Regulatory Sequence Analysis

Applying comparative genomics to detect cis-acting elements

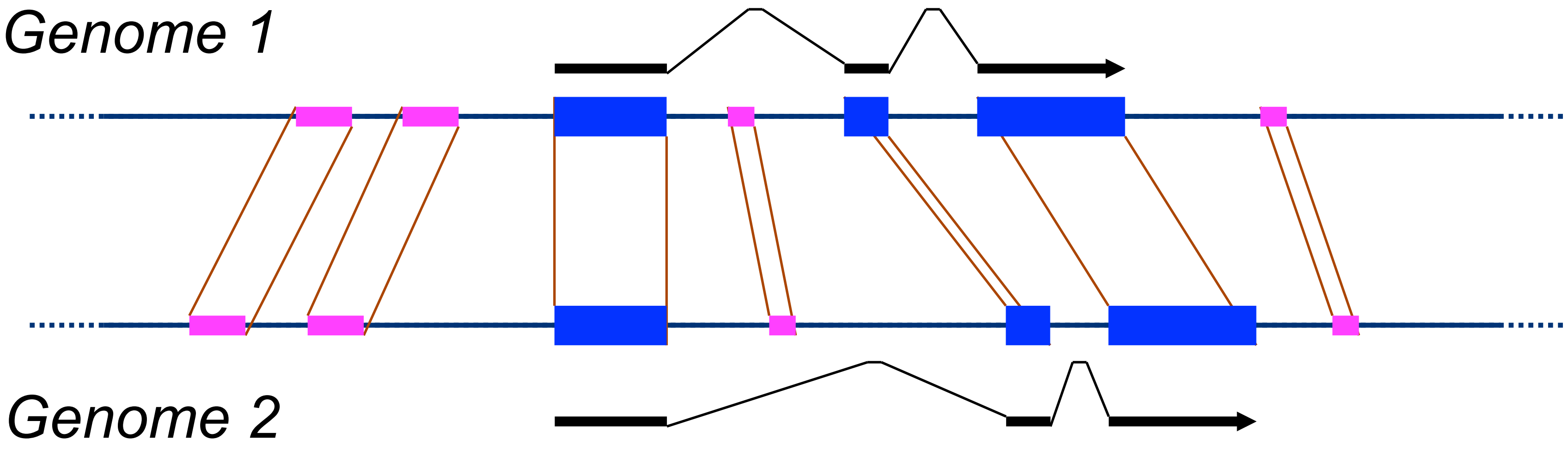
Jacques.van.Helden@ulb.ac.be

Université Libre de Bruxelles, Belgique

Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)

http://www.bigre.ulb.ac.be/

Phylogenetic footprinting to define regulatory regions



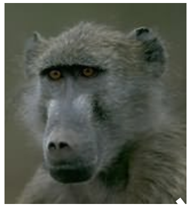
Genome 1

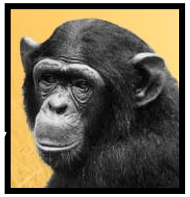
Genome 2

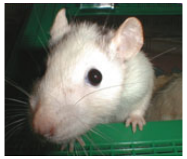
🟦 *conserved exon* 🟪 *conserved non-coding region*

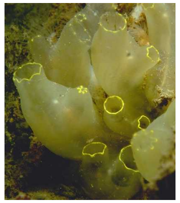
- Within non-coding sequences, regulatory elements evolve slower than their surrounding.
- Conserved non-coding sequences contain a high concentration in regulatory elements.

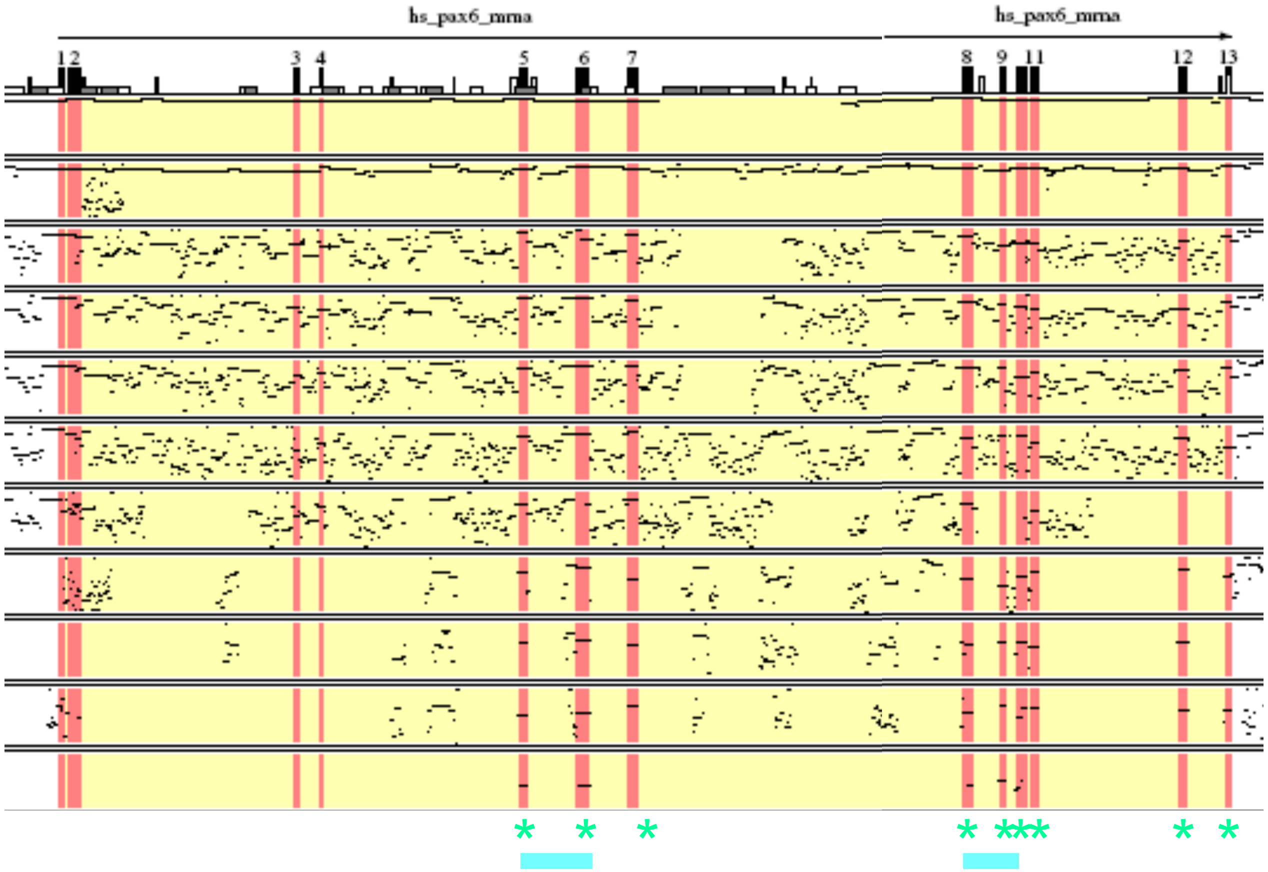
Phylogenetic footprints for the pax6 gene









