Overview

- Lesson 1
  - Comparative Genomics
  - Functional genomic elements (ENCODE)
- Lesson 2
  - Gene annotation, text-mining, ontologies, data integration, applications
- Lesson 3
  - Prediction and analysis of regulatory sequences

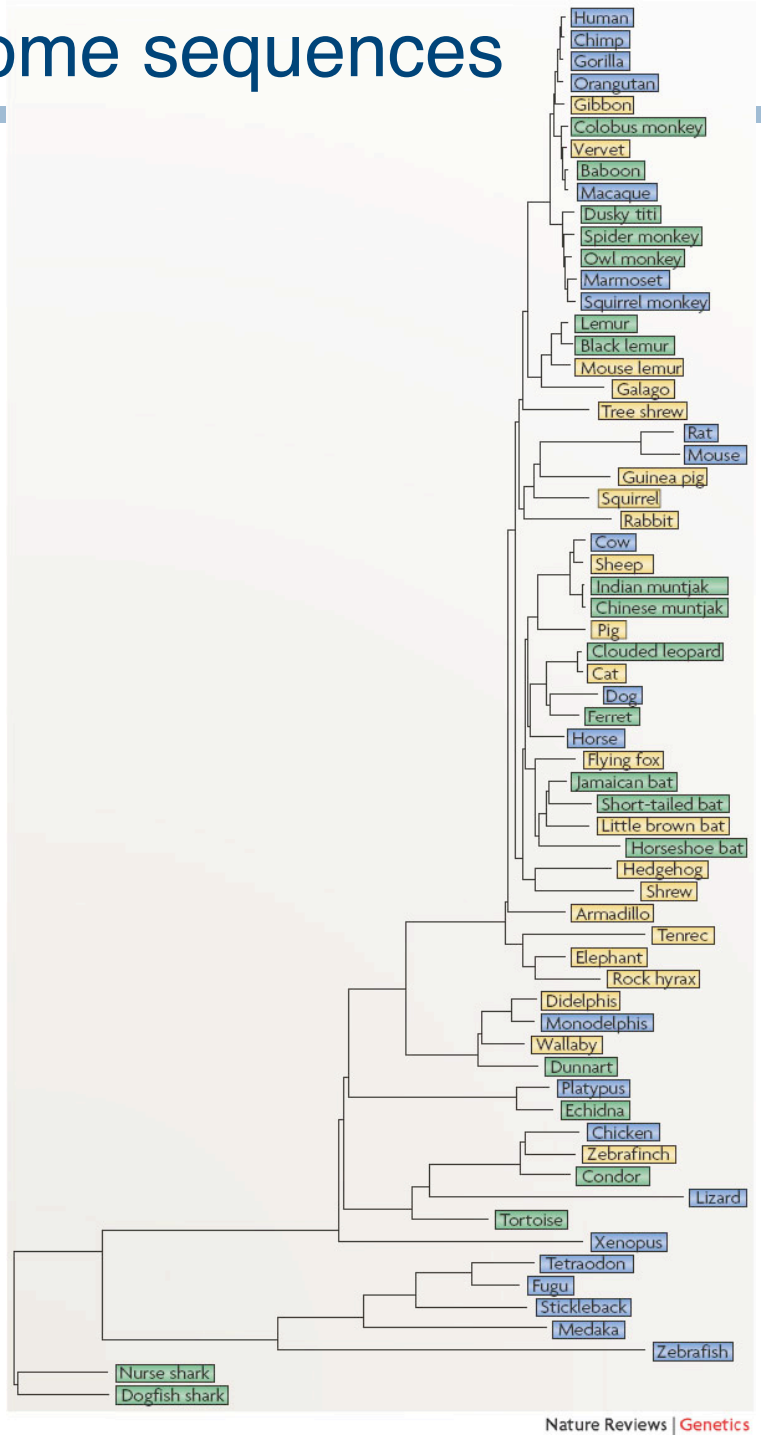
Comparative Genomics

Dr. ir. Stein Aerts
Laboratory of Neurogenetics
Department of Molecular and Developmental Genetics
VIB-KULeuven





