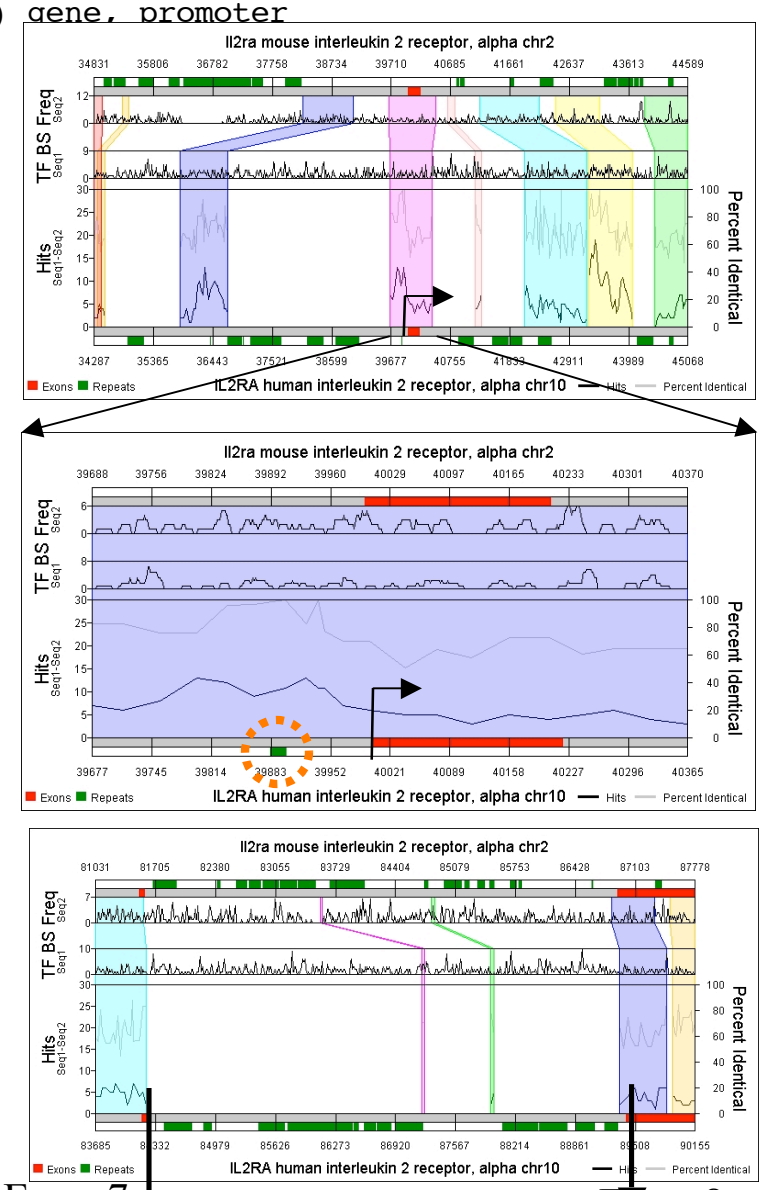
Identify relevant keywords



- Gene
- Exon
- Intron
- Repeat Elements
- TFBS (Hs, Mm, Conserved)
- Hs-Mm Alignment
- Regulatory:
  - Promoter
  - Enhancer
  - Silencer
- Non-coding SNPs