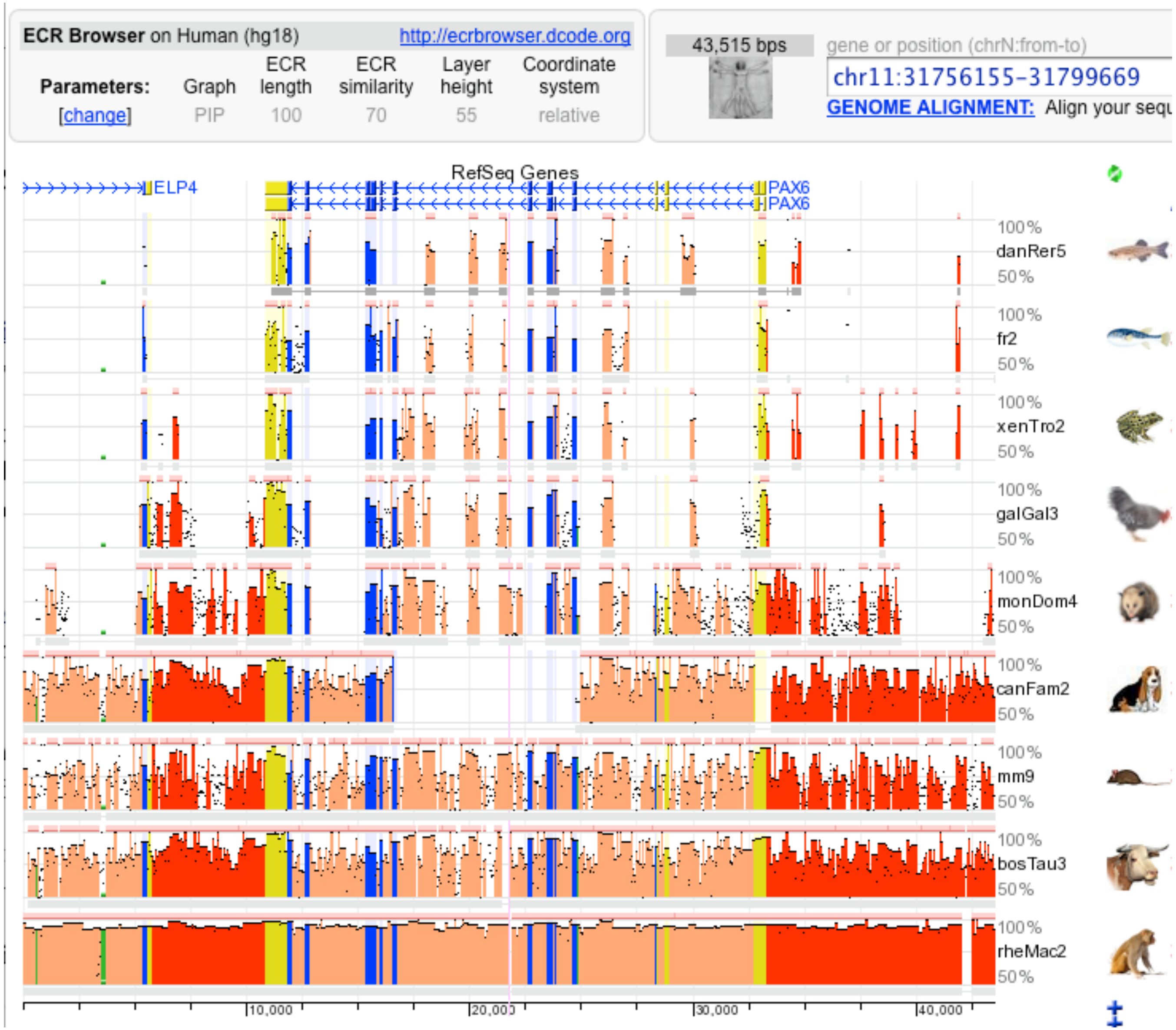


Functional domains

Coding exons

Slide from Philippe Gautier

Pourcentages de positions identiques (PIP) dans la région chromosomique de Pax6



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

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

43,515 bps

gene or position (chrN:from-to)

chr11:31756155-31799669

GENOME ALIGNMENT: Align your sequ

Image générée sur l'ECR browser (http://ecrbrowser.dcode.org/)

- La génomique comparative permet d'améliorer la localisation des gènes.
- Alignement de la région génomique contenant le gène Pax6, entre le génome humain et une série d'organismes de plus en plus distants évolutivement (de bas en haut).
- Les blocs de séquences conservées reflètent souvent la présence de fragments codants.
- Cependant, il existe également des segments conservés dans les régions non-codantes.