Genome sequences



Next challenge: distinguish functional DNA and assign a role to it

-> genome-wide predictions by comparative genomics

-> experimental tests, also on a large scale, e.g., ENCODE

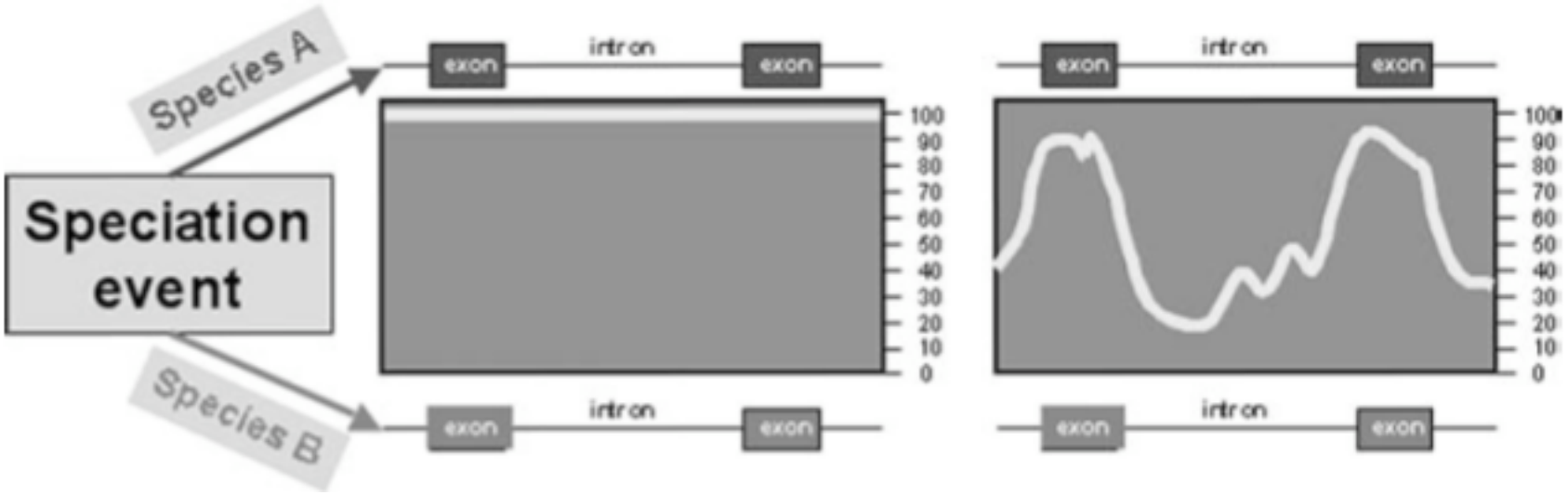
Readings

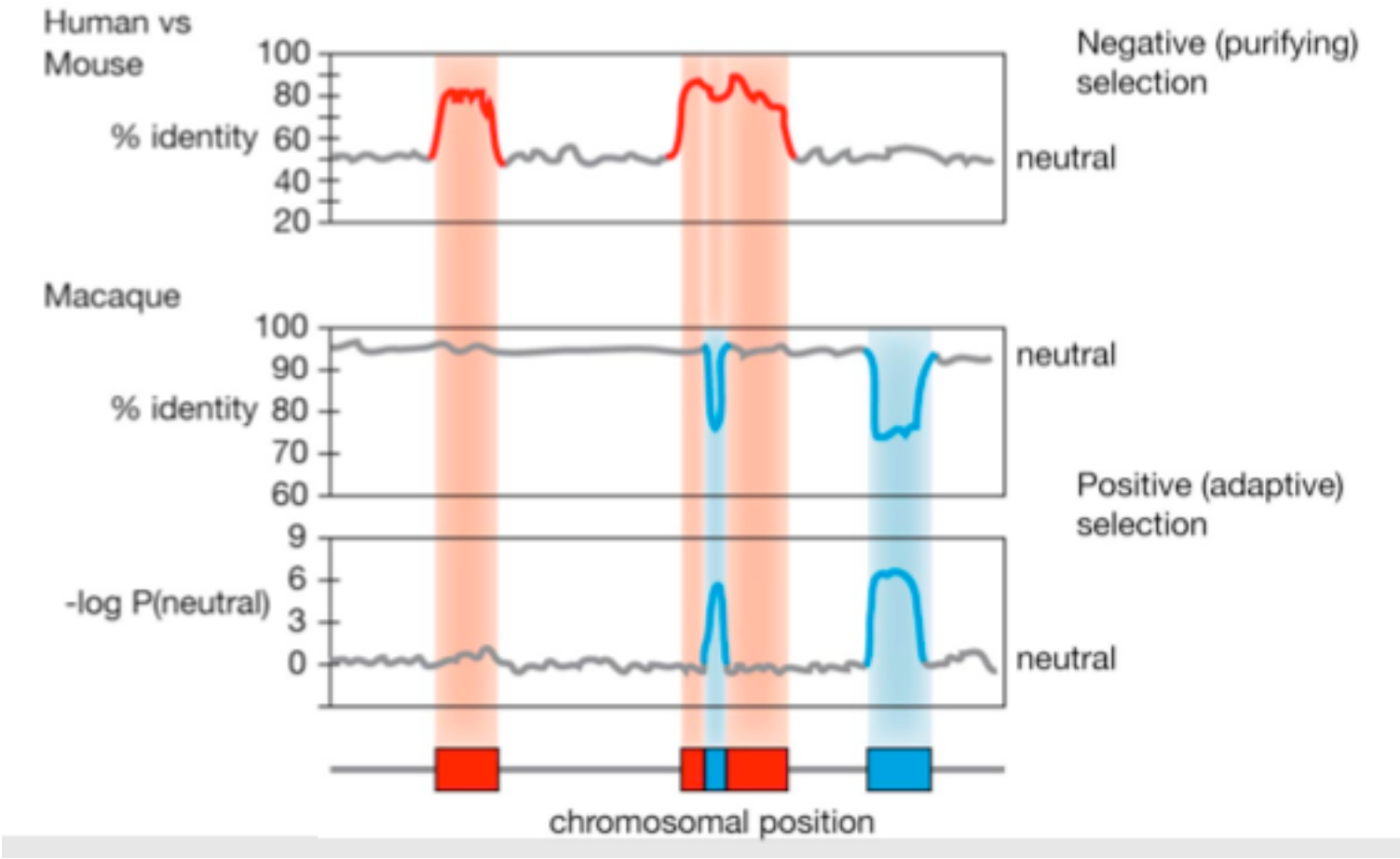
- Comparative Genomics, Chapter in Human Genetics - Principles and Approaches - 4th Edition, Ross C. Hardison
  - http://www.bx.psu.edu/~ross/ComparGeno/HardisonHumGenWFigsIntgr.pdf
- Comparative Genomics Miller et al. Annu Rev Genomics Hum Genet 2004
- Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. Margulies & Birney Nature Rev Genet 2008
- Waterston et al. Nature 2002 Mouse Genome Paper
- Kellis et al. Nature 2003: 4 yeast species
- Nature 2007 12 Drosophila species

Introduction

- Goal of comparative genomics:
  - Finding DNA sequences that show significant signs of positive or negative selection *(and infer such sequences are functional)*
    - Positive (Darwinian) selection: fixation of advantageous alleles
      - Results in adaptive evolution
      - Sequence changes more rapidly than the bulk
    - Negative (purifying) selection: removal of disadvantageous alleles
      - Sequence is constrained by its function to remain similar to its ancestor
      - Sequence changes more slowly than the bulk
      - 'constrained' sequences: similarity level is greater than expected for neutral DNA
- Mutations of single bases; indels (insertions and deletions, resulting from replication errors or recombination); chromosomal rearrangements
