Inserm
UMR 599 Centre de Recherche en Cancérologie de Marseille

INSTITUT CENTRE RÉGIONAL DE LUTTE CONTRE LE CANCER PAOLI-CALMETTES PROVENCE-ALPES-CÔTE D'AZUR

TAGC EAM 296 INSERM

icim Institut de Cancérologie et d'Immunologie de Marseille

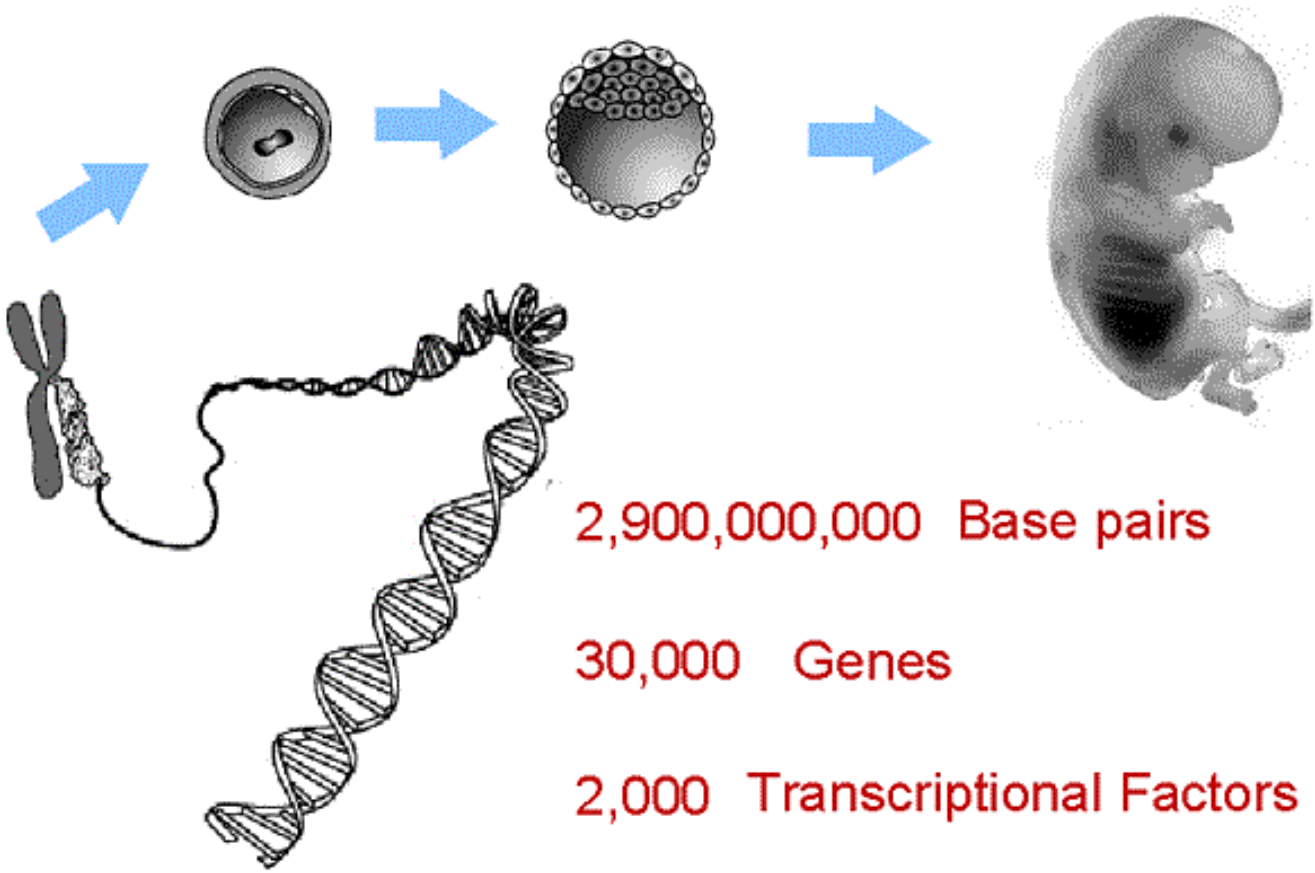
Analysis and annotation of the transcriptional regulatory sequences of higher eukaryotes: the point of view from the wet lab

Après-midi thématique séquences de régulation transcriptionnelle
IRISA – Rennes – Jeudi 11 janvier 2007

Jean Imbert
Equipe Régulations Transcriptionnelles

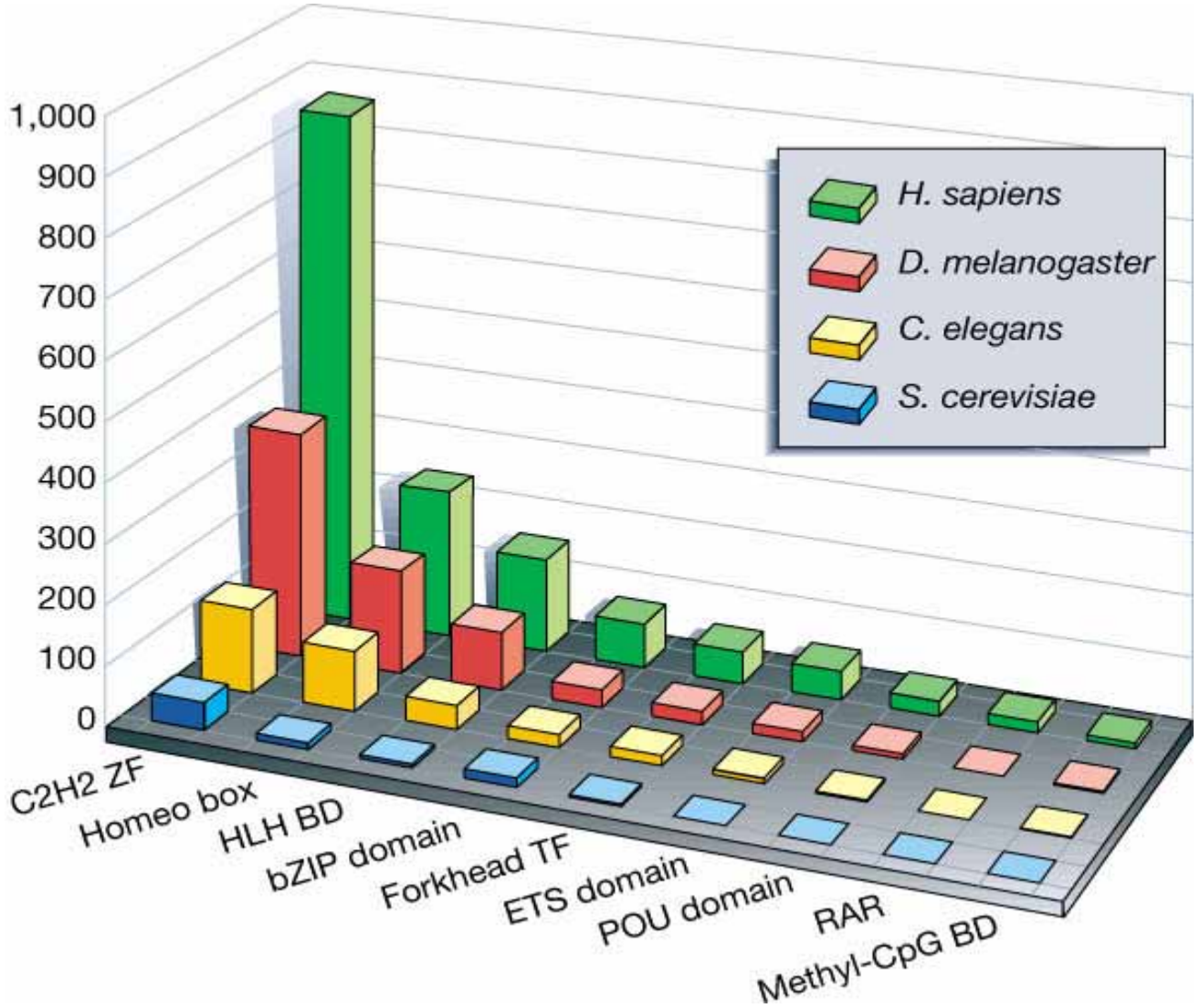
Si vous aviez des doutes sur l’intérêt de la chose…

The Human Genome



From Bing Ren, Laboratory of Gene Regulation, Ludwig Institute for Cancer Research, UCSD

Genome-wide comparison of transcriptional activator families in eukaryotes



C2H2 zinc fingers are found in 2% of all human genes, and they are by far the most abundant class of DNA-binding domains found in human transcription factors.

From Tupler R, Perini G, Green MR. Nature 409: 832-833, 2001.

Cancer Associated Transcription Factors

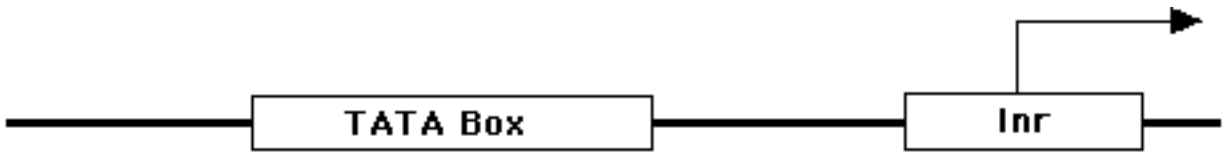
NAME	CANCER TYPE
AL5Q31	acute lymphoblastic leukemia
ASCL1	small cell lung cancer (SCLC)
BCL3	B-cell leukemia
BCL6	B cell lymphoma
CBFB	myeloid leukemia
CBL	pre-B and pro-B cell lymphomas
CTNNB1	Colon cancer
DEK	Leukemia, acute nonlymphocytic
ERG	Acute myeloid leukemia (AML)
ETS1	erythroblastosis
ETS2	erythroblastosis
ETV6	AML
FOS	murine osteosarcoma
FKHR	rhabdomyosarcoma
GAS41	Glioma
GLI	Glioma
HOX11	Leukemia
IRF2	AML
IRF4	multiple myeloma
JUN	murine osteosarcoma
LMO2	Acute T-cell leukemia
LYL1	T-cell leukemia
MAF	multiple myeloma
MLL	AML
MYB	Leukemia
C-MYC	Lymphoma, Breast Cancer, lung cancer
N-MYC	neuroblastoma

NAME	CANCER TYPE
NFKB1	Acute lymphoblastic leukemia
NFKB2	B-Cell lymphoma
PAX3	alveolar rhabdomyosarcoma
PAX5	B-Cell lymphoma
PAX7	alveolar rhabdomyosarcoma
PBX1	leukemia
RARA	AML
REL	Diffuse large cell lymphoma
RELA	Diffuse large cell lymphoma
RUNX1	Leukemia, acute myeloid
SPI1	acute murine erythroleukemia
STAT3	Leukemia
TAL1/SCL	T cell leukemia
TAL2	T cell leukemia
TCF3	Acute leukemias
BLIMP1	B-cell non-Hodgkin lymphoma
E2F1	Murine Reproductive tract sarcomas
IRF1	AML
MAXI1	prostate adenocarcinoma
PML	Acute promyelocytic leukemia
RB1	retinoblastoma
SMAD3	colorectal cancer
SMAD4/DPC4	pancreatic carcinoma; juvenile polyposis
TFE3	renal cell carcinoma
TP53	Colorectal cancer and other types
WT1	Wilms tumor
ZF9/KLF6	prostate cancer

From Bing Ren, Laboratory of Gene Regulation, Ludwig Institute for Cancer Research, UCSD

Comment définir un gène et ses séquences régulatrices ?

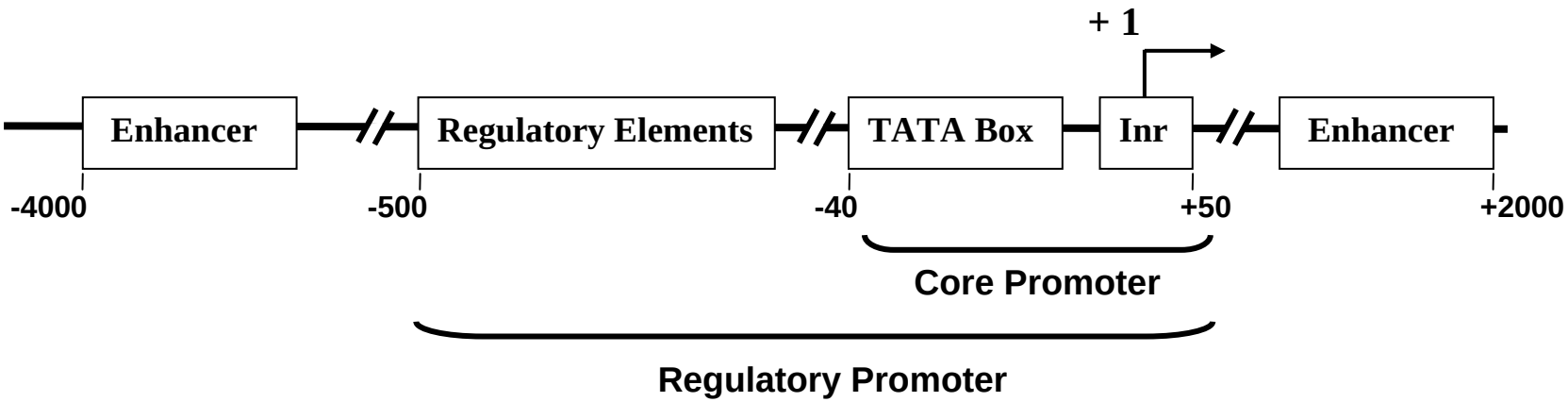
CORE PROMOTER



TATA Box: TATAAAA (about 25 base pairs upstream of the start point)

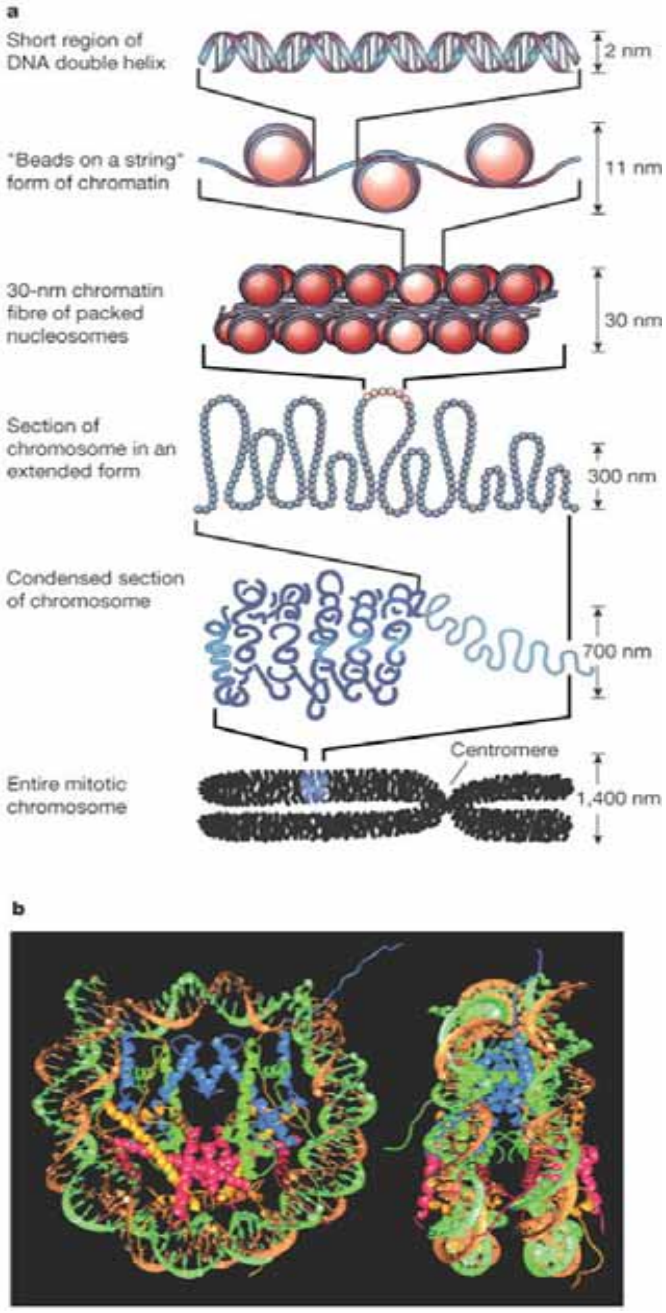
Initiator (Inr): PyPyC**A**PyPyPyPyPy

MODEL OF TYPICAL GENE PROMOTER AND REGULATORY REGIONS

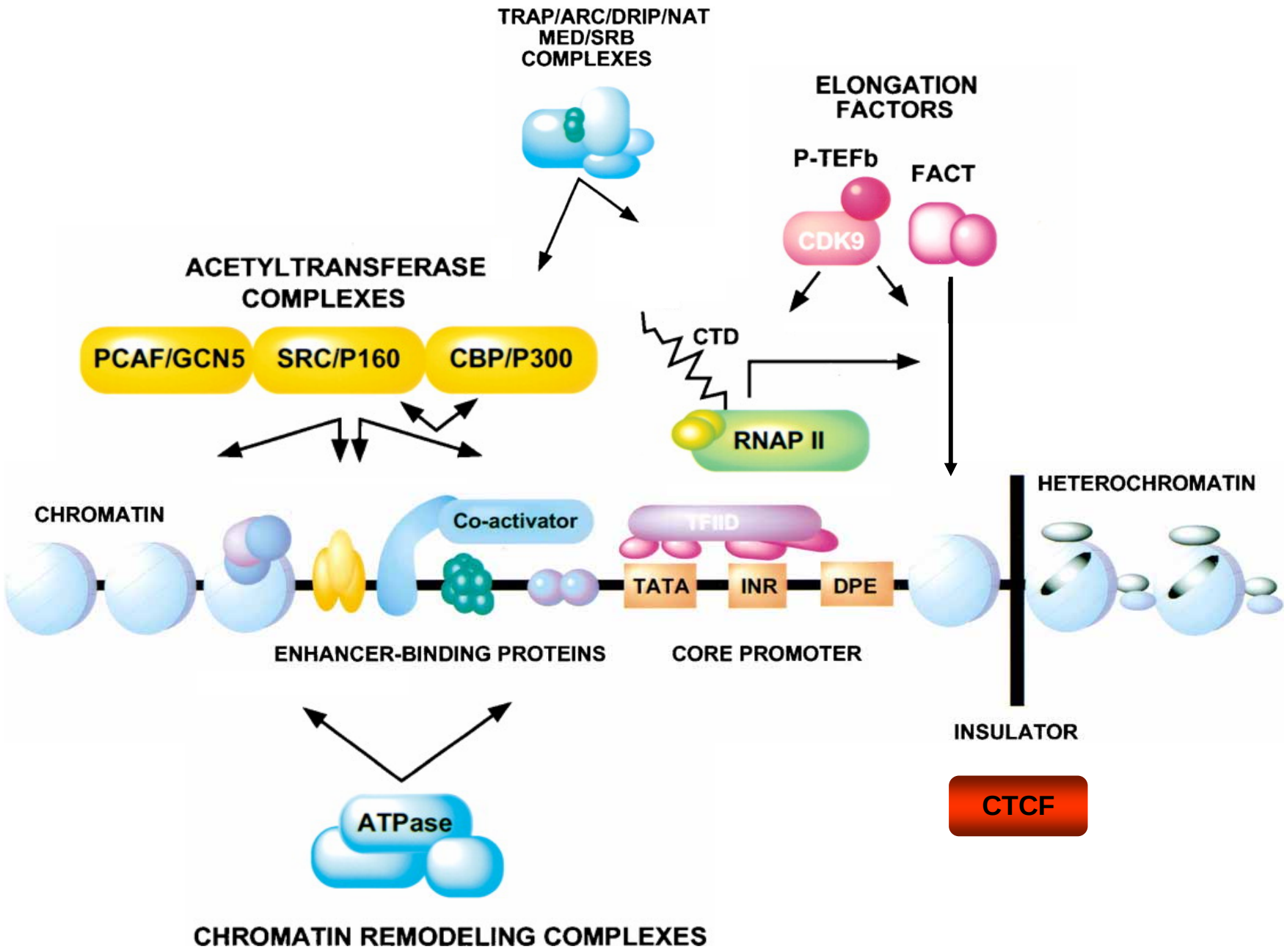


***cis*-acting regulatory elements**
From Pennacchio, L. A. and E. M. Rubin. 2001. Genomic strategies to identify mammalian regulatory sequences. Nat. Rev. Genet. 2:100-109.
***Promoter**:* Sequence of DNA near the 5' end of a gene that acts as a binding site for RNA polymerase and from which transcription is initiated.
Enhancer: Control element that elevates the levels of transcription from a promoter, independent of orientation or distance.
Locus control region (LCR): Confers tissue-specific temporally regulated expression of linked genes. LCRs function independently of position, but they are copy number dependent and open the nucleosome structure so that other factors can bind. LCRs affect replication timing and origin usage.
Boundary element/insulator: DNA sequence that prevents the activation or inactivation of transcription by blocking the effects of surrounding chromatin.
Silencer: Control element that suppresses gene expression independent of orientation or distance.
Matrix attachment region (MAR)/scaffold attachment region (SAR): DNA sequence that binds the nuclear scaffold and can affect transcription. These elements probably form higher-order looped structures within chromosomes and influence gene expression by separating chromosomes into regulatory domains.

Ça se complique encore…



From Felsenfeld and Groudine, Nature, 421:448-454, 2003



Modified from Jones and Kadonaga, Genes Dev. 14:1992-1996, 2000.

Human Genome

Lander et al. (2001) Nature 409: 860-921

Table 21 Characteristics of human genes

	Median	Mean
Internal exon	122 bp	145 bp
Exon number	7	8.8
Introns	1,023 bp	3,365 bp
3′ UTR	400 bp	770 bp
5′ UTR	240 bp	300 bp
Coding sequence	1,100 bp	1,340 bp
(CDS)	367 aa	447 aa
Genomic extent	14 kb	27 kb

x 35 k genes

Human Genome size : 3 Gb

Primary transcripts (pre-mRNA) : 1 Gb

Intergenic DNA: 2 Gb

2,000,000 probes at 1 kb resolution

Prediction of promoting regions (1st exon) ?

TOOLS FOR THE IDENTIFICATION OF REGULATORY SEQUENCES AND THEIR COGNATE TRANSCRIPTION FACTORS

1. Identification and characterization of regulatory sequences

- Gene reporter assays: transient or stable transfection (CAT, Luciférase, SEAP, GFP, β-gal, etc.)
- *In vitro* transcription assay
- Sequencing, database mining (web: TESS, Euk. Pr. Database, TRANSFAC, MATINSPECTOR, TFSEARCH, etc.)
- Animal or cellular models: enhancer trap, enhancer knock-in, minichromosomes, etc.

2. Identification and characterization of specific transcription factors (TFs)

- EMSA and sequels: UV-crosslinking, pull-down assay using biotynilated oligonucleotide
- Footprint detection:
  * *in vitro :* nucleases (DNAse I hypersensitivity, S1 nuclease, Mmase) or chemical compounds
  * *in vivo* : genomic footprinting, Chromatin ImmunoPrecipitation (ChIP) and sequels

3. TFs physical and functional interactions

- Transfection and biochemistry
- ChIP-on-chip

L’analyse des séquences….

Promoter Regulatory Elements: Features and Facts

- Degenerate sequence motifs
- Length: 6 to 20 bp
- Low complexity (8-12 bits)
- Binding sites of transcription factors
- Excess of binding sites over binding proteins in the nucleus
- Most in vitro binding sites not functional in vivo
- Some in vivo binding sites also not functional
- Regulatory potentials depends on cooperative effects between multiple elements

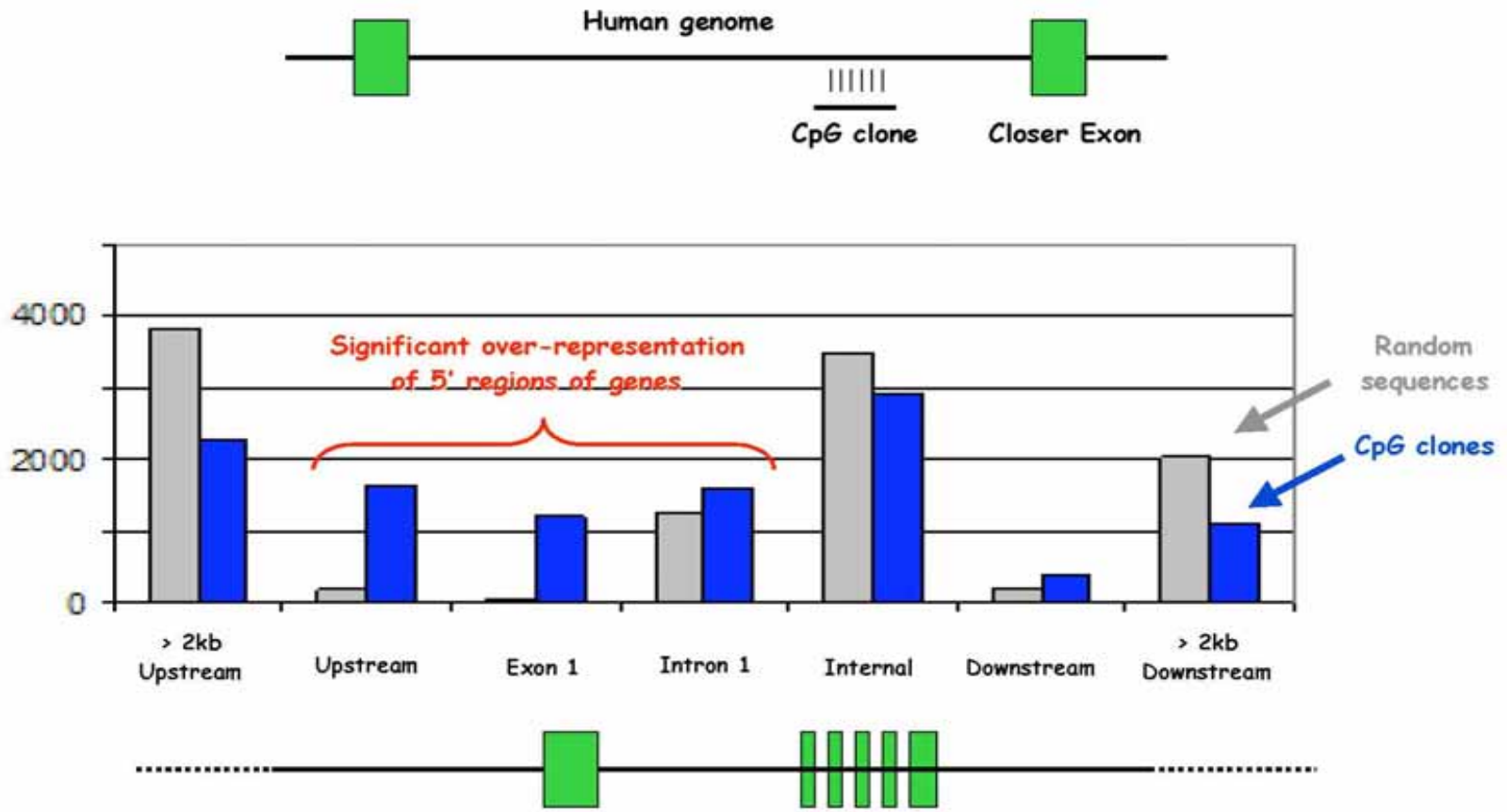
CpG islands

- CpG dinucleotides are present at 20% of predicted frequency
- CpG islands: >200 bp long, >50 %G+C, CpG >0.6 predicted
- CpG islands account for 1% of the genome
- 29,000 CpG islands are predicted in the human genome
- ~60% of known genes have a CpG island near 5' end
- CpG island microarrays are promoter- and regulatory region-enriched arrays

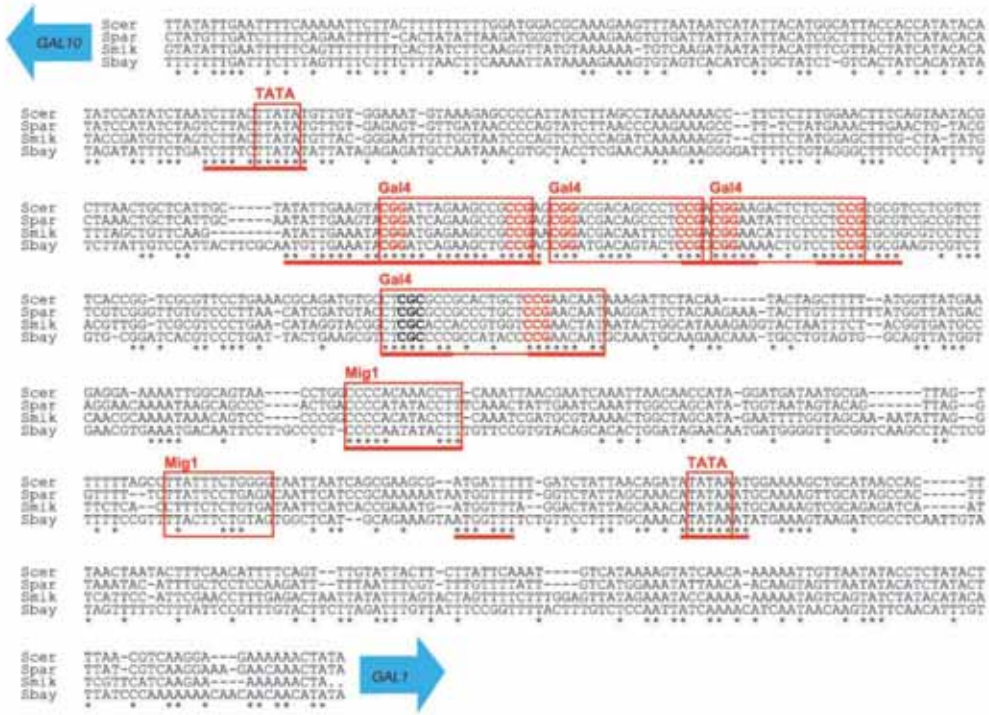
29,000 CpG islands are predicted in the human genome

Weinmann et al. (2002) Genes & Dev. 16:235-244
Oberley et al. (2004) Methods Enzymol. 376:315-334

CpG Island Localization



Regulatory information is encoded in the DNA…

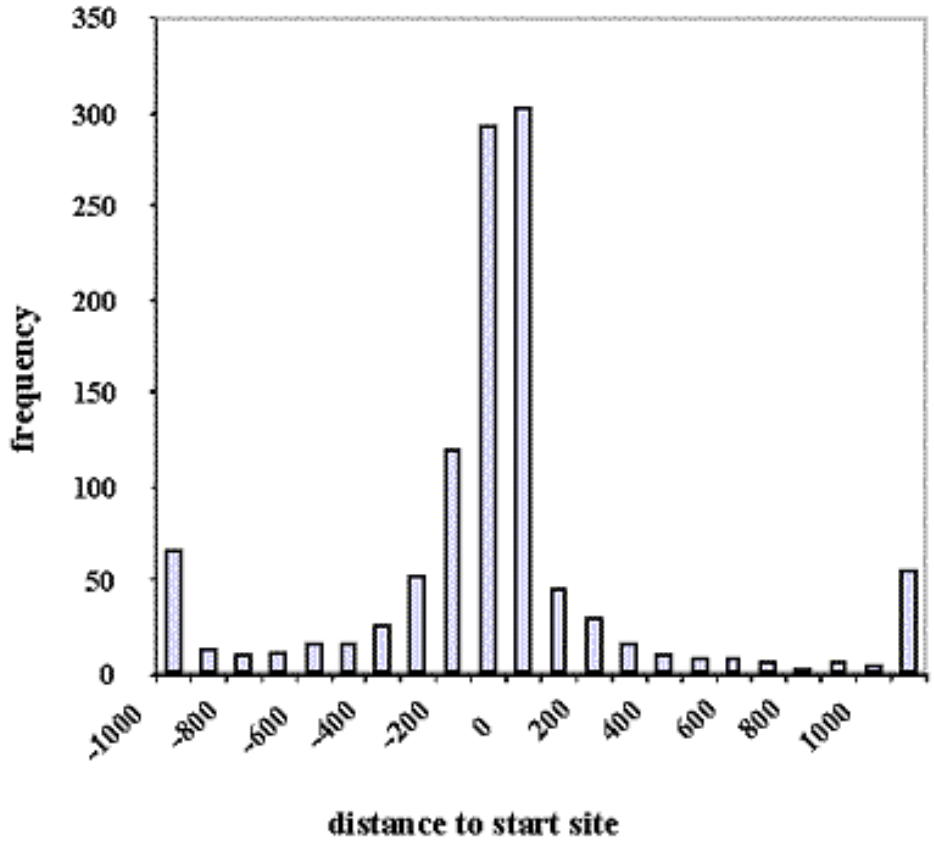


… and can be uncovered by phylogenetic footprinting

Source: Kellis et al. and Lander, Nature 423: 241-54 (2003)

From Julia Zeitlinger - Whitehead Institute, UC Davis, USA

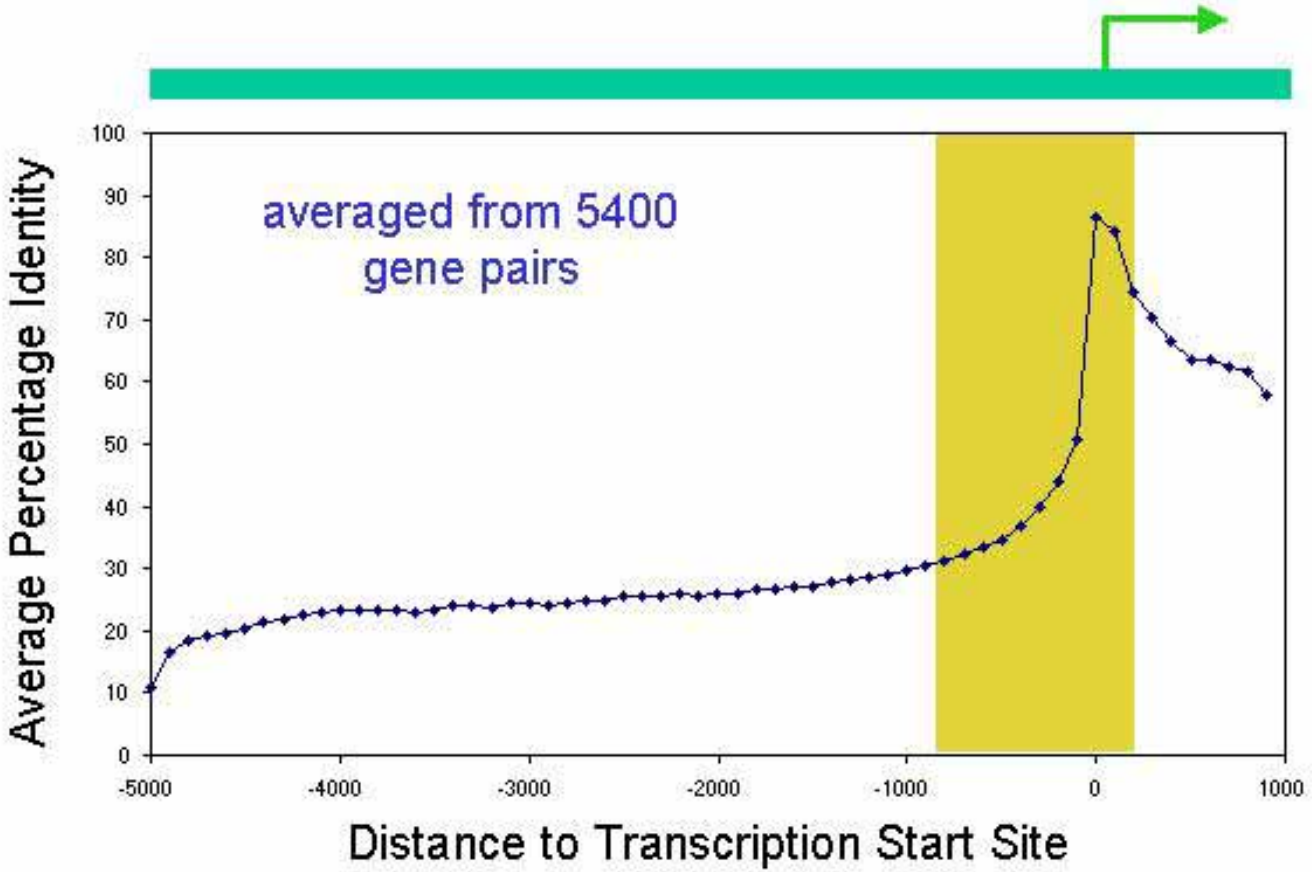
Transcription Factors Tend to Bind to Proximal Promoter Regions



