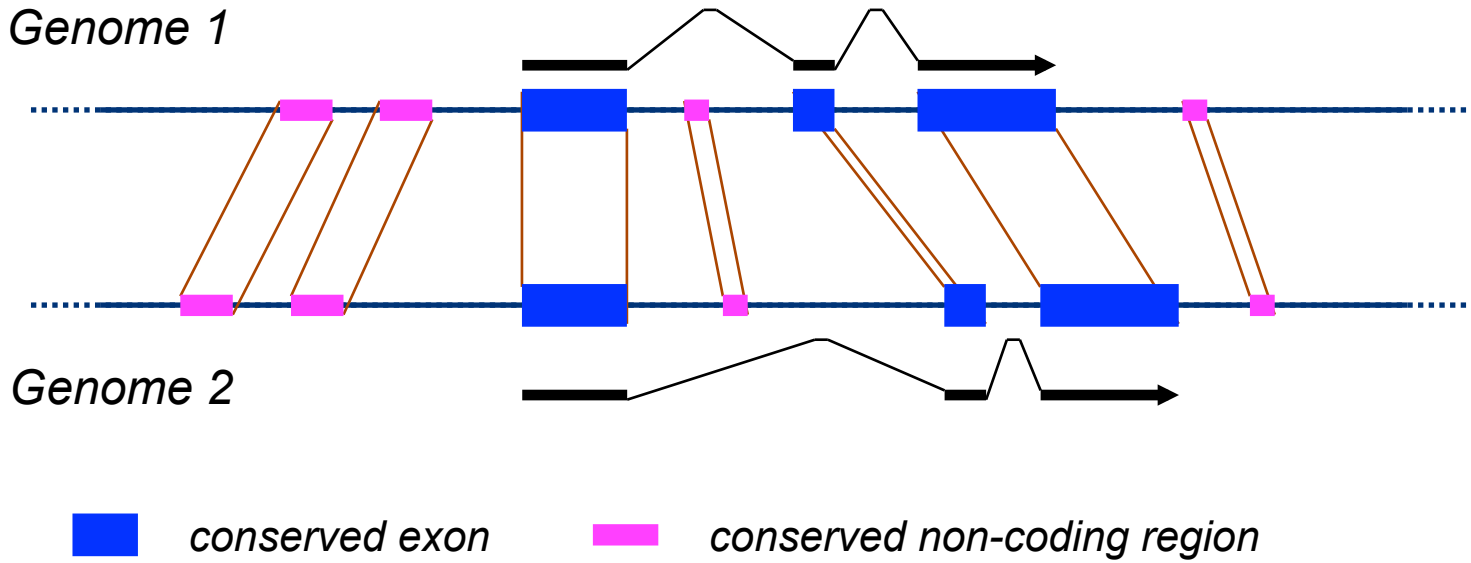
Applying comparative genomics to detect cis-acting elements

Phylogenetic footprinting to define regulatory regions

- Within non-coding sequences, regulatory elements evolve slower than their surrounding.
- Conserved non-coding sequences conserved regions, which can reveal cis-acting elements.
- Those elements were called “Phylogenetic footprints” by reference to the “DNAse footprints” method used by molecular biologists to reveal the sequence of transcription factor binding site.



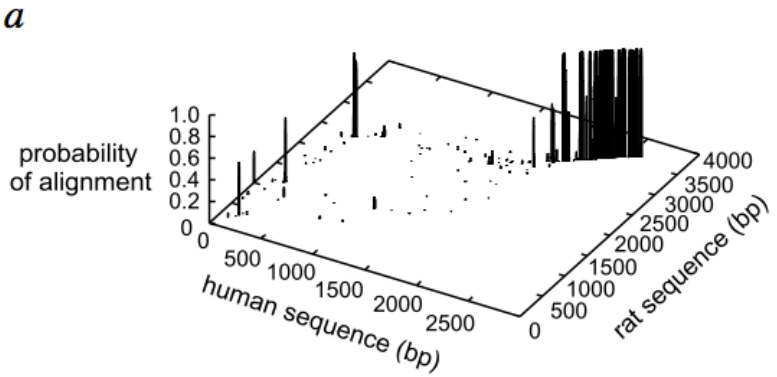
- Tagle et al. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol (1988) vol. 203 (2) pp. 439-55
- Wasserman, W. W. & Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278, 167-81.

- A 2-species comparison of the genomic region surrounding the first exon of natriuretic propeptide (NPPA) gene revealed conservation of almost all known binding sites for 3 muscle-specific transcription factors (MYF, SRF and MEF2).

Human-mouse genome comparisons to locate regulatory sites

Wyeth W. Wasserman[1,3], Michael Palumbo[2], William Thompson[2], James W. Fickett[1] & Charles E. Lawrence[2]

Elucidating the human transcriptional regulatory network[1] is a challenge of the post-genomic era. Technical progress so far is impressive, including detailed understanding of regulatory mechanisms for at least a few genes in multicellular organisms[2–4], rapid and precise localization of regulatory regions within extensive regions of DNA by means of cross-species comparison[5–7], and *de novo* determination of transcription-factor binding specificities from large-scale yeast expression data[8]. Here we address two problems involved in extending these results to the human genome: first, it has been unclear how many model organism genomes will be needed to delineate most regulatory regions; and second, the discovery of transcription-factor binding sites (response elements) from expression data has not yet been generalized from single-celled organisms to multicellular organisms. We found that 98% (74/75) of experimentally defined sequence-specific binding sites of skeletal-muscle-specific transcription factors are confined to the 19% of human sequences that are most conserved in the orthologous rodent sequences. Also we found that in using this restriction, the binding specificities of all three major muscle-specific transcription factors (MYF, SRF and MEF2) can be computationally identified.