- Conserved sequence: reliable alignment

Evolution of functionally important regions over time





- Purifying selection: rate of sequence change is slower than that of neutral DNA
- Positive (Darwinian) selection: rate of sequence change is faster than that of neutral DNA

Neutral DNA

- Fourfold synonymous sites
- Pseudogenes
- Ancestral repeats

Alignments of biological sequences

- Optimize a similarity score
  - Matching symbols favored
  - Mismatches not favored
  - Gaps penalized
    - Gap-open penalty + affine gap penalties (additional smaller penalty for each additional position included in the gap)
- Global alignment
  - Maps each symbol in one sequence onto a corresponding symbol in another sequence
  - Needleman-Wunsch 1970
  - ClustalW
  - AVID, LAGAN
- Local alignment
  - When a match of only a portion of two or more sequences is desired
  - Smith-Waterman 1981
  - Blast
  - blastZ (long genomic DNA sequences)

Aligning large genomic sequences

- Find reliable alignments within homology blocks and deduce how the various homology blocks are connected in genomes of compared species.
- Genes on same chromosome = syntenic
  - Groups of genes that are syntenic in humans are frequently also syntenic in mouse; I.e., *conserved synteny*
  - They frequently also maintain a similar order and orientation (indicates homology = similarity because of common ancestry)
  - Homologous segments rarely extend for entire chromosomes
    - One human chromosome will align with several homology blocks in mouse, many of which are on different chromosomes.