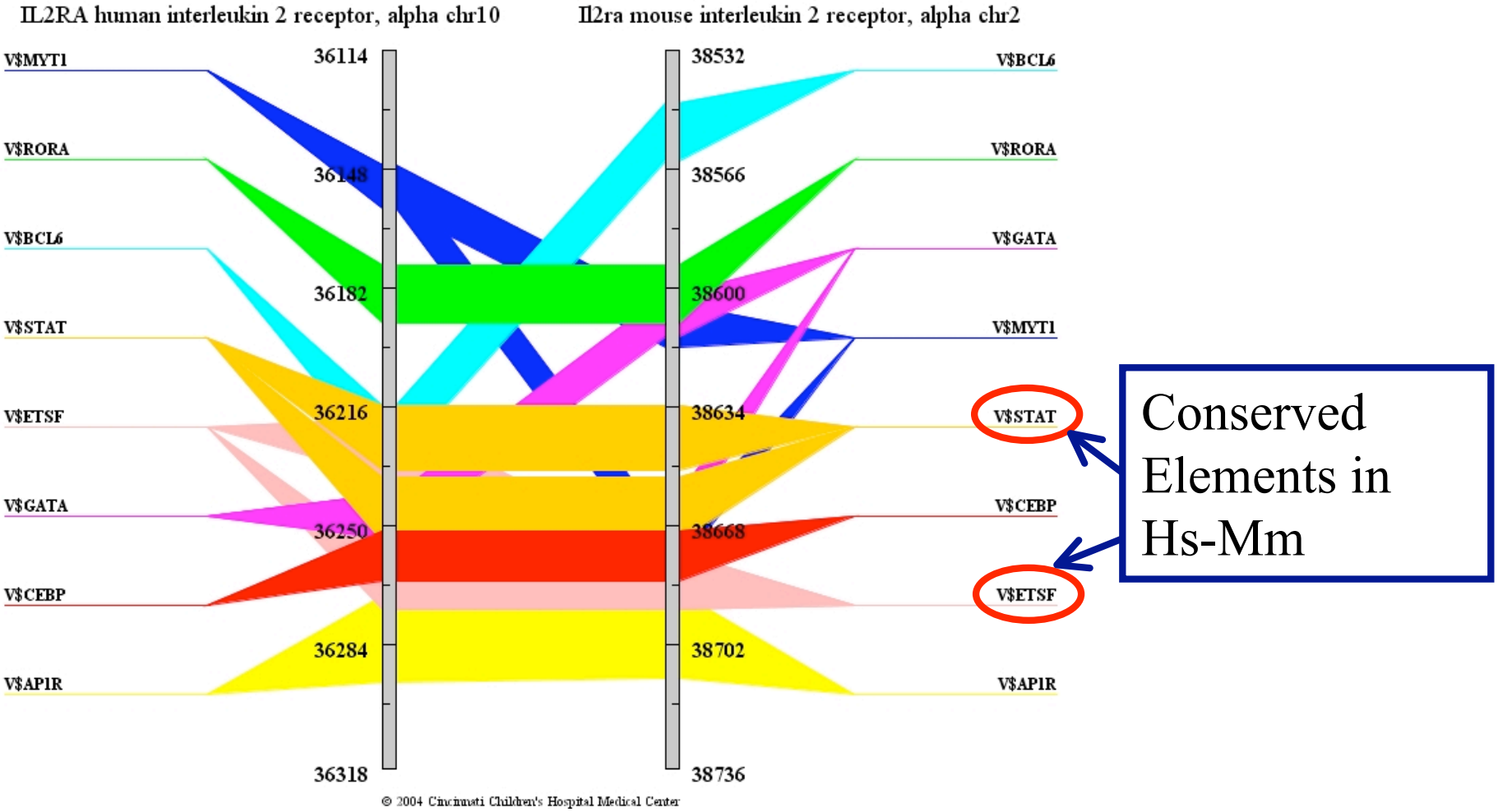
Sequence Ontology: GenomeTrafac

```
LOCUS       HSU57613                10975 bp    DNA     linear   PRI 10-OCT-2001
DEFINITION  Human interleukin-2 receptor alpha chain (IL2RA) gene, promoter
            region and exon 1.
ACCESSION   U57613 AF243502
FEATURES             Location/Qualifiers
     source          1..10975
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="10"
                     /map="10p15-14"
     gene            324..>10975
                     /gene="IL2RA"
     enhancer        324..401
                     /gene="IL2RA"
                     /note="Positive regulatory region III.
     enhancer        3828..3860
                     /gene="IL2RA"
                     /note="Positive regulatory region 1.
     enhancer        3967..4040
                     /gene="IL2RA"
                     /note="Positive regulatory region II.
```



enhancer: 324..401 (GenomeTrafac: 36208-36285)