- Wasserman et al. Human-mouse genome comparisons to locate regulatory sites. Nat Genet (2000) vol. 26 (2) pp. 225-8

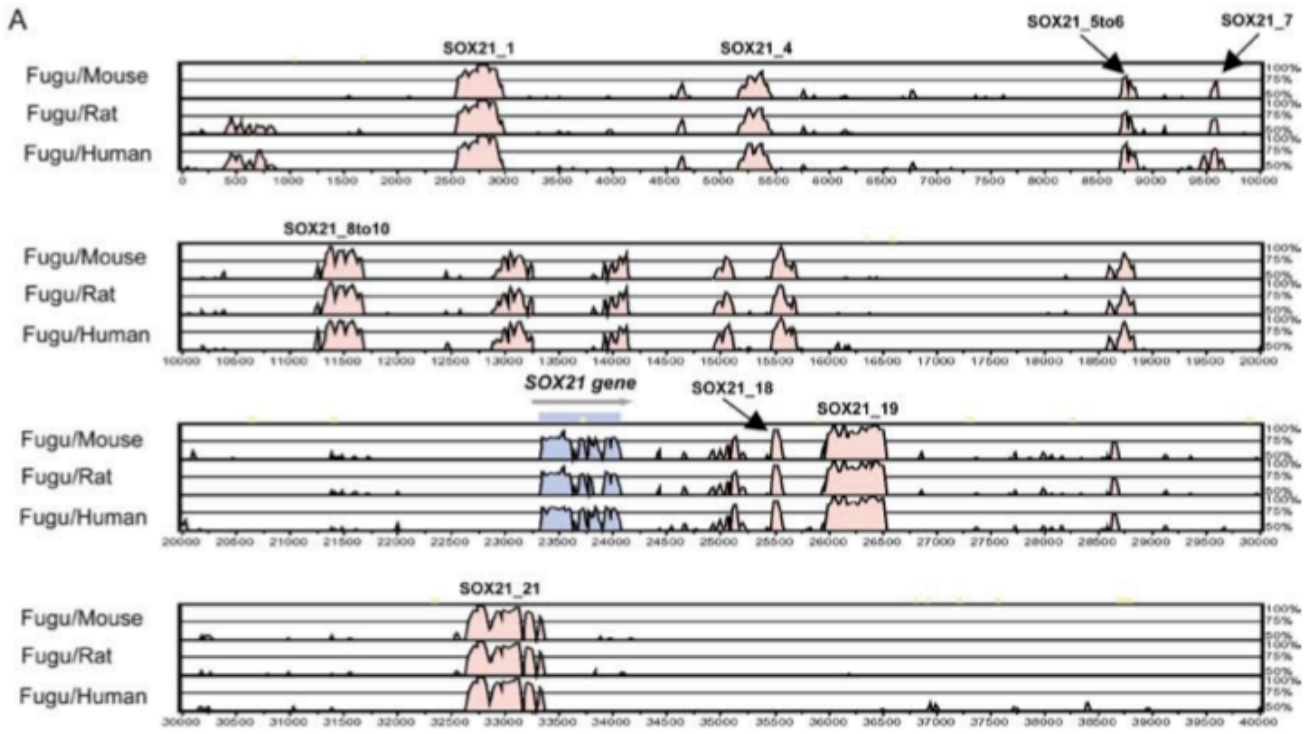
Promoter elements conserved from fish to mammal

Open access, freely available online PLoS BIOLOGY

Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development

Adam Woolfe[1], Martin Goodson[1], Debbie K. Goode[1], Phil Snell[1], Gayle K. McEwen[1], Tanya Vavouri[1], Sarah F. Smith[1], Phil North[1], Heather Callaway[1], Krys Kelly[1], Klaudia Walter[2], Irina Abnizova[2], Walter Gilks[2], Yvonne J. K. Edwards[1], Julie E. Cooke[1], Greg Elgar[1]*