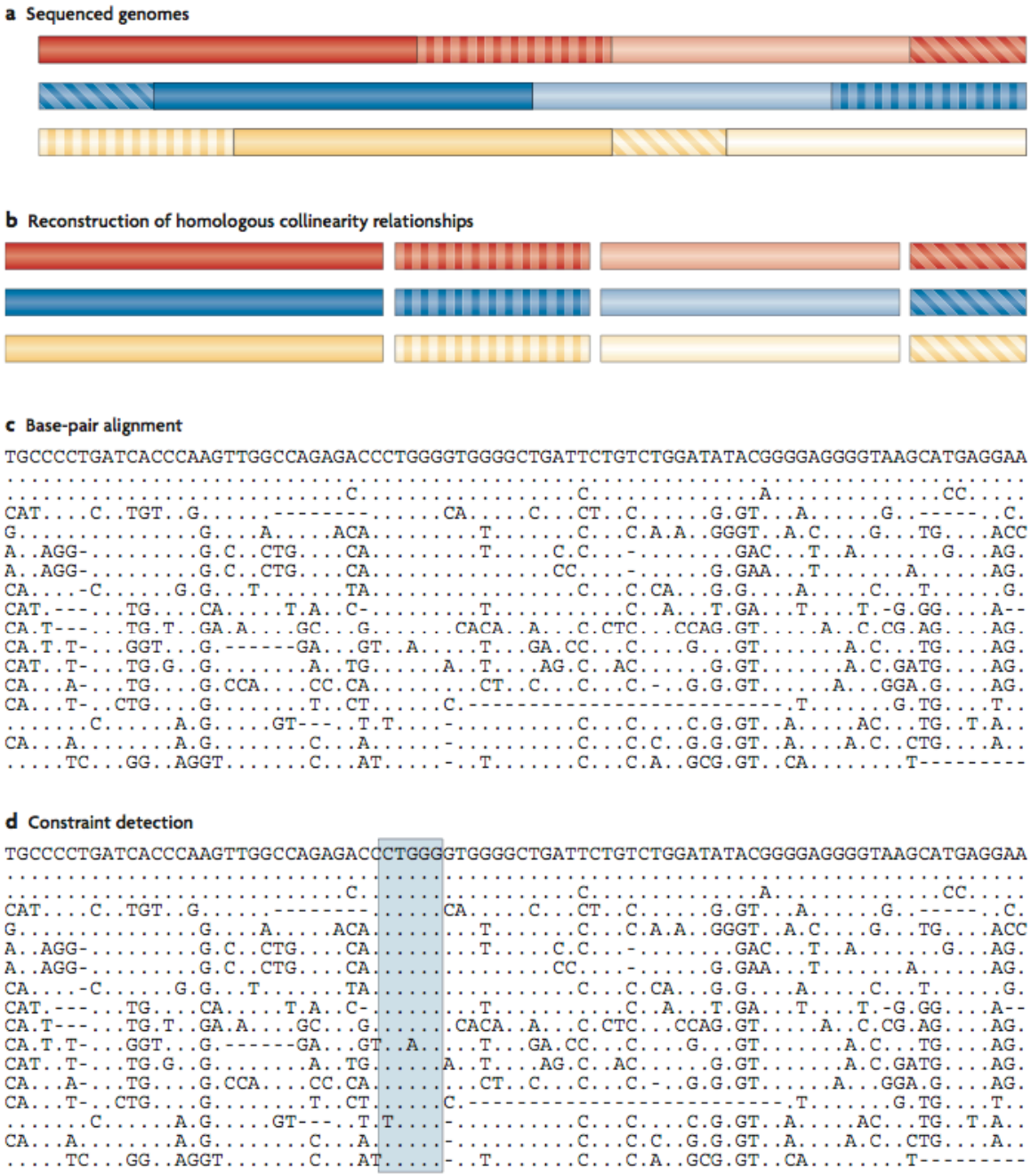


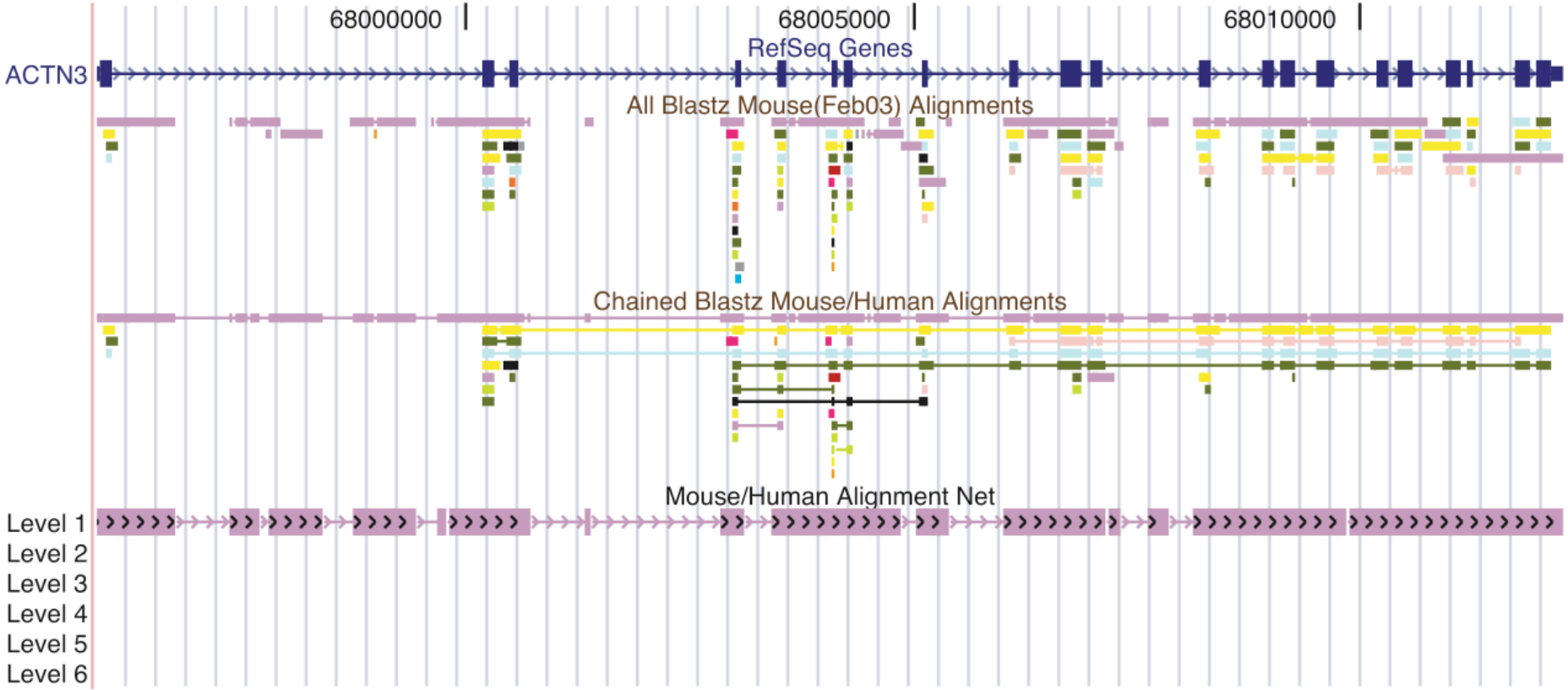
Figure 1 | **Overview of comparative sequence analysis. a** | Genomes of different species are sequenced by various strategies and assembled by computational algorithms. **b,c** | Homologous collinear segments are then identified and aligned. **d** | Finally, downstream analyses such as identifying constrained sequences can be carried out.