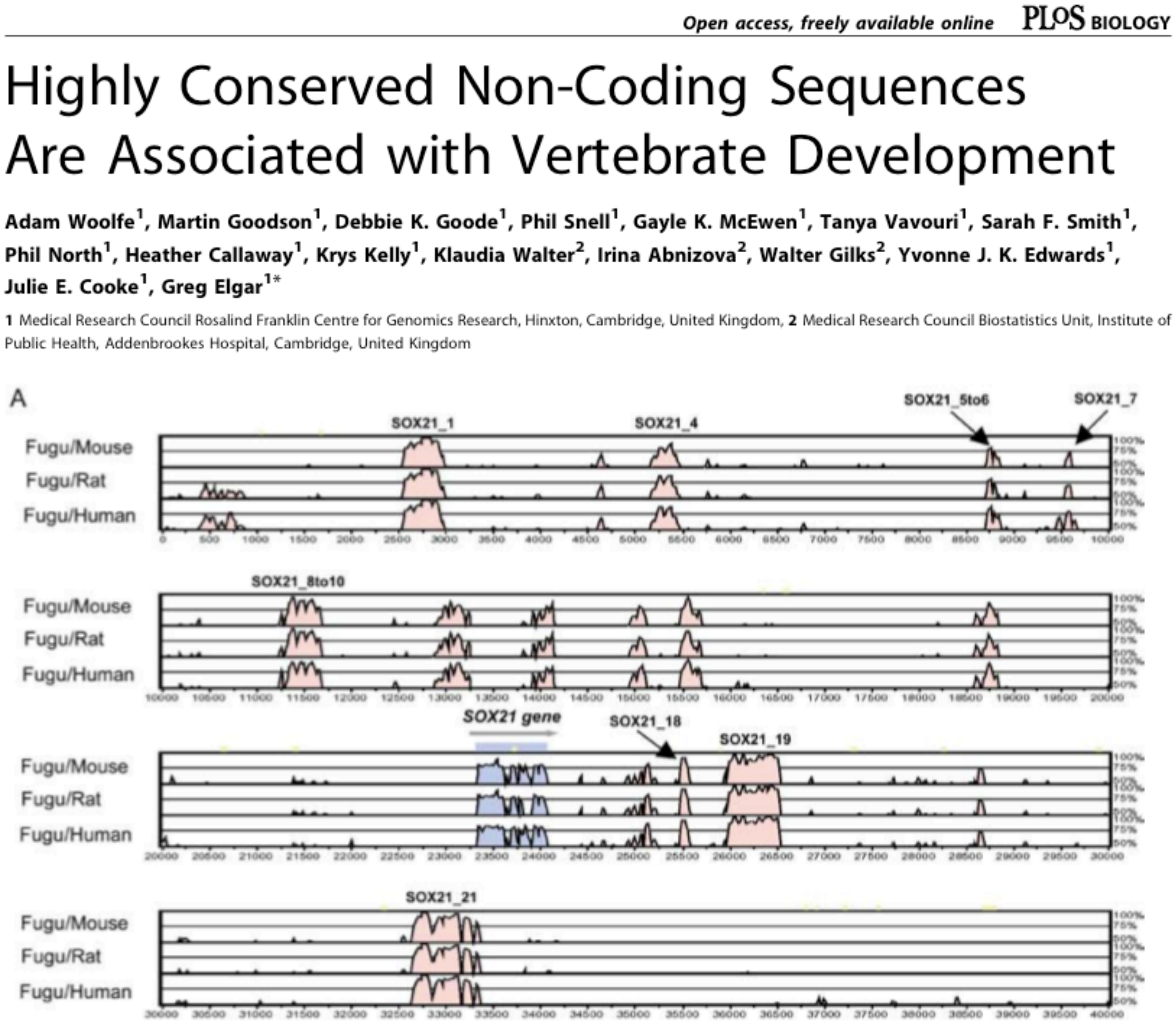
🟦 Exons

🟧 Introns

🟨 Extrémités non traduites

🟥 Régions intergéniques

Promoter elements conserved from fish to mammal



Open access, freely available online PLoS BIOLOGY

Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development

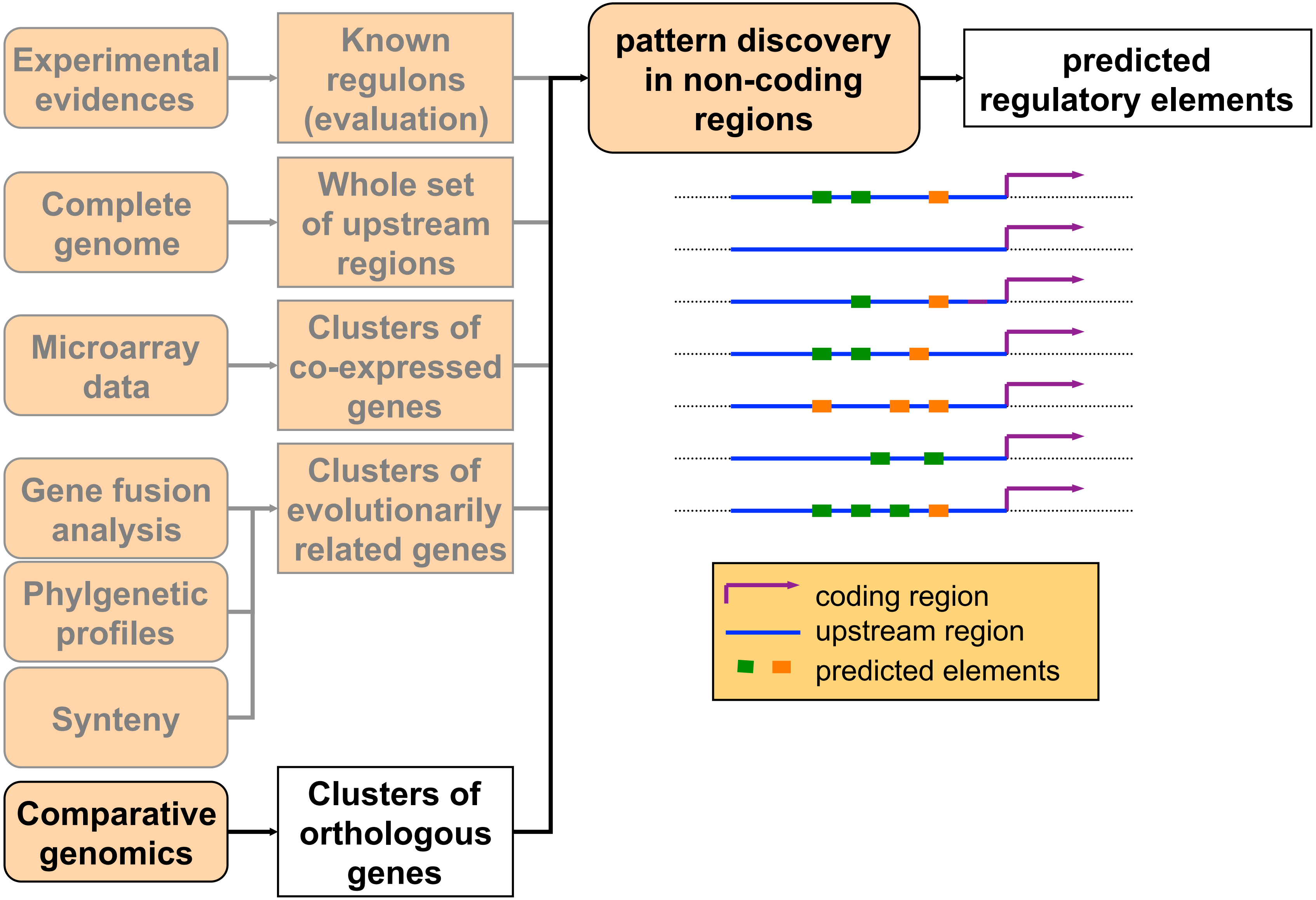
Adam Woolfe[1], Martin Goodson[1], Debbie K. Goode[1], Phil Snell[1], Gayle K. McEwen[1], Tanya Vavouri[1], Sarah F. Smith[1], Phil North[1], Heather Callaway[1], Krys Kelly[1], Klaudia Walter[2], Irina Abnizova[2], Walter Gilks[2], Yvonne J. K. Edwards[1], Julie E. Cooke[1], Greg Elgar[1*]