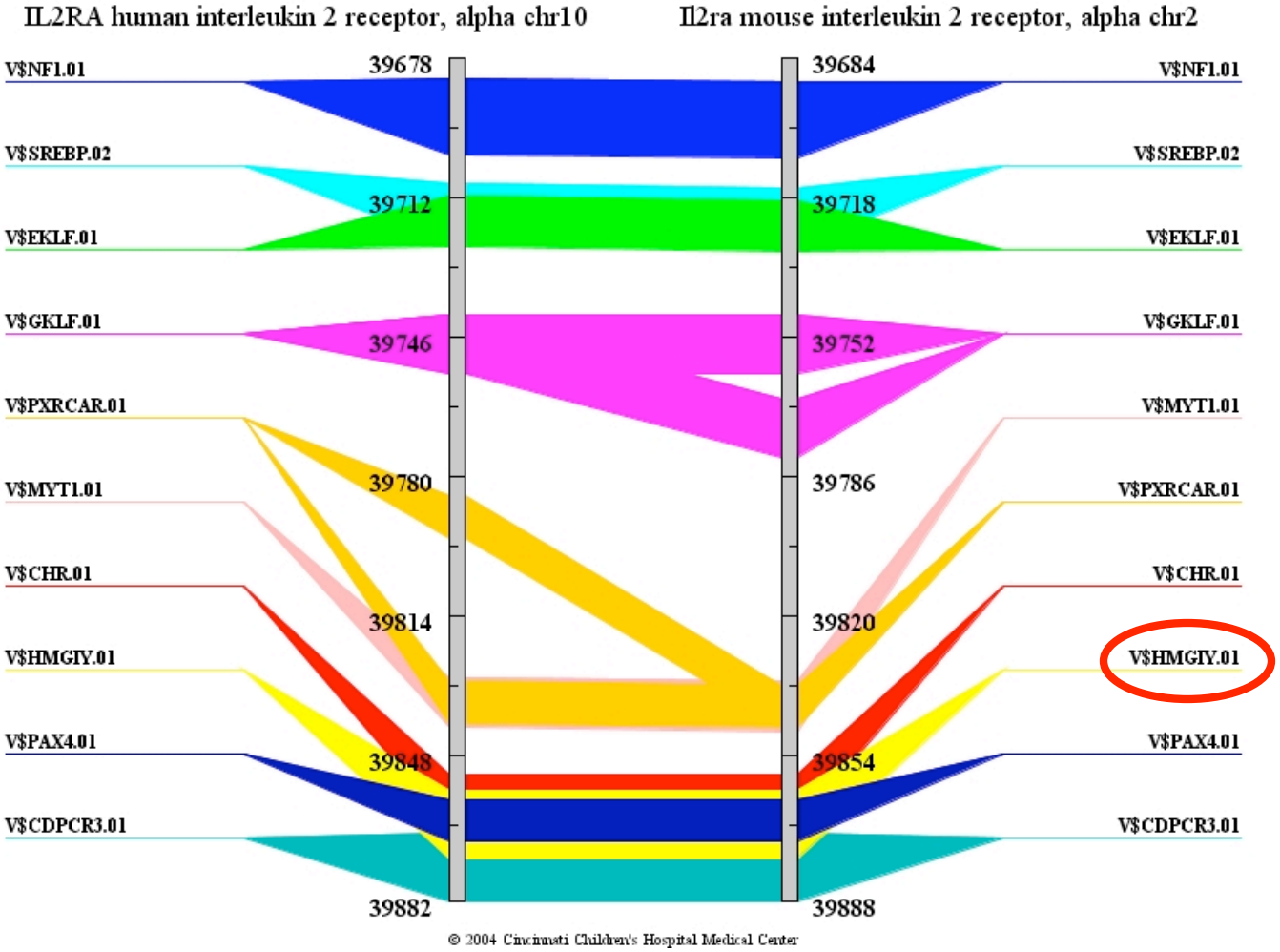
/note="Positive regulatory region III. A composite site with one consensus and one non-consensus *gamma interferon* activated sequence (GAS motif). Together these motifs mediate the binding of Stat5A and Stat5B. An Ets binding site overlaps the consensus motif, whereas a GATA motif overlaps the non-consensus GAS motif. PRRIII also contains an Ets binding site downstream of both GAS motifs; this Ets binding site can bind Elf-1. PRRIII is essential for IL-2 induced IL-2Ralpha promoter activity YT cells."