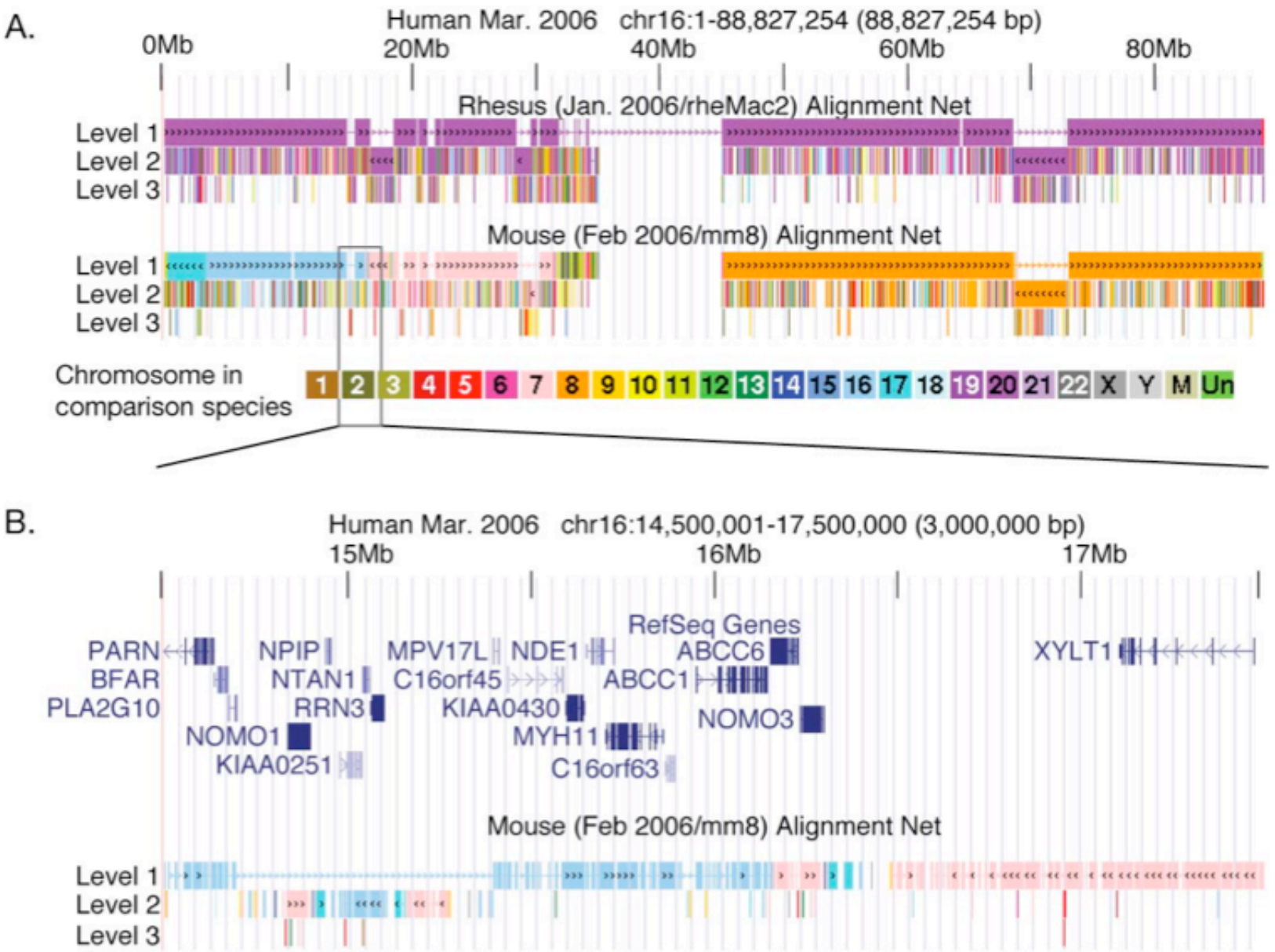
Various methods

- Ensembl
  - *Mercator* for homologous collinearity
  - *PECAN* alignments
  - *GERP* constraints
- UCSC Genome Browser
  - *Chains-and-nets* for homologous collinearity
  - *MultiZ* alignments
  - *Phastcons* constraints

- Local alignments
  - E.g., BlastZ
- Chains
  - Collection of local alignments can be organized into chains to maintain the order of DNA segments along the chromosome
- Nets
  - The chains can be nested in a group, called a net
    - These are used to navigate local alignments through rearrangements

Chains & nets (UCSC)




A.
Human Mar. 2006 chr16:1-88,827,254 (88,827,254 bp)
0Mb
20Mb
40Mb
60Mb
80Mb
Rhesus (Jan. 2006/rheMac2) Alignment Net
Level 1
Level 2
Level 3
Mouse (Feb 2006/mm8) Alignment Net
Level 1
Level 2
Level 3
Chromosome in comparison species
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y M Un
B.
Human Mar. 2006 chr16:14,500,001-17,500,000 (3,000,000 bp)
15Mb
16Mb
17Mb
RefSeq Genes
PARN
NPIP
MPV17L
NDE1
ABCC6
XYLT1
BFAR
NTAN1
C16orf45
ABCC1
PLA2G10
RRN3
KIAA0430
NOMO3
NOMO1
MYH11
KIAA0251
C16orf63
Mouse (Feb 2006/mm8) Alignment Net
Level 1
Level 2
Level 3