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

- 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

Motifs in clusters of orthologous genes (COGs)



Experimental evidences

Known regulons (evaluation)

Complete genome

Whole set of upstream regions

Microarray data

Clusters of co-expressed genes

Gene fusion analysis

Phylgenetic profiles

Synteny

Clusters of evolutionarily related genes

Comparative genomics

Clusters of orthologous genes

pattern discovery in non-coding regions

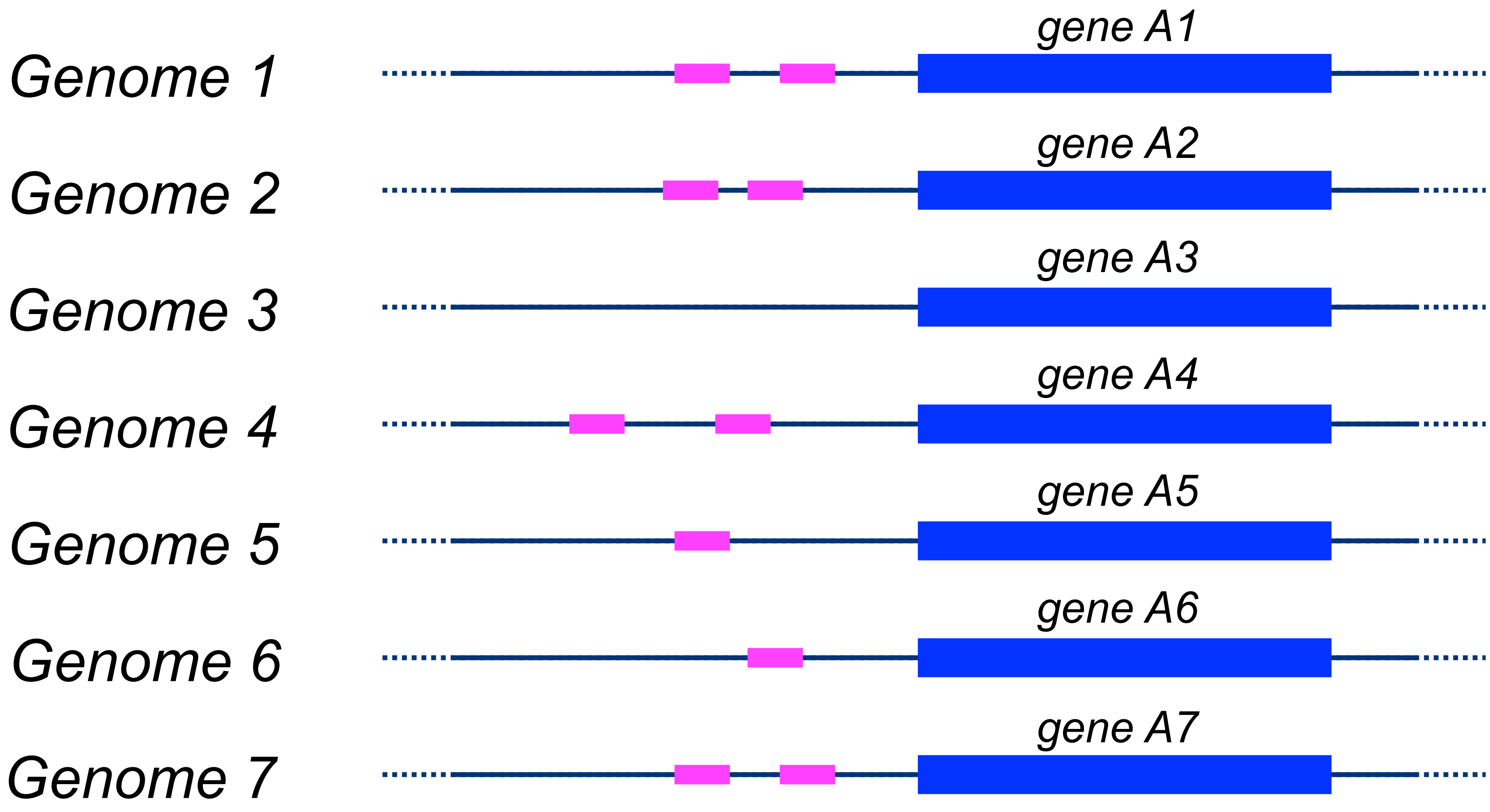
predicted regulatory elements

coding region

upstream region

🟩 🟧 predicted elements