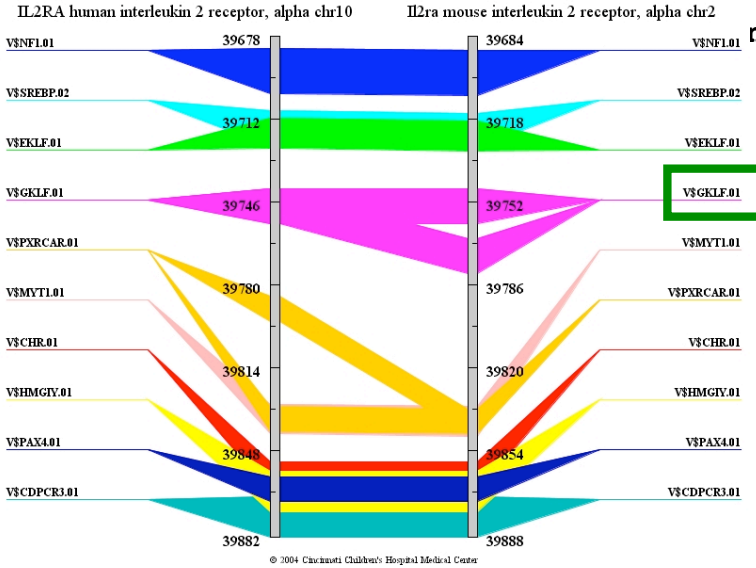


enhancer: 3967..4040 (GenomeTrafac: 39862-39882) → AAAAAAAAAAAAAAAAAA

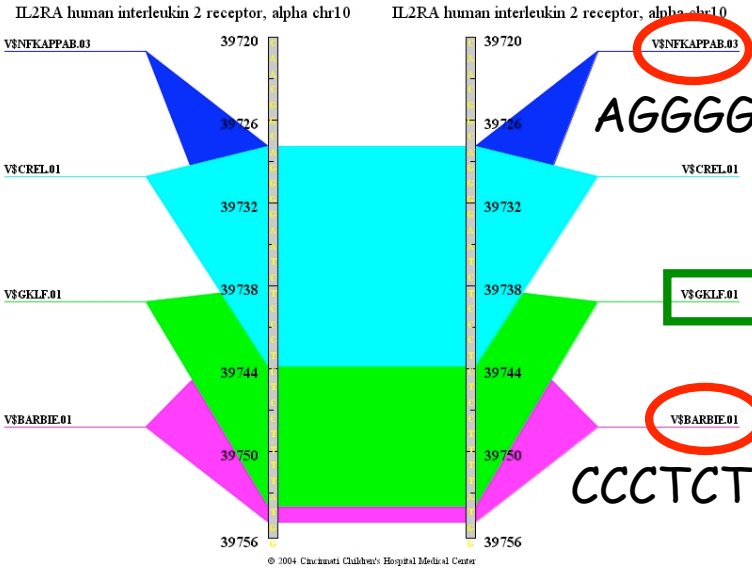
(GenomeTrafac 39904-39937)

/note="Positive regulatory region II, contains an essential binding site for Elf-1 and multiple sites for HMG-I(Y). The Elf-1 site is essential for PMA-induced IL-2Ralpha promoter activity in Jurkat T cells and IL-2 induced IL-2Ralpha promoter activity in YT natural killer like cells."






```
ncer:      3828..3860 (GenomeTrafac: 39723-39755)
/note="Positive regulatory region 1.  This spans an NF-kB
site (GGGGAATCTCCC) and a CArG motif (CCTTTTATGG).  An Sp1
site begins within PRRI and extends 3' of PRRI."
```



