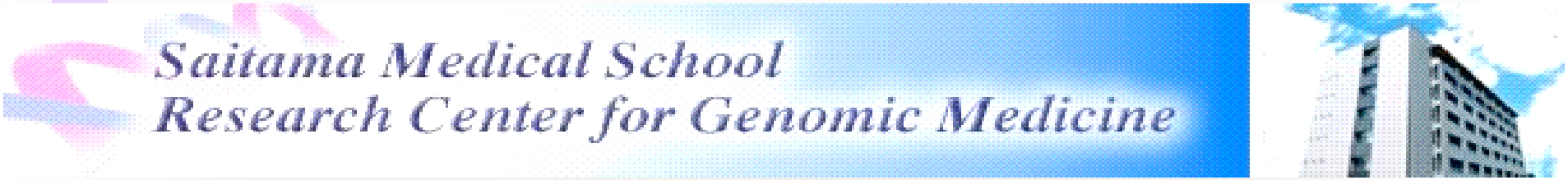
SayaMatcher （狭山茶）



Hidemasa Bono

Division of Functional Genomics & Systems Medicine
Research Center for Genomic Medicine

Saitama Medical School, JAPAN

http://kishoi.jp/SayaMatcher/

mailto: bono@saitama-med.ac.jp

Working both on bench and desk

I have studied tissue specific genes, but…

Browse genes with tissue-specific expression

09 spleen	58 thymus	06 kidney	10 heart
07 brain	15 cerebellum	12 lung	13 liver
65 cerebellum_neonate10day	16 placenta	17 testis	83 uterus
18 pancreas	20 small_intestine	22 stomach	90 colon
47 skin_neonate10day	98 bone	xm muscle	B4 adipose

Tissues colored by Tissue Ontology
immune circulator neural germ digestive skin bone muscle adipose