Source:TransFAC

From Bing Ren, Laboratory of Gene Regulation, Ludwig Institute for Cancer Research, UCSD

Gene promoter regions are highly conserved between human and mouse orthologs



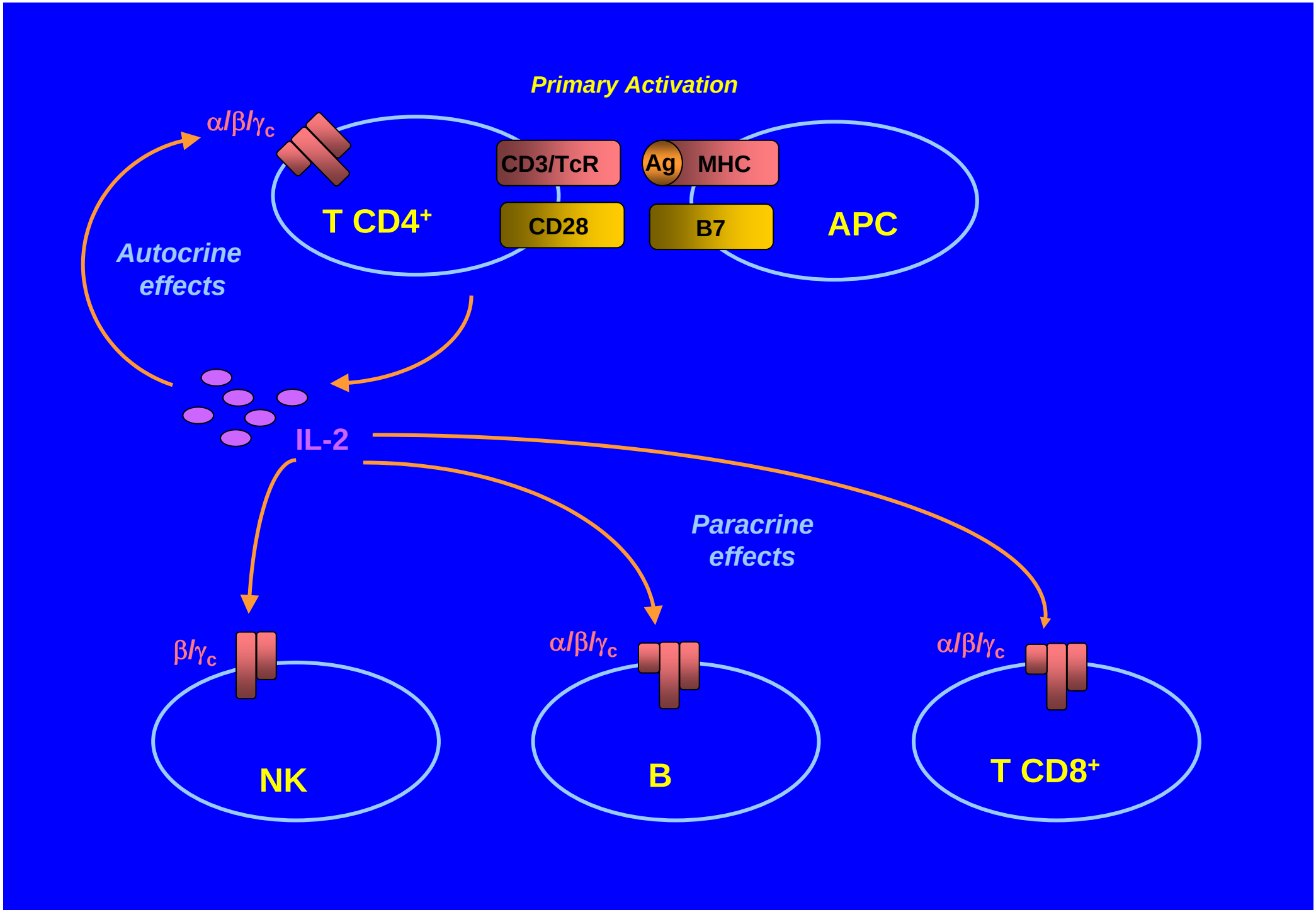
From Bing Ren, Laboratory of Gene Regulation, Ludwig Institute for Cancer Research, UCSD

Mais quelles sont les limites d’un gène ?

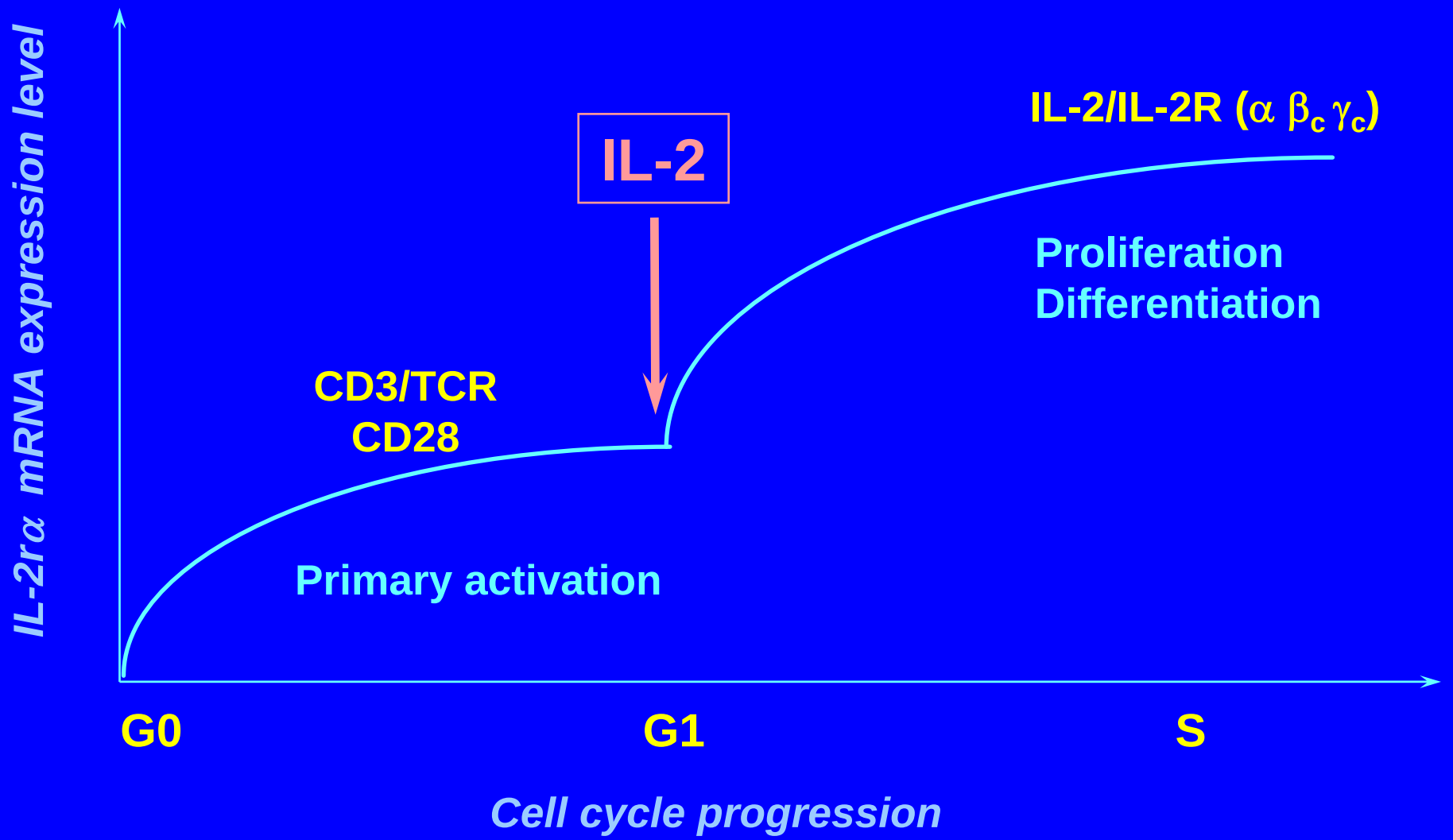
Some examples….

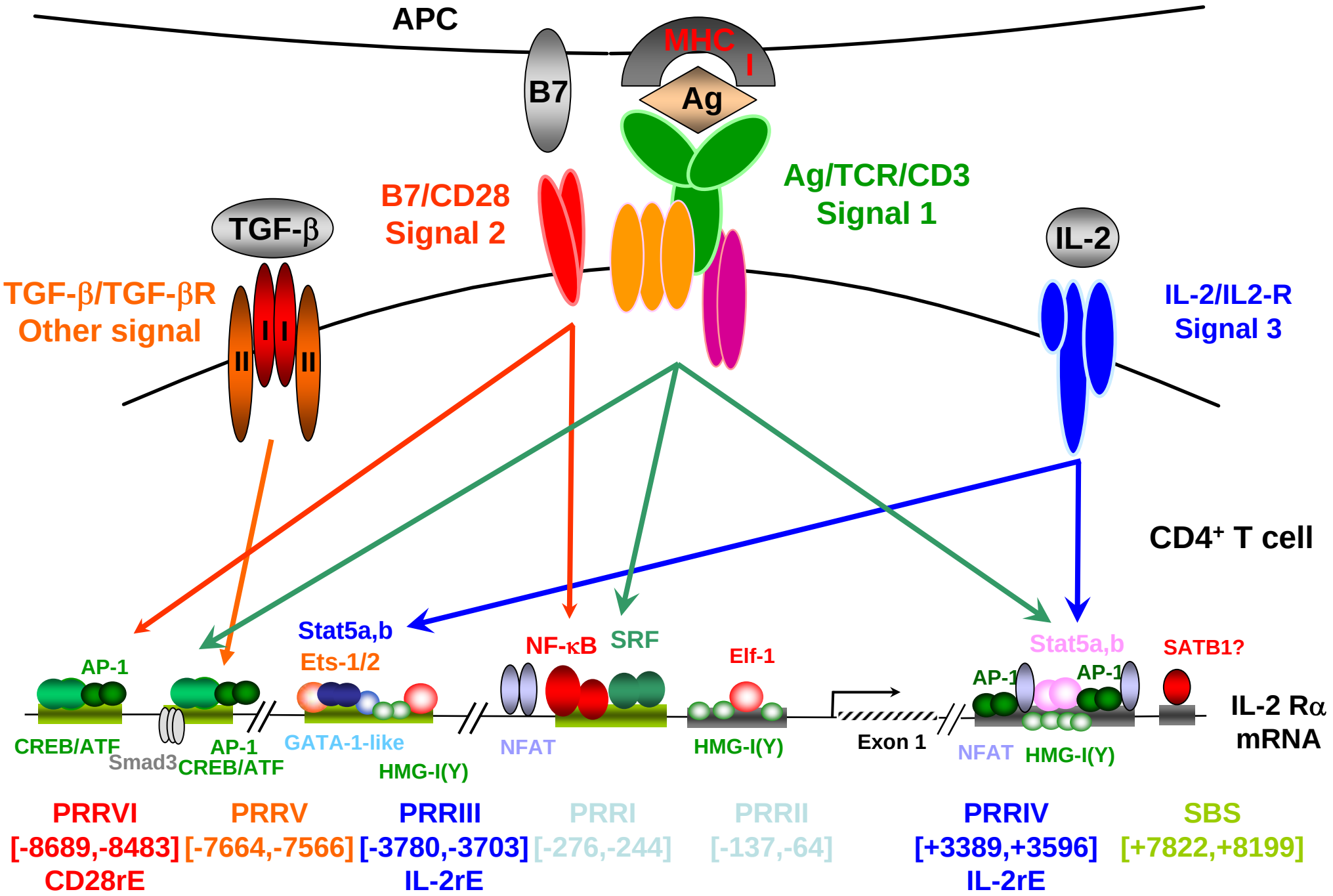
Control of CD25/IL-2Rα gene transcription

Kim, H. P., Imbert, J., and Leonard, W.J.
Both integrated and differential regulation of components of the IL-2/IL-2 receptor system.
***Cytokine Growth Factor Rev*, *online* doi:10.1016/j.cytogfr.2006.07.003, 2006.**


Primary Activation
α/β/γc
CD3/TcR
Ag
MHC
T CD4+
CD28
B7
APC
Autocrine effects
IL-2
Paracrine effects
β/γc
α/β/γc
α/β/γc
NK
B
T CD8+

CD25/IL-2Rα GENE TRANSCRIPTION DURING T CELL ACTIVATION




APC
B7
MHC I
Ag
B7/CD28
Signal 2
Ag/TCR/CD3
Signal 1
TGF-β
TGF-β/TGF-βR
Other signal
IL-2
IL-2/IL2-R
Signal 3
CD4+ T cell
Stat5a,b
Ets-1/2
NF-κB
SRF
Elf-1
Stat5a,b
SATB1?
AP-1
AP-1
AP-1
AP-1
IL-2 Rα
mRNA
CREB/ATF
Smad3
AP-1
CREB/ATF
GATA-1-like
HMG-I(Y)
NFAT
HMG-I(Y)
Exon 1
NFAT
HMG-I(Y)
PRRVI
[-8689,-8483]
CD28rE
PRRV
[-7664,-7566]
PRRIII
[-3780,-3703]
IL-2rE
PRRI
[-276,-244]
PRRII
[-137,-64]
PRRIV
[+3389,+3596]
IL-2rE
SBS
[+7822,+8199]

Available online at www.sciencedirect.com
ScienceDirect
Cytokine & Growth Factor Reviews 17 (2006) 349–366
Cytokine & Growth Factor Reviews
www.elsevier.com/locate/cytogfr

Survey

Both integrated and differential regulation of components of the IL-2/IL-2 receptor system

Hyoung Pyo Kim [a,*], Jean Imbert [b], Warren J. Leonard [a]

[a] *Laboratory of Molecular Immunology, Immunology Center, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892-1674, United States*
[b] *Centre de Recherche en Cancérologie de Marseille, UMR599 INSERM-Institut Paoli-Calmettes-Universite de la Mediterranee, 27 Boulevard Lei Roure, 13009 Marseille, France*

Available online 5 September 2006

Abstract

Interleukin-2 was discovered in 1976 as a T-cell growth factor. It was the first type I cytokine cloned and the first for which a receptor component was cloned. Its importance includes its multiple actions, therapeutic potential, and lessons for receptor biology, with three components differentially combining to form high, intermediate, and low-affinity receptors. IL-2Rα and IL-2Rβ, respectively, are markers for double-negative thymocytes and regulatory T-cells versus memory cells. γ_c, which is shared by six cytokines, is mutated in patients with X-linked severe-combined immunodeficiency. We now cover an under-reviewed area—the regulation of genes encoding IL-2 and IL-2R components, with an effort to integrate/explain this knowledge.
Published by Elsevier Ltd.

Keywords: IL-2; IL-2 receptor; Transcription; γ_c; XSCID

TPA BN000945

1: BN000945. Reports ...[gi:117414121]

Features Sequence

```
LOCUS       BN000945              78588 bp    DNA     linear   PRI 02-NOV-2006
DEFINITION  TPA_exp: Homo sapiens IL2RA gene for interleukin 2 receptor, alpha,
            promoter region and complete CDS.
ACCESSION   BN000945
VERSION     BN000945.1  GI:117414121
KEYWORDS    Third Party Annotation; TPA; IL2RA gene; interleukin 2 receptor,
            alpha; TPA:EXPERIMENTAL.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1
  AUTHORS   Cross,S.L., Feinberg,M.B., Wolf,J.B., Holbrook,N.J., Wong-Staal,F.
            and Leonard,W.J.
  TITLE     Regulation of the human interleukin-2 receptor alpha chain
            promoter: activation of a nonfunctional promoter by the
            transactivator gene of HTLV-I
  JOURNAL   Cell 49 (1), 47-56 (1987)
   PUBMED   3030566
REFERENCE   2
  AUTHORS   Cross,S.L., Halden,N.F., Lenardo,M.J. and Leonard,W.J.
  TITLE     Functionally distinct NF-kappa B binding sites in the
            immunoglobulin kappa and IL-2 receptor alpha chain genes
  JOURNAL   Science 244 (4903), 466-469 (1989)
   PUBMED   2497520
REFERENCE   3
  AUTHORS   Lin,B.B., Cross,S.L., Halden,N.F., Roman,D.G., Toledano,M.B. and
            Leonard,W.J.
  TITLE     Delineation of an enhancerlike positive regulatory element in the
            interleukin-2 receptor alpha-chain gene
  JOURNAL   Mol. Cell. Biol. 10 (2), 850-853 (1990)
   PUBMED   2153927
```

50 annotated features, including 6 regulatory regions and 44 functional regulatory elements with their cognate transcription factors

20 references (originally 28…)

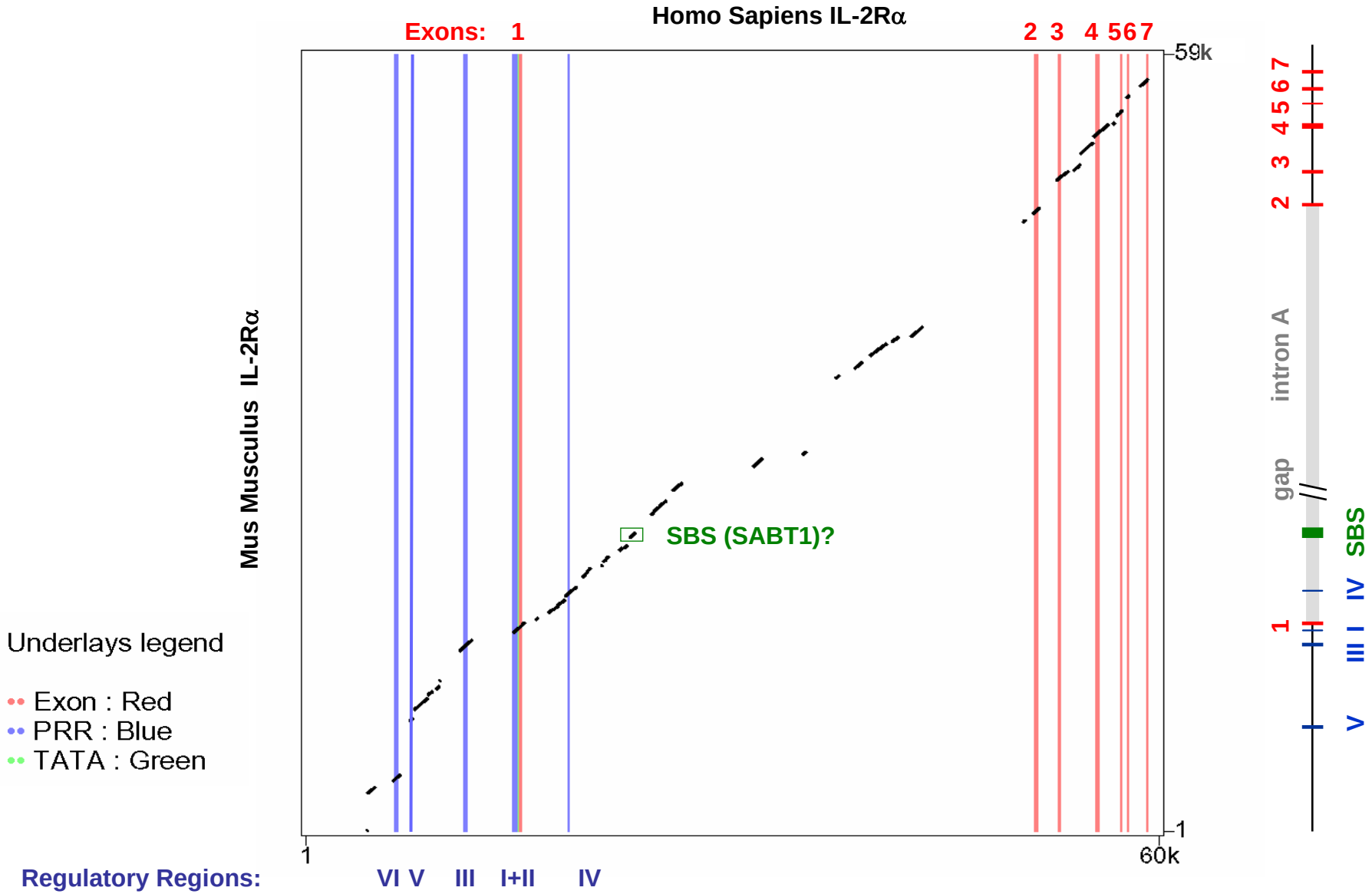
Phylogenetic footprinting: two species comparison

PipMaker

http://pipmaker.bx.psu.edu/pipmaker/

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). PipMaker: A Web Server for Aligning Two Genomic DNA Sequences. Genome Res. 10, 577-586.

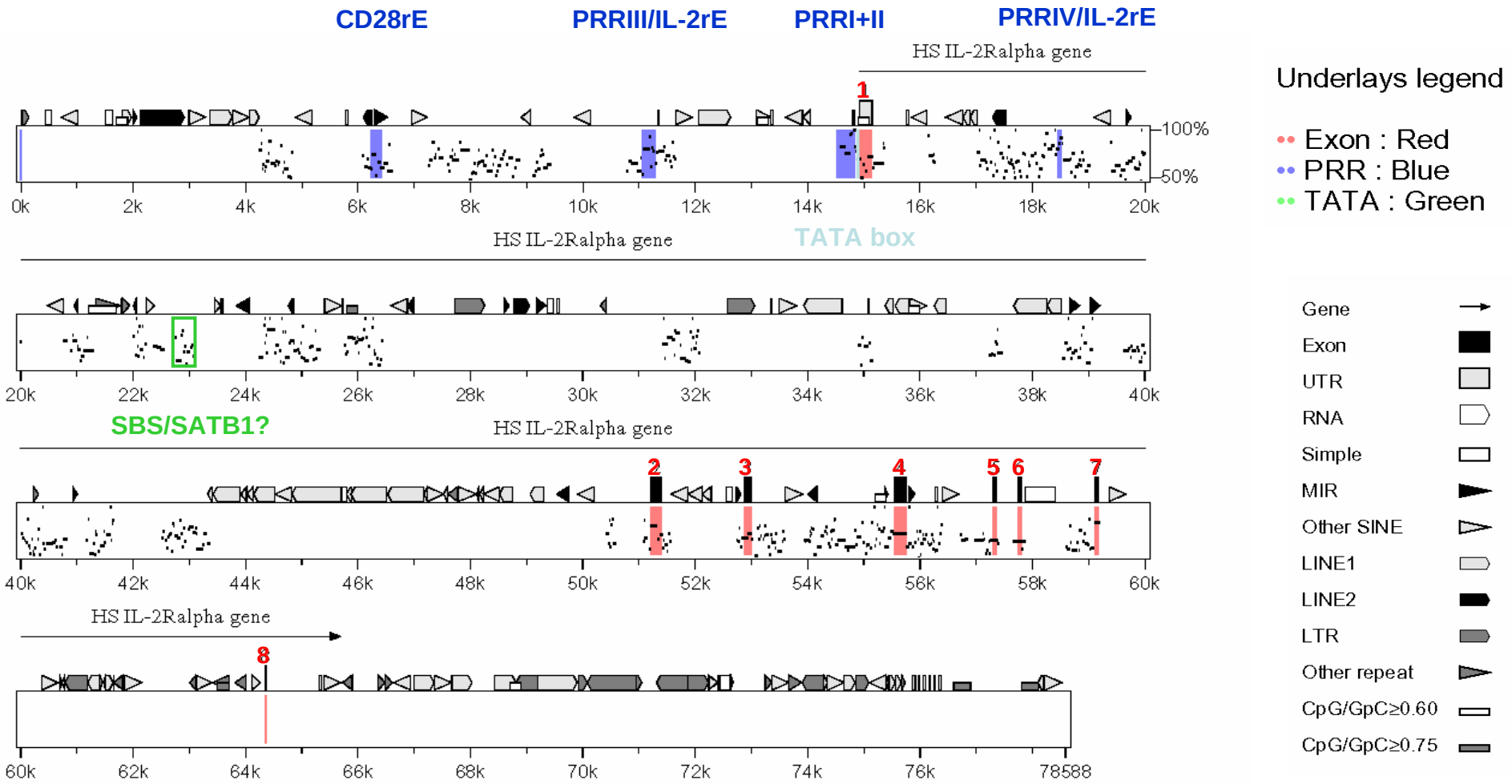
Homo Sapiens/Mus Musculus IL-2Rα locus dotplot comparison



Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). PipMaker: A Web Server for Aligning Two Genomic DNA Sequences. Genome Res. 10, 577-586.

update JI 17/10/05

Percent Identity Plot (PIP) of Homo Sapiens and Mus Musculus IL-2Rα locus



Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). PipMaker: A Web Server for Aligning Two Genomic DNA Sequences. Genome Res. 10, 577-586.

Homo Sapiens versus Mus Musculus PRRI/II conservation

Local identities (aligment n°6)

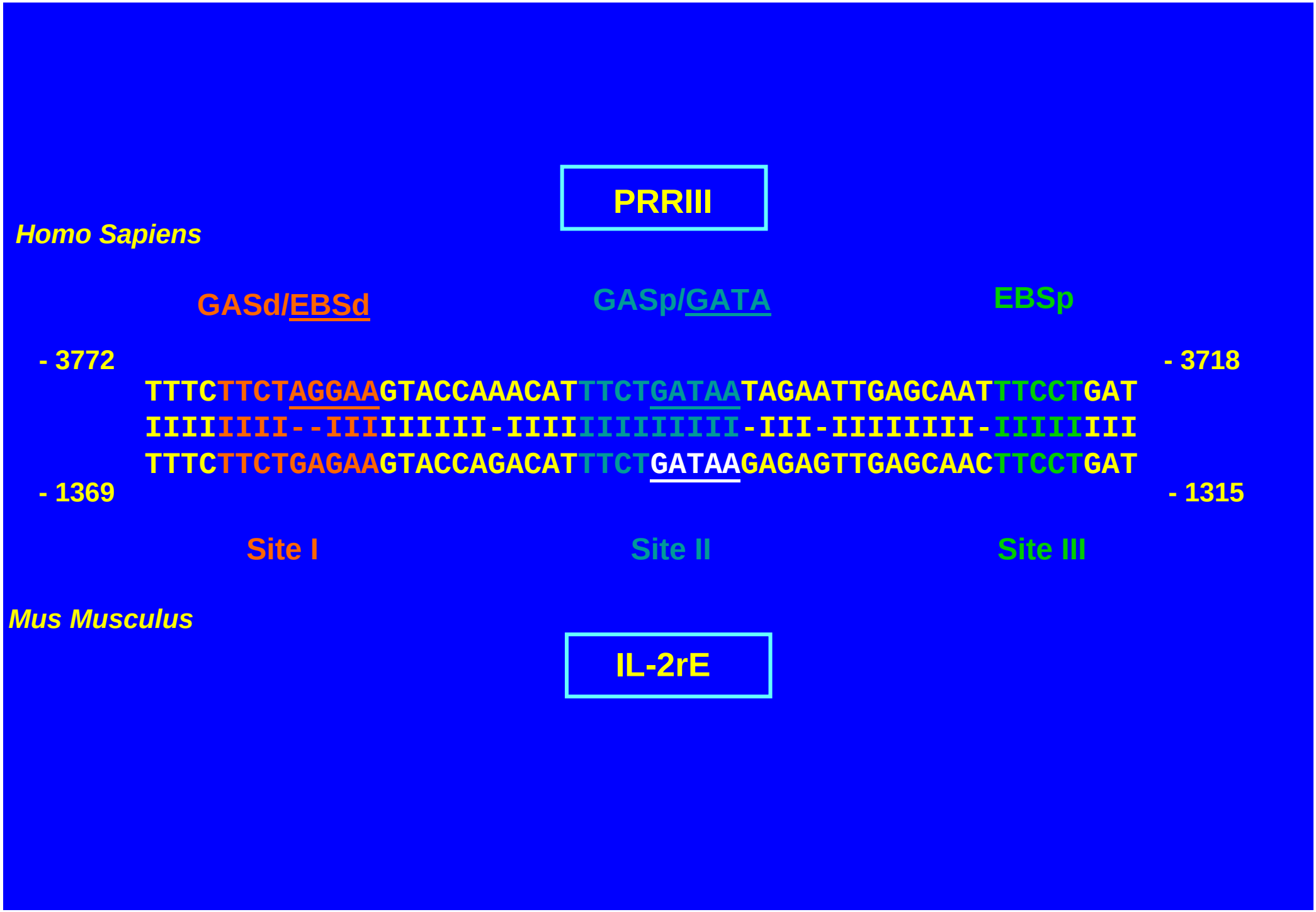
8724-8798 <--> 11629-11703 83% (75 nt)
8800-8886 <--> 11704-11790 76% (87 nt)
8887-8911 <--> 11792-11816 96% (25 nt)
8913-8946 <--> 11817-11850 97% (34 nt)
8954-8979 <--> 11851-11876 100% (26 nt)
8982-8999 <--> 11877-11894 83% (18 nt)
9003-9006 <--> 11895-11898 100% (4 nt)
9007-9015 <--> 11901-11909 78% (9 nt)
9017-9079 <--> 11910-11972 70% (63 nt)
9082-9128 <--> 11973-12019 51% (47 nt)
9129-9153 <--> 12021-12045 64% (25 nt)
9160-9202 <--> 12046-12088 58% (43 nt)
9203-9292 <--> 12090-12179 73% (90 nt)
9293-9313 <--> 12181-12201 62% (21 nt)

```
Seq 1 = ">HS_IL-2Ra_9.3KB genomic fragment from ENSG00000134460 5'->3'"
Seq 2 = ">MmIL-2Ra_12.5kb genomic fragment from phage Ch4Cl9 5'->3'"

       Local Alignment Number 6
       Similarity Score:  24894
       Match Percentage:  71 %
       Number of Matches:  425
       Number of Mismatches:  142
       Total Length of Gaps:  29
       Begins at (8724,11629) and Ends at (9313,12201)
     0     .    :    .    :    .    :    .    :    .    :
  8724 GAGGACTCAGCTTATGAAGTGCTGGGTGAGACCACTGCCAAGAAGTGCTT
       ||||||||||:|||::||:: ||::||||||||||||||||||||||||
 11629 GAGGACTCAGTTTACAAAACCCTAAGTGAGACCACTGCCAAGAAGTGCTT
    50     .    :    .    : κB site  :    .  SRE    .SP1/
  8774 GCTCACCCTACCTTCAACGGCAGGGGAATCTCCCTCTCCTTTTATGGGCG
       ||||||||: ||| | :||||||||-||||:||||:||||| ||::|||:
 11679 GCTCACCCCTCCTGCCGCGGCAGGG AATCCCCCTTTCCTTGTACAGGCA
   100 GC-box   :    .    :    .    :    .    :    .    :
  8824 TAGCTGAAGAAAGGATTCATAAATGAAGTTCAATCCTTCTCATCAACCCC
        |:|  ||:|||||||:||||||:||||::::||||||||||:||| | |
 11728 AAACACAAAAAAGGACTCATAAGTGAAGCCTGATCCTTCTCACCAAACAC
   150     .    :    .    :    .    :    .    :    .    :
  8874 AGCCCACACCTCC AGCAATTGAACTTGAAAAAAAAAACCTGGTTTGAAA
        ||||||||||||-||:||||||||||||||||||||||-||||||||||
 11778 TGCCCACACCTCCTAGTAATTGAACTTGAAAAAAAAAAC TGGTTTGAAA
   200     .    : HMGI(Y)      .HMGI(Y)  .    :    .EBS/:
  8923 AATTACCGCAAACTATATTGTCATCaaaaaaaaaaaaaaaaaaaaaaCACT
       ||||||||||||:||||||||||-------|||||||||||||||||||||
 11827 AATTACCGCAAACCATATTGTCAT       AAAAAAAAAAAAAAACACT
   250 Elf-1 HMGI(Y) .    :    .    :    .    :    .    :
  8973 TCCTATATTTGAGATGAGAGAAGAGAGTGCTAGG  CAGTTTCCTGGCTG
       ||||||||--|||||| | |||| ||||---|||||--||: |||||-||||
 11870 TCCTATA  TGAGATCACAGAACAGAG   TAGGCACAAGTTCCT GCTG
   300     .    :    .TATA box .    :    .    :    .+1  :
  9021 AACACGCCAGCCCAATACTTAAAGAGAGCAACTCCTGACTCCGATAGAGA
       |:|| ::|||||:|||:|||||| |||:|||||||||:|| : || || |
 11914 AGCAGATCAGCCTAATGCTTAAATAGAACAACTCCTGGCTGTCATTGACA
   350     .    :    .    :    .    :    .    :    .    :
  9071 CTGGATGGACCCACAAGGGTGACAGCCCAGGCGGACCGATCTTCCCATCC
       :||  |::|-- :| |:|:|||||| |:::| || |:|| |::::  ||:
 11964 TTGTCTAAA  AGCCAAGATGACAGACTGAGAGGCCTGAGCCCTTGTTCT
   400     .    :    .    :    .    :    .    :    .    :
  9121 CACATCCT CCGGCGCGATGCCAAAAAGAGGCTGACGGCAACTGGGCCTT
        :|||:||-||:| : ||||| : ||||:||:||------||: ::: |
 12012 GGCATTCTCCCAGGAAGATGCAGTAAAGGGGTTG      ACCCAATATA
   450     .    :    .    :    .    :    .    :    .    :
  9170 CTGCAGAGAAAGACCTCCGCTTCACTGCCCCGG CTGGTCCCAAGGGTCA
       ||||||||||   | |||: ||| || |:||: -|||:||||| | |:||
 12056 CTGCAGAGAATTTCATCCAGTTCCCTCCTCCATCCTGATCCCATGTGCCA
   500     .    :    .    :    .    :    .    :    .    :
  9219 GGAAGATGGATTCATACCTGCTGATGTGGGGACTGCTCACGTTCATCATG
       |||||||||| :||::|:|||||||| |||::| ||| |:|| |:||||:
 12106 GGAAGATGGAGCCACGCTTGCTGATGTTGGGGTTTCTCTCATTAACCATA
   550     .    :    .    :    .    :    .    :    .
  9269 GTGCCTGGCTGCCAGGCAGGTAAG GGCCTGTGGGTGCCCCCGGAA
       ||:||::|:||:|:|||||||||-|:|| | :: |||||: ||||
 12156 GTACCCAGTTGTCGGGCAGGTAAGAGACCAGGAACTGCCCTGGGAA
```