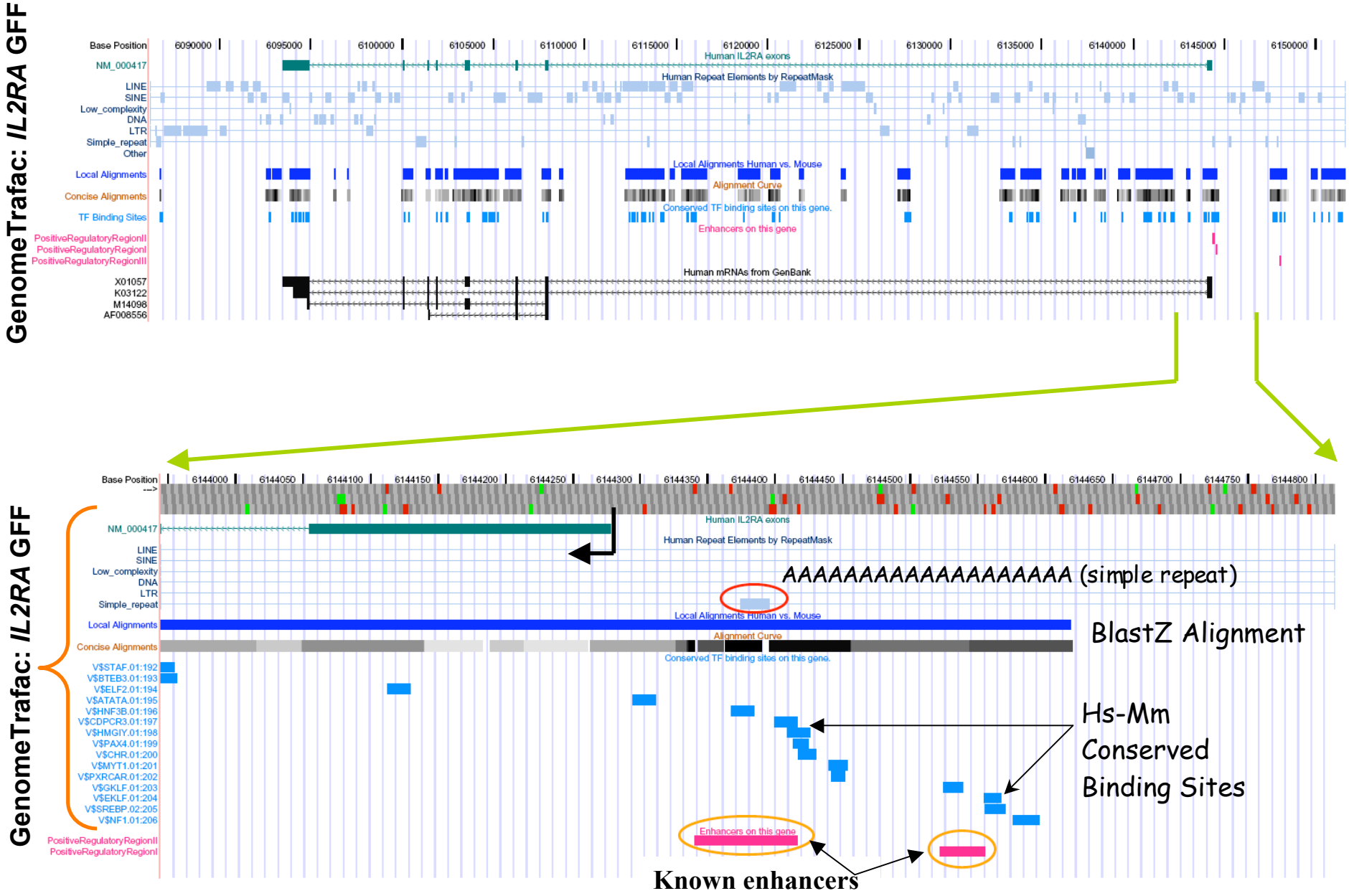
AGGGGAATCTCCCTC

Conserved Element

CCCTCTCCTTTTATG

Elements specific to human *IL2RA*

```
browser position chr10:6053512-6184278
browser pix 680
browser hide all
#browser dense ruler refGene ensGene tigrGeneIndex sanger22 uniGene genscan
# exon 1 data NM_000417
track          name="Human Exons"          description="Human IL2RA exons"          color=0,128,128          visibility=2
priority=1     useScore=0
chr10          GenomeTrafac   exon          6093511       6094865       .          -          .          NM_000417
chr10          GenomeTrafac   exon          6100021       6100088       .          -          .          NM_000417
chr10          GenomeTrafac   exon          6101396       6101468       .          -          .          NM_000417
chr10          GenomeTrafac   exon          6101838       6101910       .          -          .          NM_000417
chr10          GenomeTrafac   exon          6103446       6103662       .          -          .          NM_000417
chr10          GenomeTrafac   exon          6106212       6106323       .          -          .          NM_000417
chr10          GenomeTrafac   exon          6107802       6107994       .          -          .          NM_000417
chr10          GenomeTrafac   exon          6144056       6144278       .          -          .          NM_000417
# exon 2 data NM_008367
track          name="Mouse Exons"          description="Mouse Il2ra exons"          color=128,0,128          visibility=2
priority=2     useScore=0
chr10          GenomeTrafac   exon          6093936       6097370       .          -          .          NM_008367
chr10          GenomeTrafac   exon          6102684       6102751       .          -          .          NM_008367
chr10          GenomeTrafac   exon          6104002       6104074       .          -          .          NM_008367
chr10          GenomeTrafac   exon          6105125       6105197       .          -          .          NM_008367
chr10          GenomeTrafac   exon          6106702       6106918       .          -          .          NM_008367
chr10          GenomeTrafac   exon          6110187       6110301       .          -          .          NM_008367
chr10          GenomeTrafac   exon          6112656       6112833       .          -          .          NM_008367