READ

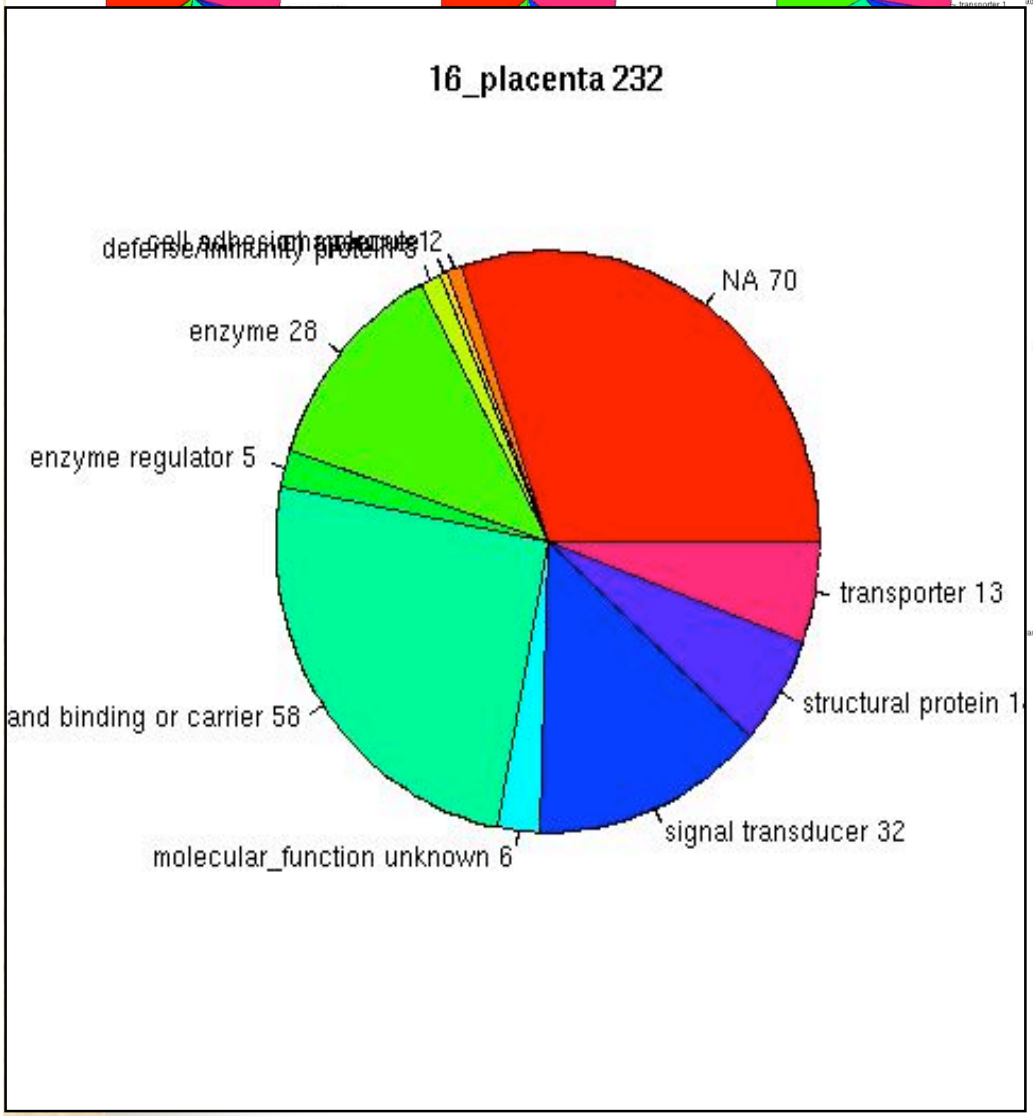
RikenExpressionArrayDatabase

READ: Log-transformed ratio data for in expression data

cloneid	09	58	06	10	12	13	07	15	65	16	17	83	18	20	22	90	47	98	xm	B4	
1600013K07	0	0.1	0.1	0.2	0.2	0.1	0.6	0.4	0.3	7.8	0.2	0.6	0.4	0	0.2	0	0	0.3	0.8	0	TB5731 Mus musculus trophoblast specific protein alpha (Tpbpa), mRNA. [RTPS]
5133400E06	0	0.8	0	0.4	0	0	0	0	0.2	6.4	0.8	0.2	0.5	0.1	0.4	0	0.1	0	0.5	0.1	Unknown EST [phase1]
1600012I11	0.2	1.4	0	0.1	0.5	0.1	0.2	0.1	0.2	7	0	0.1	0.1	0.3	0	0	0.2	0.2	0.7	0.2	Unknown EST [phase1]
1600022H16	0.5	1	0.5	0	0	0.3	0.2	0	0.1	8.2	0.3	1.7	0.5	0	0.5	0.1	0.2	0.3	0.2	0.5	TB6423 Mus musculus cathepsin J (Ctsj), mRNA. [RTPS]
1600027L19	0.3	0.3	0.4	0.2	0	0.3	0.1	0.3	1.4	7.5	0.3	0.8	0.8	0.4	0	0.5	0.2	0.4	0.2	0	TB4512 Mus musculus prolactin-like protein C (Prlpc), mRNA. [RTPS]
1600021O09	0	0.1	1	0.1	0.2	0.4	0.1	0	0.2	7.5	0.1	0.8	1.2	0.4	0	0.1	0.1	0.4	0.3	0	TB5731 Mus musculus trophoblast specific protein alpha (Tpbpa), mRNA. [RTPS]
1600019F10	0.2	0.2	0.2	0.4	0	0.1	0.3	0.2	0.7	7.1	0.3	0.2	0.4	0.1	0.3	0.3	0	0.2	0.3	1.3	similar to PROLACTIN-LIKE PROTEIN C BETA ISOFORM 1-A [Rattus norvegicus] [fantom]
1600021M17	0	1.3	0.5	0	0	0.7	0.1	0	1.4	7.7	0.1	0.8	0.7	0.7	0	0.2	0	0.1	0	0.1	TF11211 similar to CEACAM11 [Mus musculus] [RTPS]
1600019E22	0.4	0.6	0.9	0.3	0	0.3	0.1	0	0.2	7.9	0.3	1.4	0.5	0	0.6	0	0.3	0.7	0.9	0.3	Mm.43505 pregnancy-specific glycoprotein 28 [unigene] [phase1]
1600027N17	0.6	0.6	0.3	0.2	0.6	0	0.1	0.5	1.4	8	0.3	0.7	0.7	0.4	0.2	0.7	0	0	0.7	0.1	TF13031 hypothetical protein [RTPS]
1600019K02	0	0.1	0.7	0	0.1	0.2	0	0.2	0.8	7.8	0.1	1.8	0.3	0.2	0.4	0.1	0.4	0.5	0.1	0.5	TB7597 Mus musculus pregnancy-specific glycoprotein 23 (Psg23), mRNA. [RTPS]
1600025N01	0.2	1.3	0.3	0.5	0.1	0	0.2	0.2	0.8	7.8	0	0.7	0.9	0.4	0.6	0.2	0	0.3	0.6	0.1	pregnancy-specific glrcoprotein 21 [fantom]
1620401J06	0.3	0.3	0	0.2	0.2	0	0.5	0	0.3	6.6	0	0.1	1.6	0.1	0.4	0.5	0.1	0	0.3	0.1	trophoblast specific protein alpha [fantom]
1600017M18	0.2	1.9	0.3	0.5	0.1	0.7	0.2	0.4	0.6	7.2	0	0.6	0.7	0.5	0.3	0	0.3	0.1	0	0.5	Mm.43505 pregnancy-specific glycoprotein 28 [unigene] [phase1]
1600025D07	0.4	0.2	0.6	0.2	0	0	0	0.1	0.5	4.1	0	0.1	0.4	0.2	0.1	0	0.4	0.2	0.5	0	TB4361 Mus musculus chorionic somatomammotropin hormone 1 (Csh1), mRNA. [RTPS]