PRRI
[-276,-244]

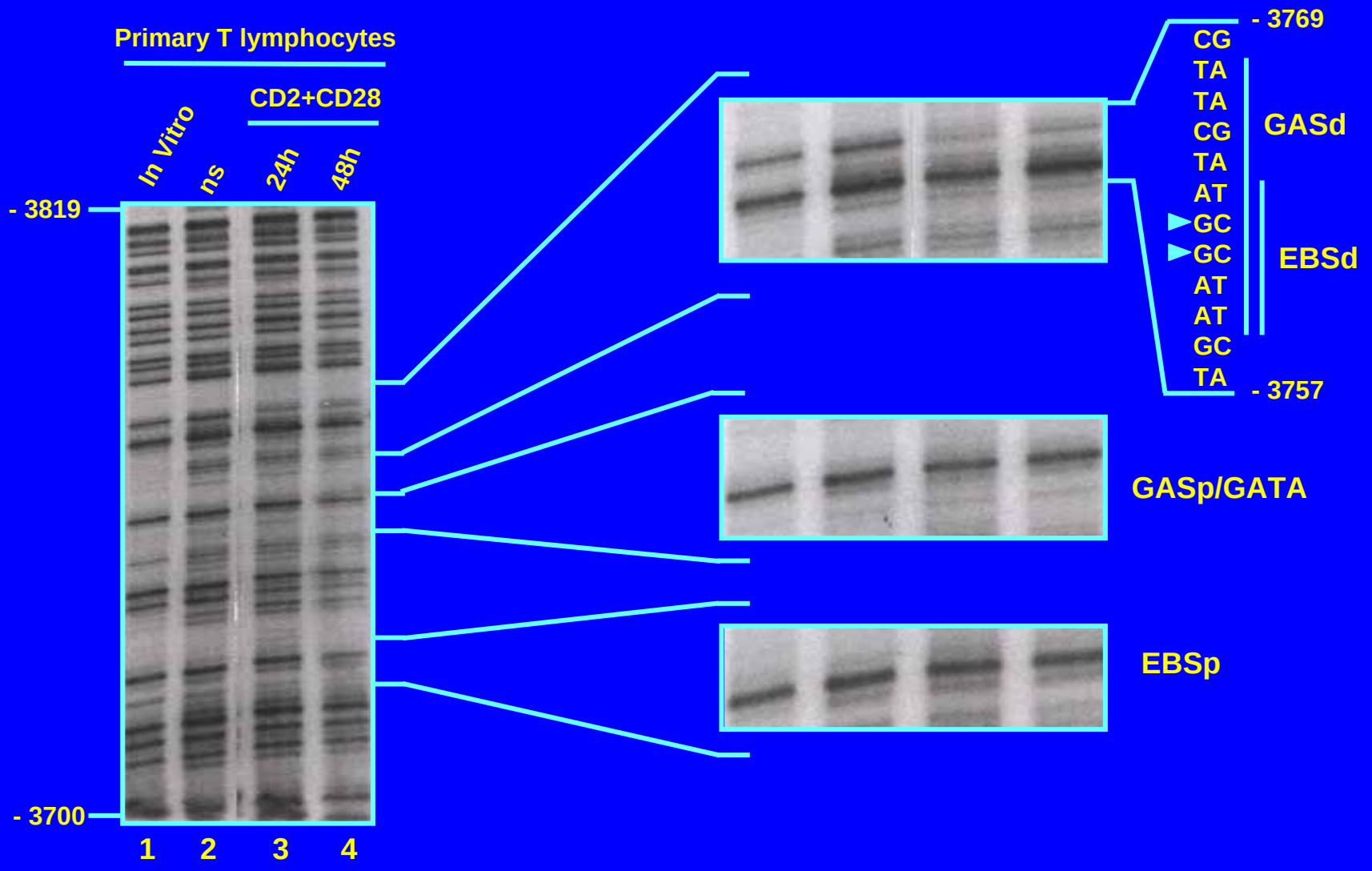
PRRII
[-137,-64]


PRRIII
Homo Sapiens
GASd/EBSd
GASp/GATA
EBSp
- 3772
- 3718
TTTCTTCTAGGAAGTACCAAACATTTCTGATAATAGAATTGAGCAATTTCCTGAT
IIIIIIII--IIIIIIIII-IIIIIIIIIIII-III-IIIIIIII-IIIIIIII
TTTCTTCTGAGAAGTACCAGACATTTCTGATAAGAGAGTTGAGCAACTTCCTGAT
- 1369
- 1315
Site I
Site II
Site III
Mus Musculus
IL-2rE

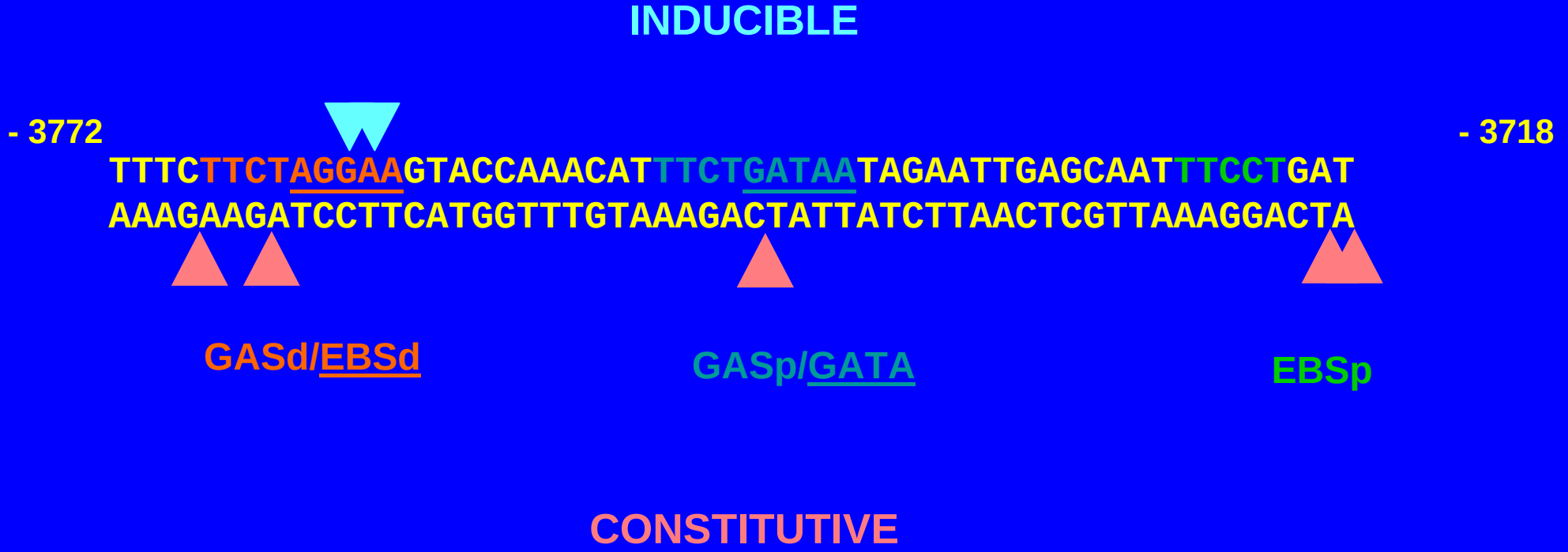
In Vivo Footprinting

- 1) Methylation of guanines (major groove) and to a lesser extent
  of adenines (minor groove) by DMS on living cells
  The level of methylation is affected by protein binding to DNA
- 2) Genomic DNA extraction
- 3) Cleavage of methylated residues by piperidine
- 4) LMP-PCR amplification of the region to be analyzed
  Last amplification cycles performed with a ^{32}P-labeled primer
- 5) Analysis of the PCR products on sequencing gel

In vivo modification of the GASd/EBSd motif occupancy in response to CD2+CD28 costimulation in purified human primary T cells

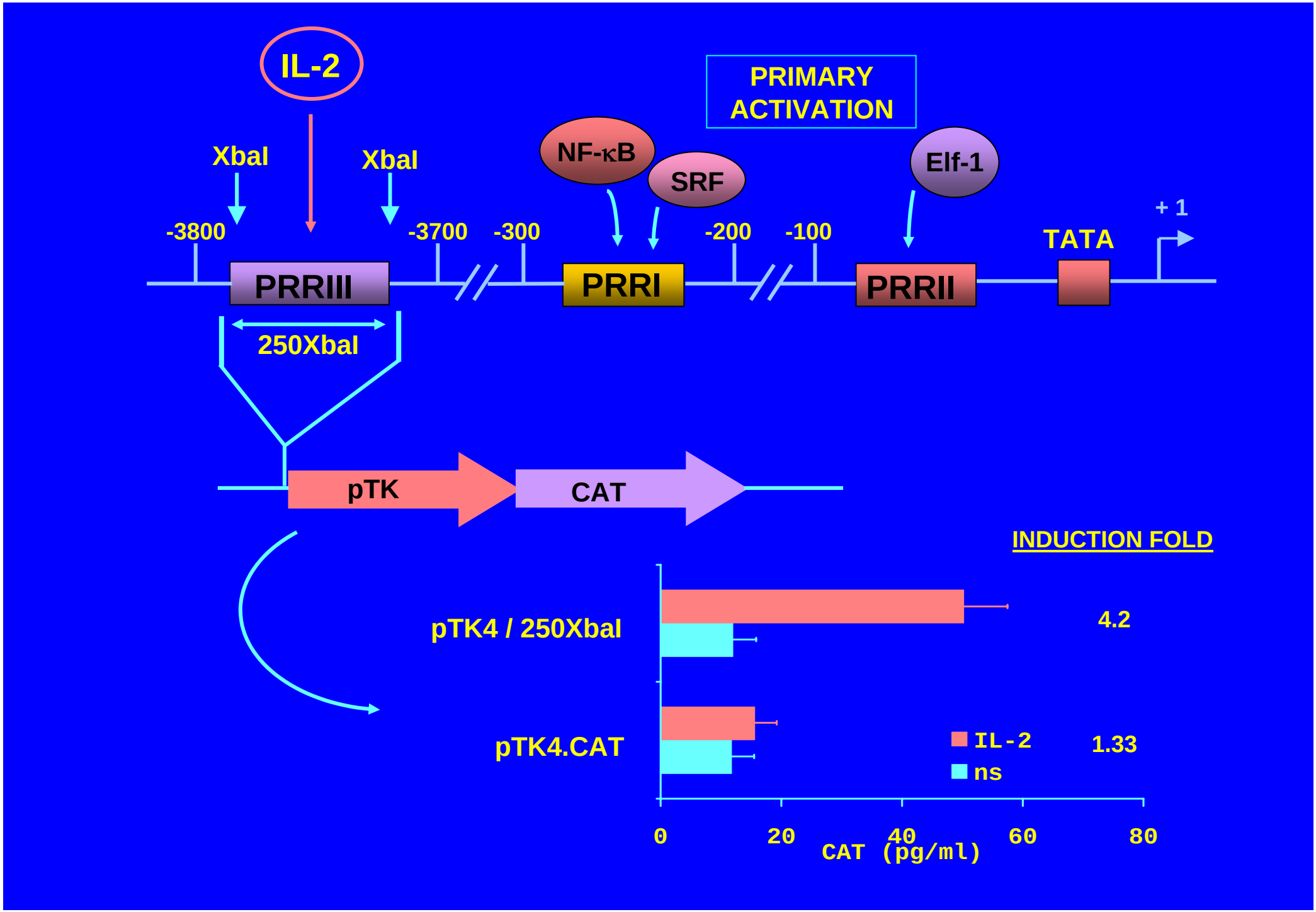


The GASd/EBSd motif is the only putative regulatory element within PRRIII modified *in vivo* in response to an IL-2-dependent induction in human T lymphocytes



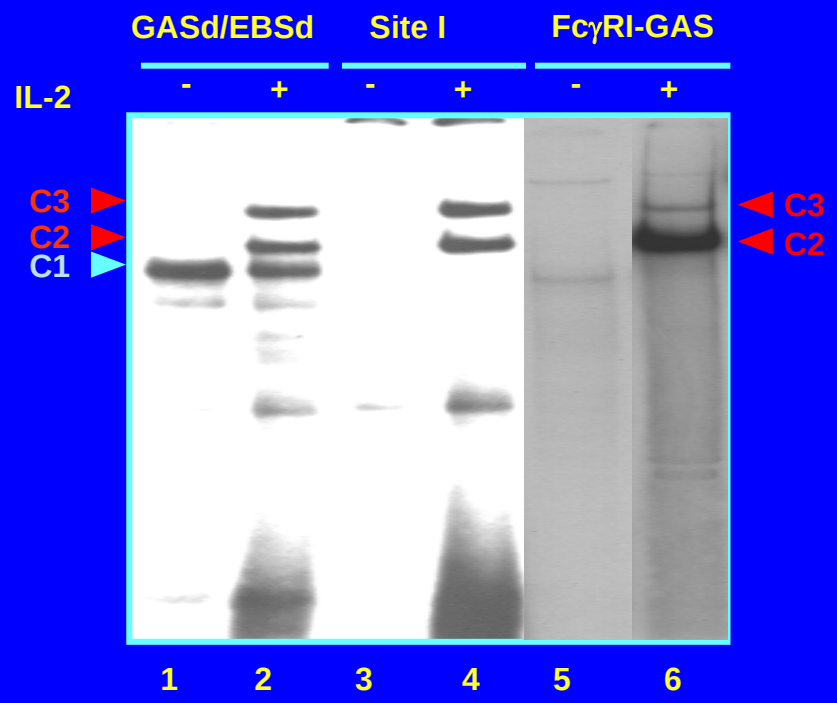
Lecine, P., Algarte, M., Rameil, P., Beadling, C., Bucher, P., Nabholz, M. and Imbert, J. Elf-1 and Stat5 bind to critical element in a new enhancer of the human interleukin-2 receptor alpha gene. Mol.Cell.Biol. 16: 6829-6840; 1996.

Gene reporter assay


IL-2
PRIMARY ACTIVATION
XbaI
XbaI
NF-κB
SRF
Elf-1
-3800
-3700
-300
-200
-100
+ 1
TATA
PRRIII
PRRI
PRRII
250XbaI
pTK
CAT
INDUCTION FOLD
pTK4 / 250XbaI
4.2
pTK4.CAT
IL-2
1.33
ns
0
20
40
60
80
CAT (pg/ml)

Electrophoretic Mobility Shift Assay (EMSA)

Inducibles complex C2 and C3 are GAS-specific
Constitutive complex C1 is EBS-specific



CD25/IL-2Rα GASd/EBSd EMSA probe:

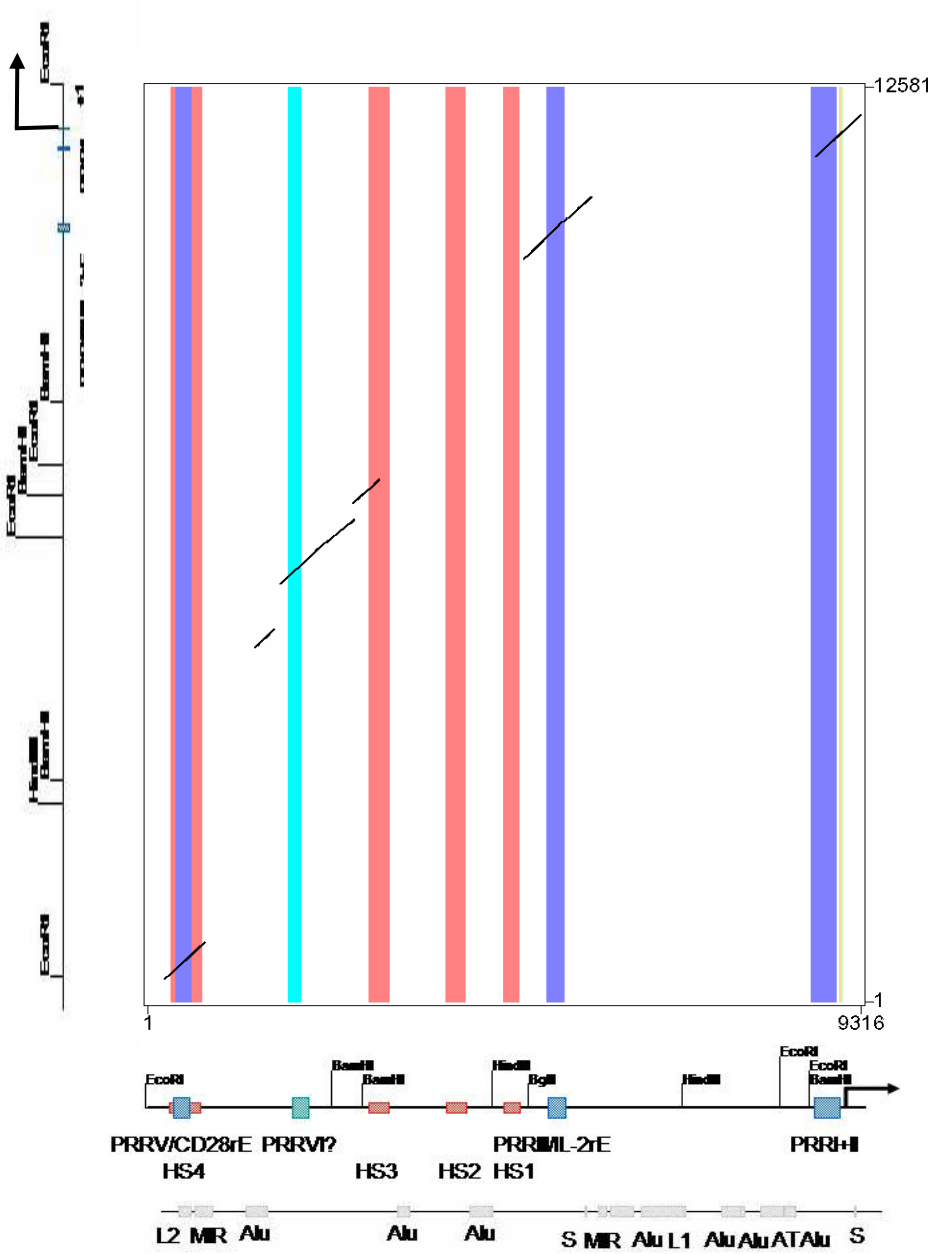
GASd
TTTCTTCTAGGAAGTACC
AAAGAAGATCCTTCATGG
EBSd

Mouse site I

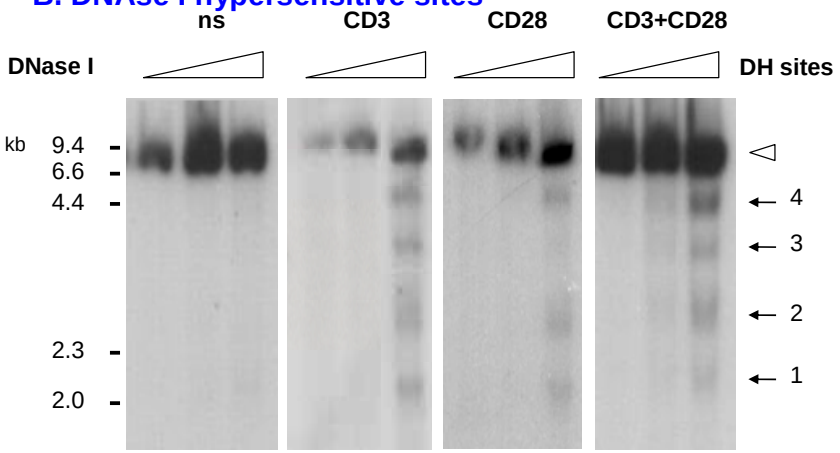
TTTCTTCTGAGAAGTACC
AAAGAAGACTCTTCATGG

Phylogenetic footprinting and DNAse I hypersensitive site mapping

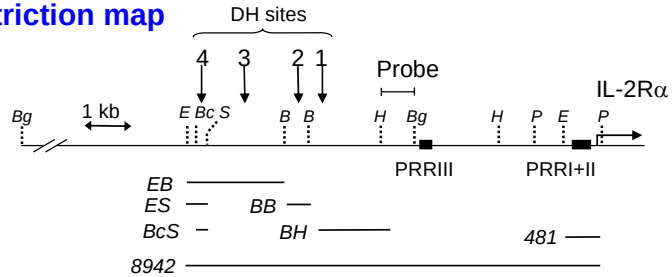
A. *Homo Sapiens/Mus musculus* CD25/IL-2Rα gene



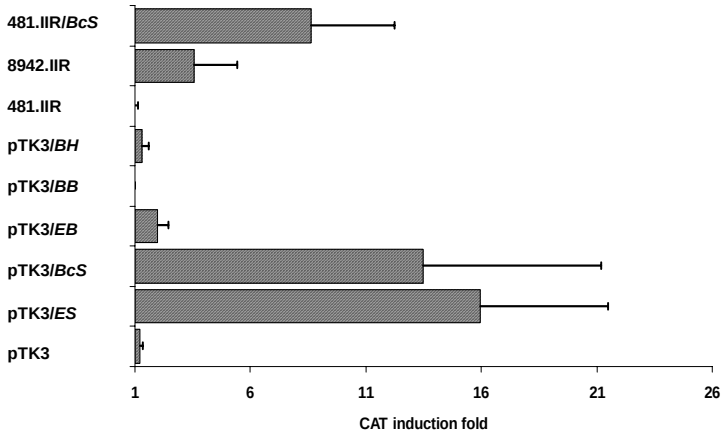
B. DNAse I hypersensitive sites



C. Restriction map

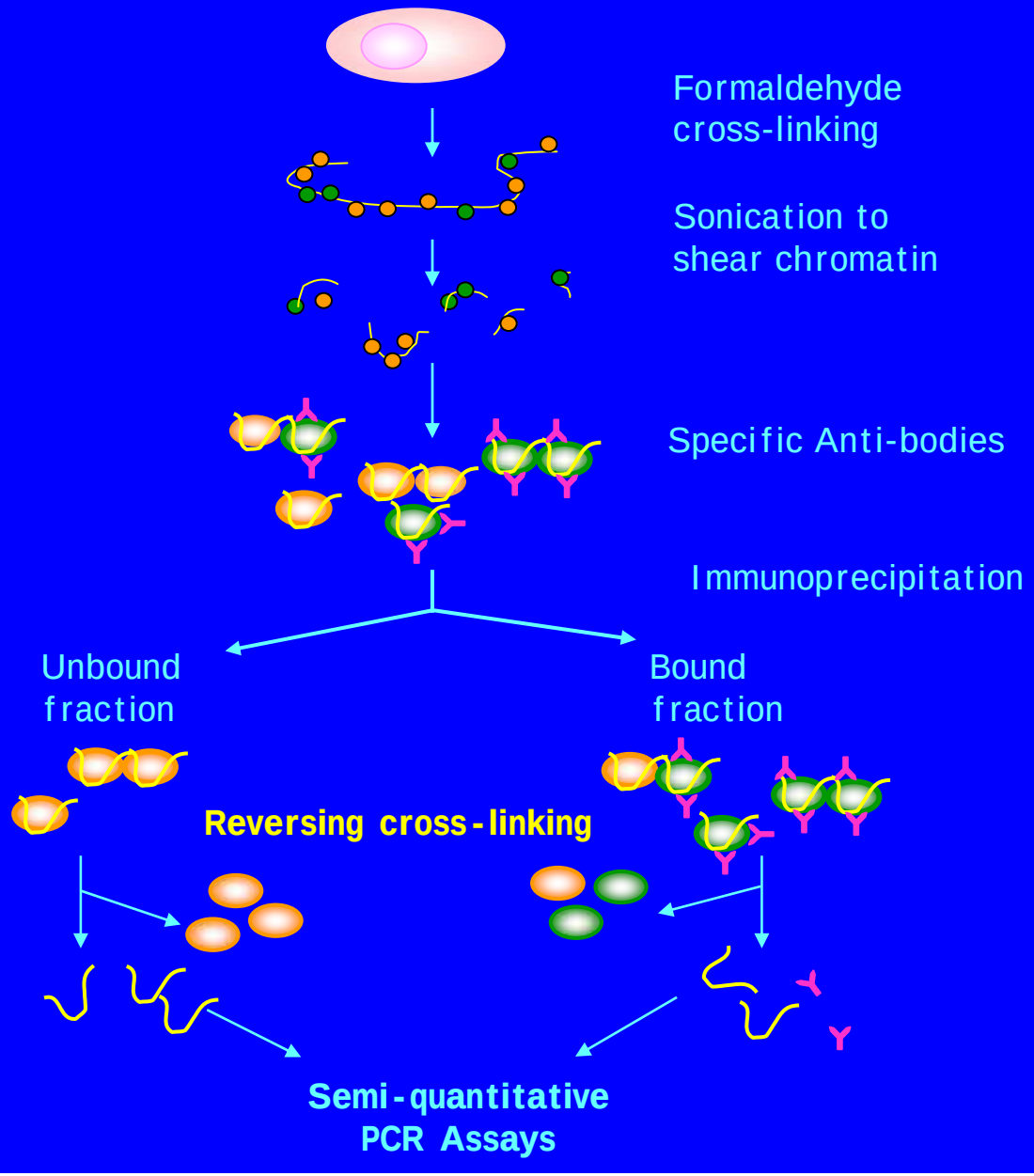


C. Gene reporter assays



Chromatin Immunoprecipitation
ChIP Assay

Chromatin immunoprecipitation (ChIP)

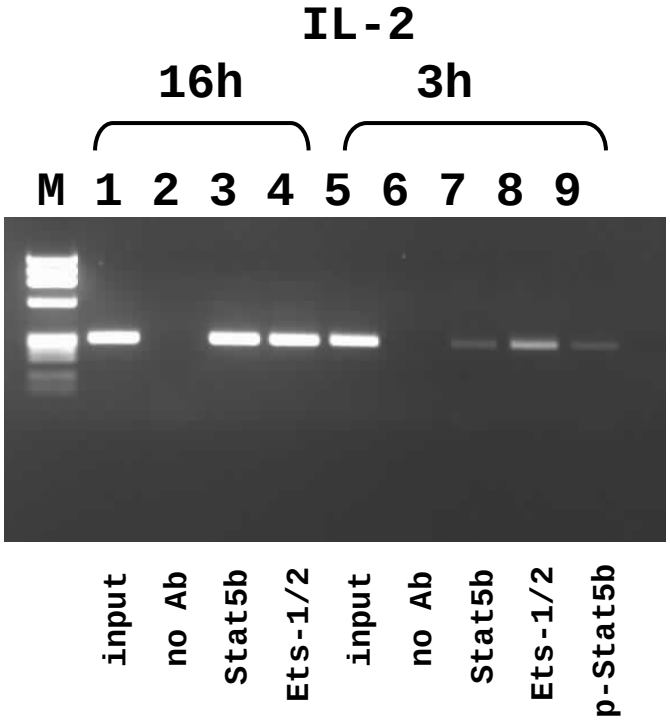


Stat5b and Ets-1/2 bind to IL-2Rα within human IL-2rE *in vivo*

PRRIII ChIP primer design

```
1 TTCTGCCCTTAGCTTCTACCCCTCTCTACTTCTGGTTAACTATGGACCACACTCTGCTTC
    BZ3-1                                              BZ3-2
61  CTCAGGAACCACCTACCAAGGCCGTATCCATCCTTCAAGGACAATACGTGGGCCTTTCCT
    >>>>>>>>>>>>>>>>>>>>                              >>>>>>>>>>>>>>>>>
121 GATCACATCAGCTCAACAACTTTTCCCTCCTACATTTCAATTGCTCTTCTTACCATAATC
    >>>
181 ATTAGTATTCACCCCACTGTACGTCTAGAAAGAAAGTGGTCTTAAACCTAAGGGAAGGCA
241 GTCTAGGTCAGAAATTTGTTGTCCGCTGTTCTGAGCAGTTTCTTCTAGGAAGTACCAAAC
                                                       BZ3-3
301 ATTTCTGATAATAGAATTGAGCAATTTCCTGATGAAGTGAGACTCAGCTTGCACTGTTGA
                                                      <<<<<<<<<<<<<<<<<<<
361 CCGGCTGTCCTGGATGAACCTAGTTACTTTTAACCAAATGTTCCTTTCTTGAACTTGTTC
    <<
421 CTTTCTTGAACTTAATCTATC
```

OLIGO	start	len	tm	gc%	3' seq
BZ3-1	62	20	59.96	55.00	TCAGGAACCACCTACCAAGG
BZ3-3	362	20	59.62	55.00	GGTCAACAGTGCAAGCTGAG
BZ3-2	104	20	60.71	50.00	ATACGTGGGCCTTTCCTGAT



Cela ne marche pas toujours…

A Well-Conserved but not Functional Candidate for a CD28 Responsive Enhancer

Comparison of two nucleotidic sequences:

Sequence 1 : cons4HS

Sequence 2 : cons4mus

```
 resetting to DNA matrix
 LALIGN finds the best local alignments between two sequences
 version 2.0u4 Feb. 1996
Please cite:
 X. Huang and W. Miller (1991) Adv. Appl. Math. 12:373-381

 Comparison of:
(A) cons4HS
(B) cons4mus
 using matrix file: DNA, gap penalties: -16/-4

  79.6% identity in 201 nt overlap; score:  612
```

```
                    NF-ATp
               10        20        30        40        50        60
cons4H CATAGTGGATTTTGGTTTTCCACGGGACCCCTGTGCCCTTGTCTAGTAGAATCTGGTGGA
       :::::::::: :::::::::: ::::::         ::::::::: :: :: :::::
cons4m CATAGTGGATTCTGGTTTTCCACAGGACCC---------TGTCTAGTAAAACCTAGTGGA
               10        20        30                  40        50
                    NF-κB/CD28RC   CREB         Ets
               70        80        90       100       110       120
cons4H AATTACAAACTGCAGAAATTCAACTCAGTGCCGCAATAACAGGATGCACCTGTAGATTTC
       :::::::: ::: ::::::::: :   ::::: :::::::::::::::::::::::::::
cons4m AATTACAAGCTG-AGAAATTCAGCCTTGTGCCACAATAACAGGATGCACCTGTAGATTTC
               60        70        80        90       100       110
       GAS
               130       140       150       160       170       180
cons4H GTAGAATTAGCAGCAGCATTCTTTCAATACCAGTTTGAGAGAAATAACCCTGTTTGCATA
         ::::::::: :: :  :::: : ::  ::: :::::::::::: :: ::::::::  :
cons4m ACAGAATTAGCTGCTGTCTTCTCTTAACGCCAATTTGAGAGAAAGAAGCCTGTTTGTCTG
               120       130       140       150       160       170

               190       200
cons4H GTGCCAACTGGGGCAGAATCT
        ::::::   ::::::::::::
cons4m CTGCCAAACAGGGCAGAATCT
               180       190
```

Et il y a des surprises…

The crucial CRE/TRE motif within CD25/IL-2Rα PRRV/CD28rE is contained in a SINE/MIR repeat and is not conserved between *Homo Sapiens* and *Mus Musculus*

Local identities (aligment n°1)

211-245 <--> 312-346 74% (35 nt)
261-312 <--> 347-398 62% (52 nt)
313-322 <--> 400-409 70% (10 nt)
323-371 <--> 414-462 61% (49 nt)
374-388 <--> 463-477 60% (15 nt)
389-444 <--> 480-535 75% (56 nt)
447-482 <--> 536-571 53% (36 nt)
491-542 <--> 572-623 67% (52 nt)
544-602 <--> 624-682 61% (59 nt)
603-664 <--> 684-745 66% (62 nt)
666-686 <--> 746-766 57% (21 nt)
690-709 <--> 767-786 85% (20 nt)
712-744 <--> 787-819 58% (33 nt)

```
Seq 1 = ">HS_IL-2Ra_9.3KB genomic fragment from ENSG00000134460 5'->3'"
Seq 2 = ">MmIL-2Ra_12.5kb genomic fragment from phage Ch4Cl9 5'->3'"

    Local Alignment Number 1
    Similarity Score:  12780
    Match Percentage:  59 %
    Number of Matches:  325
    Number of Mismatches:  175
    Total Length of Gaps:  42
    Begins at (211,312) and Ends at (744,819)

  0     .    :    .    :    .    :    .    :    .    :
211 CTCTAAAAAGGTTCTGCATACAGtcattcattcaatgcttaacgactgag
    ||:| ||: |||||| ||:| |:|||||| ||||||--------------
312 CTTTTAAGTGGTTCTCCACAGAATCATTAATTCAA

 50     .    :    .    :    .    :    .    :    .    :
261 cattaattccatgctaagtactgaactcagcactaggaataagaaggcga
    ||   ||: ::|||||:|| ||: |||||| |||::|| |||:| ||::|
347 CAGGTATCATGTGCTAGGTCCTATACTCAGAACTGAGACTAAAATGGTAA

100     .    :    .    :    .    :    .    :    .    :
311 cc tagaggcata    tcctctctctaaagatgcatagagcctcattgg
    | -||||:| | |----|:||:|:|||||||:|::||:||||| : ::|||
397 CAATAGAAGAAGACAATTTCTTTTTCTAAAAACACACAGAGCACAGCTGG

150   BclI   :    .    :    .    :    .    :    .    :
356 aatgatcagccgtgtctccagagagctacaagg  cagTTTTCAATTGGT
    |||  |::| |:||| --||||  | |: |:||--|||||||  |:|||||
447 AATTCTTGGGCATGTA  CAGATTGATGGAGGGTACAGTTTTAAGTTGGT

200     .    :    .    :    .    :    .    :    .    :
404 AAATGCCCTGAGAGTGATGGGCTTGTGGCATGTGTAAgggttagacagac
    |||||:|:||||||: |||:  | ||:|| ||| |:|||||--||::: :
495 AAATGTCTTGAGAGCTATGACATGGTAGCCTGTTTGAGGGT  GATGATT

250     .    :    .    : CRE/TRE:    .    :    .    :
454 ctgggacctagacatgacaccactcctgacgaattatgtgagtgtgggtg
     |:|:  |||  | | |||:| ||||||:--------|||||  ||||||::
543 GTAGACACTACTCCTCACATCTCTCCTGG        GTGAGAGTGGGCA

300     .    :    .    :    .    :    .    :    .    :
504 tttcacaaccacaatgagatgcaatgcctgcacttgtaacatggaaatag
    |||||:|||::|||:|:|:|::||||:|| ||:: |:||-|||||||| |
585 TTTCATAATTACAGTAAAATATAATGTCTTCATCAGCAA ATGGAAATCG

350     .SphI    .    :    .    :    .    :    .    :
554 tgatggcatgccggccccgccagattgctgtgagaagtcagcggcagag
    |||:::::||:|:||:  :|  |:|||||||| |||:||||::: | |:-
634 TGACAATGTGTCAGCTAGACATGGTTGCTGTGTGAAATCAGTAAGATAAC

400     .    :    .    :    .    :    .    :    .    :
603 acatgcaacattctcagcacagtgcttgccatgtagtaagggcctagtca
    ||||  || || : ::|||||| ||:|||:||||||:||| :::|| |||
684 ACATTAAAAATGTATGGCACAGAGCCTGCTATGTAATAAACATTTATTCA

450     .    :    .    :    .    :    .    :    .    :
653 gtgctagTGATTCCTTTCAATATTCCTAAGATGCAGATAAGGGAACAGCC
    :|| |||||:||- |||| :||:|   :| ||||---||||| ||||||:
734 ATGGTAGTGGTT ATTTCCGTACTAGAGATATGC   TAAGGCAACAGCT

500     .    :    .    :    .    :    .    :
703 CAGAGGAGGGGGAGCACTTCCAGAGGGAGGGATGCGGTGAGA
    |||||:|--|:||::||:||:| | :: ||||| :|:|||:|
780 CAGAGAA  GAGAATACCTCTACATAATGGGATCTGATGAAA
```