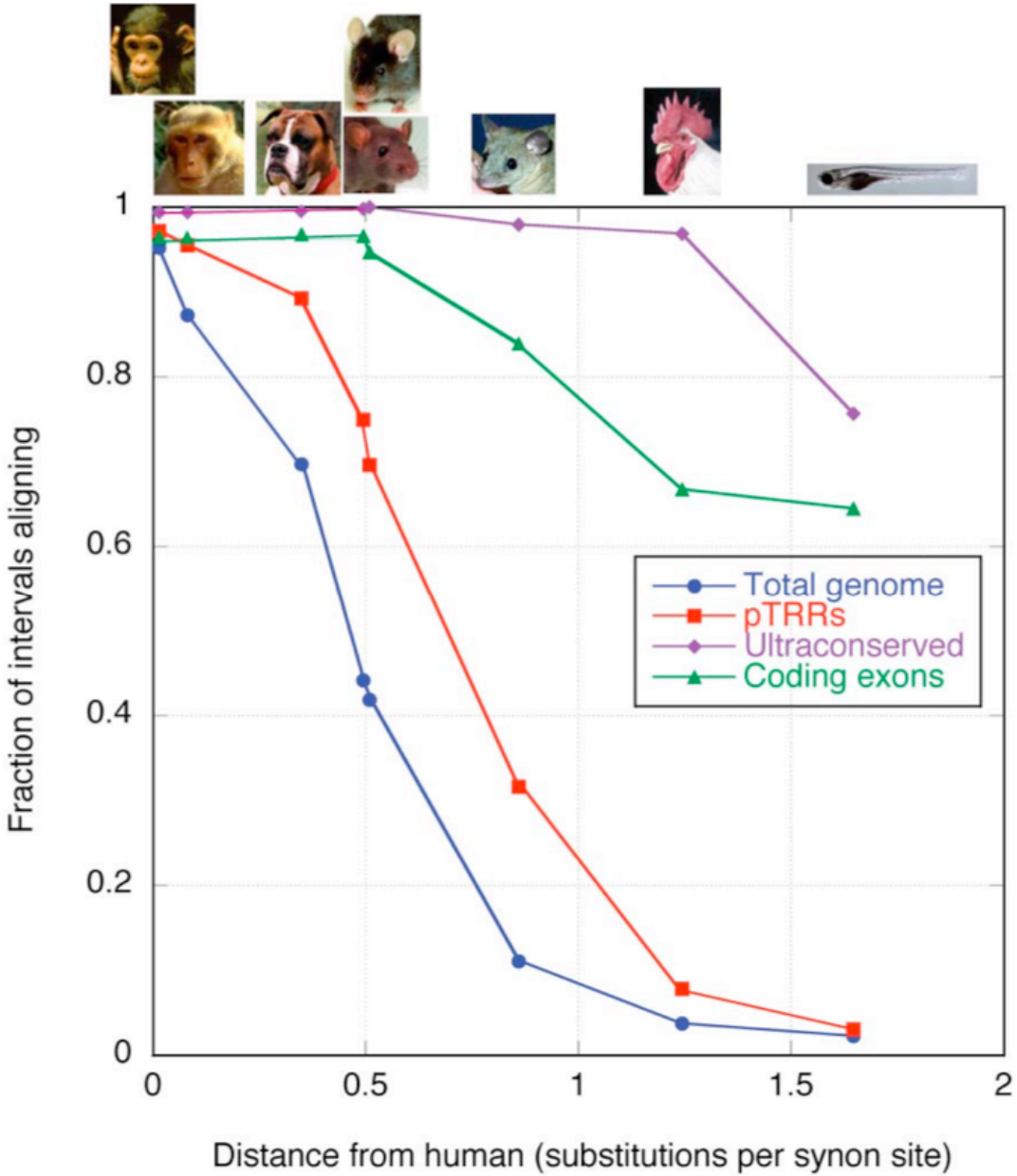
Assessment of conserved function from alignments

Phylogenetic depth of alignments

Table Q.2. Portions of the human genome conserved and constrained between various species.