chr10          GenomeTrafac   exon          6144063       6144278       .          -          .          NM_008367
# repeat 1 data NM_000417
track          name="Human RepeatMask"     description="Human Repeat Elements by RepeatMask"        color=0,60,120 visibility=2
priority=3     useScore=1
chr10          GenomeTrafac   repeat        6077386       6077435       87.66827762323152          .          .          DNA;DNA
chr10          GenomeTrafac   repeat        6076916       6077065       87.66827762323152          .          .          DNA;DNA
```


GenomeTrafac: IL2RA GFF
Base Position
6090000
6095000
6100000
6105000
6110000
6115000
6120000
6125000
6130000
6135000
6140000
6145000
6150000
NM_000417
Human IL2RA exons
Human Repeat Elements by RepeatMask
LINE
SINE
Low_complexity
DNA
LTR
Simple_repeat
Other
Local Alignments
Local Alignments Human vs. Mouse
Concise Alignments
Alignment Curve
TF Binding Sites
Conserved TF binding sites on this gene.
Enhancers on this gene
PositiveRegulatoryRegionII
PositiveRegulatoryRegionI
PositiveRegulatoryRegionIII
Human mRNAs from GenBank
X01057
K03122
M14098
AF008556
GenomeTrafac: IL2RA GFF
Base Position
6144000
6144050
6144100
6144150
6144200
6144250
6144300
6144350
6144400
6144450
6144500
6144550
6144600
6144650
6144700
6144750
6144800
Human IL2RA exons
NM_000417
Human Repeat Elements by RepeatMask
AAAAAAAAAAAAAAAAAAA (simple repeat)
Local Alignments Human vs. Mouse
BlastZ Alignment
Alignment Curve
Conserved TF binding sites on this gene.
V$STAF.01:192
V$BTEB3.01:193
V$ELF2.01:194
V$ATATA.01:195
V$HNF3B.01:196
V$CDPCR3.01:197
V$HMGIY.01:198
V$PAX4.01:199
V$CHR.01:200
V$MYT1.01:201
V$PXRCAR.01:202
V$GKLF.01:203
V$EKLF.01:204
V$SREBP.02:205
V$NF1.01:206
Hs-Mm
Conserved
Binding Sites
Enhancers on this gene
PositiveRegulatoryRegionII
PositiveRegulatoryRegionI
Known enhancers

Tasks

- Parsing all GenBank files that have regulatory information explicitly stated (as in the “Features” E.g. enhancer – 1045 sequences). Extract the sequences and map them to GenomeTrafac and UCSC.
- Text-parsing “regulatory-rich” abstracts – Identify set of keywords.
- Associating nucleotide sequences to the “regulatory-rich” abstracts.
- Incorporating the noncoding SNPs and known regulatory region info into the GenomeTrafac GFF export files.