HS4 (-8.5Kb)
CD28rE (PRRV)
[-8689,-8484]

BclI-SphI fgt

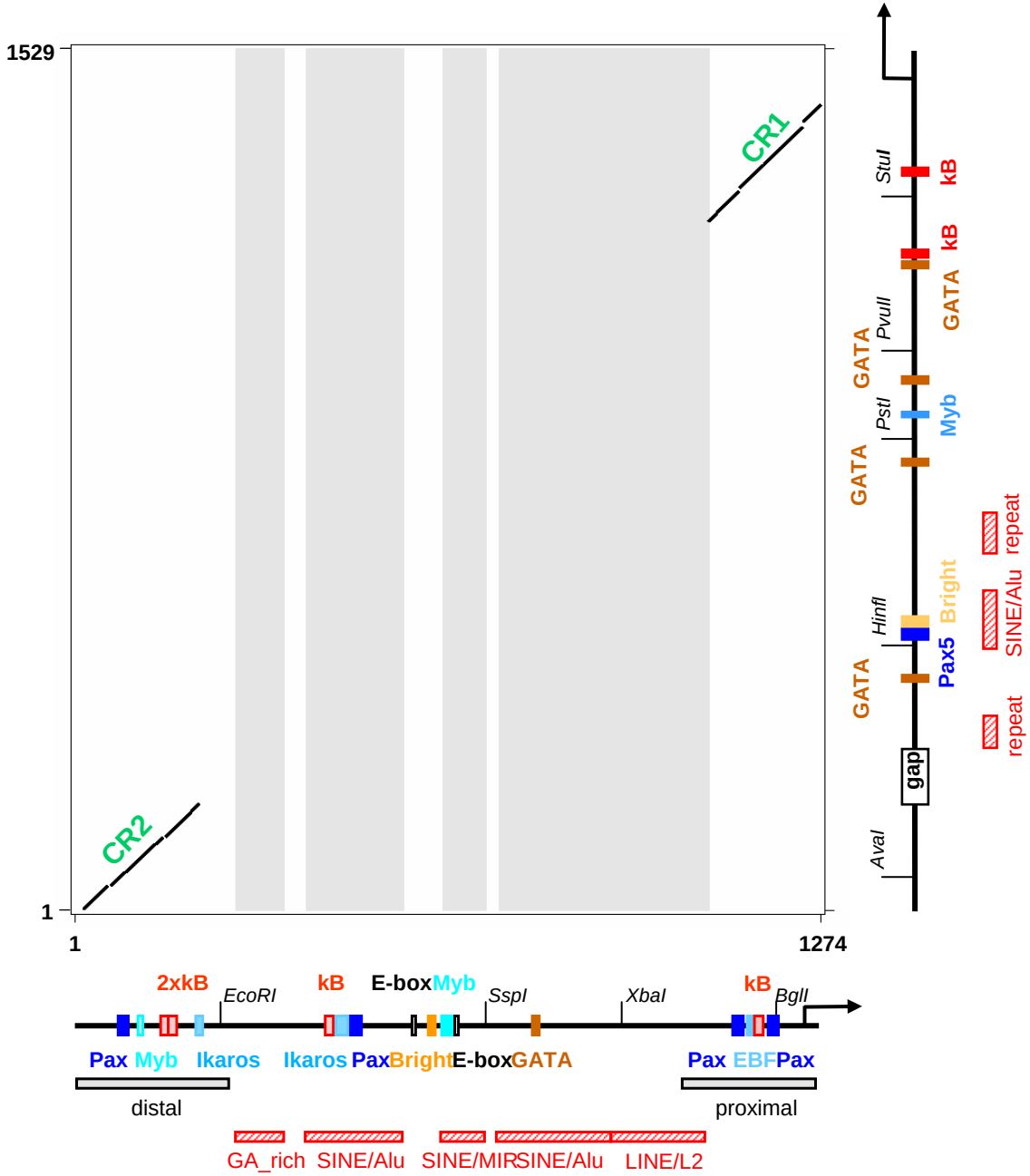
underlined:
SINE/MIR repeat

Parfois on peut aller très vite de l’ordinateur à la paillasse…

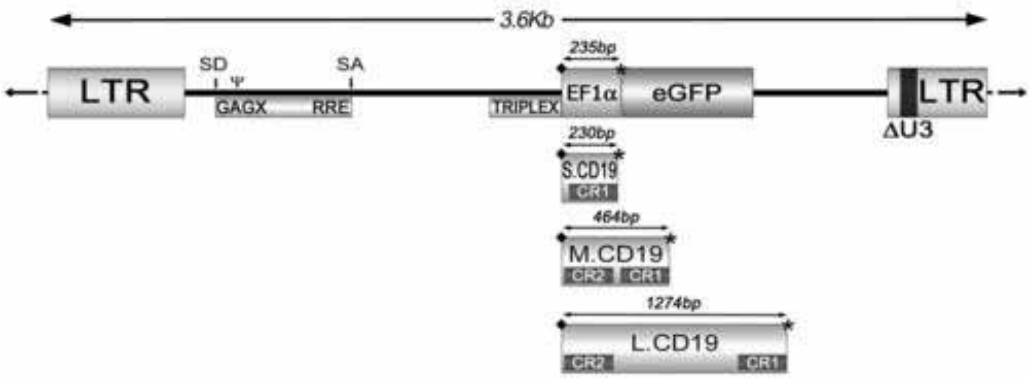
Selection of CD19 B-cell specific regulatory sequence and design of CD19-GFP lentiviral vector

Moreau, T., F. Bardin, J. Imbert, C. Chabannon, and C. Tonnelle. 2004. Restriction of transgene expression to the B-lymphoid progeny of human lentivirally transduced CD34+ cells. Mol.Ther,. 2004, in press.

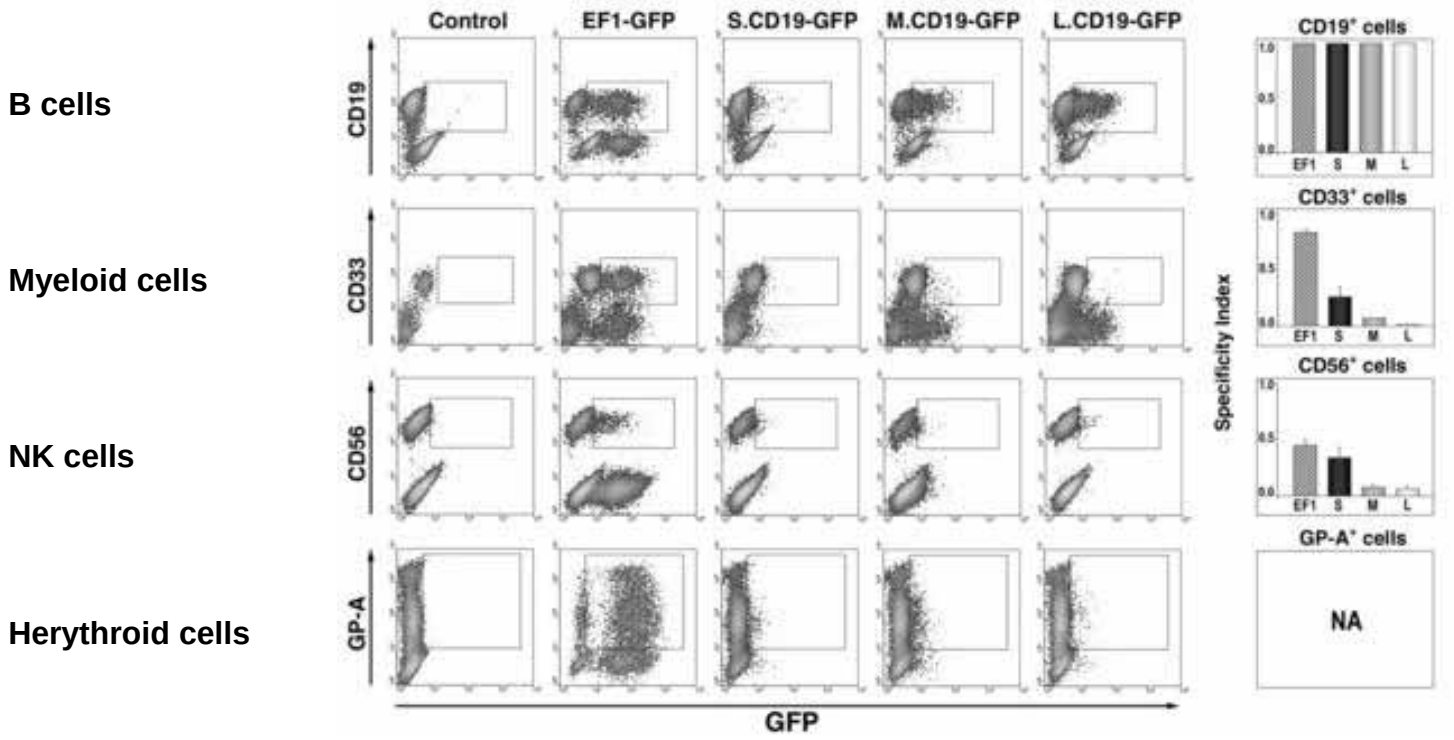
Computational analysis of homologies between Human and Mouse CD19 gene 5' regions



A. Recombinant CD19-GFP lentiviral vectors



B. GFP expression in the progeny of transduced CD34⁺ progenitor cells differentiated in vitro

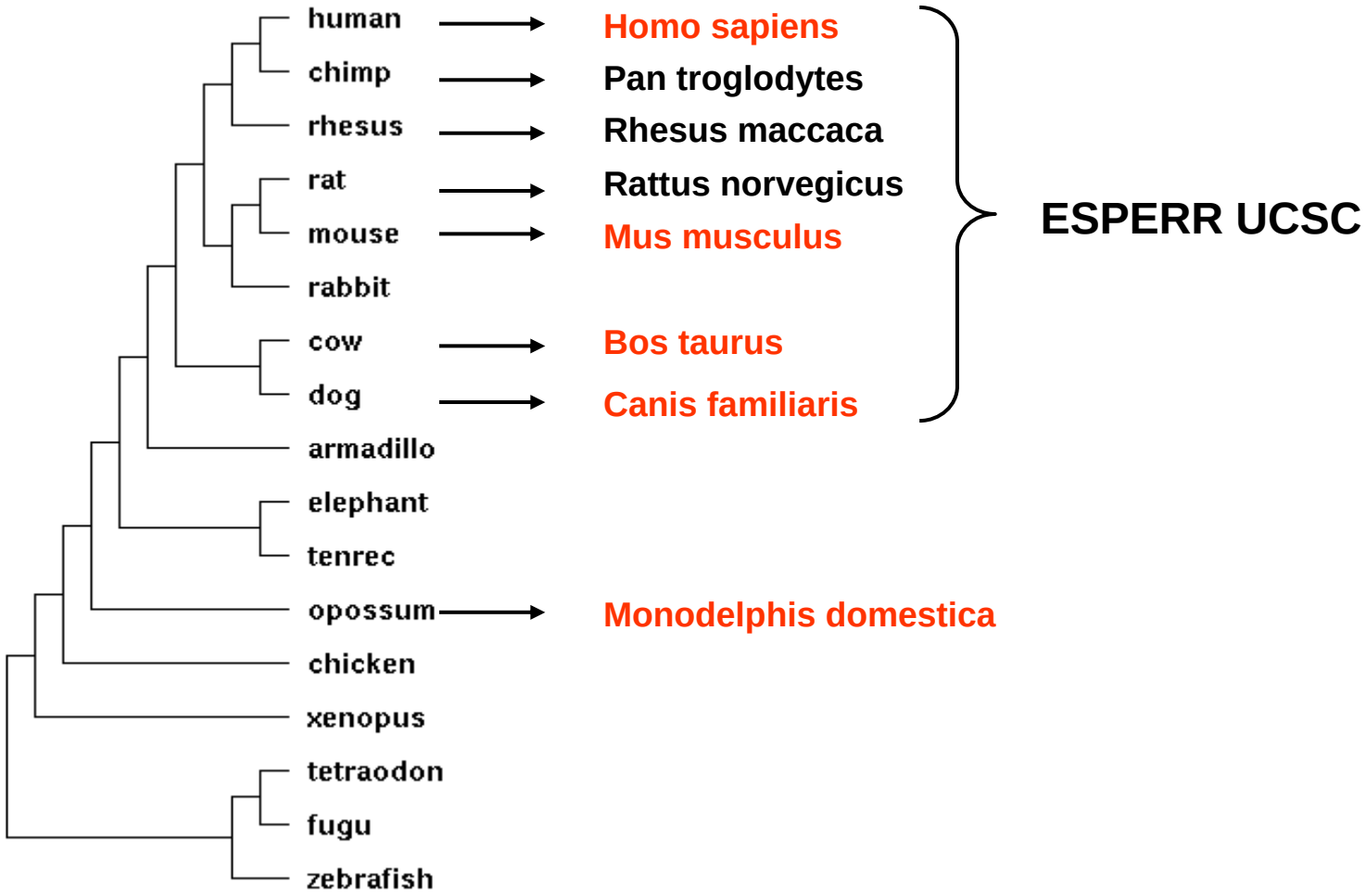


From Moreau et al., Mol.Ther, 2004.

De nouveaux outils…

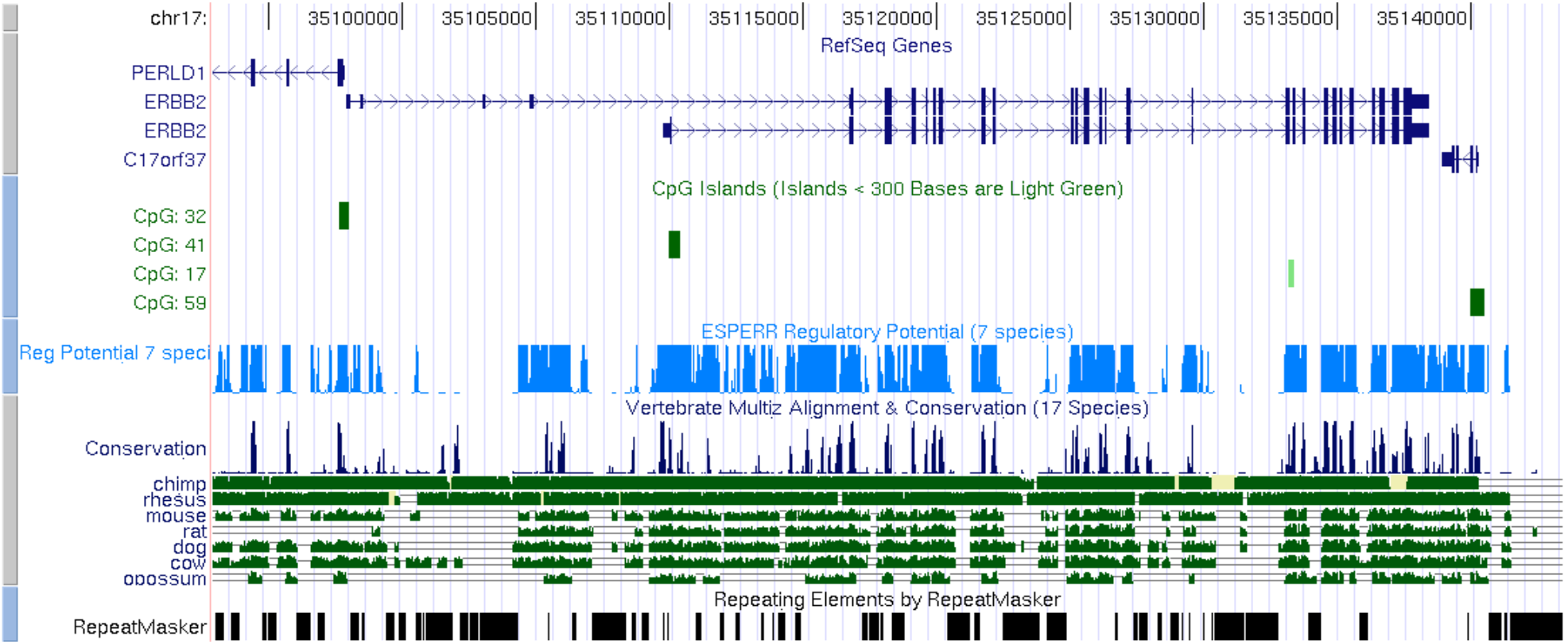
Phylogenetic footprinting: multi-species comparison

Species choice for phylogenetic footprinting analysis



Homo sapiens ERBB2 locus

chr17:35,092,919-35,143,441



UCSC Genome Bioinformatics: http://genome.ucsc.edu/

ESPERR Regulatory Potential (7 species): alignments of human, chimpanzee (panTro2), macaque (rheMac2), mouse (mm8), rat (rn4), dog (canFam2), and cow (bosTau2).

VISTA Tools: mVISTA and rVISTA

http://genome.lbl.gov/vista/

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004 Jul 1;32 (Web Server issue):W273-9

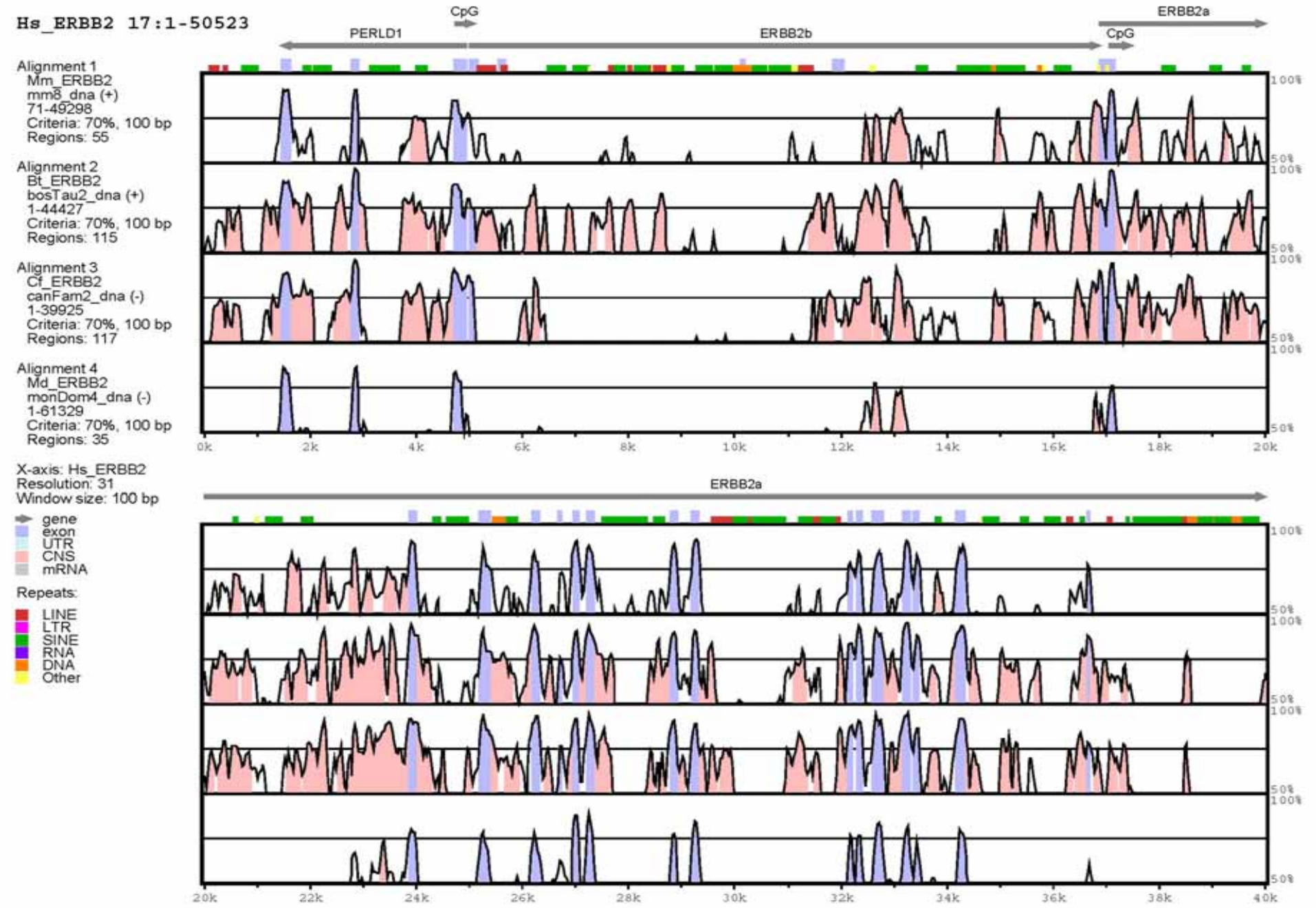
mVISTA

Limits:

- **sequence numbers: 2-100**
- **size: ?**

Mayor C., Brudno M., Schwartz J. R., Poliakov A., Rubin E. M., Frazer K. A., Pachter L. S. and Dubchak I. (2000) VISTA: Visualizing Global DNA Sequence Alignments of Arbitrary Length. Bioinformatics, 16:1046.

ERBB2 locus

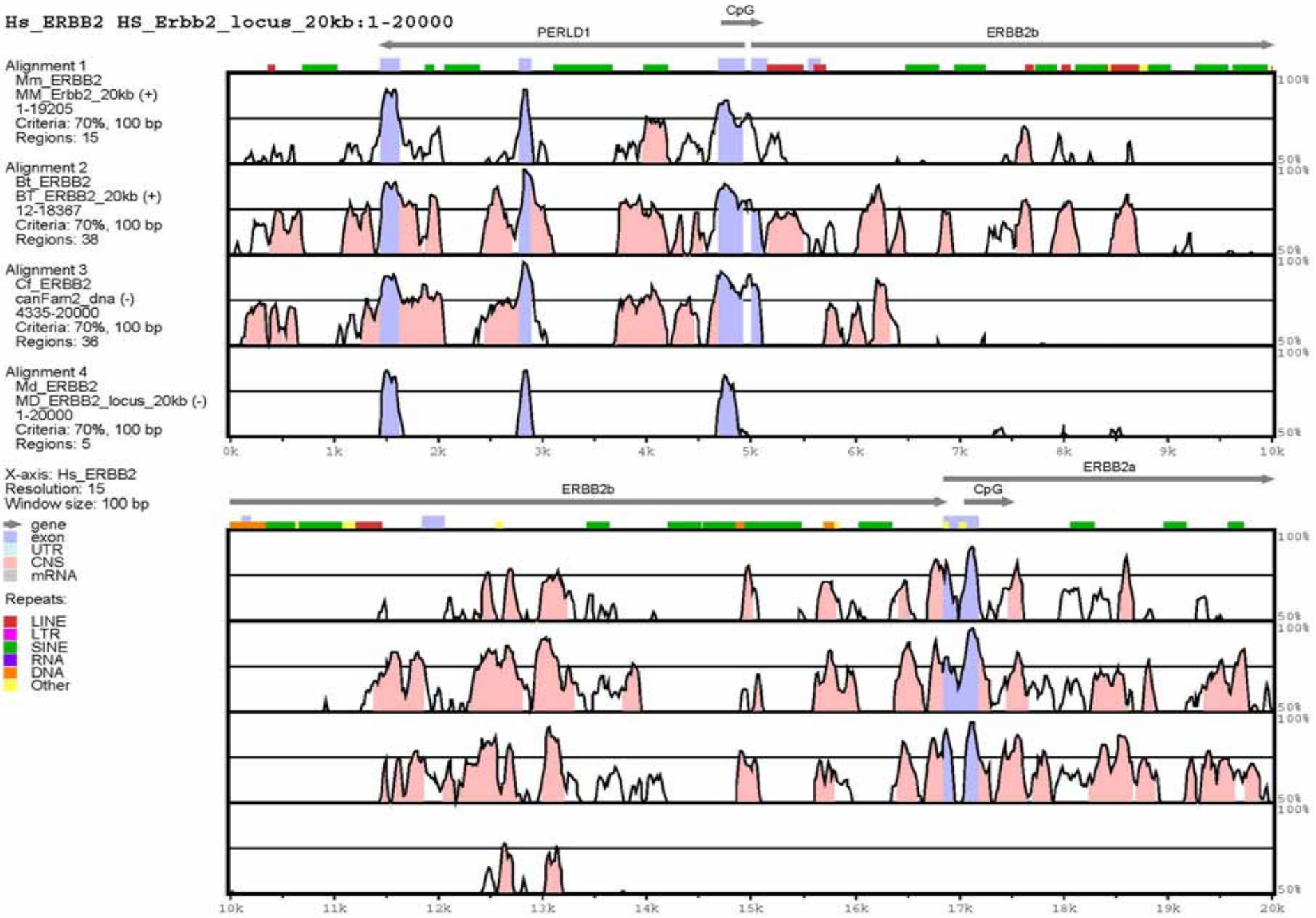

Hs_ERBB2 17:1-50523
CpG
PERLD1
ERBB2b
CpG
ERBB2a
Alignment 1
Mm_ERBB2
mm8_dna (+)
71-49298
Criteria: 70%, 100 bp
Regions: 55
Alignment 2
Bt_ERBB2
bosTau2_dna (+)
1-44427
Criteria: 70%, 100 bp
Regions: 115
Alignment 3
Cf_ERBB2
canFam2_dna (-)
1-39925
Criteria: 70%, 100 bp
Regions: 117
Alignment 4
Md_ERBB2
monDom4_dna (-)
1-61329
Criteria: 70%, 100 bp
Regions: 35
X-axis: Hs_ERBB2
Resolution: 31
Window size: 100 bp
gene
exon
UTR
CNS
mRNA
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
100%
50%
0k
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
ERBB2a
20k
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k

rVISTA

Limits:

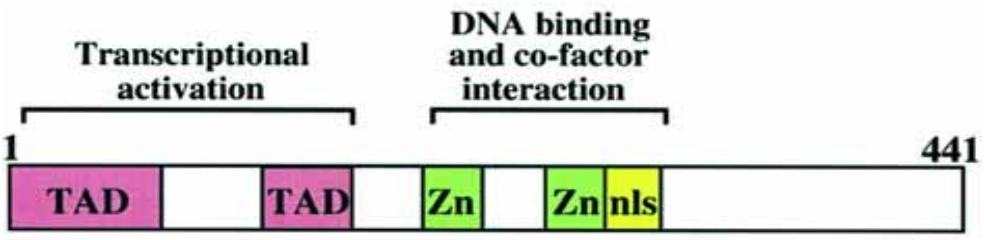
- **sequence numbers: 2-100**
- **size: 20 Kb per sequence for a total of 60 Kb**

Loots, G., Ovcharenko,I., Pachter,L., Dubchak,I., Rubin, E. rVISTA for comparative sequence-based discovery of functional transcription factor binding sites. (2002) Genome. Res. 12:832-839

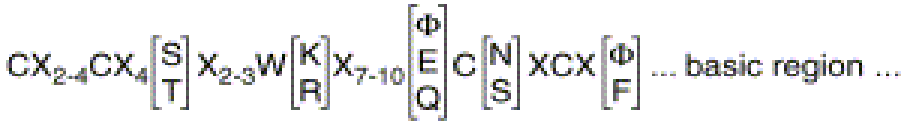

Hs_ERBB2 HS_Erbb2_locus_20kb:1-20000
CpG
PERLD1
ERBB2b
Alignment 1
Mm_ERBB2
MM_Erbb2_20kb (+)
1-19205
Criteria: 70%, 100 bp
Regions: 15
Alignment 2
Bt_ERBB2
BT_ERBB2_20kb (+)
12-18367
Criteria: 70%, 100 bp
Regions: 38
Alignment 3
Cf_ERBB2
canFam2_dna (-)
4335-20000
Criteria: 70%, 100 bp
Regions: 36
Alignment 4
Md_ERBB2
MD_ERBB2_locus_20kb (-)
1-20000
Criteria: 70%, 100 bp
Regions: 5
X-axis: Hs_ERBB2
Resolution: 15
Window size: 100 bp
gene
exon
UTR
CNS
mRNA
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
100%
50%
0k
1k
2k
3k
4k
5k
6k
7k
8k
9k
10k
ERBB2a
ERBB2b
CpG
10k
11k
12k
13k
14k
15k
16k
17k
18k
19k
20k

GATA-4: a member of the GATA family (GATA-1/6)

A. GATA-4

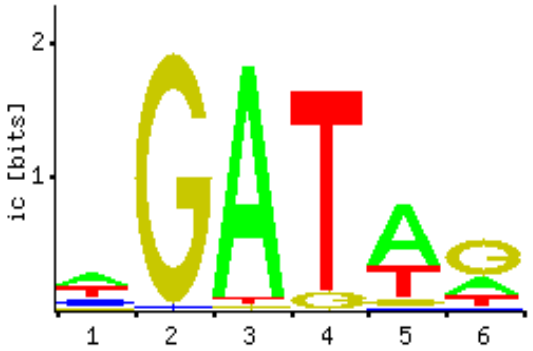


B. Consensus DNA binding domain of the GATA factor family



Current Opinion in Genetics & Development

C. GATA-3 consensus DNA binding motif



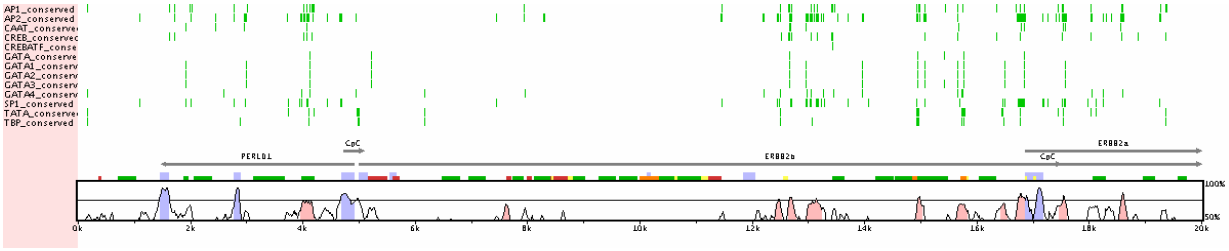

```
A [25  0 61  0 39 15 ]
C [14  1  0  0  1  3 ]
G [ 4 62  1  5  4 37 ]
T [20  0  1 58 19  8 ]
```

(A/T/C)GATA(A/G)
WGATAR

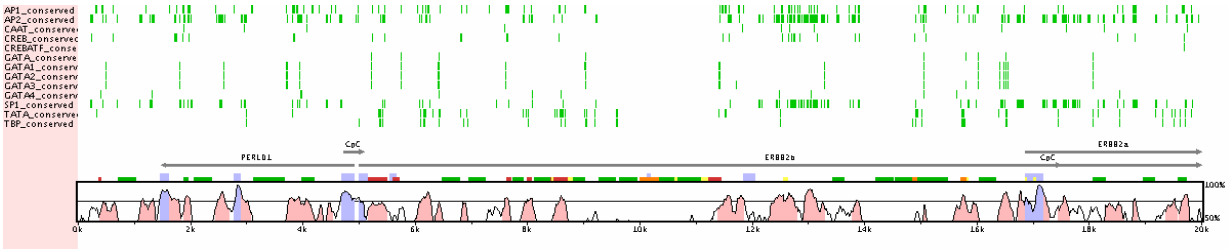
chromosomal localization
8p23.1-8p23.1*

rVISTA Homo sapiens ERBB2a-b 20kb

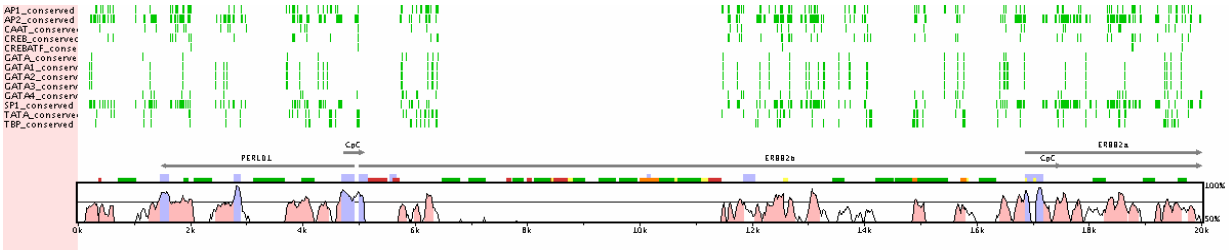
Mm



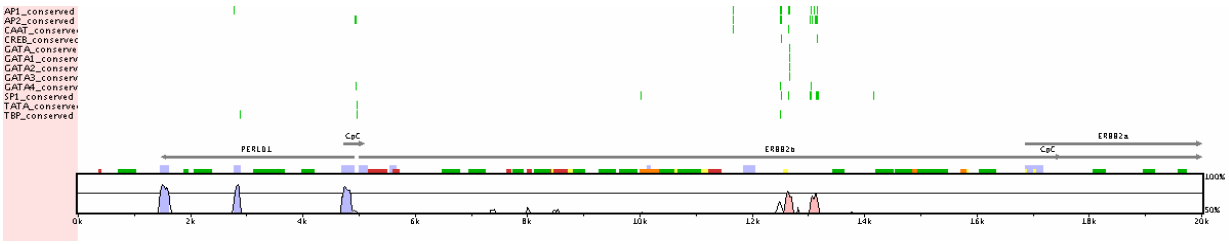
Bt



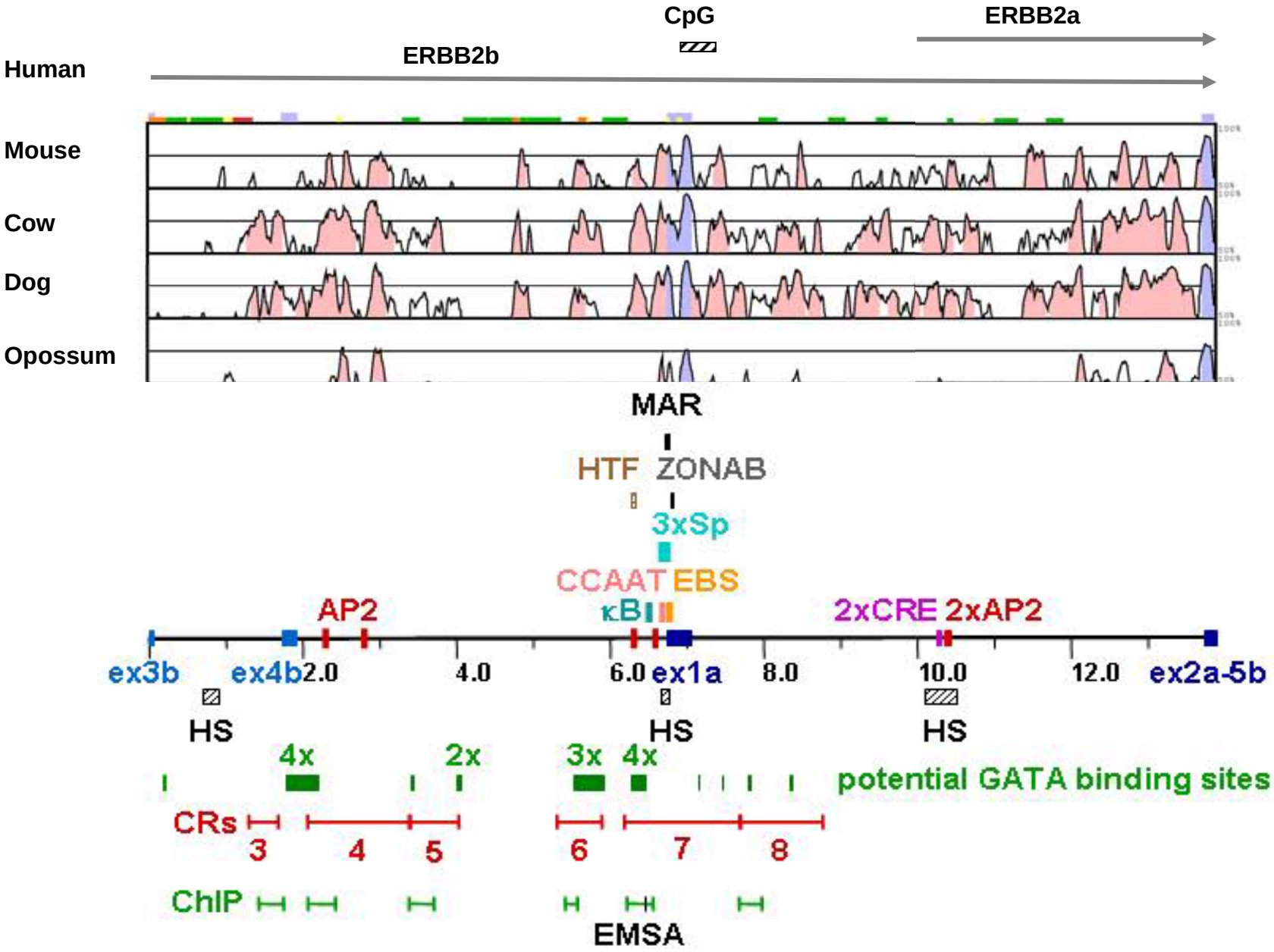
Cf



Md



Homo sapiens ERBB2a regulatory regions (13,883 bases)



http://www.dcode.org/

TOOLS
ECR Browser
ECRbase
SynoR
Mulan
zPicture
multiTF
rVista 2.0
eShadow
Creme 2.0
Array2BIO

NEWS

PUBLICATIONS

ABOUT US

MAILING LIST

LINK TO DCODE!