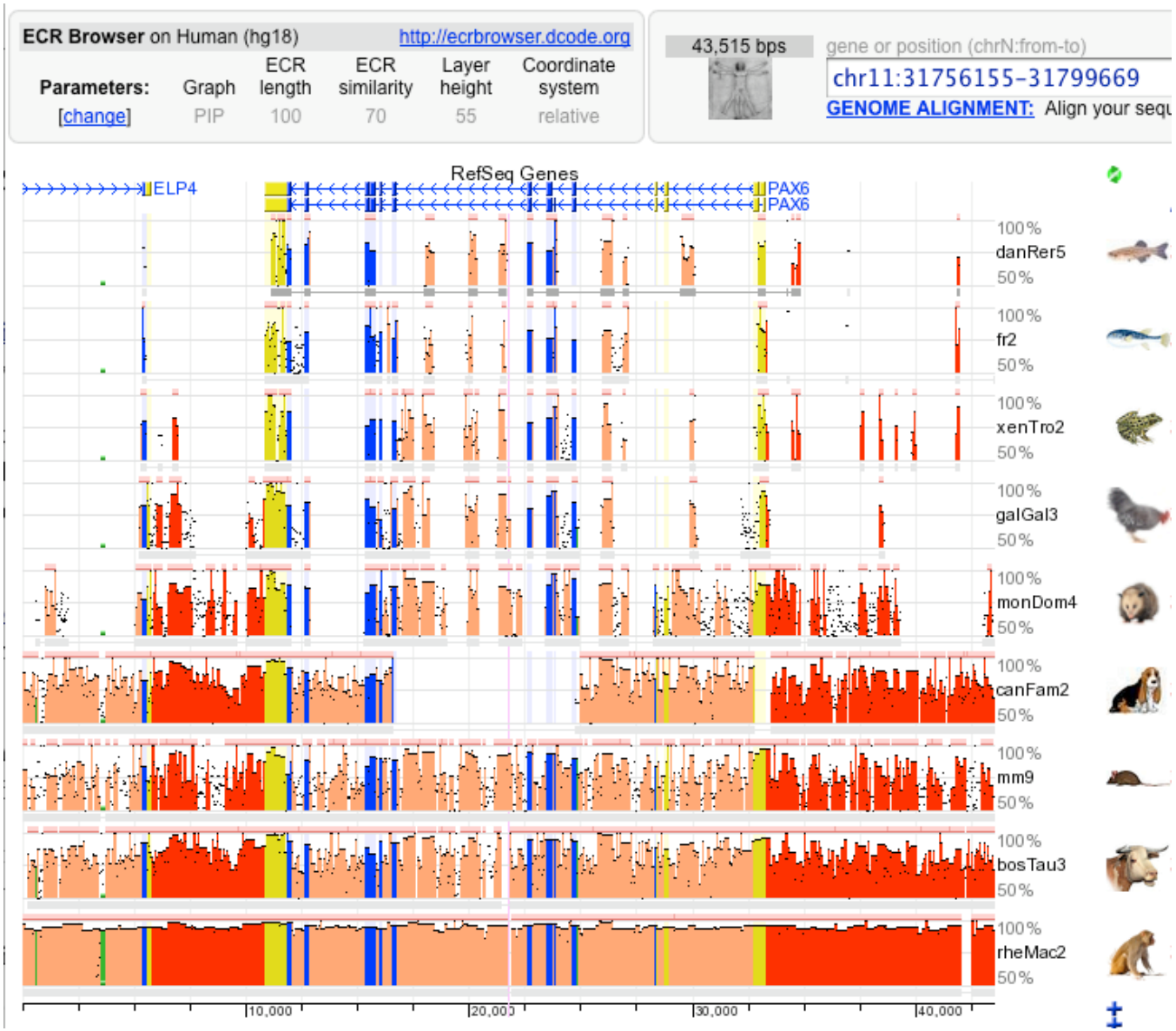
1 Medical Research Council Rosalind Franklin Centre for Genomics Research, Hinxton, Cambridge, United Kingdom, **2** Medical Research Council Biostatistics Unit, Institute of Public Health, Addenbrookes Hospital, Cambridge, United Kingdom



- Multi-species comparisons allow to analyze the relationship between degree of conservation and evolutionary distance.
- Example
  - Intergenic region upstream the gene Sox21 of Fugu, aligned with promoters of 3 mammalian species.
  - The peaks indicate highly conserved regions.

Woolfe et al. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol (2005) vol. 3 (1) pp. e7