Phylogenetic footprinting to predict regulatory sites



Genome 1

gene A1

Genome 2

gene A2

Genome 3

gene A3

Genome 4

gene A4

Genome 5

gene A5

Genome 6

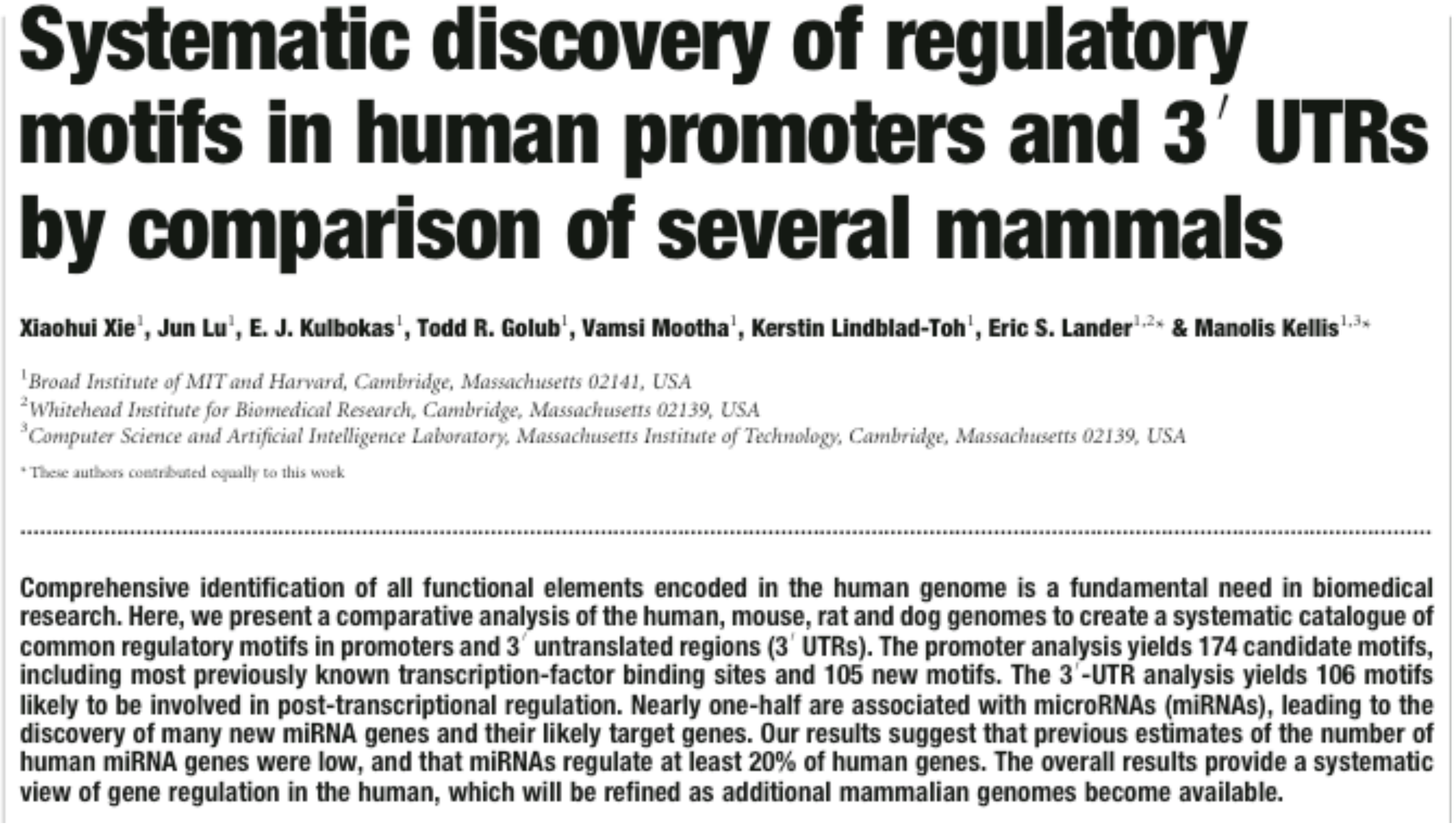
gene A6

Genome 7

gene A7

🟦 *orthologous genes* ▬ *conserved regulatory sites*

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*]

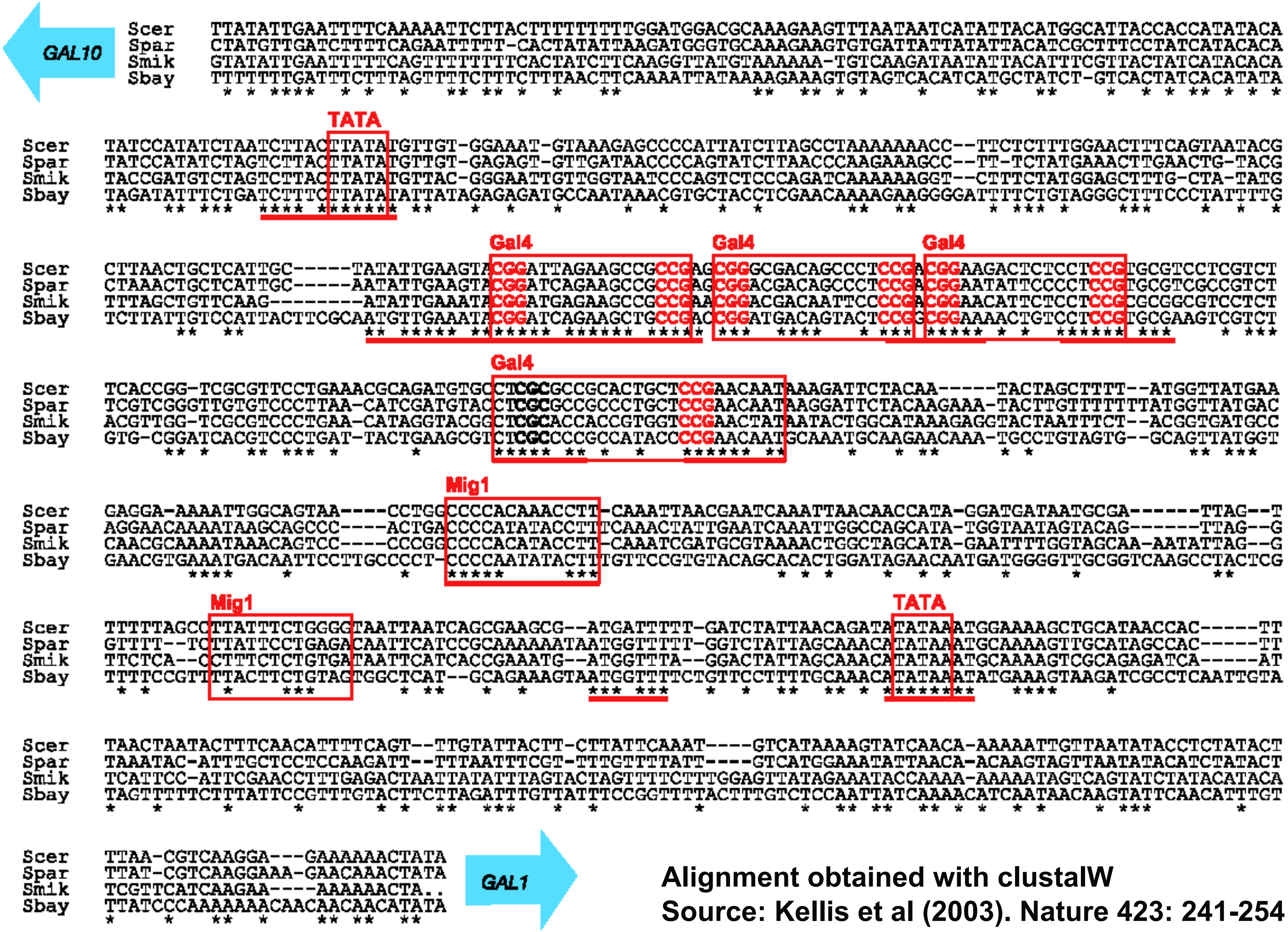
[1] *Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA*
[2] *Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA*
[3] *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[*] 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