18 visitors online

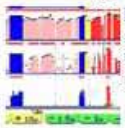
DCODE.org Comparative Genomics Center is a publicly available resourse for regulatory genome data mining. It provides tools for evolutionary comparisons, sequence alignments, and detection of functional sequence patterns.

News

November 7, 2006

"**ECRbase**: Database of Evolutionary Conserved Regions, Promoters, and Transcription Factor Binding Sites in Vertebrate Genomes" manuscript describing a new Dcode.org resource was published in Bioinformatics [**PDF**].

Whole genome alignments



ECR Browser -- Evolutionary conservation of multiple genomes. Identification and sequence analysis of regulatory elements.
Genome Alignment -- 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.

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.

Regulation of co-expressed genes

SynoR -- Prediction of synonymous regulatory elements in vertebrate genomes.
Creme 2.0 -- Identification of cTFBS modules specific to promoters of co-expressed human genes.

Additional resources

Xenomics -- *Xenopus tropicalis* computational resources
Insitu.dcode.org - *Xenopus tropicalis in situ* database
Reverse complement a nucleotide sequence
Primers/oligos selection tool
Batch **sequence retrieval** from the UCSC Genome Browser

Comparative Genomics Center, Computation Directorate
Lawrence Livermore National Laboratory

Google Health News

21-Dec-2006 US Teens Turning From Illicit Drugs... >>>
21-Dec-2006 Gut Bacteria May Determine Dieting Efficiency >>>
21-Dec-2006 New PET Tricks for Old-Age Disease >>>
21-Dec-2006 New flu pandemic 'would kill 62 million' >>>
21-Dec-2006 Biomarkers Little Help for Heart Attacks >>>
21-Dec-2006 Health Highlights: Dec. 21, 2006 >>>
21-Dec-2006 US measles outbreak tied to traveler from Romania >>>
21-Dec-2006 Indonesia claiming bird flu success as cases drop >>>

More news from **Google News**

Genetics news

21-Dec-06 Researchers make progress in studying genetic traits of India-born populations >>>
21-Dec-06 Feinstein Institute and Cold Spring Harbor Lab join forces, seek manic depressio >>>
21-Dec-06 Shotgun sequencing finds nanoorganisms >>>
21-Dec-06 Intelligent software solutions to better understand biological processes >>>
21-Dec-06 Scientists develop method to find genetic basis for plant variation >>>
21-Dec-06 What it means to be human >>>
21-Dec-06 Structure of iron regulatory protein-RNA complex solved >>>
21-Dec-06 JCI table of contents: Dec. 21, 2006 >>>

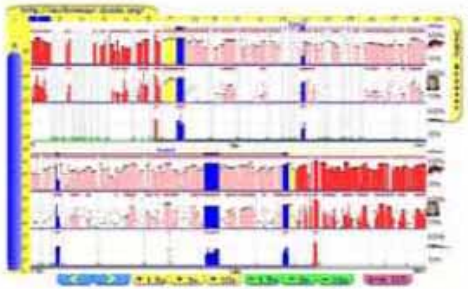
Bioinformatics news

21-Dec-06 Intelligent software solutions to better understand biological processes >>>
19-Dec-06 Plant biologist seeks molecular differences between rice and its mimic >>>
19-Dec-06 Human-chimp difference may be bigger >>>
14-Dec-06 Beyond the book: Software automates access to brain atlases >>>

More news from **EurekAlert!**

ECR Browser

http://ecrbrowser.dcode.org/



The ECR Browser is a dynamic whole-genome navigation tool for visualizing and studying evolutionary relationships between vertebrate and non-vertebrate genomes. The tool is constantly being updated to include the most recently available sequenced genomes (currently: **human**, **dog**, **mouse**, **rat**, **chicken**, **frog** two pufferfish (***Fugu*** and ***Tetraodon***), **zebrafish**, and **6 fruitflies**).

Evolutionary Conserved Regions (**ECRs**) that have been mapped within alignments of the genomes are presented in this graphical browser, which depicts and color-codes ECRs in relation to known genes that have been annotated in the base genome. '**Grab ECR**' feature allows users to rapidly extract sequences that correspond to any ECR, to visualize underlying sequence alignments and/or identify conserved transcription factor binding sites.

In addition to accessing pre-computed alignments for the available genomes, the ECR Browser can also be used as an alignment tool. It allows users to map submitted sequences to specific homologous positions within the human, mouse, rat, chicken, *Fugu* or fruit fly genomes and to create a detailed alignment using the blastz alignment program.

Please note: Not all alignments are visible by default. Click "Browser Settings" to see all available alignments

Run the ECR Browser

Select a base organism and indicate the name of a gene or a chromosomal location (*chr1:from-to* format)

Human (Mar'06) (hg18) | GATA2 | Submit

GENOME ALIGNMENT: Align your sequence to either Human, Mouse, Rat, Chicken, Fugu or Drosophila genome

Help

- Details of the alignment strategy and ECR Browser structure.
- Instructutions on how to use the ECR Browser.

Citing the ECR Browser

I. Ovcharenko, M.A. Nobrega, G.G. Loots, and L. Stubbs
ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes
Nucleic Acids Research, 32, W280-W286 (2004) [PDF]

Special features:

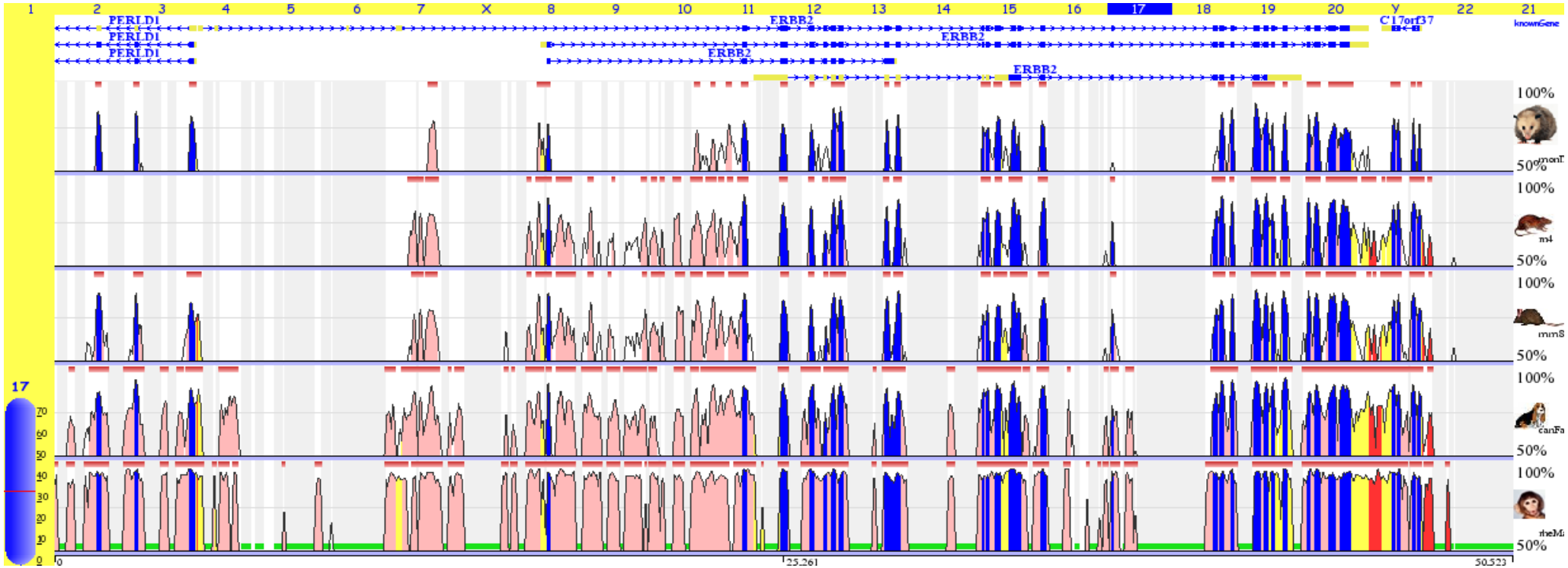
- **Identification of regions of homology and significant paralogy between the base sequence and other aligned genomes:** The user can retrieve locations of homologous sequences within aligned genomes and view the extent of sequence match to the base genome. *Details*.
- **User-determined parameters:** The browser permits the user to decide the lengths and percent identities that will be considered significant, and which will highlighted on the viewing screen and available for download and analysis. *Details*.
- **Easy retrieval and analysis of ECR sequences and ready access to underlying alignments:** In the ECR browser, conserved sequence peaks are not simply static features on a graph, but are active links that permit easy access to sequences and alignments. The user can retrieve ECR sequences and send them on for additional analysis, such as rVista. A tool to facilitate primer design for experimental manipulation of conserved exons, enhancers, or other elements is also provided. *Details*.
- **A dynamic portal to several other sequence analysis tools** (UCSC Genome Browser, GALA annotation database, Ensembl, Rat Genome Browser at MCW, NCBI, zPicture and rVista 2.0 tools) providing users with a wide range of choices for detailed evolutionary analysis and follow-up studies that might include annotating novel genes using mRNA, EST and conserved sequence data or analysis of transcription factor binding sites.

About the ECR Browser

The ECR Browser was developed by Ivan Ovcharenko with help from Marcelo Nobrega, Gabriela Loots and Lisa Stubbs.

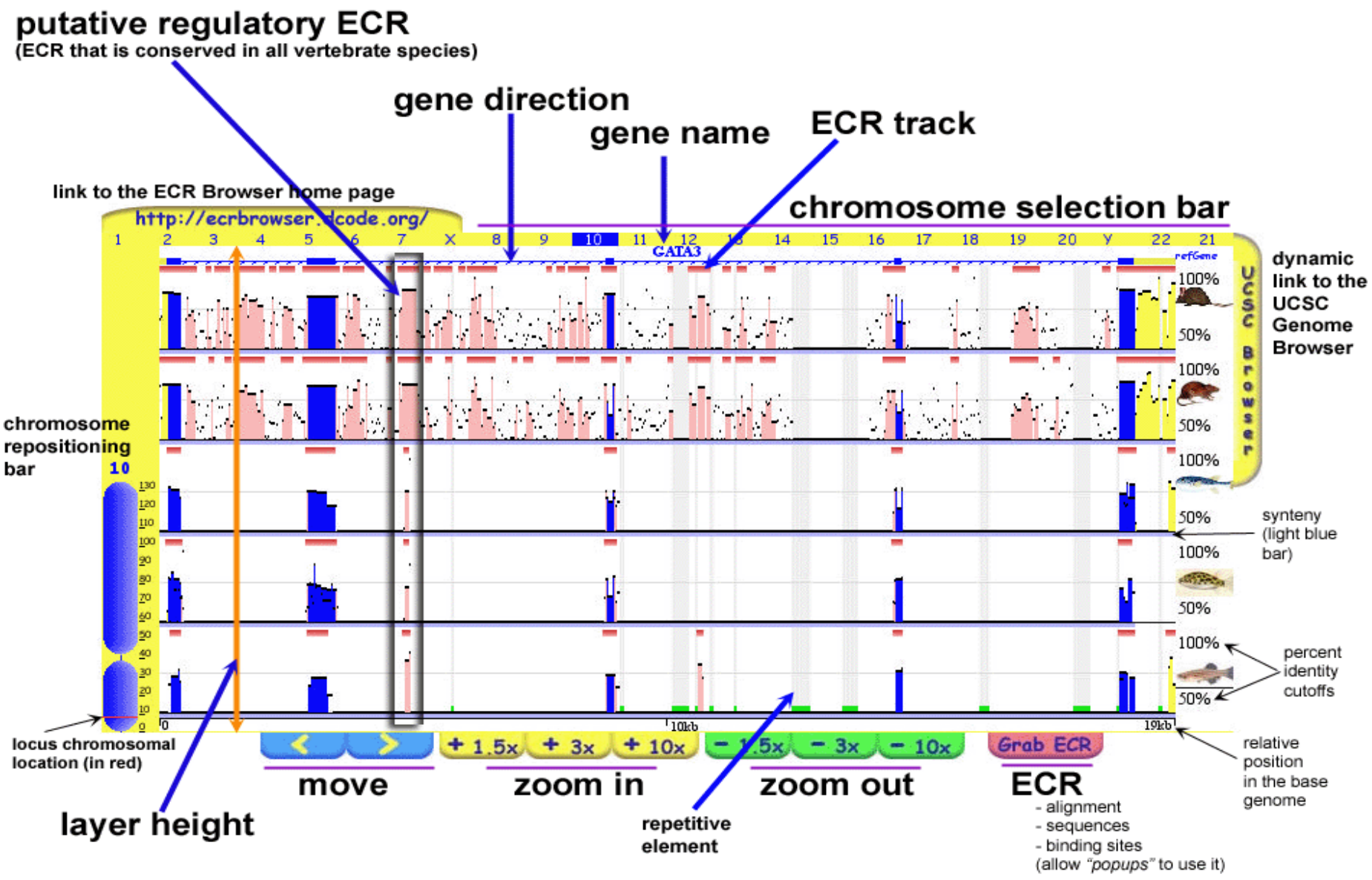
Homo sapiens ERBB2 locus

chr17:35,092,919-35,143,441



G.G. Loots and I. Ovcharenko
ECRbase: Database of Evolutionary Conserved Regions, Promoters, and Transcription Factor Binding Sites in Vertebrate Genomes. Bioinformatics, 23(1):122-4 (2007)

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

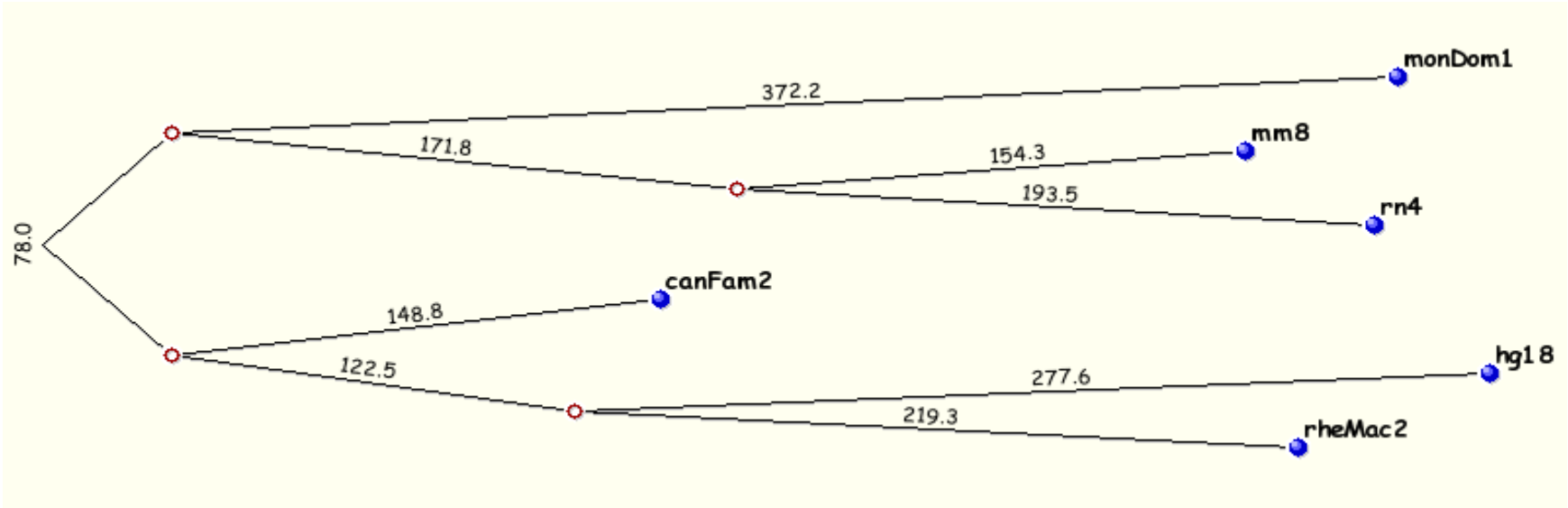


Color codes

coding exon | UTR | intron | intergenic element

Species

mouse | rat | fugu | tetraodon | zebrafish


78.0
372.2
171.8
154.3
193.5
148.8
122.5
277.6
219.3
monDom1
mm8
rn4
canFam2
hg18
rheMac2

multiTF :: **http:// multitf . dcode . org**

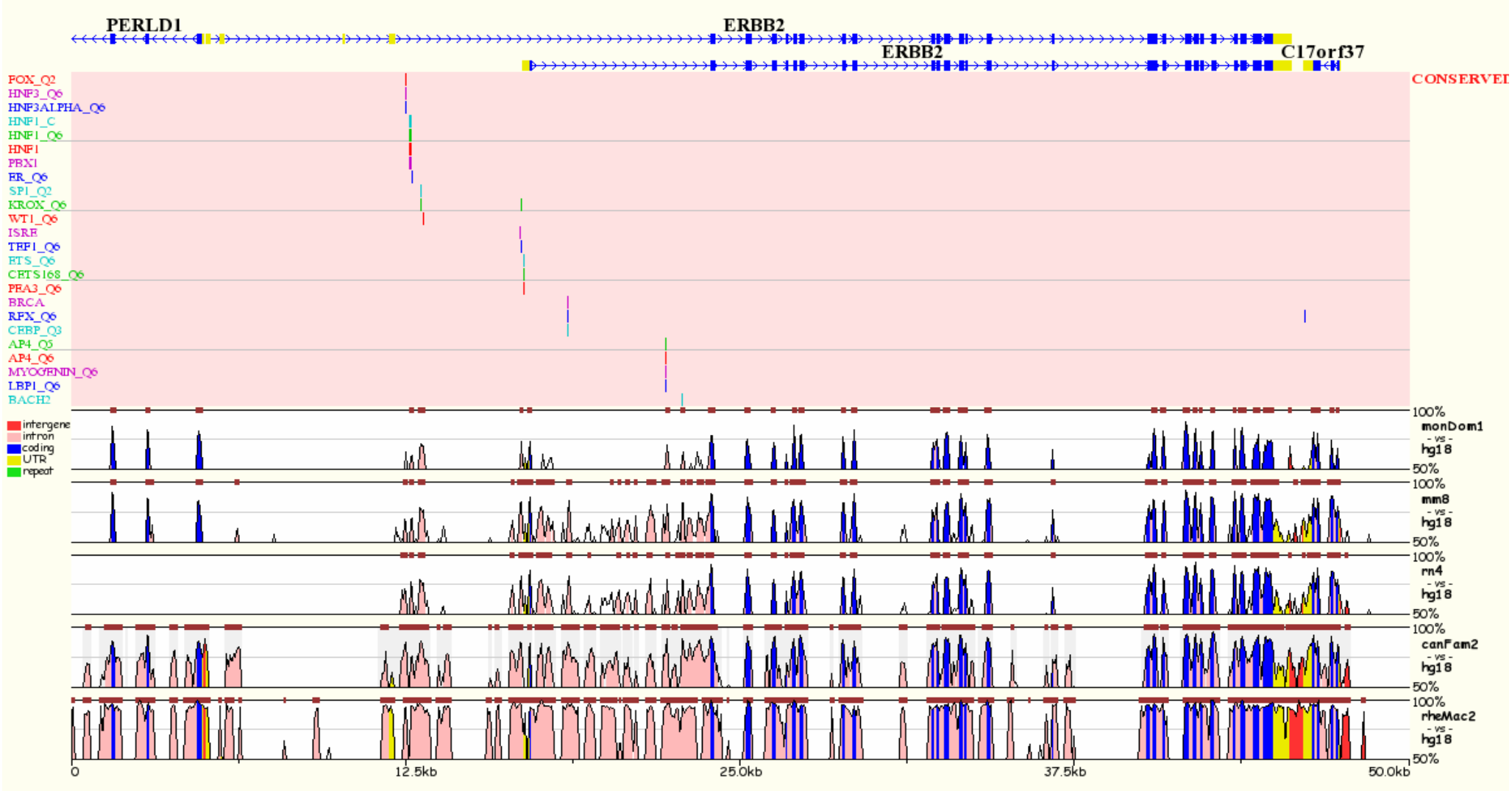
multiTF identifies transcription factor binding sites conserved across multiple species

There are 3 diffrent ways to **initiate a multiTF search:**

1. **Mulan** (mulan.dcode.org) multiple sequence alignments of FINISHED sequences can be automatically submitted to multiTF from the results web page.
2. **GALA** (Genome Alignment and Annotation Database; http://gala.cse.psu.edu/) automatically forwards genome alignment data to Mulan for the subsequent multiTF post-processing.
3. **ECR Browser** (ecrbrowser.dcode.org) multi-genome sequence data can be automatically forwarded to Mulan and then to multiTF by using the "Synteny/Alignments" feature of the tool.

Return to previously submitted request. ID: SUBMIT

multiTF: ERBB2 Hs-Md-Mm-Cf-Rm



SynoR: http://synor.dcode.org/

Synonymous gene regulation (identified by shared tissue and timepoint specificities) is established through genomic elements that contain a specific cluster of transcription factor binding sites (TFBS). SynoR performs vertebrate genome scans for evolutionary conserved clusters of TFBS in a predefined configuration to identify synonymous regulatory elements (SREs). SynoR searches for SREs by utilizing known patterns of TFBS in active regulatory elements (REs) as seeds for genome scans. Multiple-species conservation layers allow the use of differential conservation filters in the search for SREs and SynoR results provide an extensive annotation of identified genes that contain detected REs. Gene Ontology classifiers are utilized to perform functional classification of the identified genes, and integrated GNF Expression Atlas data permits on-the-fly analysis of tissue-specificities of predicted SREs.

I. Ovcharenko and M.A. Nobrega. Identifying synonymous regulatory elements in vertebrate genomes. *Nucleic Acids Research*, 33, W403-7, 2005.

SynoR :: Genome Miner for Synonymous Regulation - Windows Internet Explorer

http://synor.dcode.org/

Fichier Edition Affichage Favoris Outils ?

http://synor.dcode.or... | http://synor.dcode.or... | SynoR :: Genome ... | Genomatix: Matrix Lib... | ECR Browser

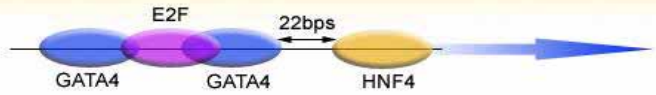
Page Outils

Dcode.org Mulan ECR Browser ECR Base SynoR zPicture multiTF rVista 2.0 Creme 2.0 eShadow array2BIO

SynoR [Genome Miner for Synonymous Regulation]

http://synor.dcode.org

Details ... Instructions ... Citing SynoR



1. List transcription factors

TFBS cluster specifications:
(Format: <name> <count> <strand>)

STAT5A
STAT5B
ETS
HMGIY

Example:
(see Instructions for details)
GATA4 2 +
HNF4_01

☐ Fix the order of TFBS

List of available TFBS

2. Specify distance limitations between neighboring binding sites

at least 4 bps, but not more than 100 bps
(Note: The minimal distance is 3 bps, the maximal is 1000 bps.)

3. Select genomes

base genome: human (hg18)
comparison genome(s): mouse (mm8)

SUBMIT

Internet 100%

démarrer | Boîte de r... | SynoR :: ... | 2 Explor... | 2 Bloc-n... | Microsoft ... | 2 Micro... | SECentral... | Adobe Ac... | FR | 18:02

Results

Request ID: **s01080816575494**

132 TFBS module instances identified in **human (hg18)** -vs- **mouse (mm8)** genome alignments.

Detailed modules' annotation: [**table**] -or- [text]
[location, length, sequence, gene type, neighboring genes, cross-species conservation analysis]

Module types:

intergenic	68 (52%)
intron	44 (33%)
cds	13 (10%)
utr	4 (3%)
promoter	3 (2%)

Functional annotation of genes corresponding to noncoding modules:
- Enrichment in Gene Ontology categories: [**table**] and [text]
- Tissue-specificity of identified genes: **all noncoding** -or- **excluding distant (over 50k) elements/genes**
(Using expression data from GNF Expression Atlas2. GNF disclaimer on data distribution.)

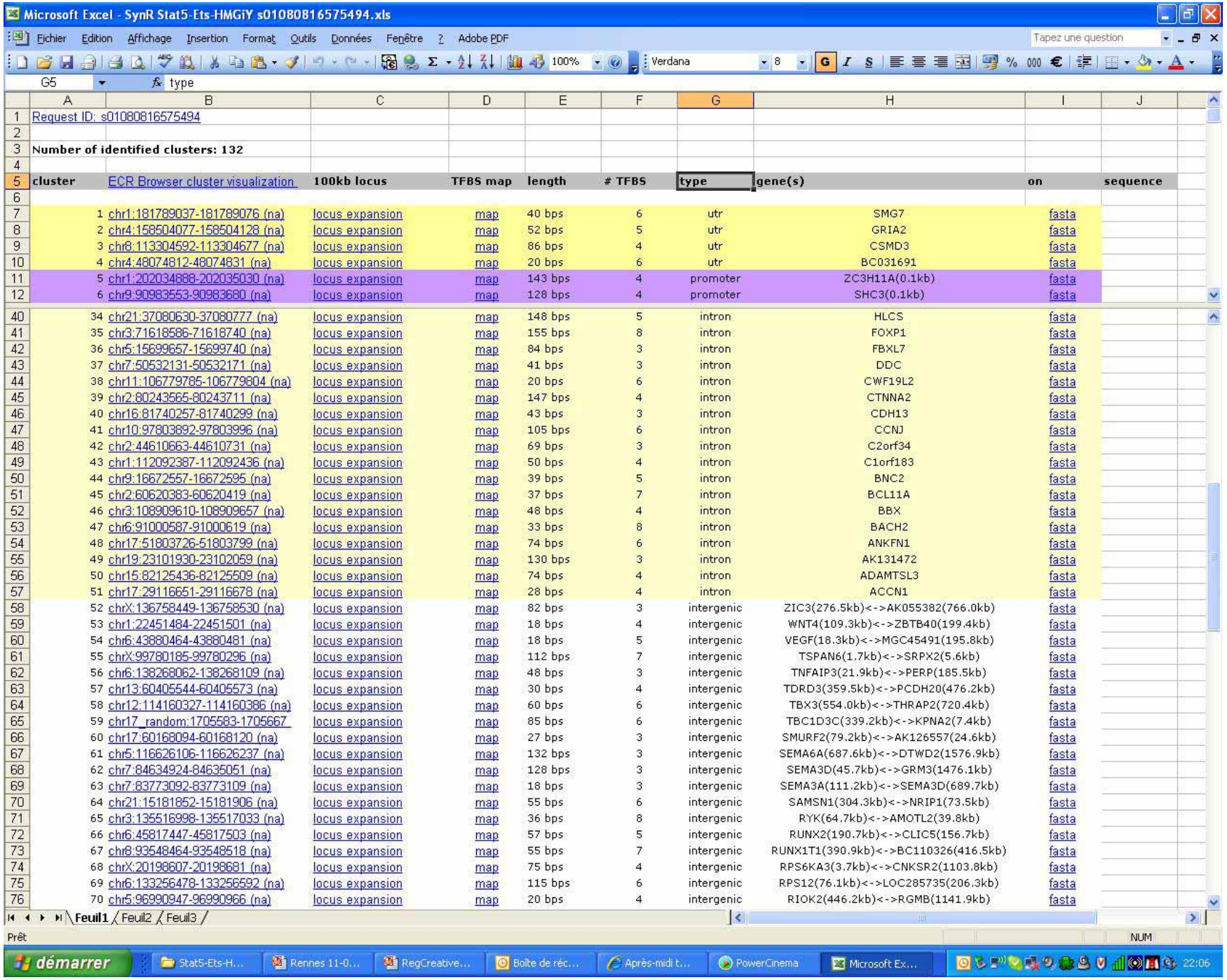
Chromosomal distribution (click on the image to enlarge it):

Search seed:

group	strand	count	TFBS
1	+-	1	STAT5A_01 or STAT5B_01 or STAT5A_02 or STAT5A_03
2	+-	1	ETS1_B or ETS2_B or ETS_Q4
3	+-	1	HMGIY_Q6

distance limitations: 4 .. 100

Contact Ivan Ovcharenko (ovcharenko1@llnl.gov) if you have any questions or suggestions



Microsoft Excel - SynR Stat5-Ets-HMGiY s01080816575494.xls

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ? Adobe PDF

Tapez une question

Verdana 8

G5 type

Request ID: s01080816575494

Number of identified clusters: 132

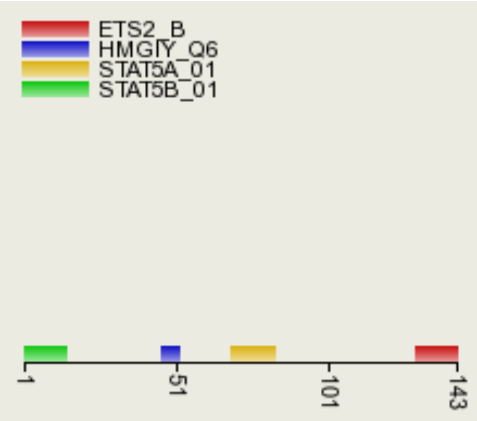
cluster	ECR Browser cluster visualization	100kb locus	TFBS map	length	# TFBS	type	gene(s)	on	sequence
1	chr1:181789037-181789076 (na)	locus expansion	map	40 bps	6	utr	SMG7	fasta	
2	chr4:158504077-158504128 (na)	locus expansion	map	52 bps	5	utr	GRIA2	fasta	
3	chr8:113304592-113304677 (na)	locus expansion	map	86 bps	4	utr	CSMD3	fasta	
4	chr4:48074812-48074831 (na)	locus expansion	map	20 bps	6	utr	BC031691	fasta	
5	chr1:202034888-202035030 (na)	locus expansion	map	143 bps	4	promoter	ZC3H11A(0.1kb)	fasta	
6	chr9:90983553-90983680 (na)	locus expansion	map	128 bps	4	promoter	SHC3(0.1kb)	fasta	
34	chr21:37080630-37080777 (na)	locus expansion	map	148 bps	5	intron	HLCS	fasta	
35	chr3:71618586-71618740 (na)	locus expansion	map	155 bps	8	intron	FOXP1	fasta	
36	chr5:15699657-15699740 (na)	locus expansion	map	84 bps	3	intron	FBXL7	fasta	
37	chr7:50532131-50532171 (na)	locus expansion	map	41 bps	3	intron	DDC	fasta	
38	chr11:106779785-106779804 (na)	locus expansion	map	20 bps	6	intron	CWF19L2	fasta	
39	chr2:80243565-80243711 (na)	locus expansion	map	147 bps	4	intron	CTNNA2	fasta	
40	chr16:81740257-81740299 (na)	locus expansion	map	43 bps	3	intron	CDH13	fasta	
41	chr10:97803892-97803996 (na)	locus expansion	map	105 bps	6	intron	CCNJ	fasta	
42	chr2:44610663-44610731 (na)	locus expansion	map	69 bps	3	intron	C2orf34	fasta	
43	chr1:112092387-112092436 (na)	locus expansion	map	50 bps	4	intron	C1orf183	fasta	
44	chr9:16672557-16672595 (na)	locus expansion	map	39 bps	5	intron	BNC2	fasta	
45	chr2:60620383-60620419 (na)	locus expansion	map	37 bps	7	intron	BCL11A	fasta	
46	chr3:108909610-108909657 (na)	locus expansion	map	48 bps	4	intron	BBX	fasta	
47	chr6:91000587-91000619 (na)	locus expansion	map	33 bps	8	intron	BACH2	fasta	
48	chr17:51803726-51803799 (na)	locus expansion	map	74 bps	6	intron	ANKFN1	fasta	
49	chr19:23101930-23102059 (na)	locus expansion	map	130 bps	3	intron	AK131472	fasta	
50	chr15:82125436-82125509 (na)	locus expansion	map	74 bps	4	intron	ADAMTSL3	fasta	
51	chr17:29116651-29116678 (na)	locus expansion	map	28 bps	4	intron	ACCN1	fasta	
52	chrX:136758449-136758530 (na)	locus expansion	map	82 bps	3	intergenic	ZIC3(276.5kb)<->AK055382(766.0kb)	fasta	
53	chr1:22451484-22451501 (na)	locus expansion	map	18 bps	4	intergenic	WNT4(109.3kb)<->ZBTB40(199.4kb)	fasta	
54	chr6:43880464-43880481 (na)	locus expansion	map	18 bps	5	intergenic	VEGF(18.3kb)<->MGC45491(195.8kb)	fasta	
55	chrX:99780185-99780296 (na)	locus expansion	map	112 bps	7	intergenic	TSPAN6(1.7kb)<->SRPX2(5.6kb)	fasta	
56	chr6:138268062-138268109 (na)	locus expansion	map	48 bps	3	intergenic	TNFAIP3(21.9kb)<->PERP(185.5kb)	fasta	
57	chr13:60405544-60405573 (na)	locus expansion	map	30 bps	4	intergenic	TDRD3(359.5kb)<->PCDH20(476.2kb)	fasta	
58	chr12:114160327-114160386 (na)	locus expansion	map	60 bps	6	intergenic	TBX3(554.0kb)<->THRAP2(720.4kb)	fasta	
59	chr17_random:1705583-1705667	locus expansion	map	85 bps	6	intergenic	TBC1D3C(339.2kb)<->KPNA2(7.4kb)	fasta	
60	chr17:60168094-60168120 (na)	locus expansion	map	27 bps	3	intergenic	SMURF2(79.2kb)<->AK126557(24.6kb)	fasta	
61	chr5:116626106-116626237 (na)	locus expansion	map	132 bps	3	intergenic	SEMA6A(687.6kb)<->DTWD2(1576.9kb)	fasta	
62	chr7:84634924-84635051 (na)	locus expansion	map	128 bps	3	intergenic	SEMA3D(45.7kb)<->GRM3(1476.1kb)	fasta	
63	chr7:83773092-83773109 (na)	locus expansion	map	18 bps	3	intergenic	SEMA3A(111.2kb)<->SEMA3D(689.7kb)	fasta	
64	chr21:15181852-15181906 (na)	locus expansion	map	55 bps	6	intergenic	SAMSN1(304.3kb)<->NRIP1(73.5kb)	fasta	
65	chr3:135516998-135517033 (na)	locus expansion	map	36 bps	8	intergenic	RYK(64.7kb)<->AMOTL2(39.8kb)	fasta	
66	chr6:45817447-45817503 (na)	locus expansion	map	57 bps	5	intergenic	RUNX2(190.7kb)<->CLIC5(156.7kb)	fasta	
67	chr8:93548464-93548518 (na)	locus expansion	map	55 bps	7	intergenic	RUNX1T1(390.9kb)<->BC110326(416.5kb)	fasta	
68	chrX:20198607-20198681 (na)	locus expansion	map	75 bps	4	intergenic	RPS6KA3(3.7kb)<->CNKSR2(1103.8kb)	fasta	
69	chr6:133256478-133256592 (na)	locus expansion	map	115 bps	6	intergenic	RPS12(76.1kb)<->LOC285735(206.3kb)	fasta	
70	chr5:96990947-96990966 (na)	locus expansion	map	20 bps	4	intergenic	RIOK2(446.2kb)<->RGMB(1141.9kb)	fasta	