Percentages of identical positions (PIP) in Pax6 chromosomal region

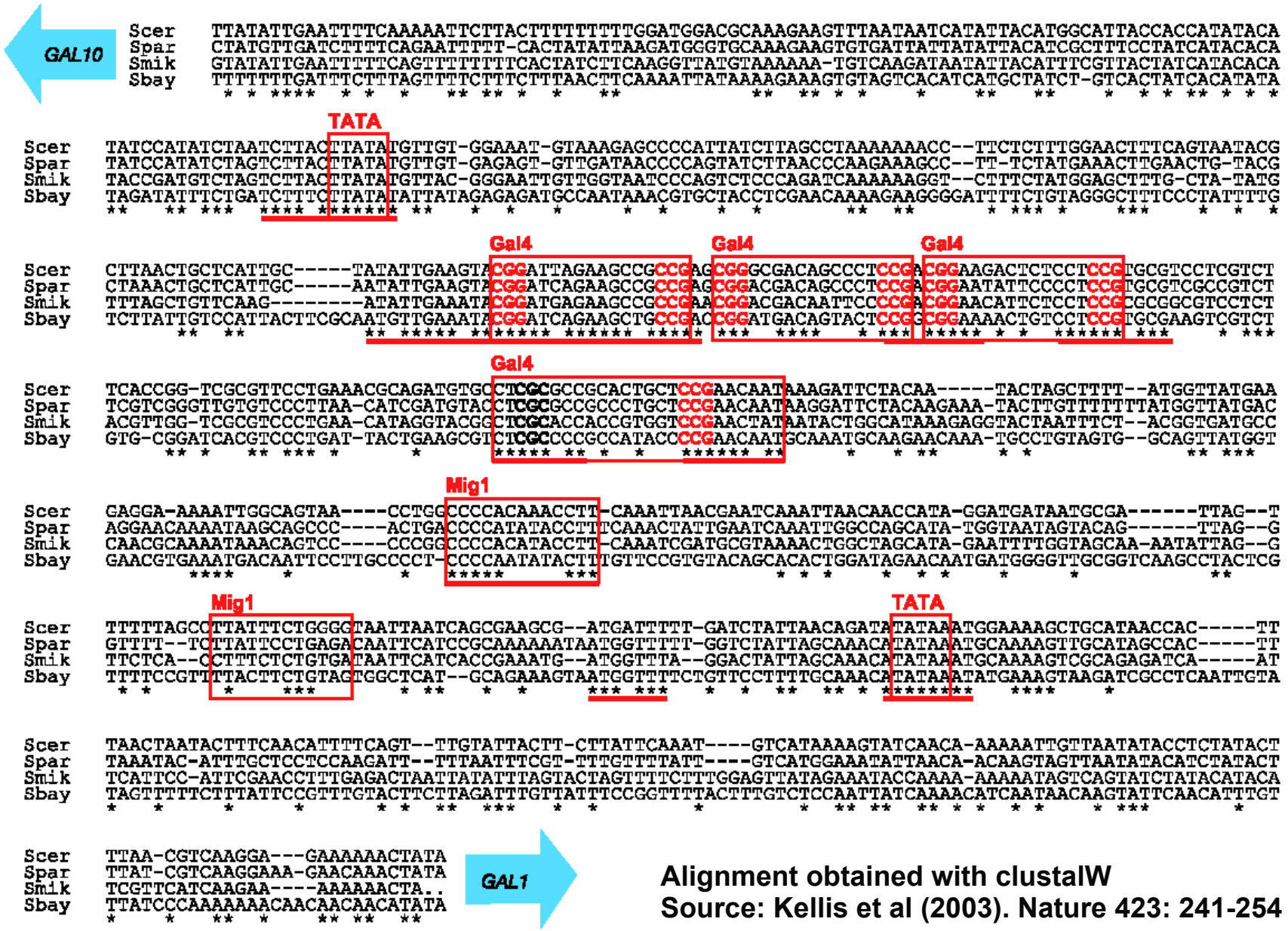


- Multi-species comparisons allow to analyze the relationship between degree of conservation and evolutionary distance.
- The plot shows the conservation of the genomic regions containing the Pax6 gene, in a series of vertebrates.
- Blocks with the highest conservation reflect exonic fragments (coding = blue; UTR = yellow).
- There are also conserved elements in the introns and intergenic region. These are likely to indicate cis-acting regulatory elements.

Coding fragments (exonic)
Introns
Utranslated regions (exonic)
Intergenic regios

Image generated on ECR browser (http://ecrbrowser.dcode.org/)

Global alignment of intergenic regions



Alignment obtained with clustalW
Source: Kellis et al (2003). Nature 423: 241-254

Another alignment in the same genomes

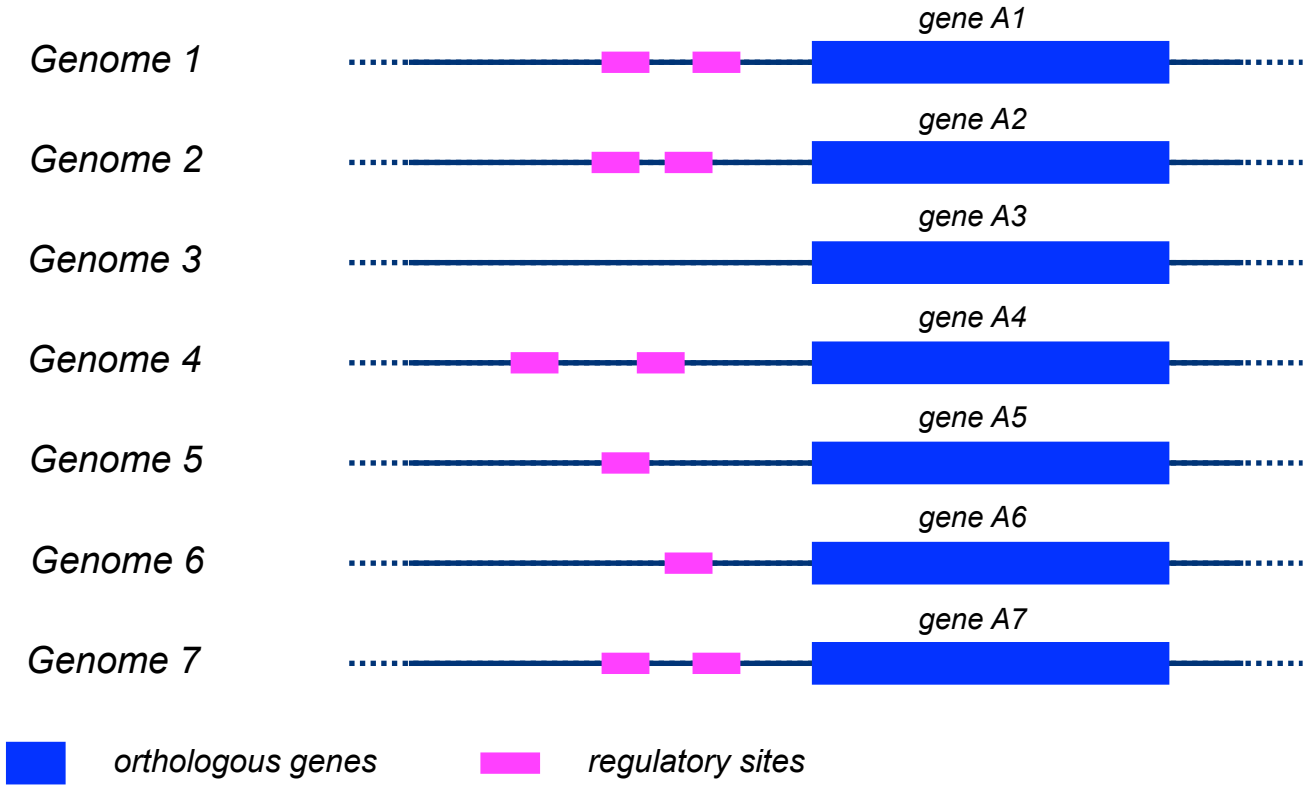
```
GAL80 (YML051W) upstream regions
Scer            ATGGCGCAAGTTTTCCGCTTTGTAATATATATTTATACCCCTTTCTTCTCTCCCCTGCAA
Spar            AGGGGCCAAAGCTCCCGCTCTGTAAAATATATTTATATCCCTTCCTTCTCTCCCCTGCAA
Smik            TAGGGACAAAGCCCGCCTTTTGTAATATATACTTATACCCTCTCCTTCTCTCCCCTGCAA
Sbay            ............................................................
                  **  ***      *  * ***** ***** ***** **  * ****************

Scer            TATAATAGTTTAATTCTAATATTAATAATA---TCCTATATTTTCTTCATTTACCGGCGC