Global alignment of intergenic regions



```
Scer   TTATATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATACA
Spar   CTATGTTGATCTTTTCAGAATTTTT-CACTATATTAAGATGGGTGCAAAGAAGTGTGATTATTATATTACATCGCTTTCCTATCATACACA
Smik   GTATATTGAATTTTTCAGTTTTTTTTCACTATCTTCAAGGTTATGTAAAAAA-TGTCAAGATAATATTACATTTCGTTACTATCATACACA
Sbay   TTTTTTTGATTTCTTTAGTTTTCTTTCTTTAACTTCAAAATTATAAAAGAAAGTGTAGTCACATCATGCTATCT-GTCACTATCACATATA
        * * ****  * *  *   ** ** *  *   **           **  ** * *    *    **   **    *  * * ** * * *

Scer   TATCCATATCTAATCTTACTTATATGTTGT-GGAAAT-GTAAAGAGCCCCATTATCTTAGCCTAAAAAAACC--TTCTCTTTGGAACTTTCAGTAATACG
Spar   TATCCATATCTAGTCTTACTTATATGTTGT-GAGAGT-GTTGATAACCCCAGTATCTTAACCCAAGAAAGCC--TT-TCTATGAAACTTGAACTG-TACG
Smik   TACCGATGTCTAGTCTTACTTATATGTTAC-GGGAATTGTTGGTAATCCCAGTCTCCCAGATCAAAAAAGGT--CTTTCTATGGAGCTTTG-CTA-TATG
Sbay   TAGATATTTCTGATCTTTCTTATATATTATAGAGAGATGCCAATAAACGTGCTACCTCGAACAAAAGAAGGGGATTTTCTGTAGGGCTTTCCCTATTTTG
       **   ** ***  **** ******** **   *  *   *     *  *    *  *       **  **      * *** *    ***     *  *  *

Scer   CTTAACTGCTCATTGC-----TATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTCCGTGCGTCCTCGTCT
Spar   CTAAACTGCTCATTGC-----AATATTGAAGTACGGATCAGAAGCCGCCGAGCGGACGACAGCCCTCCGACGGAATATTCCCCTCCGTGCGTCGCCGTCT
Smik   TTTAGCTGTTCAAG--------ATATTGAAATACGGATGAGAAGCCGCCGAACGGACGACAATTCCCCGACGGAACATTCTCCTCCGCGCGGCGTCCTCT
Sbay   TCTTATTGTCCATTACTTCGCAATGTTGAAATACGGATCAGAAGCTGCCGACCGGATGACAGTACTCCGGCGGAAAACTGTCCTCCGTGCGAAGTCGTCT
             ****           ** ***** ******* ****** ***** ***  ****   * ****** * *  ****** ****    * ***

Scer   TCACCGG-TCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAA-----TACTAGCTTTT--ATGGTTATGAA
Spar   TCGTCGGGTTGTGTCCCTTAA-CATCGATGTACCTCGCGCCGCCCTGCTCCGAACAATAAGGATTCTACAAGAAA-TACTTGTTTTTTTATGGTTATGAC
Smik   ACGTTGG-TCGCGTCCCTGAA-CATAGGTACGGCTCGCACCACCGTGGTCCGAACTATAATACTGGCATAAAGAGGTACTAATTTCT--ACGGTGATGCC
Sbay   GTG-CGGATCACGTCCCTGAT-TACTGAAGCGTCTCGCCCCGCCATACCCCGAACAATGCAAATGCAAGAACAAA-TGCCTGTAGTG--GCAGTTATGGT
            ** *  ** *** *      *      ***** ** *  *   ****** **    *   * **     * *            ** ***

Scer   GAGGA-AAAATTGGCAGTAA----CCTGGCCCCACAAACCTT-CAAATTAACGAATCAAATTAACAACCATA-GGATGATAATGCGA------TTAG--T
Spar   AGGAACAAAATAAGCAGCCC----ACTGACCCCATATACCTTTCAAACTATTGAATCAAATTGGCCAGCATA-TGGTAATAGTACAG------TTAG--G
Smik   CAACGCAAAATAAACAGTCC----CCCGGCCCCACATACCTT-CAAATCGATGCGTAAAACTGGCTAGCATA-GAATTTTGGTAGCAA-AATATTAG--G
Sbay   GAACGTGAAATGACAATTCCTTGCCCCT-CCCCAATATACTTTGTTCCGTGTACAGCACACTGGATAGAACAATGATGGGGTTGCGGTCAAGCCTACTCG
              ****    *         *   *****     ***               * *  *    *  *  *    *     *           **

Scer   TTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCG--ATGATTTTT-GATCTATTAACAGATATATAAATGGAAAAGCTGCATAACCAC-----TT
Spar   GTTTT--TCTTATTCCTGAGACAATTCATCCGCAAAAAATAATGGTTTTT-GGTCTATTAGCAAACATATAAATGCAAAAGTTGCATAGCCAC-----TT
Smik   TTCTCA--CCTTTCTCTGTGATAATTCATCACCGAAATG--ATGGTTTA--GGACTATTAGCAAACATATAAATGCAAAAGTCGCAGAGATCA-----AT
Sbay   TTTTCCGTTTTACTTCTGTAGTGGCTCAT--GCAGAAAGTAATGGTTTTCTGTTCCTTTTGCAAACATATAAATATGAAAGTAAGATCGCCTCAATTGTA
        * *      *    ***       * **   *  *     *** ***   *  *  **  ** * ********   ***    *

Scer   TAACTAATACTTTCAACATTTTCAGT--TTGTATTACTT-CTTATTCAAAT----GTCATAAAAGTATCAACA-AAAAATTGTTAATATACCTCTATACT
Spar   TAAATAC-ATTTGCTCCTCCAAGATT--TTTAATTTCGT-TTTGTTTTATT----GTCATGGAAATATTAACA-ACAAGTAGTTAATATACATCTATACT
Smik   TCATTCC-ATTCGAACCTTTGAGACTAATTATATTTAGTACTAGTTTTCTTTGGAGTTATAGAAATACCAAAA-AAAAATAGTCAGTATCTATACATACA
Sbay   TAGTTTTTCTTTATTCCGTTTGTACTTCTTAGATTTGTTATTTCCGGTTTTACTTTGTCTCCAATTATCAAAACATCAATAACAAGTATTCAACATTTGT
       *   *     *     *      * *  **  ***   *   *        *        *  ** **  ** * *  * *    * ***       *

Scer   TTAA-CGTCAAGGA---GAAAAAACTATA
Spar   TTAT-CGTCAAGGAAA-GAACAAACTATA
Smik   TCGTTCATCAAGAA----AAAAAACTA..
Sbay   TTATCCCAAAAAAACAACAACAACATATA
       *    *   **  *    ** **  **
```