Feuil1 Feuil2 Feuil3

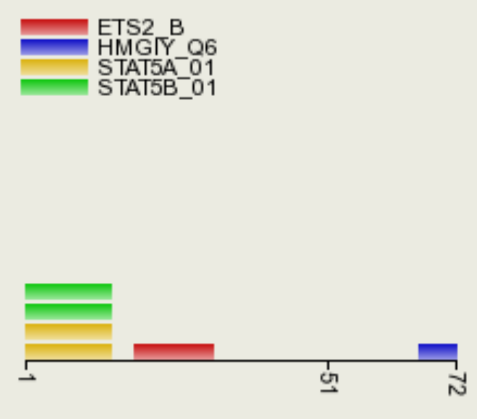
Prêt NUM

démarrer Stat5-Ets-H... Rennes 11-0... RegCreative... Boîte de réc... Après-midi t... PowerCinema Microsoft Ex... 22:06

ZC3H11A(0.1kb)



IL2RA(7.0kb)<->RBM17(19.6kb)



SynoR :: Tissue specificity of genes with noncoding elements

Request ID: s01080816575494

List of tissues, in which genes are overexpressed or suppressed:

Some overexpression

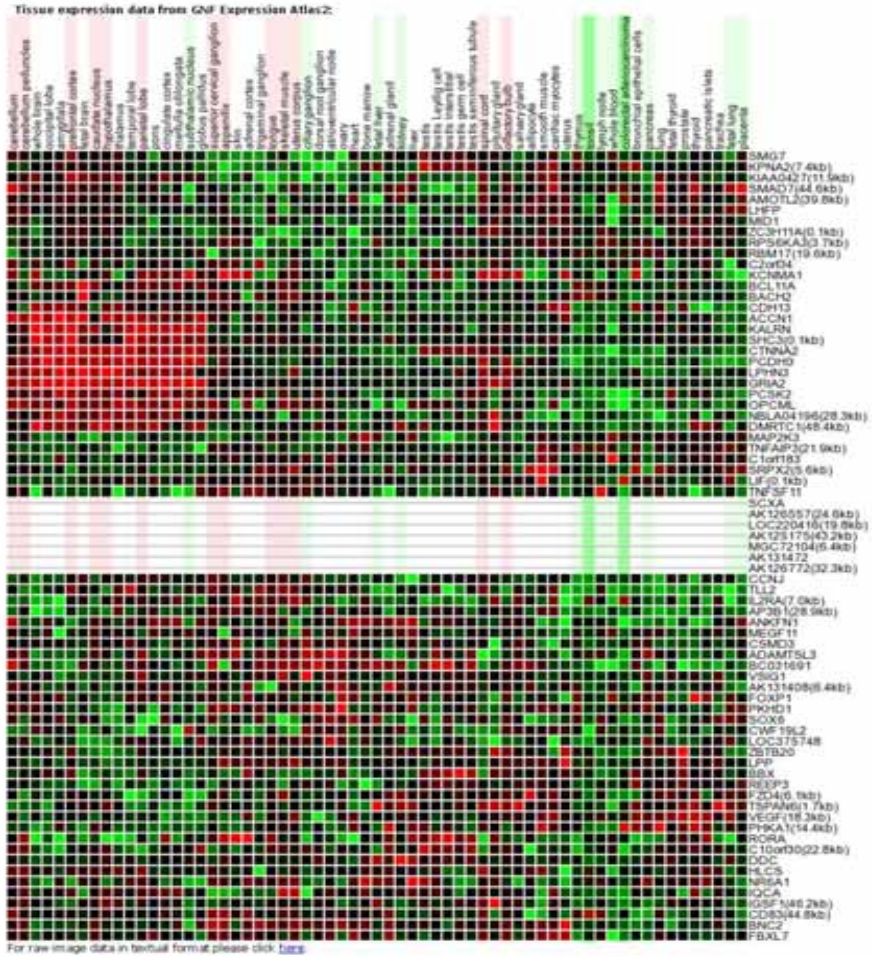
olfactory bulb, spinal cord, superior cervical ganglion, uterus corpus, cerebellum peduncles, parietal lobe, cerebellum, caudate nucleus, appendix, tongue, prefrontal cortex, hypothalamus, skeletal muscle

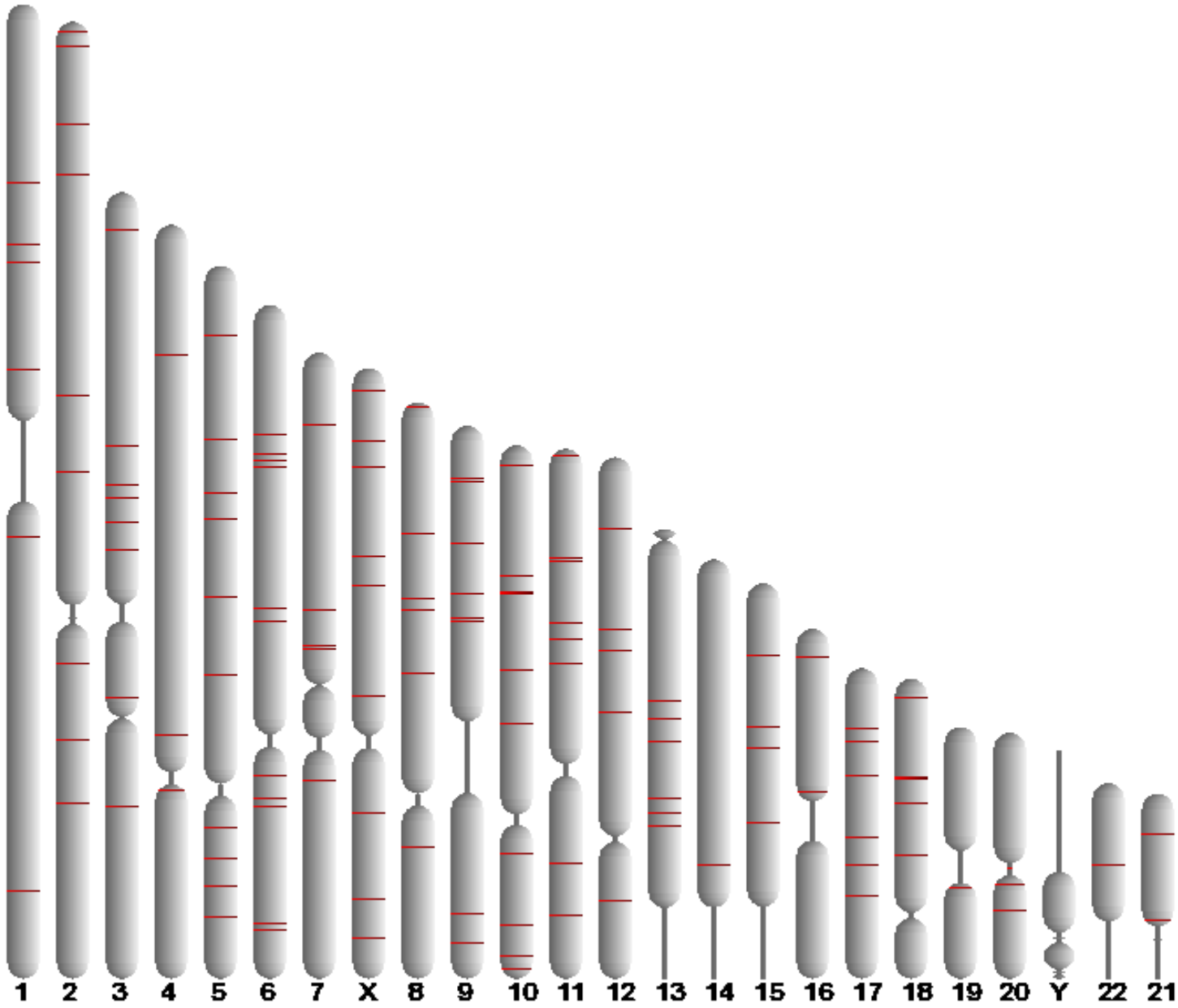
Some suppression

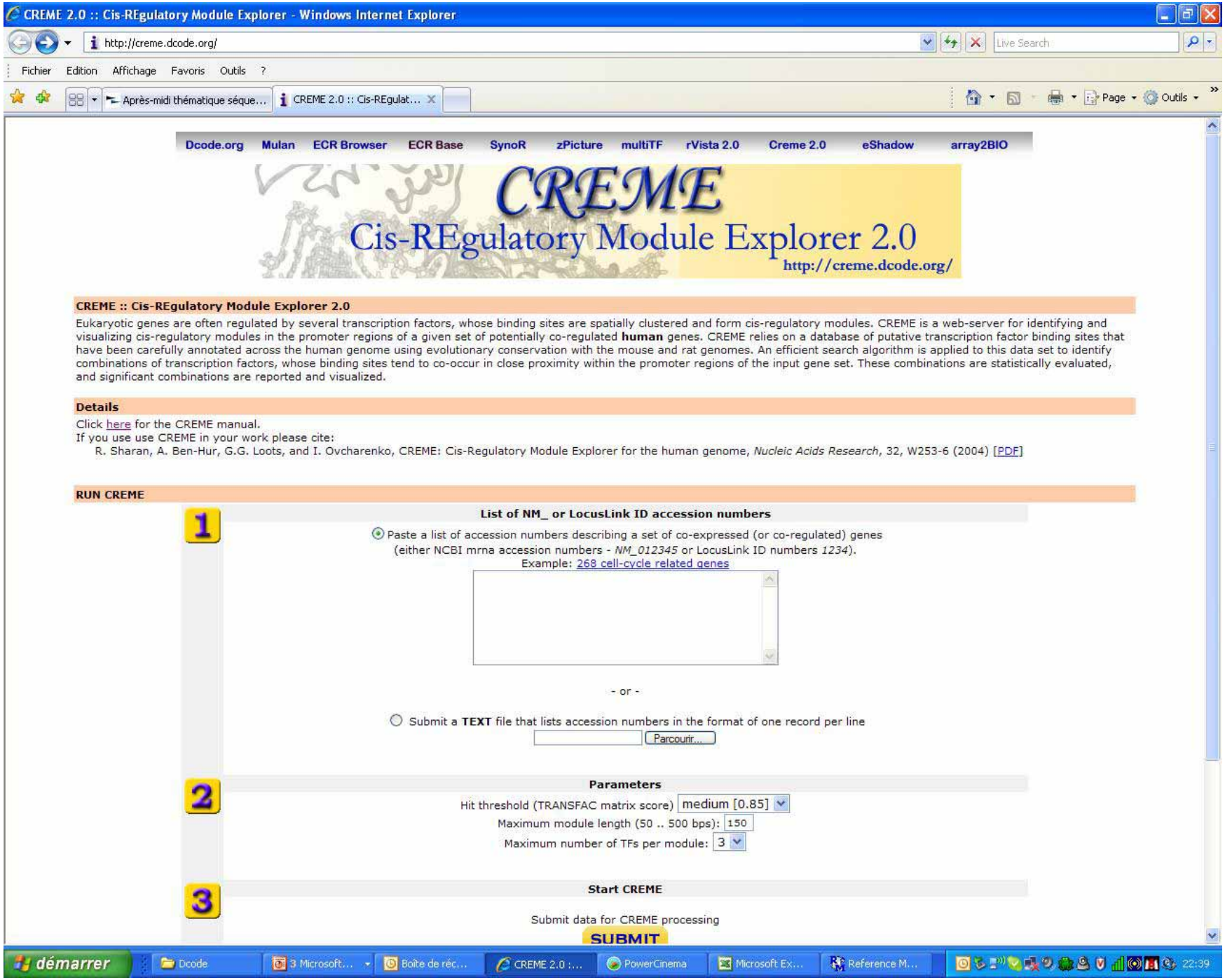
subthalamic nucleus, ciliary ganglion, thymus, lymph node, fetal lung, kidney, whole blood, pancreas, fetal liver, placenta

Significant suppression

tonsil, colorectal adenocarcinoma




1
2
3
4
5
6
7
X
8
9
10
11
12
13
14
15
16
17
18
19
20
Y
22
21


CREME 2.0 :: Cis-REgulatory Module Explorer - Windows Internet Explorer
http://creme.dcode.org/
Live Search
Fichier Edition Affichage Favoris Outils ?
Après-midi thématique séque...
CREME 2.0 :: Cis-REgulat...
Page
Outils
Dcode.org Mulan ECR Browser ECR Base SynoR zPicture multiTF rVista 2.0 Creme 2.0 eShadow array2BIO
CREME
Cis-REgulatory Module Explorer 2.0
http://creme.dcode.org/
CREME :: Cis-REgulatory Module Explorer 2.0
Eukaryotic genes are often regulated by several transcription factors, whose binding sites are spatially clustered and form cis-regulatory modules. CREME is a web-server for identifying and visualizing cis-regulatory modules in the promoter regions of a given set of potentially co-regulated human genes. CREME relies on a database of putative transcription factor binding sites that have been carefully annotated across the human genome using evolutionary conservation with the mouse and rat genomes. An efficient search algorithm is applied to this data set to identify combinations of transcription factors, whose binding sites tend to co-occur in close proximity within the promoter regions of the input gene set. These combinations are statistically evaluated, and significant combinations are reported and visualized.
Details
Click here for the CREME manual.
If you use use CREME in your work please cite:
R. Sharan, A. Ben-Hur, G.G. Loots, and I. Ovcharenko, CREME: Cis-Regulatory Module Explorer for the human genome, Nucleic Acids Research, 32, W253-6 (2004) [PDF]
RUN CREME
1
List of NM_ or LocusLink ID accession numbers
Paste a list of accession numbers describing a set of co-expressed (or co-regulated) genes
(either NCBI mrna accession numbers - NM_012345 or LocusLink ID numbers 1234).
Example: 268 cell-cycle related genes
- or -
Submit a TEXT file that lists accession numbers in the format of one record per line
Parcourir...
2
Parameters
Hit threshold (TRANSFAC matrix score) medium [0.85]
Maximum module length (50 .. 500 bps): 150
Maximum number of TFs per module: 3
3
Start CREME
Submit data for CREME processing
SUBMIT
démarrer
Dcode
3 Microsoft...
Boîte de réc...
CREME 2.0 :...
PowerCinema
Microsoft Ex...
Reference M...
22:39

MAR-Wiz

http://futuresoft.org/MAR-Wiz/

The MAR-Wiz tool aims at discovering the presence of Matrix Association Regions, or MARs, within DNA sequences. MARs constitute a significant functional block within sequences and facilitate the processes of differential gene expression and DNA replication. Our computational approach to finding MARs is based upon the co-occurrence of 20 DNA patterns that have been known to occur in the neighborhood of MARs. These motifs are used to define higher order rules defined using the various combinations in which the patterns have been known to co-occur. The mathematical density of the rule occurrences iin a regions is assumed to imply the presence of a MAR in that region

For examples:

-- Cai S, Lee CC, Kohwi-Shigematsu T. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. Nat Genet.38:1278-88, 2006.

- P. PK, O. Bischof, P. K. Purbey, D. Notani, H. Urlaub, A. Dejean, and S. Galande. Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. *Nat Cell Biol.* 9 (1):45-56, 2007.

Cai S, Lee CC, Kohwi-Shigematsu T. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. Nat Genet.38:1278-88, 2006.

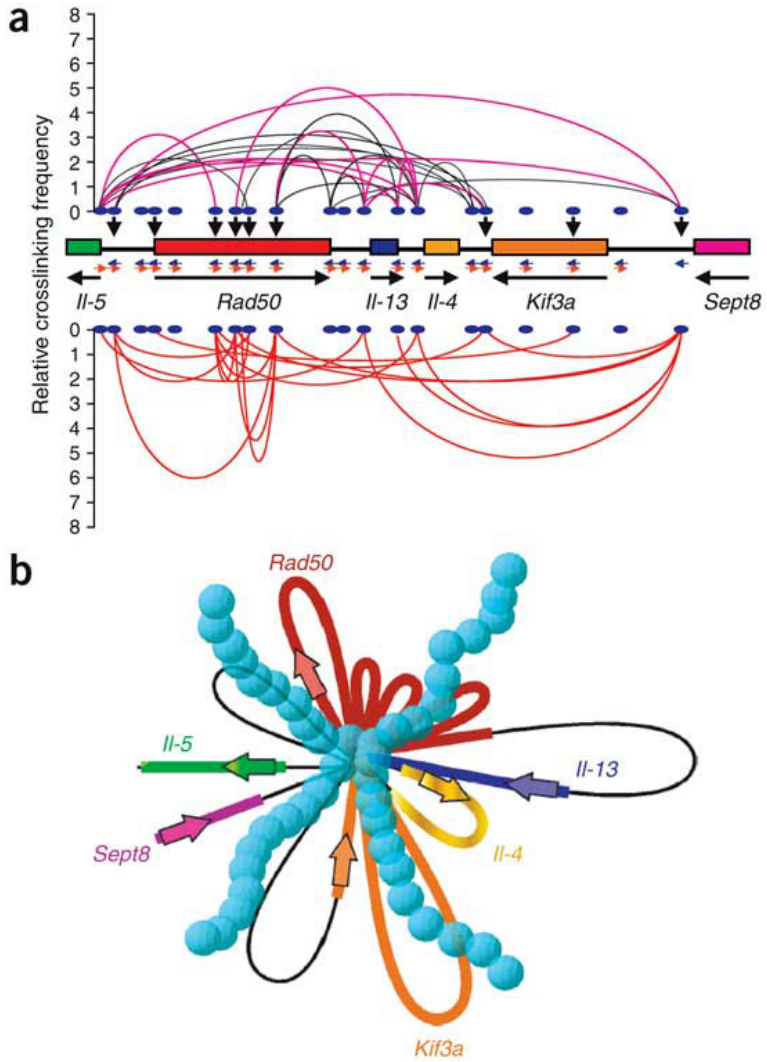


Figure 8. Summary of activation-dependent looping events and a model of transcriptionally active chromatin.

P. PK, O. Bischof, P. K. Purbey, D. Notani, H. Urlaub, A. Dejean, and S. Galande. Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. *Nat Cell Biol.* 9 (1):45-56, 2007.

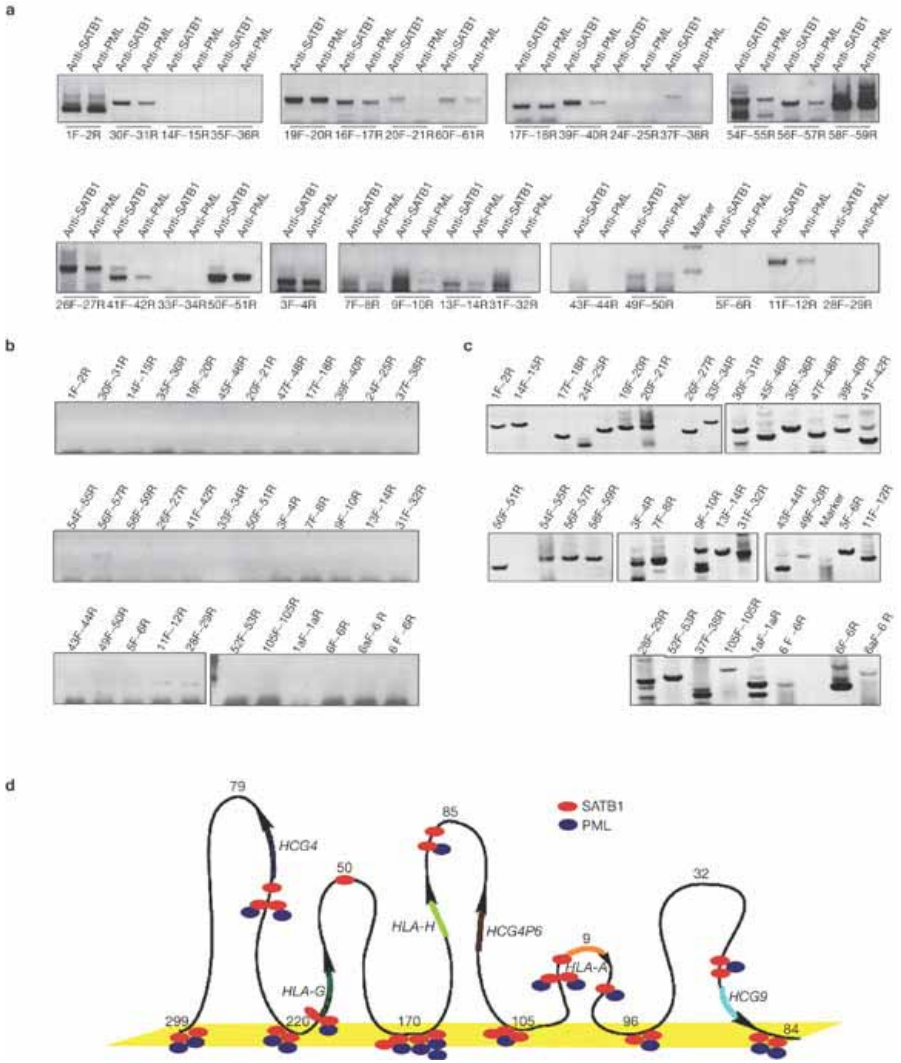


Figure 4. SATB1 and PML directly associate with specific genomic regions of MHC class I locus *in vivo*.

MAR-Analysis Summary Report

Sequence Description: Hs IL2RA
Sequence Length: 78588
Maximum and Minimum Potential = [0.0125264 .. 43.1201]

High Scoring Regions with threshold = 0.6

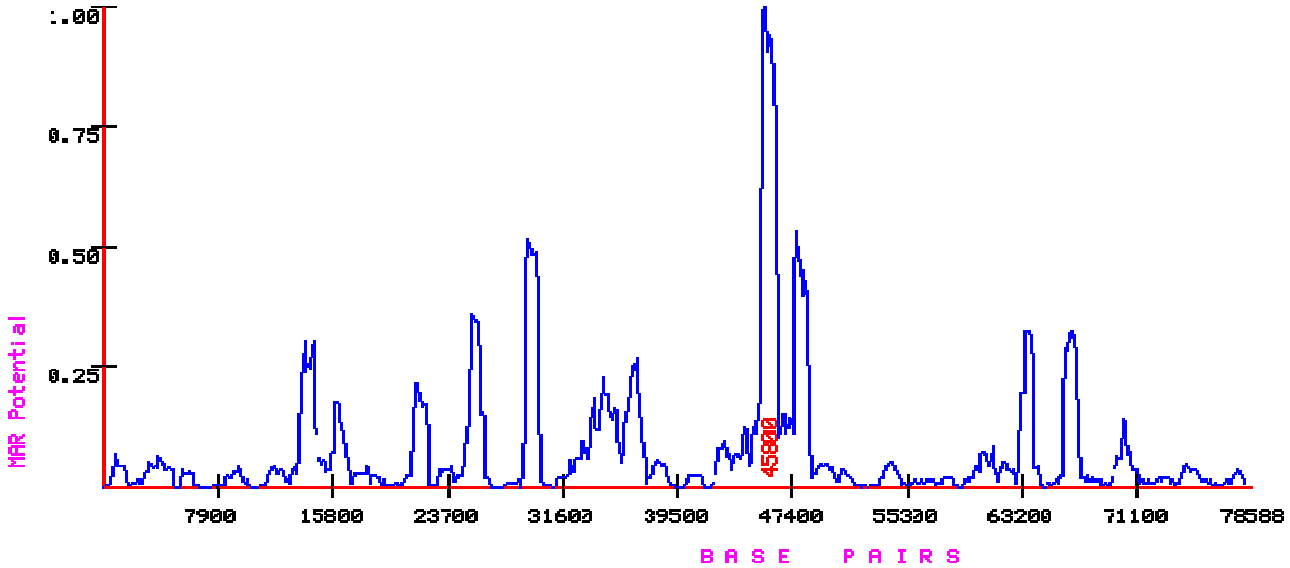
Region	Average Strength	Integrated Strength
45300 ... 46200	0.890511	802.35

Frequency of Rule Matches

Number of Rules Selected = 6

Rule Name	Forward Strand Match Count	Reverse Strand Match Count	
ORI Pattern	512	505	View Detail Locations
TG-Richness Pattern	18	20	View Detail Locations
Curved DNA Pattern	106	106	View Detail Locations
Kinked DNA Pattern	110	110	View Detail Locations
Topo II Pattern	217	210	View Detail Locations
AT-Richness Pattern	872	849	View Detail Locations

Questions or Comments...Please send E-mail to
Dr. Gautam B. Singh, *gbs@futuresoft.org*



Clone/Gene ID Converter

Bioinformatics Unit - CNIO

Version 2.0

Help / IDs relations:

Input

- Organism: Human
- Identifier: Gene Name (HUGO)
- List of Identifiers:
  * Put only one id by line.

IL2RA

Get your results on a spreadsheet file.
Get your results on a text file (tab separated).

Output

The information shown is extracted from:

Ensembl (version Hs #40.36b (Mar.#40.36a (Dec.#40.34))
UCSC (Build Hs hg17).
UniGene (Build Hs #204 (Mar.#155 (Dec.#154).
PubMed (Date: 21/08/06).
KEGG Pathway (Rel 39.0 (Date : 08/06).
Reactome (Rel. 10 (Date : 02/08/06).

Choose the information that you would like to include in your report:

Gene Level

- UniGene Cluster — Taken from UniGene
- Gene Description — Taken from UniGene
- Gene Name — Taken from UniGene
- EntrezGene — Taken from UniGene
- Ensembl Gene — Taken from Ensembl
- RefSeq_RNA — Taken from Ensembl
- RefSeq_peptide — Taken from Ensembl
- CCDS — Taken from Ensembl

Gene Location

- Ensembl
- UCSC
- UCSC*
- Ensembl* — Equivalent, uses UCSC
- UCSC* — Equivalent, uses Ensembl

Clone Level

- Clone ID — Taken from UniGene
- GenBank Accession — Taken from UniGene
- Affymetrix ID — Taken from Ensembl

Protein Level

- SwissProt — Taken from Ensembl
- Embl — Taken from Ensembl
- PDB id — Taken from Ensembl
- IPI — Taken from Ensembl
- OMIM* — Taken from Ensembl

Functional Level

- GO — Taken from Ensembl
- KEGG Pathways — Taken from KEGG
- Reactome Pathway — Taken from Reactome
- Reactome Reaction — Taken from Reactome
- PubMed id — Taken from NCBI

Convert

Please note that while working with thousands of ids, the process may take several minutes to conclude.

* Information available for Human only

	Gene Level						Clone Level		Functional Level
Gene Name	Ensembl Gene	Unigene Cluster	EntrezGene	Ensembl Location Chr. start-end (strand)	UCSC Location Chr. start-end (strand)	Ensembl* Location Chr. start-end (strand)	GenBank Accession	Clone Id	KEGG
IL2RA	ENSG00000134460	Hs.251367	3559	Chromosome 10 6056305bp - 6144204bp (-1)		Chromosome 10 6056305bp - 6144204bp (-1)	...	...	hsa04060 hsa04630 hsa04640

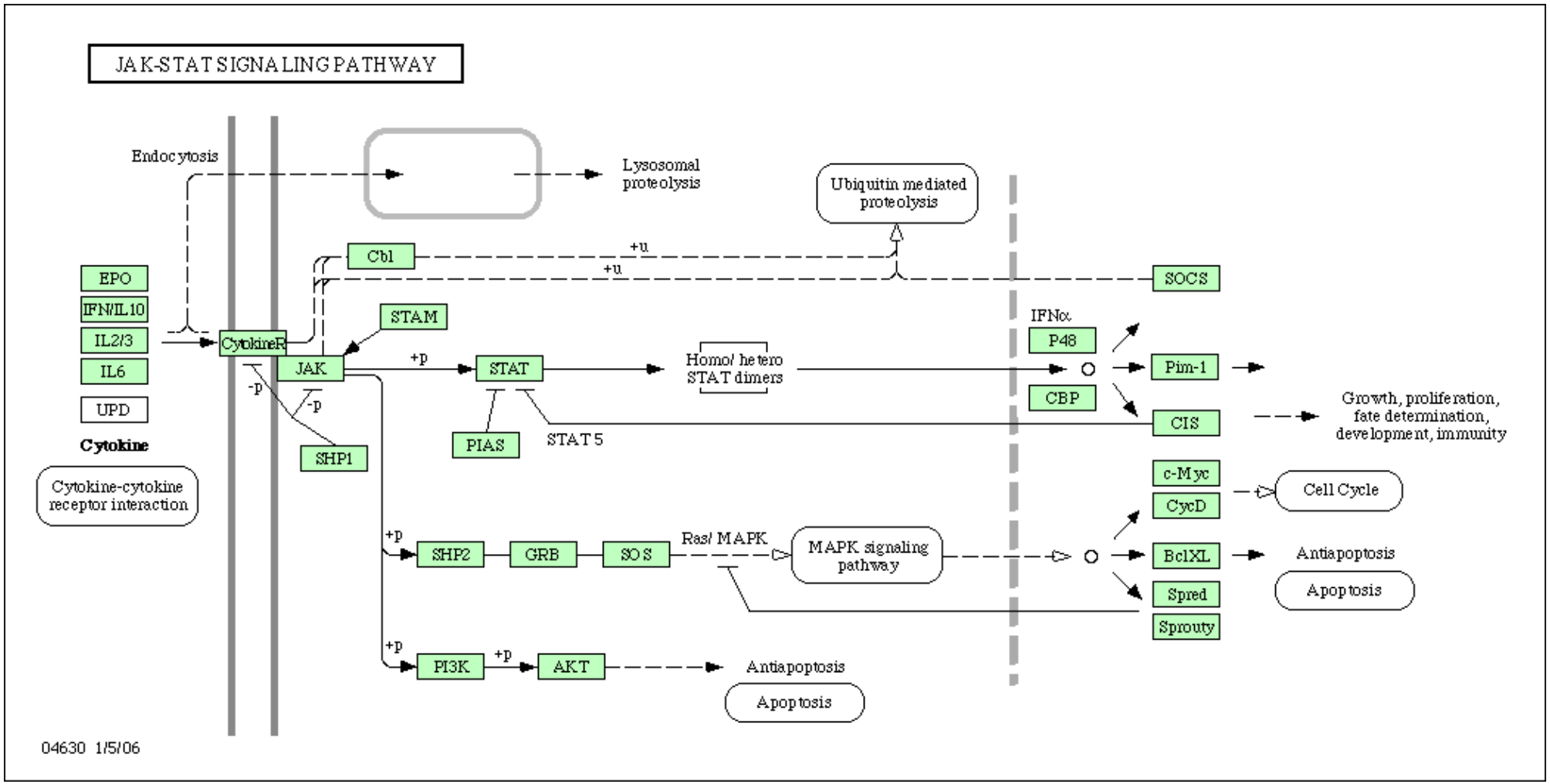
* Click on the different IDs for more information.
* The protein information displayed is the one corresponding to the Gene Name shown.
* Only upto 20 Clone Ids or GenBank Accessions are listed, to see the full list get your results on a spreadsheet file or follow the (...) link.
* If entering UniGene cluster identifiers, keep in mind that UniGene retires identifiers quite often. Click on the link on the UniGene column to know the new identifiers.

this application is part of the suite ASTERIAS

the rest of the program is logged

send comments to The Webmaster - last mod: August 31st, 2006

KEGG: hsa04630



Galaxy at Penn State University

Center for Comparative Genomics and Bioinformatics

Galaxy features connections to UCSC Table Browser, EnsMart, and contains hundreds of tools.

Some examples:

- **Extraction of multiple alignments corresponding to a genomic region,**
- **Finding exons overlapping SNPs, compute phastCons scores for a set of genomic ranges,**
- **Building histograms, compute correlations, draw scatterplots, search using simple conditional statements**
- **And much much more...**

Galaxy website

http://www.bx.psu.edu/cgi-bin/trac.cgi

Genomatix

http://www.genomatix.de/

DiAlign TF: Multiple alignment plus TF sites

GEMS Launcher Task: *DiAlign TF*: Multiple alignment plus TF sites
working on Bt_Erbb2_distal_promoter.seq, Hs_Erbb2_distal_promoter.seq, Mm_Erbb2_distal_promoter.seq, i~Cf_Erbb2_distal_promoter.seq, i~Md_Erbb2_distal_promoter.seq (5 seq.)