Spar            TATAATAGTTTAATTCTAATATTAATAATA---TCCTATATTTTCCTTACC-ACCGGCGC
Smik            CATAATAGTTAACTCCTAATATTAATAATAATATCCTACAATTTCCTTAGC-ACCGGGGC
Sbay            ............................................................
                 ********* * * ***************   ***** * **** * *   ***** **

Scer            ACTCTCGCCCGAACGACCTCAAAATGTCTGCTACATTCATAATAACCAAAAGCTCATAAC
Spar            ACTCTCGCCCGAACGACCTCAAAATGCTTGCTACATTCATAATAATCAAAAGCTTATAAC
Smik            ACTCTCGCCCGAACGACCTCAAAACGCTTGCTACATCCATAATATTCAGAACTACATCAC
Sbay            ............................................................
                ************************ *  ******** *******  ** **    ** **

Scer            TTTTTTTTT----TGAACCTGAATATATATACATCACATATCACTGCTGGTCCTTGCCGA
Spar            TTTTTTTTTCCTTTGTACCTGAATATATATACATCTCATGTCACTGCTGGTCCTTGCCGG
Smik            TTTTTTTTT-----GTACATAAAAATATATAC--CACATGTCACTGCTGATCCTTGCTGA
Sbay            ............................................................
                *********     * ** * ** ********  * *** ********* ******* *

Scer            CCAGCGTATACAATCTCGATAGTTGGTTT-C-CCGTTCTTTCCACTCCCGTCATGGACTA
Spar            CCAGCGTATACAACCTCGATAGCTGGTTTTC-CCGTTCTTCCCACTCCTGTCATGGACTA
Smik            CGAGCGTATACAAGCTCGATAGCTGGTCTTTACCGTGCCATTCCCTGCCGTCATGGACTA
Sbay            ............................................................
                * ********** ******** **** *   **** *    * ** * ***********
```