1600029O21	0.7	0.7	0.8	0.6	0.2	0	0.3	0.4	0.1	7.3	0.5	1.5	0.6	0	0.6	0.4	0	0.8	0.4	0.1	Mm.37203 chorionic somatomammotropin hormone 2 [unigene] [phase1]
1600017E04	0.2	1.2	0.7	0.4	0.1	0.3	0.1	0.3	0.7	7.6	0.4	1.7	0.7	0.1	0.3	0.1	0	0.9	0.2	0	PROLACTIN-LIKE PROTEIN C [fantom]
1600029A13	0	1.2	0.7	0.7	0.7	0.5	0.2	0.3	0.6	7.7	0.4	1.3	0.6	0.6	0.7	0.3	0.2	0.5	0.3	0.1	Mm.46091 prolactin-like protein C 2 [unigene] [phase1]
1600026F07	0.4	0.4	1	0.7	0.2	0.5	0.1	0.3	1.2	7.1	0.4	1.4	1	0.1	0.2	0.2	0.2	0	0.4	0	Unknown EST [phase1]
3830418E20	1.1	0.1	0.8	0.9	0.4	0.6	0.3	0.2	0.3	6.9	0	0.6	1.1	0.2	0	0.1	0	0.3	1.2	0.3	trophoblast specific protein alpha [fantom]
1620401H06	0.1	0.6	0.4	0.6	0.5	0.8	0.5	0.5	0.1	7.7	0.1	0.7	1.6	0.3	1	0.5	0.5	0.2	1.2	0.7	TB7972 Mus musculus cathepsin M (Ctsm), mRNA. [RTPS]
1600019A08	0.7	1	0.2	0.4	0	0	0.1	0	1.3	7.2	0.5	0.6	0.9	0.1	0.3	0.4	0.3	0.3	1.2	0.2	TB5731 Mus musculus trophoblast specific protein alpha (Tpbpa), mRNA. [RTPS]
1600021E16	0.4	0	0.5	0.2	0	0.3	0.1	0.3	0.6	7.3	0.5	0.7	0.2	0.1	0.2	0	0.1	0.7	0.5		TB6423 Mus musculus cathepsin J (Ctsj), mRNA. [RTPS]
1600029O14	0.3		0.5	0.1	0.3	0.2	0.7	0.4	0.1	7.6	0.3	0	0.2	0.1	0.1	0	0.1	0	1.2	0.1	TB9156 Mus musculus CE
9430076M13	1	0.7	1.3	1	0.6	0.6	0.5	0.6	0.1	4.5	1.7	0	0.4	0.7	0.7	0.5	0	0.8	0.4	0.5	unknown EST [fantom]
1600016O08	0.3	0.6	0	0	0.3	0.6	0.1	0.5	0	6.9	0.4		0.7	0.3	0.3	0.4	0.4	0	1.3	0.4	TB6362 Mouse pregnancy
1600016D15	0.3	0.2	1.2	0		0.1	0.6	0	0.1	6.5	0.1	0	0.1	0.4	0.1	0.1	0.1	0.3	0.8	0.2	TB7972 Mus musculus ca
1600020L08	0.8	0.2	0.4	0.2	0.3	0.2	0.2	0.2		6.1	0.1	0.4	1	0.2	0.3	0	0.1	0.1	0.6	0.2	prolactin-like protein F [f
1600029H12	0.1	0.9	1.2	0.4	0	0.3	0	0	1	7.3	0.5	1	0.7	0.1	0.1	0	0.6	0.4	0.8	0.4	weakly similar to CARCIN
1600019K19	0.6		0.6	0.4	0.2	0.3	0	0.3	0.7	8.1	0.2	1.6	0.5	0.4	0.6	0.3	0.1	0.6	0.5	0.2	Unknown EST [phase1]
1600027J17	0.6	0.8	1	0.3	0.1	0	0.2	0	0.2	7.1	0.6	1.6	1	0.1	0.5	0.1	0.1	0.5	1.3	0	cathepsin M [fantom]
1600009O20	0.1	0	0.3	0.4	0.1	0.5	0.4	0.1	0.5	7.8	0	0.6	0.5	0.5		0.4	0.4	1.7	0.7	0	similar to PROLACTIN-LIK
1600021L09	0.6	2.6	0.3	0.3	0	0	0.1	0	0.6	8	0.1	0.7	0.7	0.3	0.4	0.4	0.1	0.3	0.6	0.2	Mm.43505 pregnancy-spe
5730550L03	0.4	0.5	0.9	0.2	0.5	0.2	0.5	0.3	0.5	4.2	0.6	0.4	0.8	0.2	0.5	0.5	0.5	0.7	0.7	0.6	TB3523 Mus musculus myristoylated alanine rich protein kinase C substrate (Macs), mRNA. [RTPS]