[Alignment] [Pairwise similarities]

DiAlign professional TF Release 3.1.1 September 2005 **Fri Aug 18 15:14:04 2006**

Solution parameters:

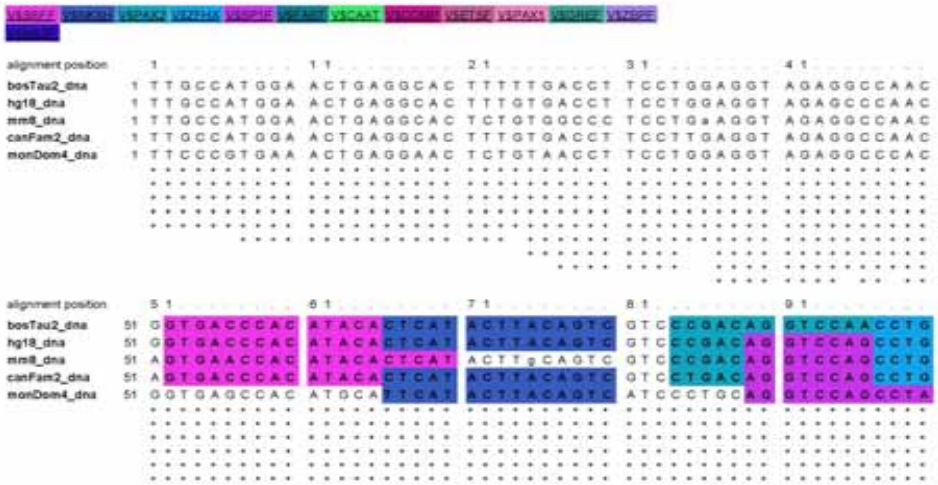
Sequence file:	Bt_Erbb2_distal_promoter.seq, Hs_Erbb2_distal_promoter.seq, Mm_Erbb2_distal_promoter.seq, i~Cf_Erbb2_distal_promoter.seq, i~Md_Erbb2_distal_promoter.seq (5 seq.)
Type of sequences:	nucleotide sequences
Threshold T:	0.00
"" signs below alignment denote:	diagonal similarity (max. similarity is represented by 10 "" signs)
Alignment output:	complete alignment is shown # nucleic acids per line: 50
TF output:	common TF matches located in aligned regions (common to 4 (80.0 %) sequences)
Family matches:	yes
MatInspector library:	Matrix Family Library Version 6.1 (June 2006)
Selected groups (core/matrix sim)	+ ALL vertebrates.lib (0.75/Optimized)

Aligned Sequences:

No.	Sequence Name	Sequence Description	Length
1	bosTau2_dna	bosTau2_dna range=chr19:32329406~32331721 5'pad=0 3'pad=0 revComp=FALSE strand=? repeatMasking=none	2316 bp
2	hg18_dna	hg18_dna range=chr17:35095702~35098580 5'pad=0 3'pad=0 revComp=FALSE strand=? repeatMasking=none	2879 bp
3	mm8_dna	mm8_dna range=chr11:98214976~98217649 5'pad=0 3'pad=0 revComp=FALSE strand=? repeatMasking=none	2674 bp
4	canFam2_dna	canFam2_dna range=chr9:28120323~28125000 5'pad=0 3'pad=0 revComp=FALSE strand=? repeatMasking=none|comm=reverse complement	2678 bp
5	monDom4_dna	monDom4_dna range=chr2:198108435~198125765 5'pad=0 3'pad=0 revComp=FALSE strand=? repeatMasking=none|comm=reverse complement	16331 bp

Alignment (DiAlign format):

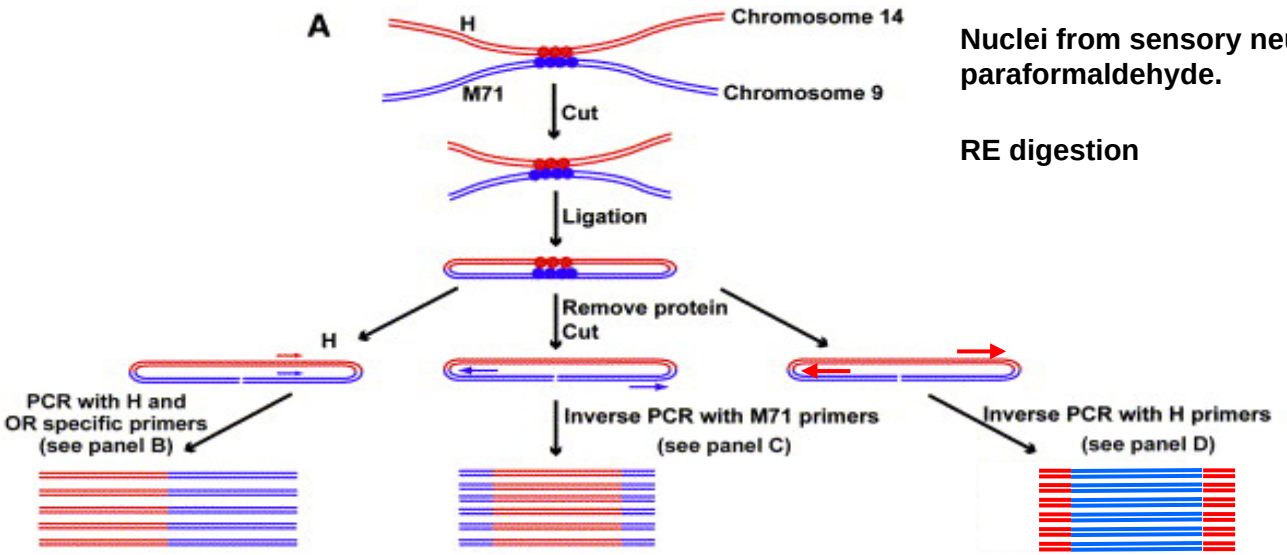
Please note that only **upper~case** letters are considered to be **aligned**. For more information, please have a look at the user guide.



S. Lomvardas, G. Barnea, D. J. Pisapia, M. Mendelsohn, J. Kirkland, and R. Axel. Interchromosomal interactions and olfactory receptor choice. *Cell* 126 (2):403-413, 2006.

F. Savarese and R. Grosschedl. Blurring cis and trans in gene regulation. *Cell* 126 (2):248-250, 2006.

Chromosome Conformation Capture (3C) Reveals the Association of the H Enhancer with Olfactory Receptor Genes

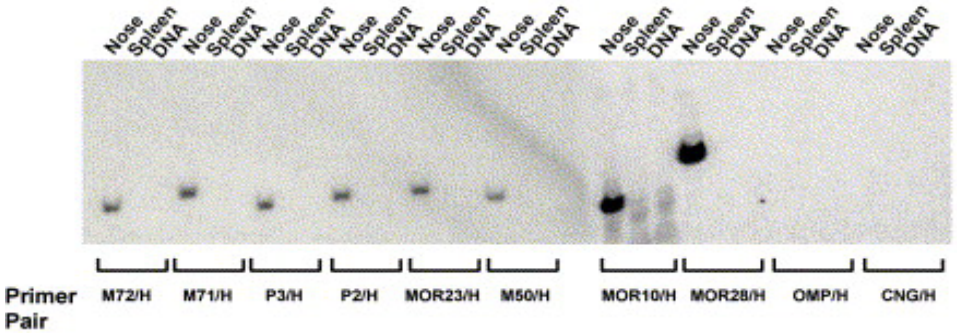


Nuclei from sensory neurons treated with paraformaldehyde.

RE digestion

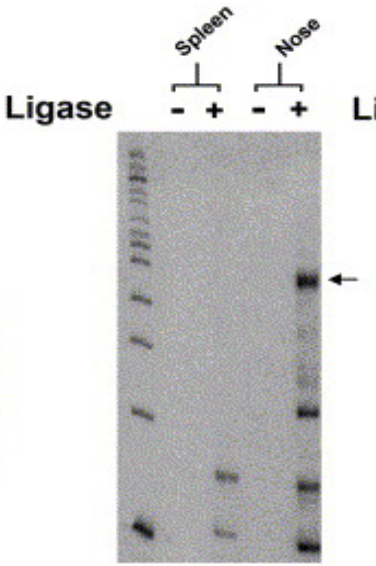
B

Chr. 14: *H, MOR28, MOR10*
Chr. 7: *Mor50, P6*
Chr. 9: *M71*



OMP and CNG: 2 genes highly expressed in olfactory sensory neurons

C



M71/H 2.3 kb product

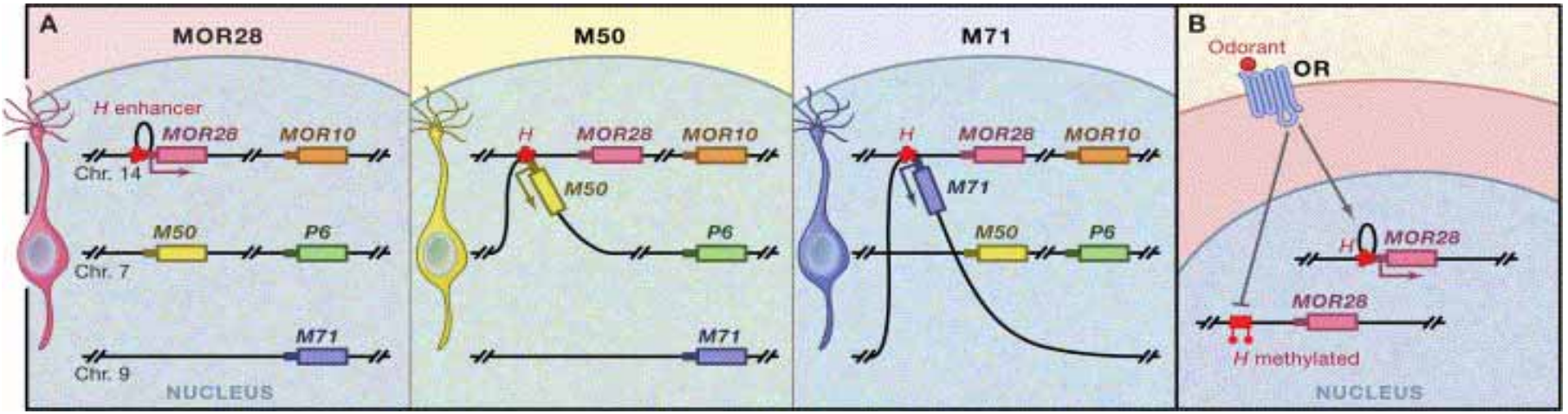
D

Products sequencing

320 PCR products:
- **84 *H/5'end OR genes***
- **236 *H/repeat elements***
- **48/84 *H/MOR48***
- **16/84 *H/MOR10***
- **20/84 *H/other OR genes***

From Lomvardas, S. et al. Cell, 126:403-413, 2006.

Monoallelic OR Gene Expression in Sensory Neurons



(A) Association of the *H* enhancer with a single OR promoter in cis or in trans mediates monoallelic OR gene expression in sensory neurons. Depicted are three different olfactory neurons in which the H enhancer, located on chromosome 14, interacts either with an OR gene in cis (for example, MOR28 on chromosome 14) or, alternatively, with an OR gene in trans (for example, M50 on chromosome 7 or M71 on chromosome 9). The enhancer is shown as a multiprotein complex, and the OR gene promoters are shown as colored boxes; active transcription is represented by an arrow.

(B) Model for the feedback regulation elicited by the expression of a functional OR gene. Upon interaction with an odorant, the OR transmits either a positive signal that strengthens the association between the *H* enhancer and the OR gene promoter or a negative signal that prevents the activation of the second H allele, possibly by DNA methylation of the H enhancer. Currently, it is unclear whether DNA methylation inactivates or activates the H enhancer.

The *H* enhancer, a single copy DNA element, is located 75 kb upstream of the *MOR28* gene cluster on chromosome 14. It can cooperate with OR gene promoter as little as 161 bp of 5'-flanking DNA sequences.

From Savarese, F. and Grosschedl, R. Cell, 126:248-250, 2006.

Few other published examples of physical and functional evidences for nonallelic interaction between chromosomes:

- **An LCR in IFN-γ locus associated with IL-4 locus on a different chromosome in committed naive T (*Spilianakis et al., Nature, 435:637-645, 2005*).**

- **The imprinting control region of the Igf2/H19 locus and the Wsb1/Nf1 gene (*Ling et al., Science, 312:269-272, 2006*).**

These nonallelic interchromosomal interactions appear relatively infrequent and transient, and their biological role is still somewhat unclear...

Regulatory region databases

PromoSer: The mammalian promoter service
http://biowulf.bu.edu/zlab/PromoSer

ORegAnno Open Regulatory Annotation:
http://www.oreganno.org

PAZAR: A public database of transcription factor and regulatory sequence annotation
http://www.pazar.info

Regulatory region analysis

PIPMaker:
http://pipmaker.bx.psu.edu/pipmaker

VISTA Tools: mVISTA and rVISTA
http://genome.lbl.gov/vista/

DCODE.org Comparative Genomics Center: Comparing genomes to decipher the code of gene regulation
http://www.dcode.org

Genomatix software GmbH:
http://www.genomatix.de

General databases and tools

UCSC Genome Browser:
http://genome.ucsc.edu

Galaxy website:
http://www.bx.psu.edu/cgi-bin/trac.cgi

IdConvert:

http://idconverter.bioinfo.cnio.es

En forme de conclusion…

Hypothesis-driven approach

- **Avoid fishing expeditions**
  - Do not expect to find a needle in a haystack
  - Do not expect the answer to pop out of the target gene list
  - Do not expect mechanistic insights

- **Envision specific outcomes and test them directly**
  - Use literature knowledge
  - Does target gene binding change between conditions, mutant backgrounds, or subunits?

- **Do not expect perfect outcomes**
  - Binding to a gene does not always correlate with expression level

From Julia Zeitlinger - Whitehead Institute, UC Davis, USA

INSERM UMR599 Cancer Research Center of Marseille - *Director : Françoise Birg*

Transcription Team

Michèle Algarté
Pierre Cauchy
Régis Costello
Brigitte Kahn-Perlès
Guoqiang Hua
Patrick Lécine
Carol Lipcey
Pascal Rameil
Frédéric Rosa
Jung-Hua Yeh
Bing Zhu
Jean Imbert

Immunology Team

Chantal Cerdan
Jacques Nunès
Daniel Olive

Gene & Cellular Therapy Team

Thomas Moreau
Christian Chabannon
Cécile Tonnelle

Oncology Team

José Adelaïde
Daniel Birnbaum

Immunology Center of Marseille-Luminy

Salvatore Spicuglia
Sanjeev Kumar
Pierre Ferrier

INSERM U363, Paris

Fabrice Gouilleux

Curie Institute, Orsay

Jacques Ghysdael

ISREC, Lausanne

Philip Bücher
Markus Nabholz

ICRF, London

Carol Beadling
Doreen Cantrell

Genzentrum, Martinsried

Patrick Baeuerle

TAGC, INSERM EMI206, Marseille

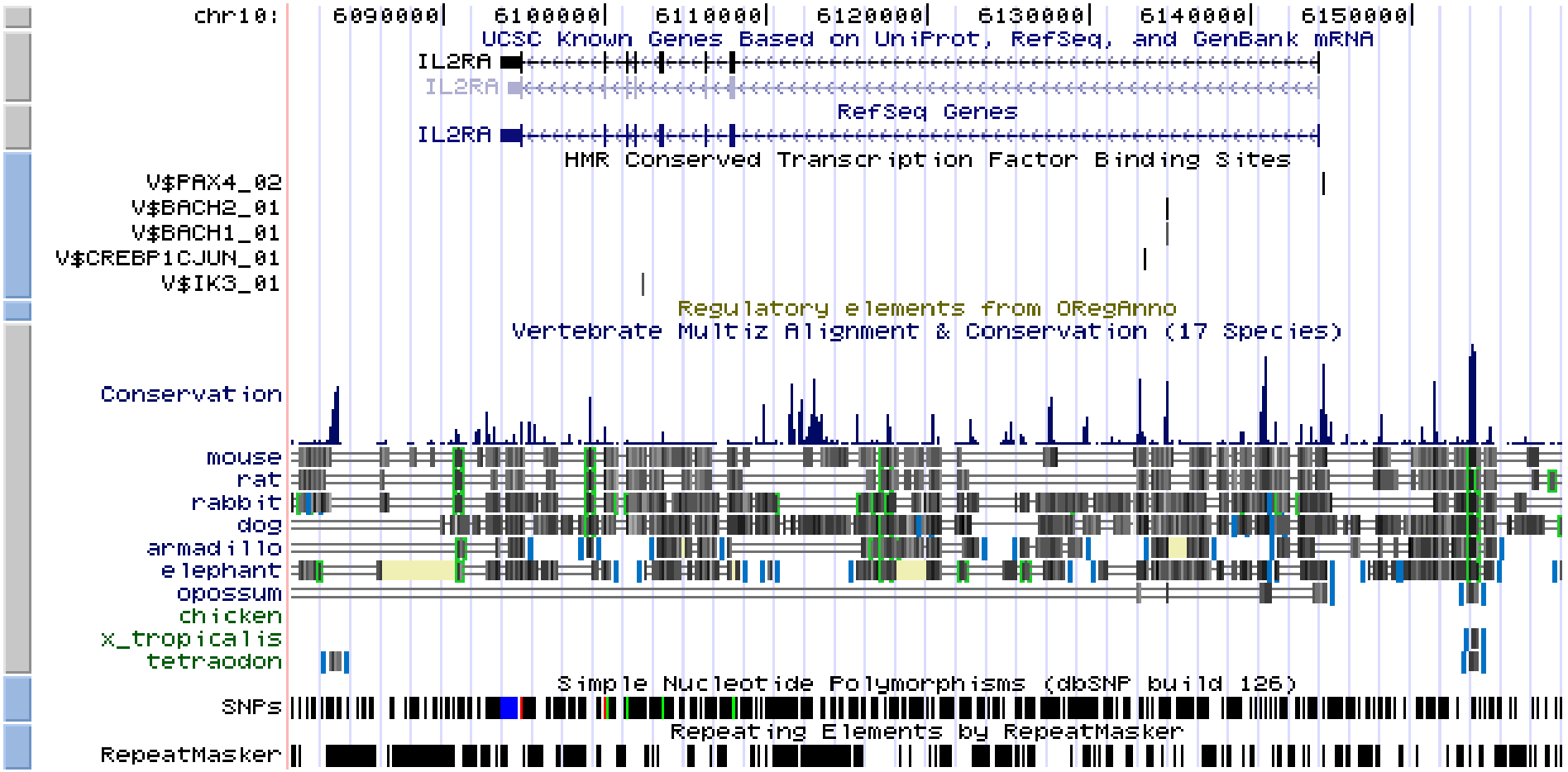
Béatrice Loriod
Benoît Ballester
Rémi Hougatte

NHLBI/NIH

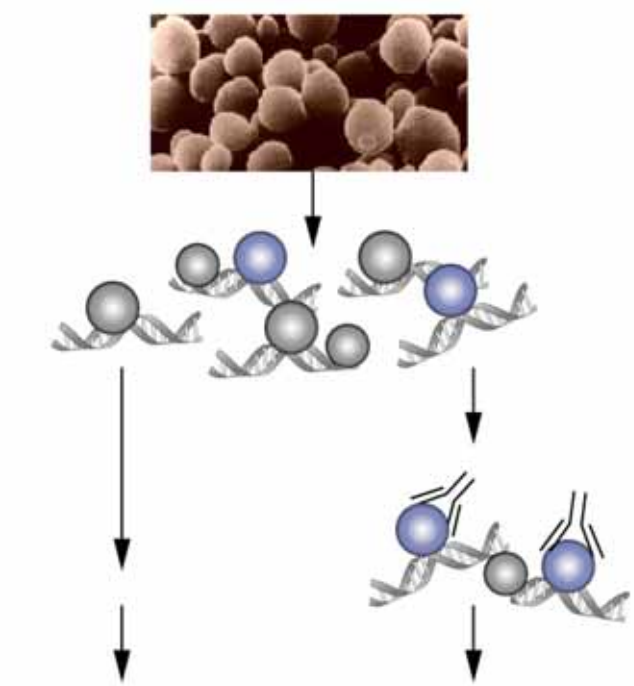
Hyoung Pio Kim
Warren Leonard

Special Bonus

UCSC Genome Browser on Human Mar. 2006 Assembly
position/search: chr10:6,080,616-6,159,203



Genomic Targets Identification of Specific Transcription Factors Using Chromatin Immunoprecipitation



Crosslink protein to binding sites in living cells

Harvest cells and fragment DNA

Enrich for protein-bound DNA fragments with antibodies

Gene-specific PCR ChIP

Cloning & Sequencing

SAGE-like GMAT/SACO

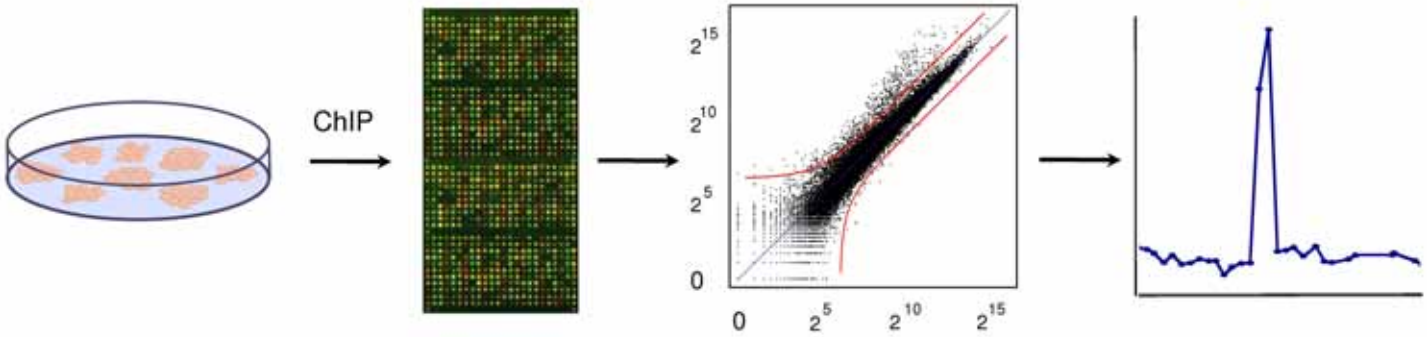
Microarray Hybridization ChIP-on-chip

Gene-specific approaches (low throughput)

Genome-wide approaches (high throughput)

ChIP-on-chip

ChIP-on-chip with whole-genome oligo arrays



Human cells | **Whole-genome oligo arrays** | **Scatter plot (ChIP / reference)** | **Identification of binding sites**

- **Whole-genome arrays**
  - 60mer every < 300 bp
  - Oligo selected by uniqueness, GC content, complexity, self-binding potential
- **Peak finding algorithm**
  - Based on expected peak curve
  - Voting of three consecutive probes

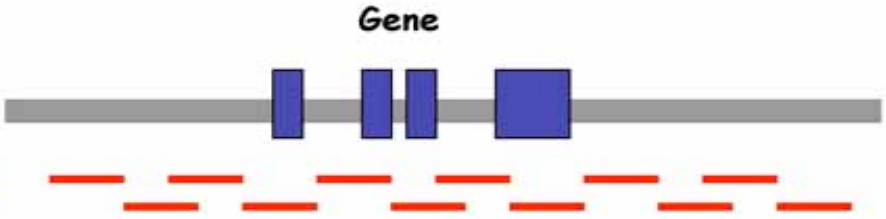
From Julia Zeitlinger - Whitehead Institute, UC Davis, USA

Strategies for the Design of Microarrays for the Human Genome

Single gene or Small Region

Horak et al. (2002) PNAS 99: 2924-2929

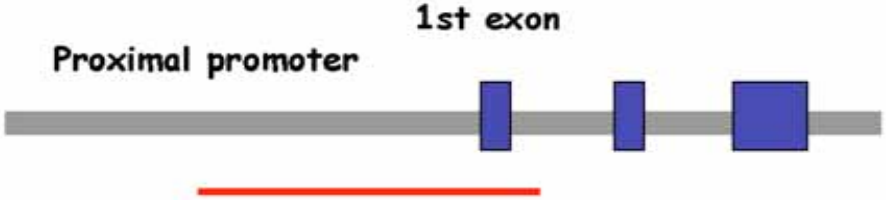
Overlapping 1kb PCR products
Tiling path microarray



Gene collection

(eg. Refseq collection)
Blais et al. (2005) Genes & Dev. 19: 1-17

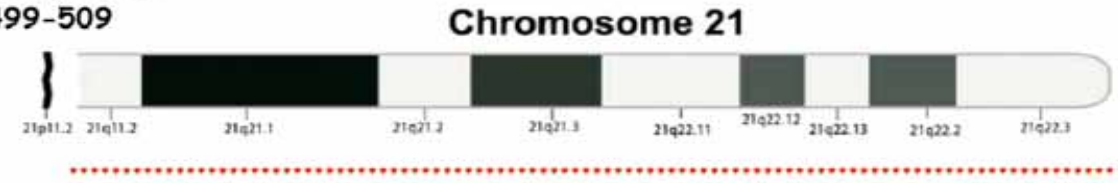
1kb PCR products



Small chromosomes

Kaparnov et al. (2002) Science 296: 916-919
Cawley et al. (2004) Cell 116:499-509

1 oligo each 35 bases



CpG islands

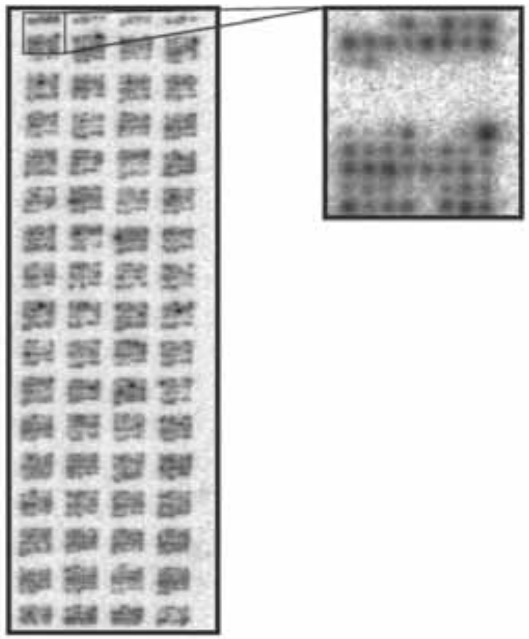
Weinmann et al. (2002) Genes & Dev. 16:235-244
Oberley et al. (2004) Methods Enzymol. 376:315-334

- CpG dinucleotides are present at 20% of predicted frequency
- CpG islands: >200 bp long, >50 %G+C, CpG >0.6 predicted
- CpG islands account for 1% of the genome
- 29,000 CpG islands are predicted in the human genome
- ~60% of known genes have a CpG island near 5' end
- CpG island microarrays are promoter- and regulatory region-enriched arrays

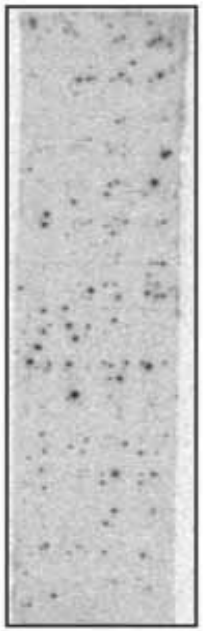
29,000 CpG islands are predicted in the human genome

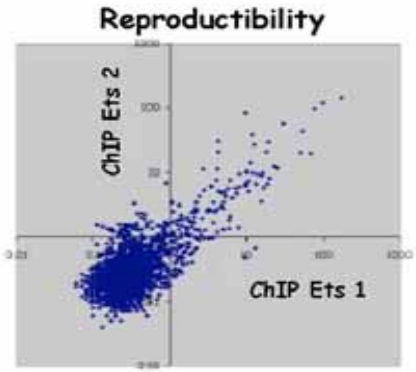
In progress analysis of a chip with ~200 human promoter regions and ~4 000 CpG islands

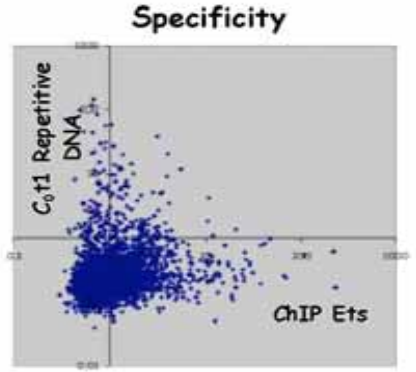
ChIP anti-Ets-1/2



Microarray CpG+3
200 selected genes
4.000 CpG clones



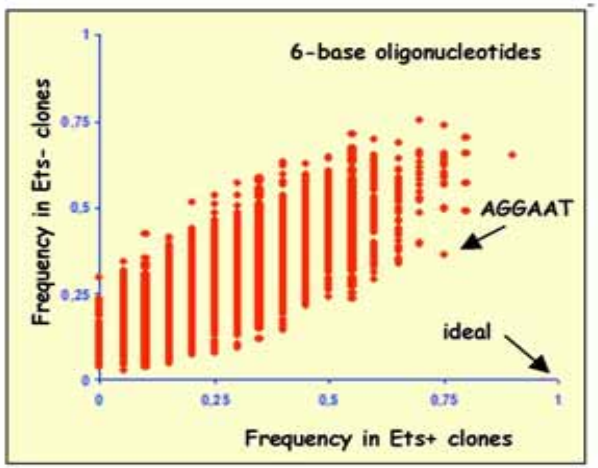




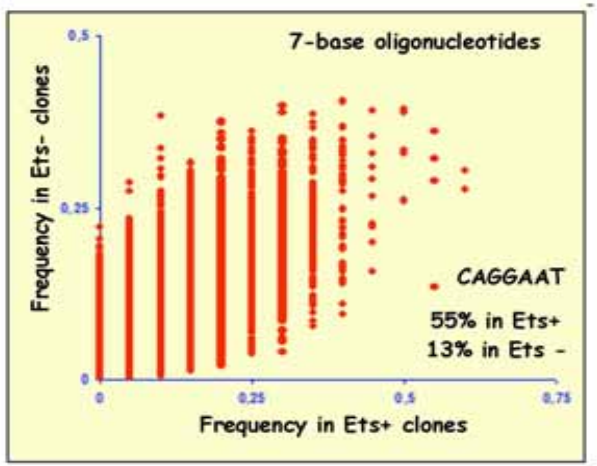
La Londe les Maures 12 May 2005. Rémi Houlgatte « Chromatin Profiling of Human Pathologies »

From Rémi Houlgatte – INSERM ERM206, Marseille, France

Ets Factor Binding Site (EBS)

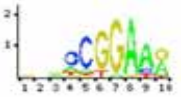
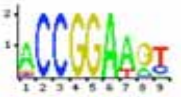


Expected frequency of a 6-base motif in 1.5 kb sequences: 37%



Expected frequency of a 7-base motif in 1.5 kb sequences: 9%

3rd best score using Gibbs sampling:

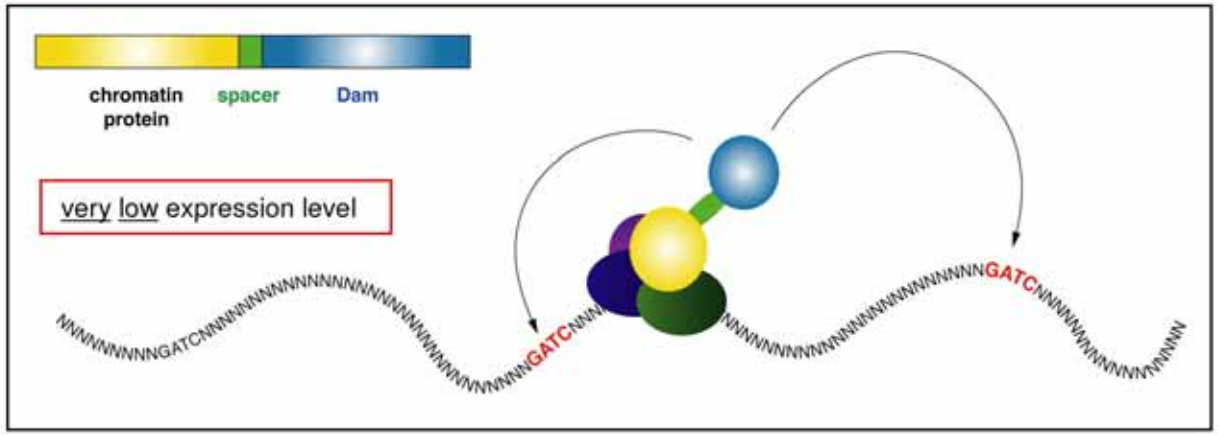
SPECIES	CLASS	LOGO
Homo sapiens	ETS	
Homo sapiens	ETS	
Homo sapiens	ETS	
Homo sapiens	ETS	
Homo sapiens	ETS	

La Londe les Maures 12 May 2005. Rémi Houlgatte « Chromatin Profiling of Human Pathologies »

From Rémi Houlgatte – INSERM ERM206, Marseille, France

Alternative Approaches…

DamID: in vivo mapping of protein-genome interactions



E. coli Dam methyltransferase:
- methylates A in sequence **GATC**
- active and non-toxic
- fairly small (32 kDa)

Adenine methylation (^{m6}A):
- absent in DNA of most eukaryotes

From Bas van Steensel, Netherland Cancer Institute

DamID versus ChIP-on-chip

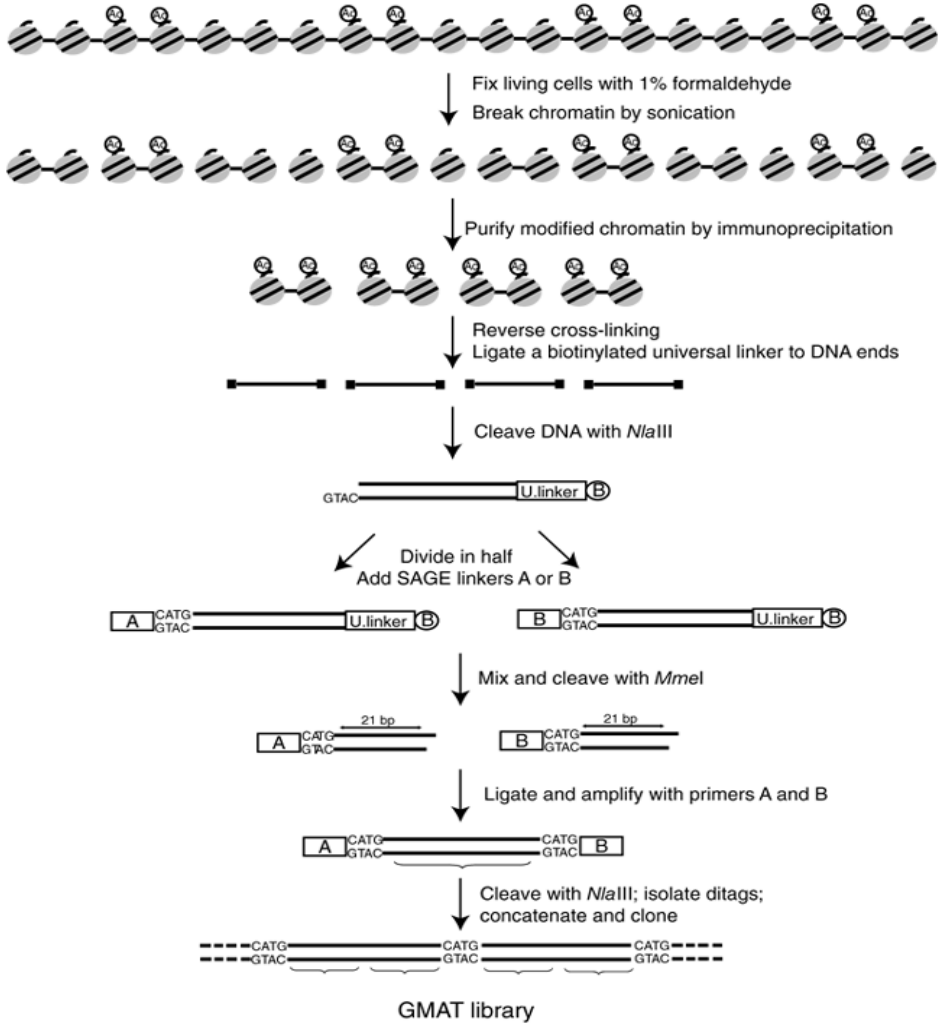
DamID	ChIP-on-chip
• No antibody required	• Highly specific antibody needed
• Works for >90% of all proteins tested	• Success rate?? – Depends on antibody, epitope exposure and "crosslinkability"
• Fusion protein may not be fully functional	• Crosslinking artefacts?
• Cannot map post-translational modifications	• Mapping of e.g. histone modifications
• Studies of mutated proteins	• More difficult, need epitope tagging and precisely controlled expression
• One fruitfly or one 35mm dish of cells is enough	• Amount of material required?
• Low time-resolution (hrs)	• High time-resolution (min)
• Mapping resolution ~1-2kb?? – may depend on local chromatin conformation	• Mapping resolution 0.5-1kb??? – may depend on local chromatin conformation

But: which one is "better" in practice?
More importantly: do the two methods give the same results?