Exercise

- Open connections
  - To ECR browser http://ecrbrowser.dcode.org/
  - to the UCSC genome browser http://genome.ucsc.edu
- Analyze the conservation of the following genes
  - Pax6
  - NPPA
  - Sox21
  - PCK1
  - Adh1A
- Tricks
  - In the UCSC genome browser, right-click on the conservation track and select the option “Configure conservation track set” to display additional species.
- Questions
  - Do you observe a correspondence between conservation profiles and gene structure (coding fragments, UTRs, introns) ?
  - Do you observe conserved elements in non-exonic regions (introns, upstream, downstream) ?
  - What are the optimal evolutionary distance to discriminate non-conserved from conserved regions (exonic and non-exonic, respectively) ?

Phylogenetic footprinting to predict regulatory sites

- Cross-species conservation profiles resulting from alignment of genomic regions are generally not sufficient to identify cis-regulatory elements.
- Multi-species sequences can be scanned to detect transcription factor binding sites, in order to analyze their conservation.



Cross-species identification of binding sites

- Schmidt et al (2010) use the ChIP-seq technology to identify the binding sites of two transcription factors (CEBPA and HNF4A) in 5 vertebrates.
- This analysis reveals the important rate of ***site turn-over*** for these factors: many sites are species-specific, or conserved between a few species only.