Gene expression profile of mouse 20tissues from

http://READ.gsc.riken.jp/fantom2/



GO slim classification of tissue specific genes in mouse 20 tissues from

http://READ.gsc.riken.jp/fantom2/

not so easy to interpret functional property of these to extract gene regulation network!

Depicting gene regulation networks from microarray data

1. Co-regulated genes
   - Any DNA motifs in upstream regions?

   ⇒ hard to decipher them in mammal

2. Known transcription factor binding sites (TFBS)
   - Any features in gene expression profiles for genes with particular TFBS in upstream region?

What is SayaMatcher?

- System(pipeline) to get coordinates of transcription factor binding sites (TFBS) in the genome
  - TFBS pattern to be found is too short to blast/blat/ssaha, and have too many hits for genome sequence
  - Not interpretable only in the text output
    - export the annotation to genome browsers (Ensembl, UCSC)

Why is the pipeline needed?

- Genome sequences are updating
  - ⇒ not isolated task; searches need to be a ‘pipeline ’
- Too many iteration for one time job
  - All chromosome and its reverse complement
    - around 40 (around 20 x 2)
  - Species to be used (3-5)
  - Several kinds of NRE(around10)

Thousands of iterative operation needed for update

Methods for searching TFBS

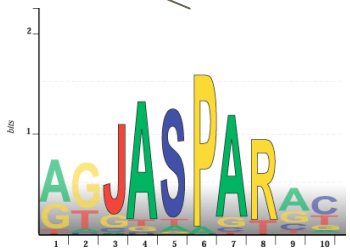
1. Regular expression
   - When the‘consensus sequence’ is reported
   - `dreg, fuzznuc` in EMBOSS
2. Position specific scoring matrix(PSSM)
   - When consensus is weak, but multiple alignment is available
   - `prophecy/profit` in EMBOSS, HMMER

SayaMatcher

Patterns of interest
(from biologist)

dreg
fuzznuc
emboss
profit
prophecy

mfa
(from
papers)