From Bas van Steensel, Netherlands Cancer Institute, Amsterdam

Alternative approaches

Genome-Wide Mapping Technique (GMAT)
Or Serial Analysis of Chromatin Occupancy (SACO)



From T. Y. Roh, W. C. Ngau, K. Cui, D. Landsman, and K. Zhao. High-resolution genome-wide mapping of histone modifications. Nat Biotechnol. 22 (8):1013-1016, 2004.

Functional Effects of Histone Modifications on Transcription

A Few Examples in Mammals

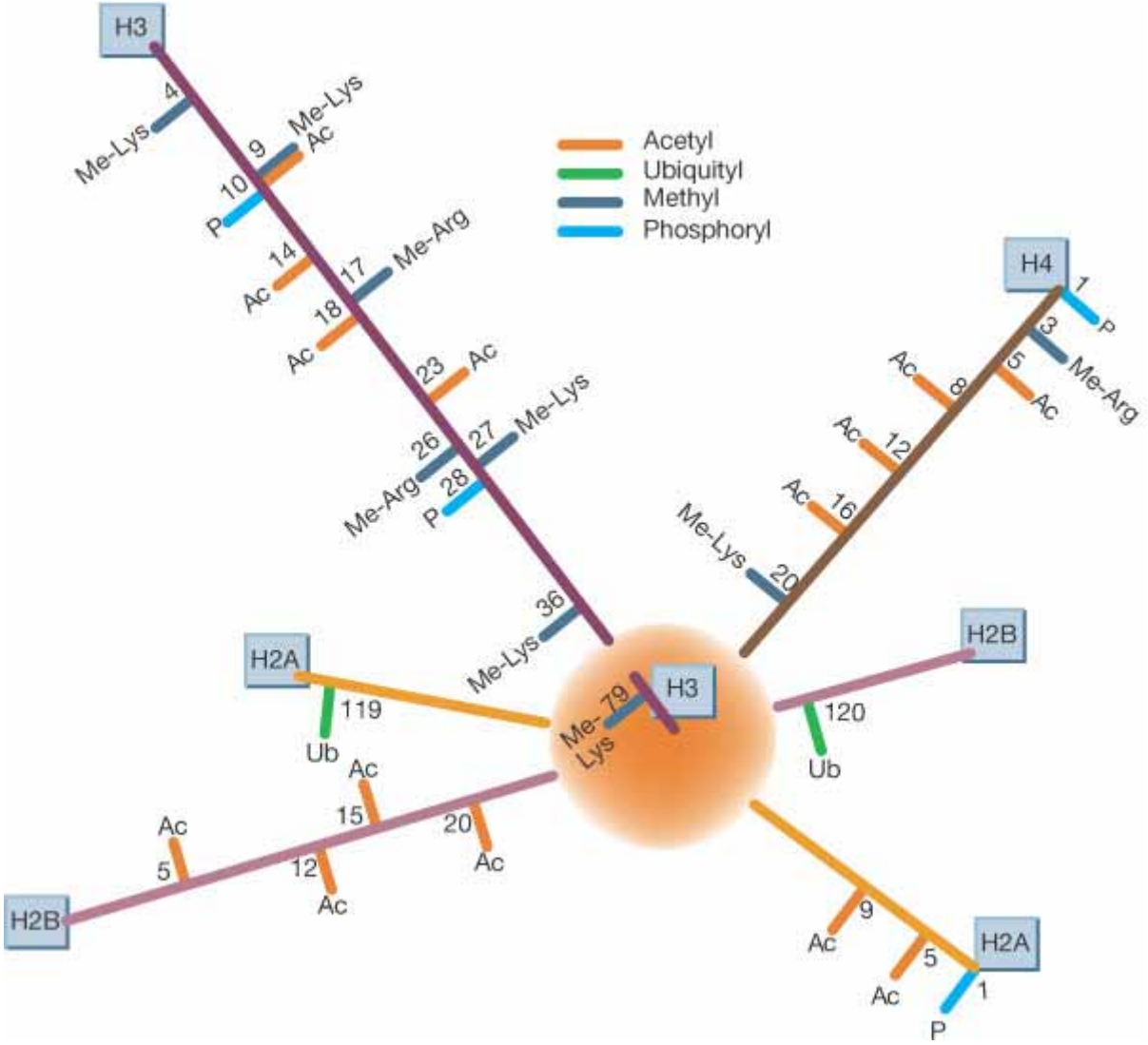
- Acetylation at lysine residues is highly associated with transcriptional **activation** (H2AK5, H2BK12, H3K9/K14, etc.)
- Methylation at lysine or arginine residues is associated with either transcriptional **activation** (H3K4me3, H4R3, etc.) or **repression** (H3K9, H3K27me3, etc.)
- Phosphorylation at serine or threonine residues is associated with either transcriptional **activation** (H3S28, etc.) or **repression** (H2AS1, etc.)

Ubiquitylation and sumoylation have been associated with mitosis, meiosis,etc.

For a continuous update: www.histone.com

From C. L. Peterson and M. A. Laniel. Histones and histone modifications. Curr.Biol. 14 (14):R546-R551, 2004.

Major Nucleosomal Histone Modification Mapping



From G. Felsenfeld and M. Groudine. Controlling the double helix. Nature 421 (6921):448-453, 2003.

Quelques infos utiles

Site web de l'atelier INSERM « Epigénomique : analyse à grande échelle des modifications de la chromatine et des interactions entre génomes et facteurs régulateurs de la transcription »
12-13 mai 2005, La Londe Les Maures et 16-20 mai 2005, Marseille
http://gin.univ-mrs.fr/~denis/Epigenomics/

Grange, T., Imbert, J. and Thieffry, D. Epigenomics: large scale analysis of chromatin modifications and transcription factors/genome interactions. *BioEssays* 27:1203-1205, 2005.

Site web « ChIP-on-chip » à partir des travaux de Bing Ren, Laboratory of Gene Regulation, Ludwig Institute for Cancer Research, UCSD
http://www.chiponchip.org/

Vous y trouverez des protocoles, une liste d'anticorps validés pour ChIP et ChIP-on-chip, des références, etc.

Site web Ecole-Chercheurs INRA « La phylogénomique : une aide à l'étude des grandes fonctions du vivant »
12-14 décembre 2006, Carry-le-Rouet
http://www.inra.fr/internet/Projets/agroBI/PHYLO.html

Resources

Companies offering genome-wide or customize oligo and/or promoter chips

1) **Agilent: www.home.agilent.com**

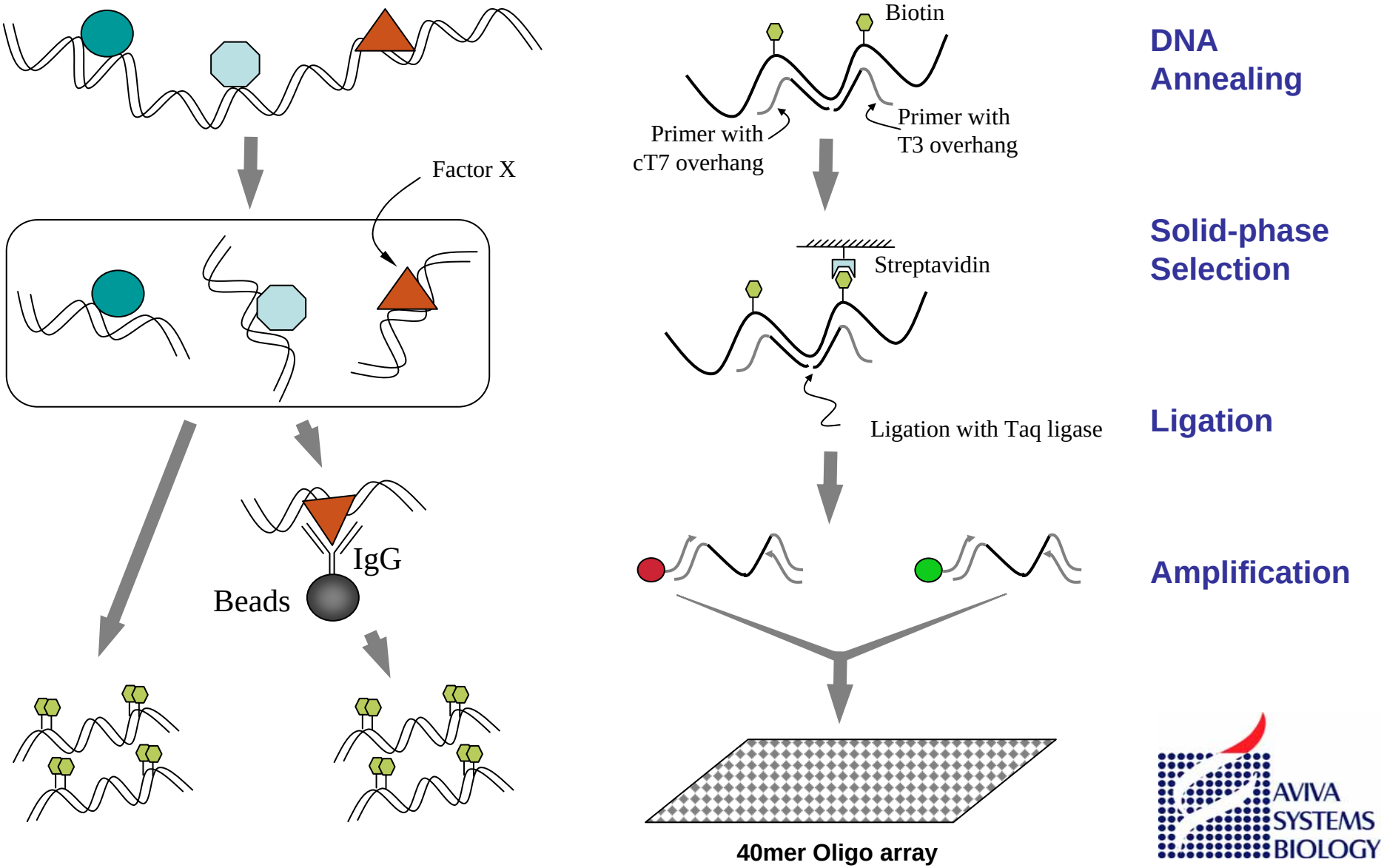
2) **Affymetrix: www.affymetrix.com**

3) **Nimblegen Systems: www.nimblegen.com**

4) **Aviva Systems Biology: www.avivasysbio.com**

5) **human and mouse CpG islands:**
Microarray Centre, University Health Network, Toronto, www.microarrays.ca

The ChIP-GLAS Technology



articles Nature 434 (7031):338-345, 2005.

Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals

Xiaohui Xie[1], Jun Lu[1], E. J. Kulbokas[1], Todd R. Golub[1], Vamsi Mootha[1], Kerstin Lindblad-Toh[1], Eric S. Lander[1,2*] & Manolis Kellis[1,3*]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA
[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

*These authors contributed equally to this work

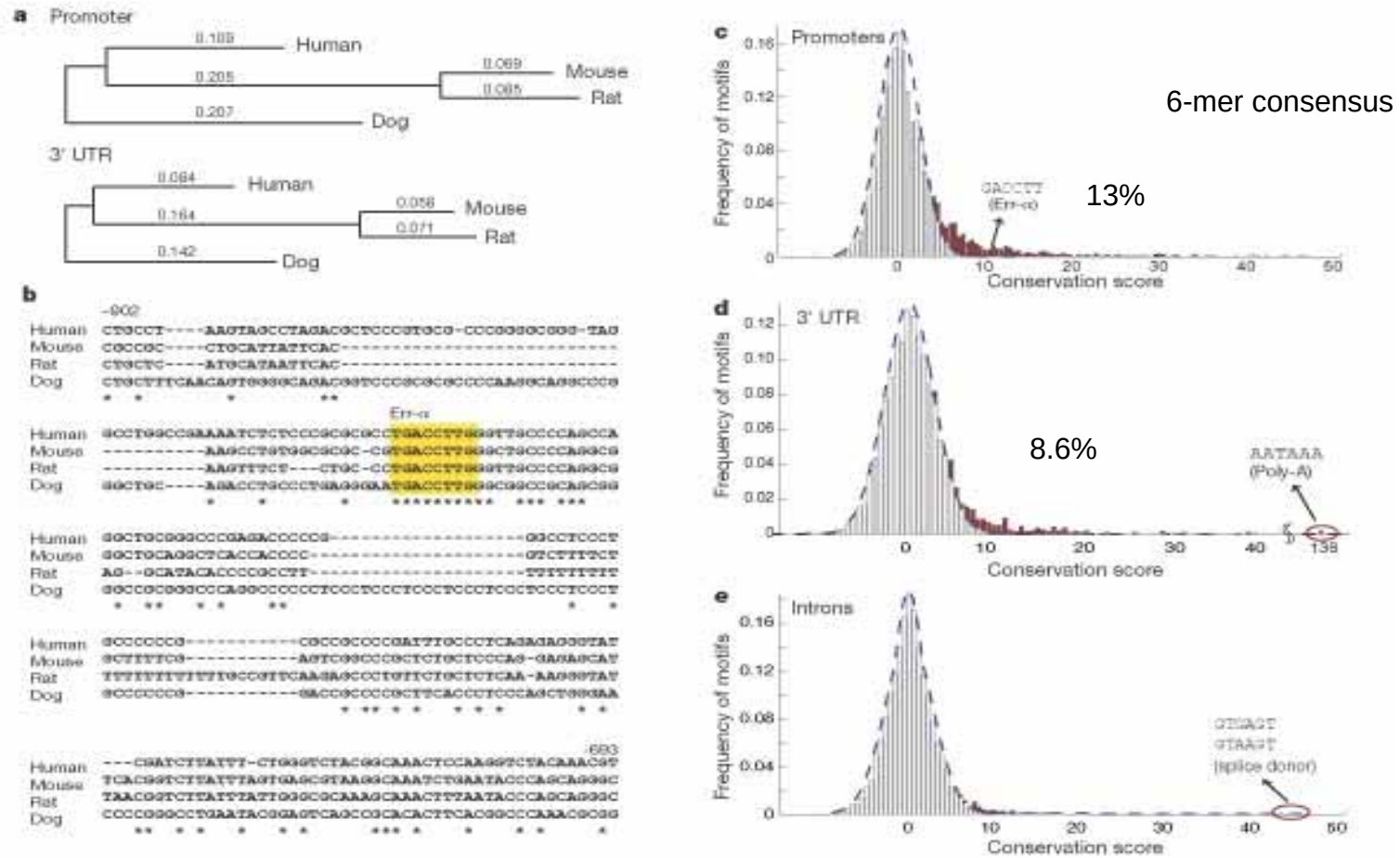
Comprehensive identification of all functional elements encoded in the human genome is a fundamental need in biomedical research. Here, we present a comparative analysis of the human, mouse, rat and dog genomes to create a systematic catalogue of common regulatory motifs in promoters and 3′ untranslated regions (3′ UTRs). The promoter analysis yields 174 candidate motifs, including most previously known transcription-factor binding sites and 105 new motifs. The 3′-UTR analysis yields 106 motifs likely to be involved in post-transcriptional regulation. Nearly one-half are associated with microRNAs (miRNAs), leading to the discovery of many new miRNA genes and their likely target genes. Our results suggest that previous estimates of the number of human miRNA genes were low, and that miRNAs regulate at least 20% of human genes. The overall results provide a systematic view of gene regulation in the human, which will be refined as additional mammalian genomes become available.

1. Whole genome alignment of 4 mammalian genomes: human, mouse, rat, dog

2. Extraction aligned promoter and 3’ UTR using RefSeq annotations

    17 700 Refseq
    68 Mb promoter dataset
    15 Mb 3’ UTR
    123 Mb intronic sequence as control dataset: last 2 introns

Proportion of aligned bases between 4 species:
    Promoters 51% (44% upstream, 58% downstream TSS)
    3’ UTR 73%
    Control dataset 34%
    Whole genome 28%



8-mer TGACCTTG: Err-α binding site (estrogen-related receptor alpha)

434 occurrences in human promoter regions, 162 conserved across all 4 species

37% conservation rate compared to 8% for a random 8-mer motif in promoter dataset and 6.2% in control dataset

MCS: motif conservation score, represents the number of standard deviations (s.d.) by which the observed conserved rate of a motif exceeds the expected conservation rate of comparable motifs (Ex.: MCS for Err-α is 25.2 s.d.

446 mammalian motifs in TRANSFAC clustered in 123 motif clusters based on sequence similarity

63% MCS>3, including 50% MCS>5, compared to only 1.6% with MCS>3 control set

3' UTR few motifs defined: polyadenylation signal AATAAA, 6 617 conservation versus 14 266 occurrences

46%, MCS 135 compared to 10% for a control random motif in 3' UTR and 6% in intronic controls

Screening for **H**ighly **C**onserved 6-18-mer **M**otifs with MCS>6 (10^{12} possible occurrences)

Promoters: **174** HCM including 59 strong matches with known consensus and 10 weaker
72% of the 123 TRANSFAC clusters (5% matches for a random set)
105 potential new regulatory motifs, for example M_4 (ACTAYRNNNCCCR)
occurs 520 times, 317 conserved across 4 species (61%)

Table 1 **Top 60 of 174 discovered motifs in human promoters**

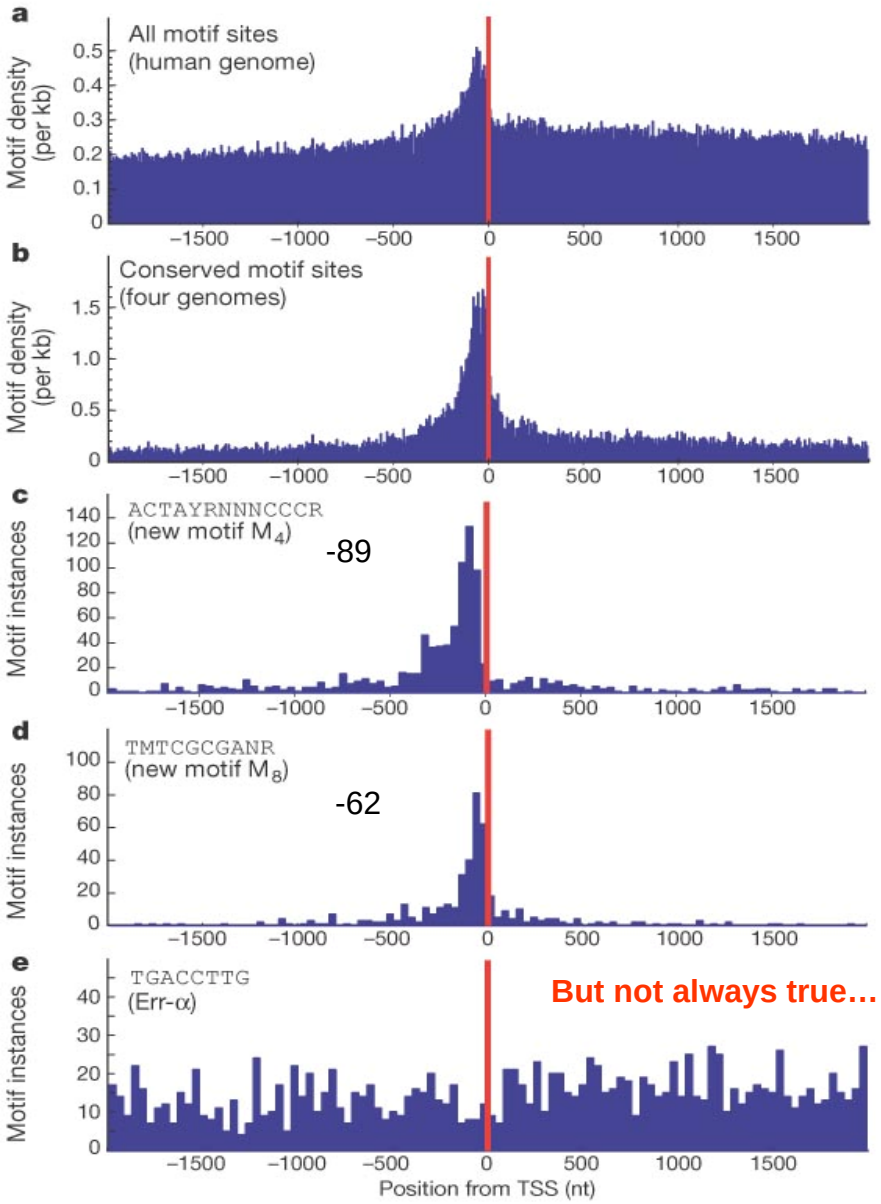
No.	Discovered motif	MCS	Known factor*	Conservation rate†	Tissue enrichment‡	Position bias§
1	RCGCAnGCGY	107.8	NRF-1	0.49	15.0	−62
2	CACGTG	85.3	MYC	0.47	8.8	−62
3	SCGGAAGY	80.4	ELK-1	0.44	22.4	−24
4	ACTAYRnnnCCCR	69.5	–	0.61	8.1	−89
5	GATTGGY	64.6	NF-Y	0.51	9.8	−63
6	GGGCGGR	63.9	SP1	0.21	11.4	−63
7	TGAnTCA	62.8	AP-1	0.38	6.5	–
8	TMTCGCGAnR	55.7	–	0.64	9.4	−62
9	TGAYRTCA	55.7	ATF3	0.50	6.1	−66
10	GCCATnTTG	54.7	YY1	0.72	12.2	–
11	MGGAAGTG	51.6	GABP	0.43	13.9	−23
12	CAGGTG	47.6	E12	0.26	9.9	–
13	CTTTGT	46.0	LEF1	0.42	13.6	–
14	TGACGTCA	44.8	ATF3	0.44	4.2	−22
15	CAGCTG	43.9	AP-4	0.27	8.9	–
16	RYTTCCTG	43.0	C-ETS-2	0.32	7.4	−24
17	AACTTT	42.1	*IRF1*	0.43	11.1	–
18	TCAnnTGAY	40.4	SREBP-1	0.47	4.9	−64
19	GKCGCn(7)TGAYG	40.1	–	0.35	5.6	−62
20	GTGACGY	38.4	E4F1	0.34	6.6	−56
21	GGAAnCGGAAnY	37.7	–	0.68	7.0	−33
22	TGCGCAnK	37.4	–	0.24	8.2	−17
23	TAATTA	37.3	CHX10	0.29	7.1	–
24	GGGAGGRR	33.5	MAZ	0.16	9.4	–
25	TGACCTY	33.4	ESRRA	0.30	7.7	–
26	TTAYRTAA	32.6	E4BP4	0.34	6.1	–
27	TGGn(6)KCCAR	32.3	–	0.27	4.5	–
28	CTAWWWATA	32.3	RSRFC4	0.36	7.6	–
29	CTTTAAR	30.8	–	0.43	5.4	–
30	YGCGYRCGC	30.5	–	0.19	5.2	−31
31	GGGYGTGnY	30.0	–	0.24	5.4	−63
32	TGASTMAGC	27.2	NF-E2	0.39	5.4	−66
33	YTATTTTnR	26.4	MEF-2	0.21	7.1	–
34	CYTAGCAAY	26.1	–	0.50	5.2	−142
35	GCAnCTGnY	25.7	MYOD	0.25	8.2	–
36	RTAAACA	25.6	FREAC-2	0.46	7.0	–
37	GTTRYCATRR	25.3	–	0.54	7.6	−56
38	TGACCTTG	25.2	ERR-α	0.37	8.1	–
39	TCCCRnnRTGC	24.3	–	0.30	6.8	−60
40	TTCYnRGAA	24.3	STAT5A	0.19	–	–
41	TGACAGnY	24.1	MEIS1	0.27	6.9	–
42	TGACATY	23.8	–	0.23	5.8	–
43	GTTGnYnnRGnAAC	23.7	–	0.47	4.7	−57
44	YATGnWAAT	23.5	OCT-X	0.53	6.9	–
45	CCAnnAGRKGGC	23.4	–	0.47	–	−101
46	WTTGKCTG	23.0	–	0.25	5.0	−63
47	TGCCAAR	22.9	NF-1	0.25	7.0	–
48	GCGnnAnTTCC	22.8	*C-REL*	0.30	6.0	−12
49	CATTGTYY	22.5	SOX-9	0.43	5.8	–
50	RGAGGAARY	22.4	PU.1	0.22	4.0	–

*Name of the best-matching motif in TRANSFAC database, if any. Weak matches are indicated in italics.
†The percentage of human motif occurrences that match the motif consensus across all four species.
‡A measure of the maximum enrichment of conserved motif occurrences upstream of genes expressed in a compendium of 75 human tissues.
§For motifs with strong positional bias (score > 5 s.d.), the mode of the distance distribution upstream of the TSS.

Biological significance

2. Positional bias of the motif relative to TSS: consistent with a role in transcription initiation
   28% for known motifs and 35% for new

Overall 89% of known and 69% of new motifs show tissue specificity, positional bias or both



3’ UTR

1. Strong directional bias (mRNA regulatory elements)

2. Unusual length distribution (8-mer preference)

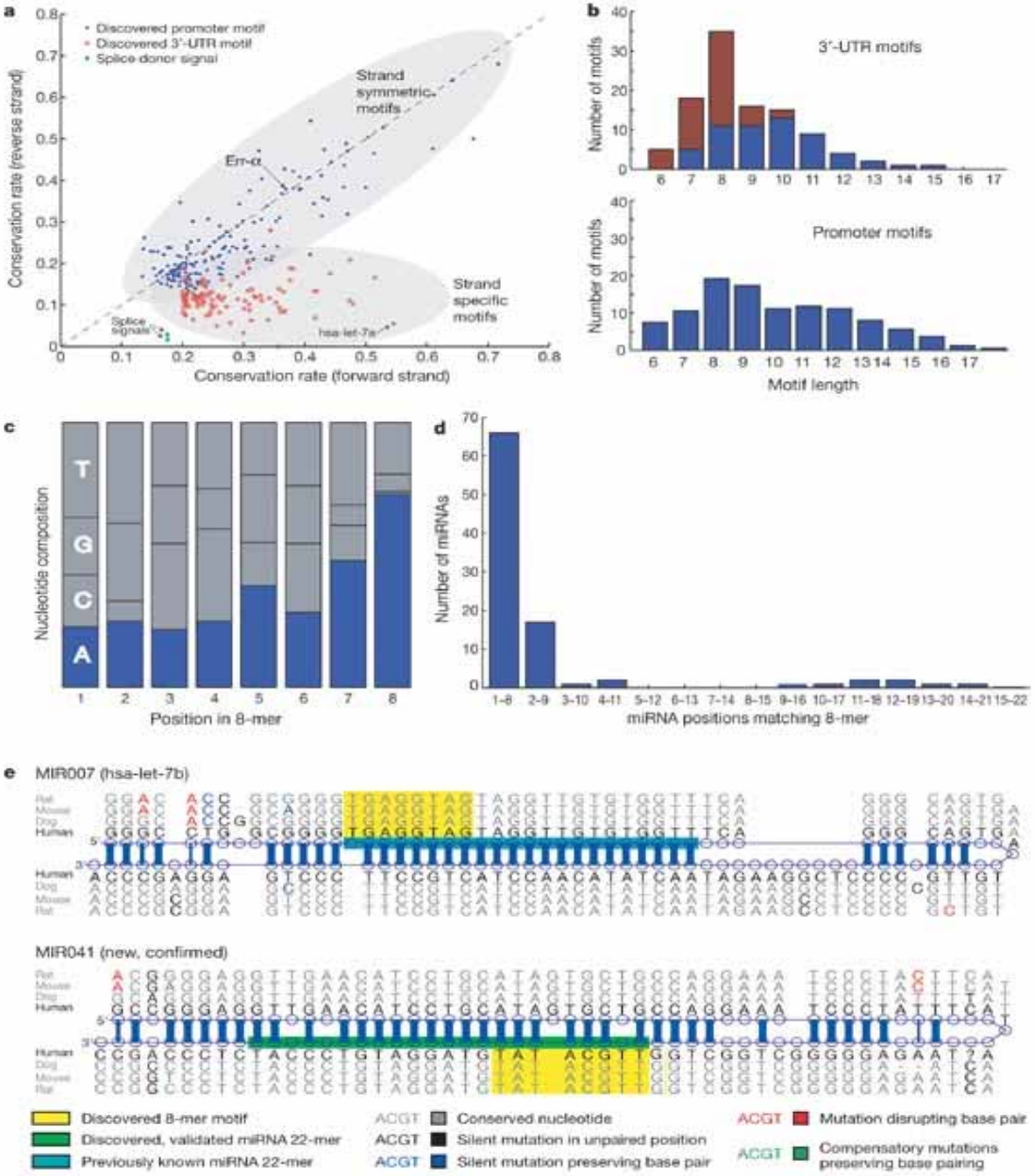

a
Discovered promoter motif
Discovered 3′-UTR motif
Splice donor signal
Strand symmetric motifs
Em-α
Strand specific motifs
Splice signals
hsa-let-7a
Conservation rate (reverse strand)
Conservation rate (forward strand)
b
3′-UTR motifs
Promoter motifs
Number of motifs
Motif length
c
Nucleotide composition
T
G
C
A
Position in 8-mer
d
Number of miRNAs
1–8 2–9 3–10 4–11 5–12 6–13 7–14 8–15 9–16 10–17 11–18 12–19 13–20 14–21 15–22
miRNA positions matching 8-mer
e
MIR007 (hsa-let-7b)
MIR041 (new, confirmed)
Rat
Mouse
Dog
Human
Discovered 8-mer motif
Discovered, validated miRNA 22-mer
Previously known miRNA 22-mer
ACGT Conserved nucleotide
ACGT Silent mutation in unpaired position
ACGT Silent mutation preserving base pair
ACGT Mutation disrupting base pair
ACGT Compensatory mutations preserving base pairing

Table 2 **Top 50 conserved 8-mers in 3′ UTRs and corresponding miRNAs**

No.	Motif	Conservation rate	MiRNA*
1	GTGCAATA	0.55	miR-92, miR-32, miR-137, miR-367, miR-25, miR-217(†), new(12)
2	GTGCCTTA	0.54	miR-124a, miR-224(†), miR-208(†), miR-34b(†), miR-9*(†), miR-34c(†), miR-330(†), new(6)
3	CTACCTCA	0.53	miR-98, let-7i, let-7g, let-7f, let-7e, let-7c, let-7b, let-7a, let-7d, miR-196b, miR-196a, new(4)
4	ACCAAAGA	0.49	miR-9, new(11)
5	TGTTTACA	0.48	miR-30e-5p, miR-30d, miR-30c, miR-30b, miR-30a-5p, new(4)
6	GCACTTTA	0.48	miR-20, miR-106b, miR-18(†), miR-93, miR-372, miR-17-5p, miR-106a, miR-302d, miR-302c, miR-302b, miR-302a, miR-373, new(4)
7	TGGTGCTA	0.43	miR-29c, miR-29b, miR-29a, miR-107(†), miR-103(†), new(6)
8	CTATGCAA	0.42	miR-153, new(9)
9	TACTTGAA	0.42	miR-26b, miR-26a, new(4)
10	CGCAAAAA	0.42	New(2)
11	GTGCCAAA	0.41	miR-96, miR-182, miR-183, new(16)
12	GTACTGTA	0.40	miR-101, miR-199a‡, miR-144, new(2)
13	ATACGGGT	0.40	miR-99a, miR-100, miR-99b(†)
14	AAGCACAA	0.40	miR-218, new(8)
15	TTTGCACT	0.37	miR-19b, miR-19a, miR-301, miR-130b, miR-130a, miR-152, miR-148b, miR-148a, miR-139, new(10)
16	TGTACATA	0.36	–
17	AAGCCATA	0.35	miR-135b, miR-135a
18	ACTGTGAA	0.35	miR-27b, miR-27a, miR-128b, miR-128a, miR-23b(†), miR-23a(†), new(5)
19	AGACAATC	0.33	miR-219, new(2)
20	TGCTGCTA	0.33	miR-195, miR-16, miR-15b, miR-15a, miR-338(†), miR-424, new(5)
21	TTTTGTAC	0.32	New(1)
22	ACATTCCA	0.32	miR-206, miR-1, miR-122a(†)
23	TGAATGTA	0.31	miR-181b(†), miR-181c, miR-181a
24	ACGGTACA	0.30	–
25	CAGTATTA	0.30	miR-200c, miR-200b, new(1)
26	TTGCATGT	0.29	New(7)
27	TCGCATGA	0.29	New(1)
28	CTCAGGGA	0.29	miR-125b, miR-125a, new(6)
29	CAAGTGCC	0.28	New(2)
30	ACTACTGA	0.28	–
31	TGGACCAA	0.28	miR-133b, miR-133a, new(3)
32	GTAAATAG	0.28	New(1)
33	TGTAGATA	0.28	–
34	ACACTACA	0.27	miR-142-3p, new(3)
35	GTACAGTT	0.26	New(1)
36	CACCAGCA	0.26	miR-138(†), new(4)
37	GGTACGAA	0.25	miR-126(†)
38	TGTATAGT	0.24	miR-381
39	AAGGGCTA	0.24	New(1)
40	AGCTTTAA	0.24	New(1)
41	ATTTATCG	0.23	–
42	GGCAGCTA	0.23	miR-22(†), new(1)
43	GCTGTAAA	0.23	New(4)
44	GCACTAAT	0.22	–
45	AAAGGTGC	0.22	–
46	ATGTAGCA	0.22	miR-221(†), miR-222(†)
47	ACACTGGA	0.21	miR-199b(†), miR-199a(†), miR-145(†), new(2)
48	GTATATAG	0.21	–
49	TTTGATAA	0.21	miR-361(†)
50	AAGCATGC	0.21	New(1)

Top 50 of 72 highly conserved 8-mer motifs in 3′ UTRs and their corresponding miRNAs.
*Known human miRNAs matching complements of each 8-mer and its variants grouped in the same cluster. The dagger symbol (in parentheses) indicates miRNAs with one mismatch. When multiple new miRNAs are discovered using the 8-mer motifs as seeds, their number is indicated in parentheses. (See Supplementary Information for full alignments of known and discovered miRNAs.)
‡The miRNA mature product comes from the 3′ arm of the stem loop.