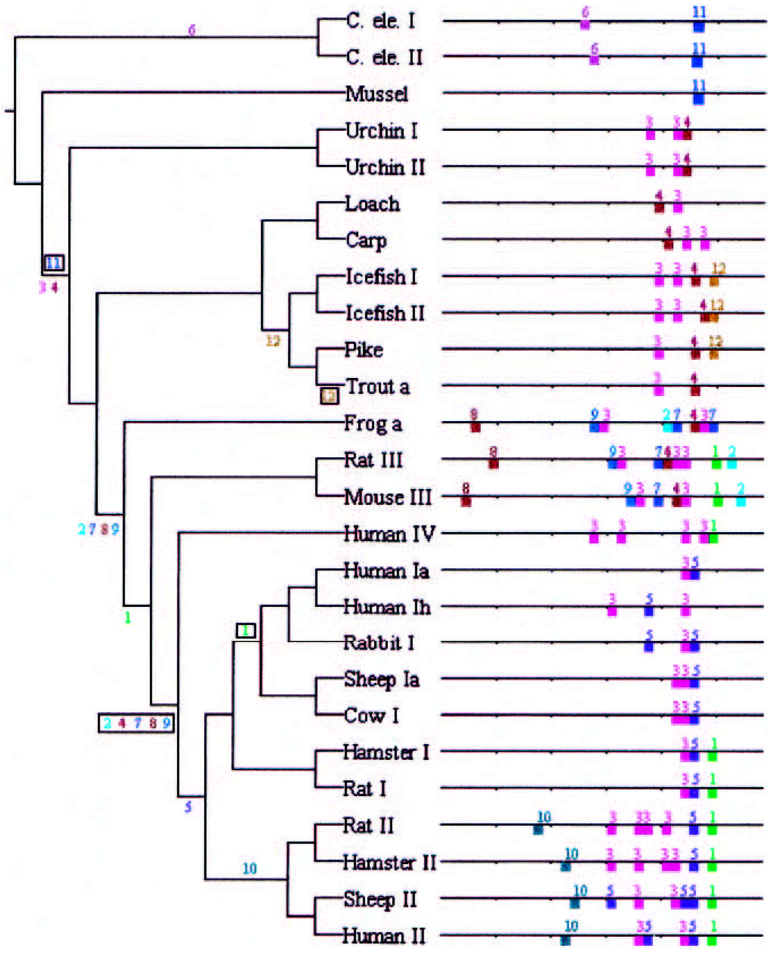
Alignment obtained with clustalW

Source: Kellis et al (2003). Nature 423: 241-254

Another alignment in the same genomes

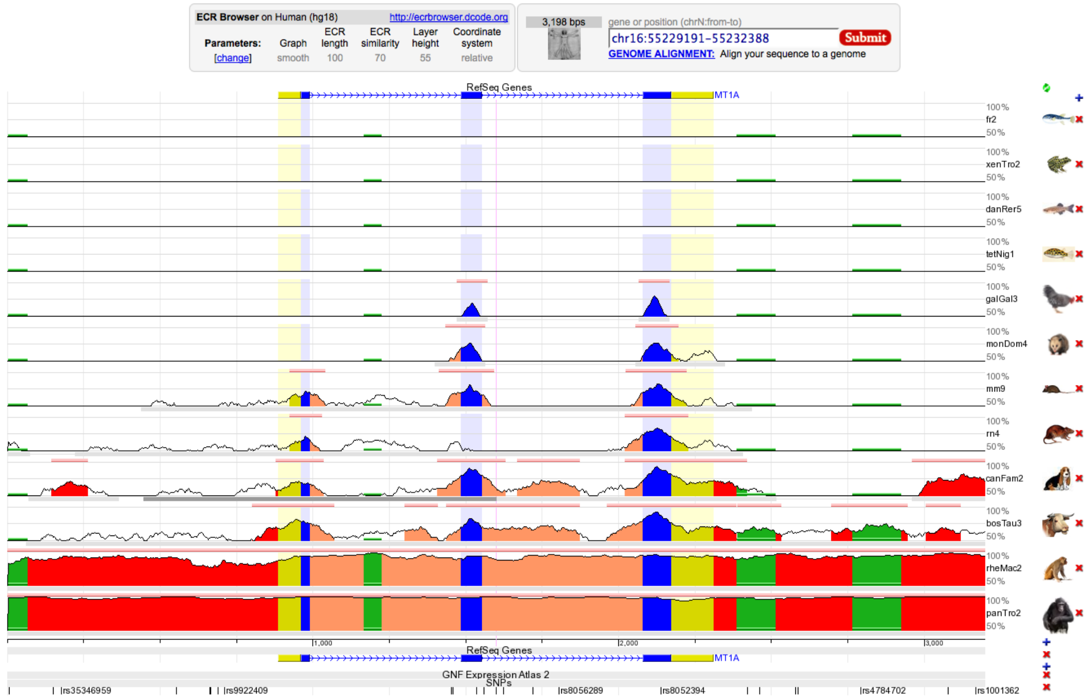
```
GAL80 (YML051W) upstream regions
Scer            ATGGCGCAAGTTTTCCGCTTTGTAATATATATTTATACCCCTTTCTTCTCTCCCCTGCAA
Spar            AGGGGCCAAAGCTCCCGCTCTGTAAAATATATTTATATCCCTTCCTTCTCTCCCCTGCAA
Smik            TAGGGACAAAGCCCGCCTTTTGTAATATATACTTATACCCTCTCCTTCTCTCCCCTGCAA
Sbay            ............................................................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

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            ............................................................
                * *********** ******** **** *   **** *    * ** * ***********
```

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.