LDAS → candidate → HTML

BED

Genome
annotation

Transcriptome
annotation

UCSC Genome Browser

DAS

project Ensembl

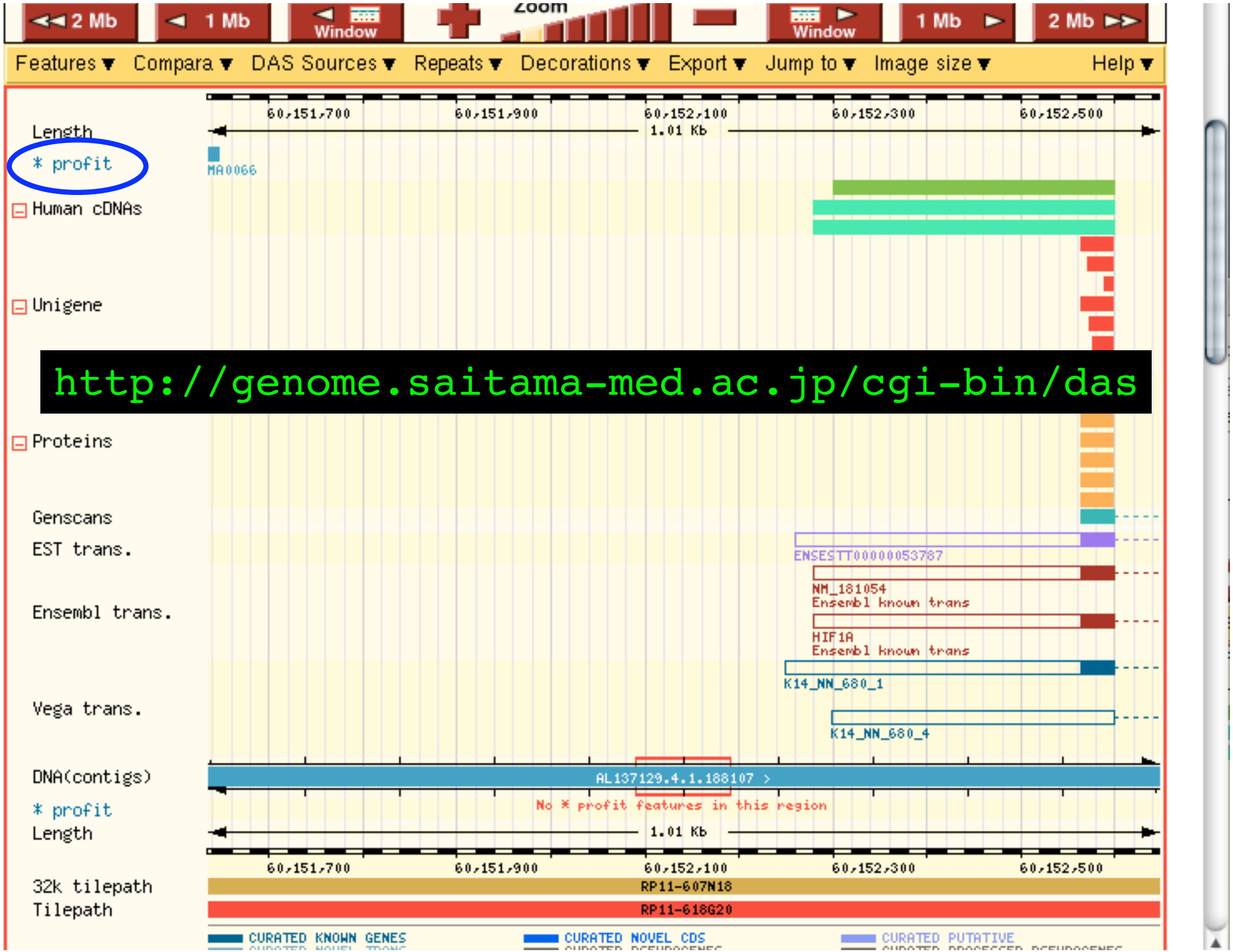
Genes with predicted ERE in upstream region (-10kb)