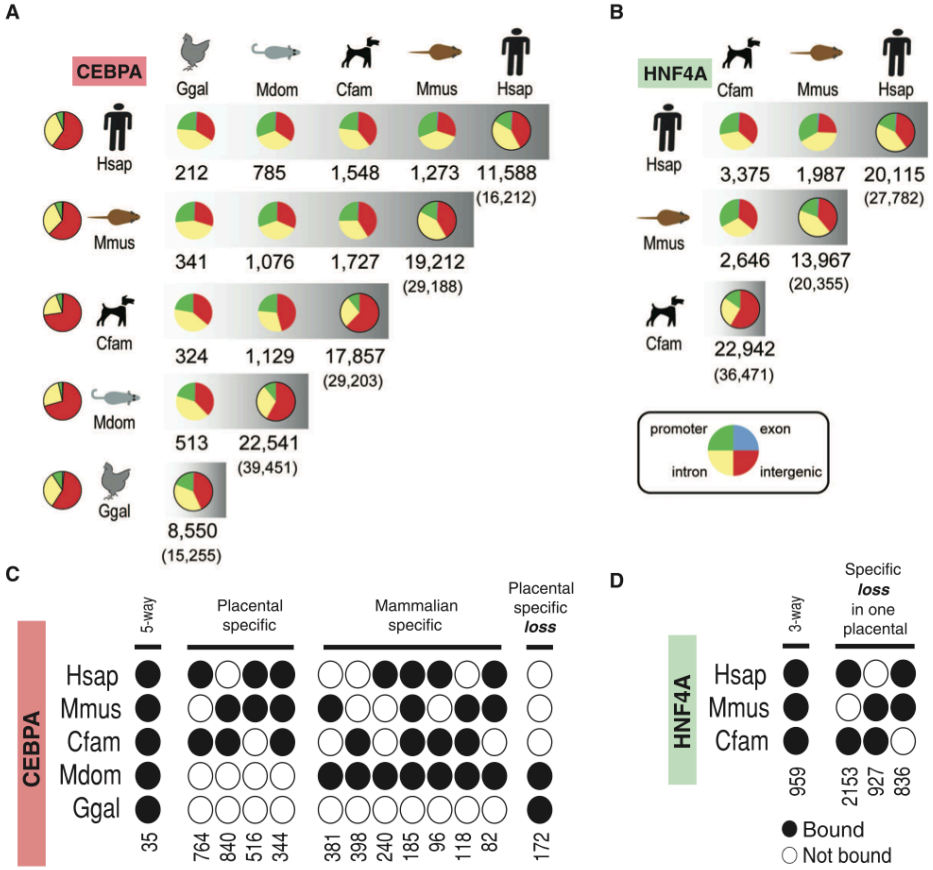


Fig. 2. Conservation and divergence of TF binding. For (**A**) CEBPA and (**B**) HNF4A, the pairwise distribution and numbers of binding events are shown as a pie chart distributed into the following segments: intergenic (red), intronic (yellow), exonic (blue), and promoter (TSS ±3 kb) (green) regions. The left-most column contains the distributions of the bulk genomes. The right-most pie chart represents all binding events in each species, with the total number of alignable peaks above the total peaks (in parentheses). (**C** and **D**) Multispecies CEBPA and HNF4A binding event analysis, where black circles indicate binding in a given species. For instance, there are 764 regions bound by CEBPA only in dog and human (see also figs. S6, S7, and S17 and tables S2 and S6). (**E**) The DNA sequence constraint beneath binding events was measured by average GERP (*20*) scores for peaks found: in all five species (5-way), among all the placental mammals (3-way), bound in any two species (shared), within 10 kb of the TSS of functional targets (functional), and all peaks.

- Schmidt et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science (2010) vol. 328 (5981) pp. 1036-40

Pattern discovery in promoters of orthologous genes

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*

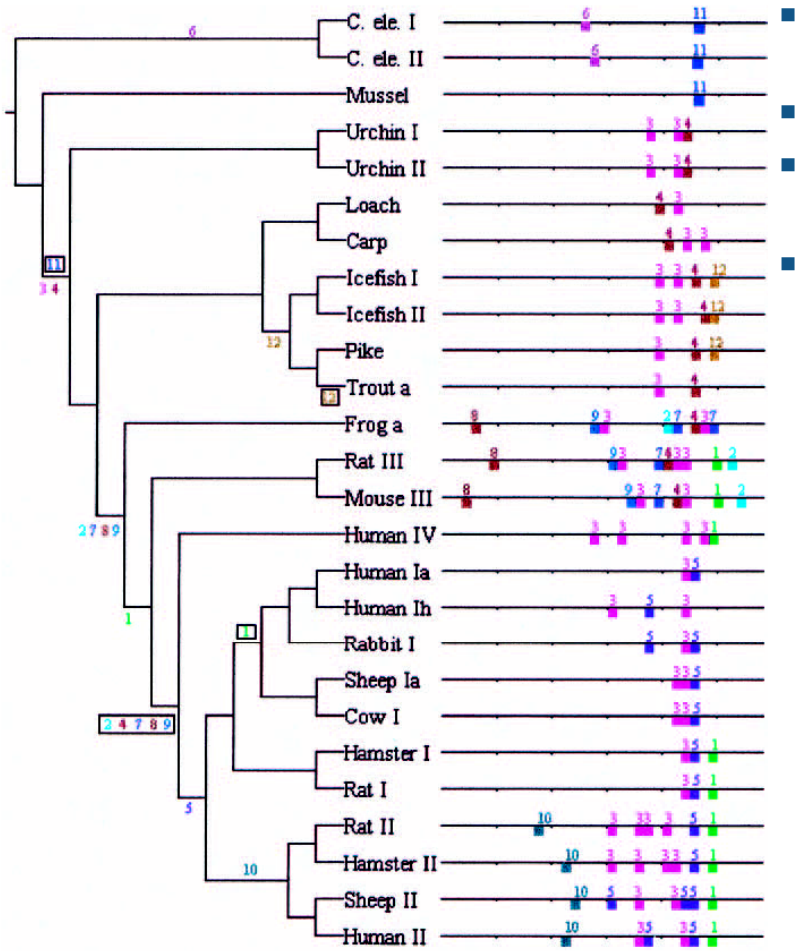
* These authors contributed equally to this work

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.

Xie et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature (2005) vol. 434 (7031) pp. 338-45

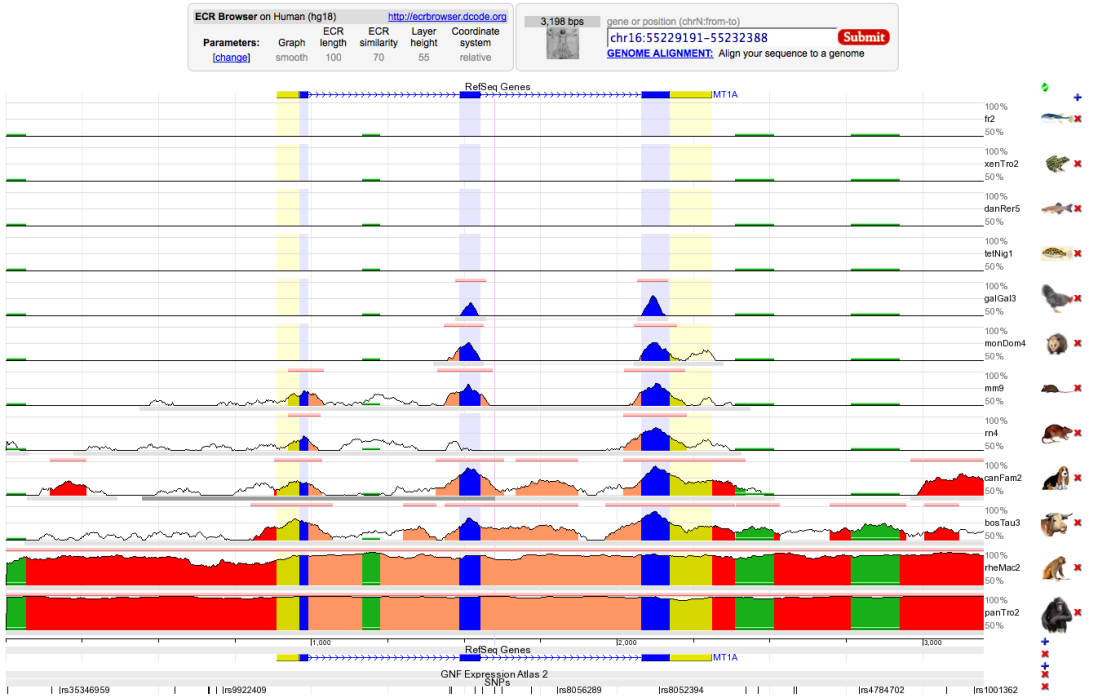
Xie et al. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci USA (2007) vol. 104 (17) pp. 7145-50