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

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

3,198 bps

gene or position (chrN:from-to): chr16:55229191-55232388 Submit

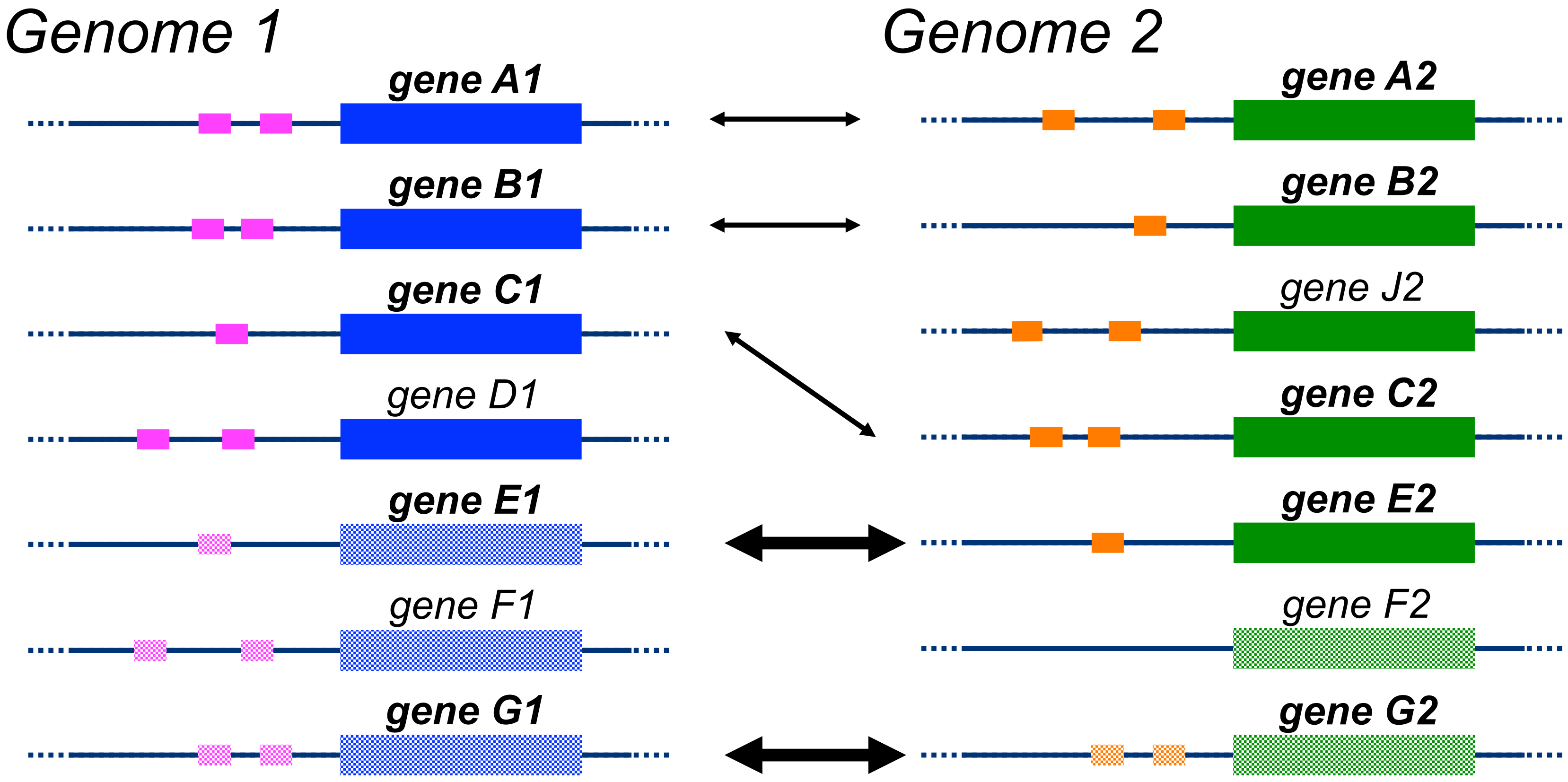
GENOME ALIGNMENT: Align your sequence to a genome

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

Cross-validation of genome-scale pattern matching

- Genome-scale pattern matching raises many false positive
- Cross-validation :
  - 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



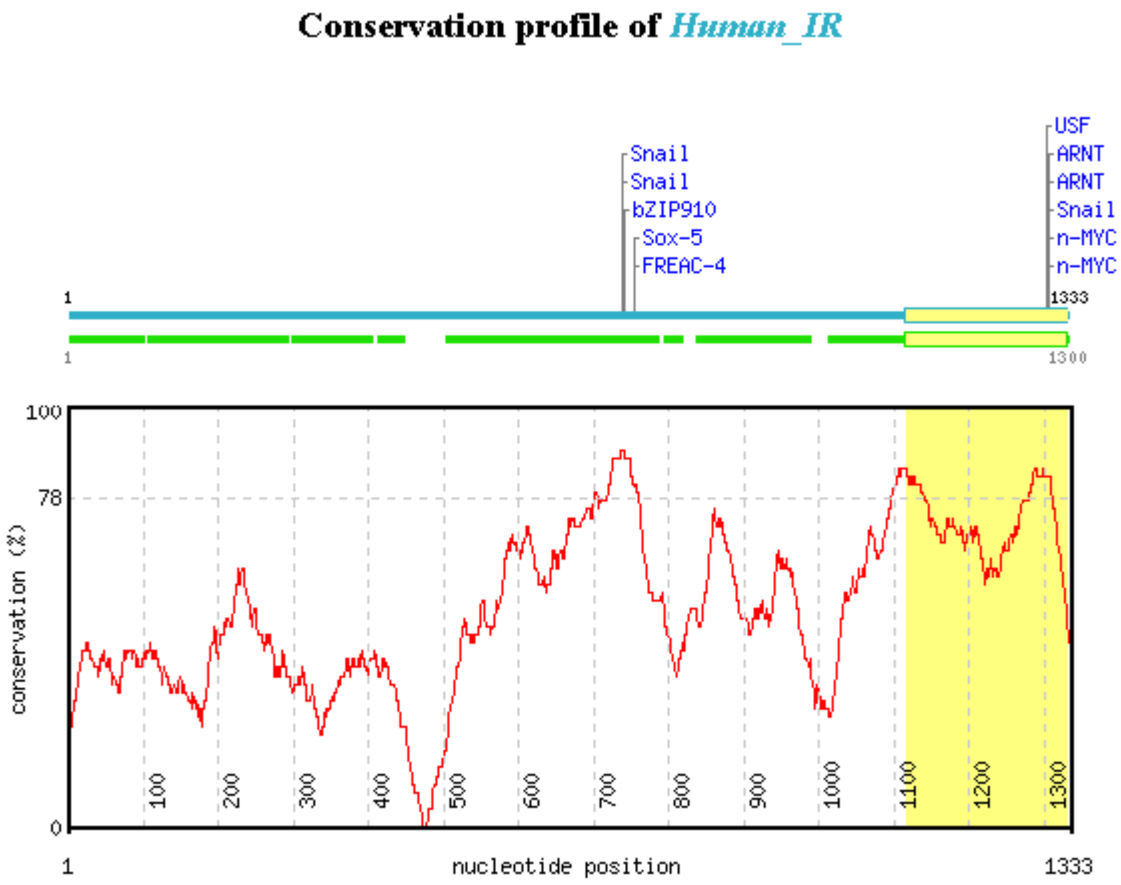
⟷ *orthology*

⬌ *cross-validation*

factor X1		*factor X2*
🟦	*known targets*	🟩
▬	*known sites*	▬
▨	*predicted targets*	▨
▨	*predicted sites*	▨

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/


Conservation profile of *Human_IR*

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.