http://10.53.95.9/~bono/ERE_human_10k.html

genes with ERE

Link to genome browser

X:18843340	-1	ENSG00000188310	\N
11:3399598	1	ENSG00000182139	\N
11:5192437	-1	ENSG00000176742	Odorant receptor HOR3'beta1. [Source:SPTREMBL;Acc:Q9H2C8]
4:56728996	1	ENSG00000174799	\N
8:8132003	1	ENSG00000173295	\N
8:11684731	1	ENSG00000177907	\N
15:62106866	-1	ENSG00000166797	\N
5:55043428	1	ENSG00000152670	DEAD-box protein 4 (VASA homolog). [Source:SWISSPROT;Acc:Q9NQI0]
17:32838007	-1	ENSG00000108702	Small inducible cytokine A1 precursor (CCL1) (T lymphocyte-secreted protein I-309). [Source:S
15:73377925	-1	ENSG00000140400	Alpha-mannosidase 2C1 (EC 3.2.1.24) (Alpha-D-mannoside mannohydrolase) (Mannosidase alpha cla
15:76143692	-1	ENSG00000136425	Kinase interacting protein 2 (KIP 2). [Source:SWISSPROT;Acc:O75838]
11:128276628	-1	ENSG00000151704	ATP-sensitive inward rectifier potassium channel 1 (Potassium channel, inwardly rectifying, s
16:30124321	1	ENSG00000149923	Serine/threonine protein phosphatase 4 (EC 3.1.3.16) (Pp4) (Protein phosphatase X) (PP-X). [S
17:67057071	-1	ENSG00000070540	\N
6:27875107	1	ENSG00000124518	Histone H2A.c/d/i/n/p (H2A.1) (H2A/c) (H2A/d) (H2A/i) (H2A/n) (H2A/p) (H2A.1b). [Source:SWISS
6:32026499	1	ENSG00000166291	Helicase SKI2W (Helicase-like protein) (HLP). [Source:SWISSPROT;Acc:Q15477]
12:48484210	-1	ENSG00000167566	\N
12:53273432	-1	ENSG00000135447	Protein phosphatase inhibitor 1 (IPP-1) (I-1). [Source:SWISSPROT;Acc:Q13522]
12:54674057	1	ENSG00000139531	Sulfite oxidase, mitochondrial precursor (EC 1.8.3.1). [Source:SWISSPROT;Acc:P51687]
12:58484579	1	ENSG00000189370	\N
2:11686471	1	ENSG00000174934	GREB1 protein isoform a; gene regulated by estrogen in breast cancer protein. [Source:RefSeq;
12:94750169	1	ENSG00000139343	Small nuclear ribonucleoprotein F (snRNP-F) (Sm protein F) (Sm-F) (SmF). [Source:SWISSPROT;Ac
2:74014735	1	ENSG00000187833	\N
2:132402869	-1	ENSG00000182126	\N
20:10403543	1	ENSG00000149346	DJ1099D15.3.1 (Novel protein (Isoform 1)) (Fragment). [Source:SPTREMBL;Acc:Q9BR42]


2 Mb
1 Mb
Window
Zoom
Window
1 Mb
2 Mb
Features
Compara
DAS Sources
Repeats
Decorations
Export
Jump to
Image size
Help
Length
* profit
Human cDNAs
Unigene
http://genome.saitama-med.ac.jp/cgi-bin/das
Proteins
Genscans
EST trans.
Ensembl trans.
Vega trans.
DNA(contigs)
* profit
Length
32k tilepath
Tilepath
60,151,700
60,151,900
60,152,100
60,152,300
60,152,500
1.01 Kb
MA0066
ENSESTT00000053787
NM_181054
Ensembl known trans
HIF1A
Ensembl known trans
K14_NN_680_1
K14_NN_680_4
AL137129.4.1.188107 >
No * profit features in this region
RP11-607N18
RP11-618G20
CURATED KNOWN GENES
CURATED NOVEL CDS
CURATED PUTATIVE

Acknowledgement

- Grant-in-Aid for Scientific Research on Priority Areas (C) ”Genome Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan
- Grant-in-Aid for Development of New Technology from The Promotion and Mutual Aid Corporation for Private Schools of Japan

poster: D-4