Footprinter example metallothionein



Source: Blanchette and Tompa (2002). Genome Research. 12, 739–748.

- 590 bp upstream of the same gene (methallothionein) in different species.
- 12 highly conserved motifs are detected.
- Each motif can be associated to a given internal node of the phylogenetic tree.
- Note:
  - Blanchette & Tompa analyzed promoters of the whole metallothionein family (orthologs + paralogs).
  - The conservation cannot be detected on simple ECR plots.
  - The pattern discovery program allows to detect the conserved elements (small sites) even though the regions are not conserved.

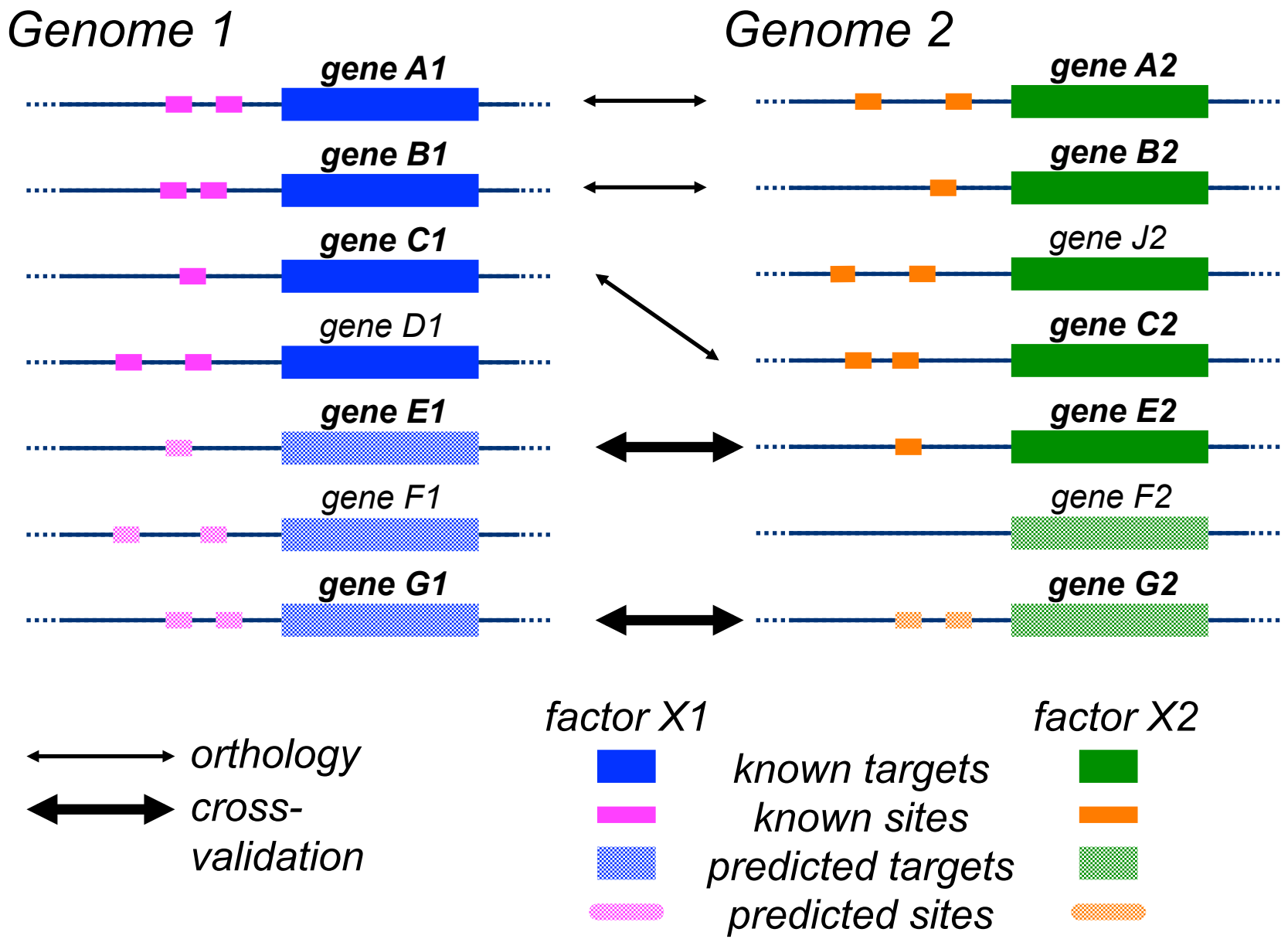


http://ecrbrowser.dcode.org/xB.php?db=hg18&location=chr11:31762916-31789455

Cross-validation of genome-scale pattern matching

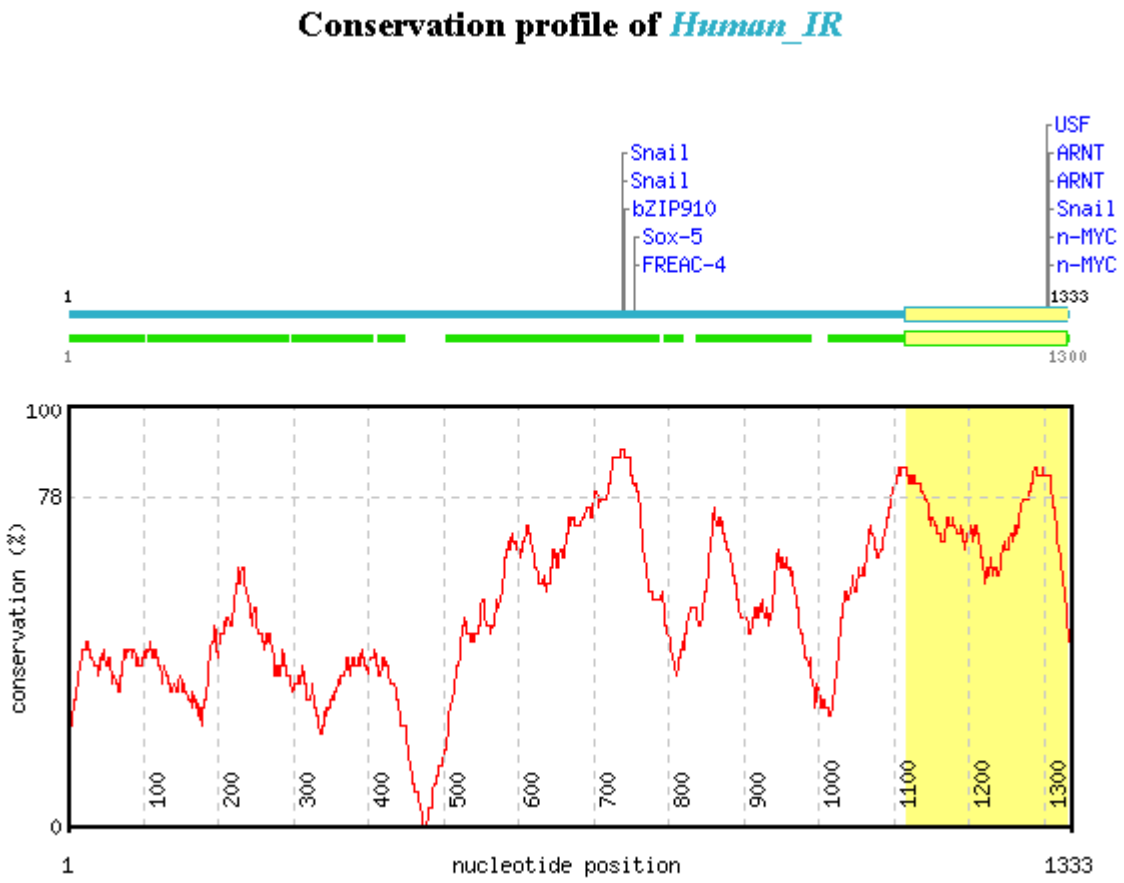
- Single-genome pattern matching raises many false positive
- Cross-validation : cross-species comparisons can be done to evaluate the reliability of the predictions.
  - gene A from genome X has a good match in its upstream sequence
  - ortholog A' from genome Y has a good match in its upstream sequence

Cross-validation of pattern matching



Cross-matches in promoters of orthologous genes

- Lenhard et al. (2003). J.Biology 2:13.
- 100 PSSM for known mammal transcription factors
- Searching for conserved matches in Human and mouse increases the selectivity by 85%.
- Consite: http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/



Phylogenetic footprinting resources

- CORG: a database for COmparative Regulatory Genomics
  - Dieterich et al. (2003), Nucleic Acids Res. 31:55-57.
  - http://corg.molgen.mpg.de
  - Systematic alignment of 15Kb upstream regions for each pair of mouse-human homologous genes (18.674 pairs).
  - 10.793 significant alignments ($P < 0.001$), containing 293.503 conserved non-coding blocks (CNB), covering 8% of the upstream sequences (http://corg.molgen.mpg.de/stats.html).

Phylogenetic footprint detection tools

- CONSITE
  - Web site: http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite
  - Explore transcription factor binding sites shared by two genomic sequences
  - Relies on a library of TF binding motifs.
- PhyloCon
  - http://ural.wustl.edu/~twang/PhyloCon/
  - Patern discovery algorithm (consensus) applied to promoters of orthologs.
  - Unix executable.
- PhyloGibbs
  - http://www.phylogibbs.unibas.ch/cgi-bin/phylogibbs.pl
  - Siddharthan R, Siggia ED, van Nimwegen E. PLoS Comput Biol 1(7): e67 (2005)
  - A Gibbs sampling adapted to search conserved motifs (positional windows of conservation across species).
- footprint-discovery (RSAT suite)
  - Web site: http://rsat.ulb.ac.be/rsat/

Summary – phylogenetic footprint detection

- Phylogenetic footprints can be detected by different approaches
  - Global alignment of promoters of orthologous genes
    - clustalW
    - e.g.: Kellis et al (2003). Nature 423: 241-254.
  - Pattern discovery in promoters of orthologous genes
    - Footprinter: http://bio.cs.washington.edu/software.html
    - Blanchette and Tompa (2002). Genome Research. 12, 739–748.
  - Matching known motifs in different species and selecting conserved sites
    - Consite: http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/
    - Lenhard et al. (2003). J.Biology 2:13.
  - Pattern matching restricted to conserved regions (detected by whole-genome alignments)

- Those methods can help in restricting the number predicted elements and increasing their likelihood to be functional, but they are still error-prone, especially in metazoan genomes.