	Distance from human		Fraction of human intervals aligning to comparison species [d]			
Comparison species [a]	Divergence time (Myr) [b]	Substitutions per synonymous site [c]	Total genome [e]	Coding exons [f]	Regulatory regions [g]	UCEs [h]
chimpanzee	5.40	0.015	0.95	0.96	0.97	0.99
macaque	25.0	0.081	0.87	0.96	0.96	0.99
dog	92.0	0.35	0.67	0.97	0.87	0.99
mouse	91.0	0.49	0.43	0.97	0.75	1.00
rat	91.0	0.51	0.41	0.95	0.70	1.00
opossum	173	0.86	0.10	0.82	0.32	0.95
chicken	310	1.2	0.037	0.67	0.06	0.95
zebrafish	450	1.6	0.023	0.65	0.03	0.76
Number			2.858×10^9 nucleotides	250,607	1369	481

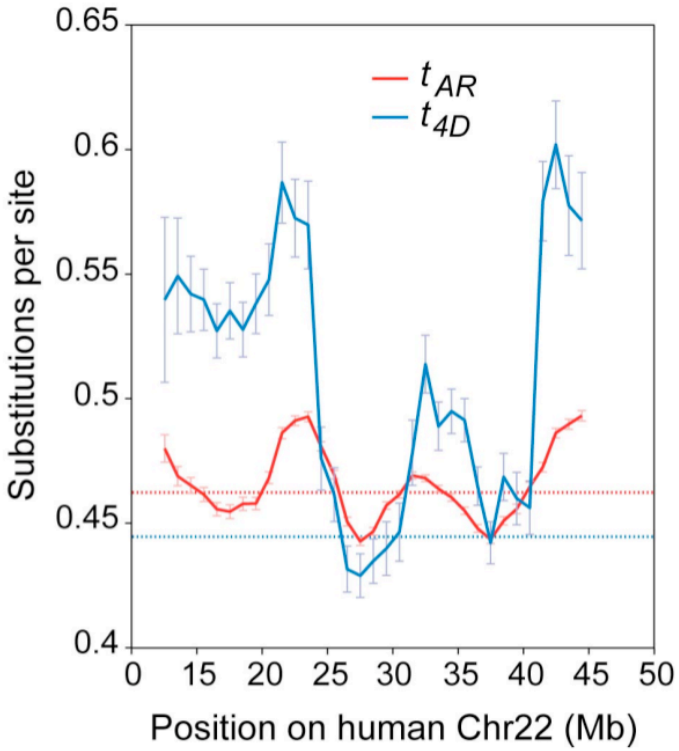
Alignments with distantly related species: indicator of constraint


Fraction of intervals aligning
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
Distance from human (substitutions per synon site)
Total genome
pTRRs
Ultraconserved
Coding exons

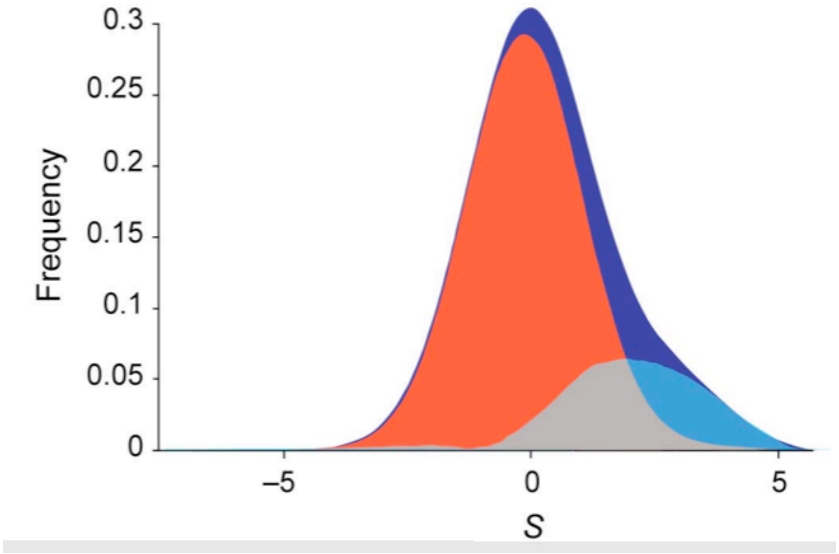
Portion of human genome under constraint

- All segments that align with a comparison species for a level of similarity higher than that seen for neutral DNA
  - But we don't know all the neutral DNA
  - Estimated neutal sites show substantial local variation across the human genome

fig= human-mouse



- Distribution of all similarity scores (blue) in human genome (compared with mouse)
  - S is adjusted for variation in the local substitution rate (similarity exceeds the similarity expected from the amount of divergence in ancestral repeats in the vicinity)
  - 50bp windows
- Distribution of similarity scores in ancestral repeats (red)
- Inferred distribution of scores for regions under constraint
  - 5% of human genome under continuous purifying selection since divergence of primates from carnivores and rodents (85-500 Mio yr)



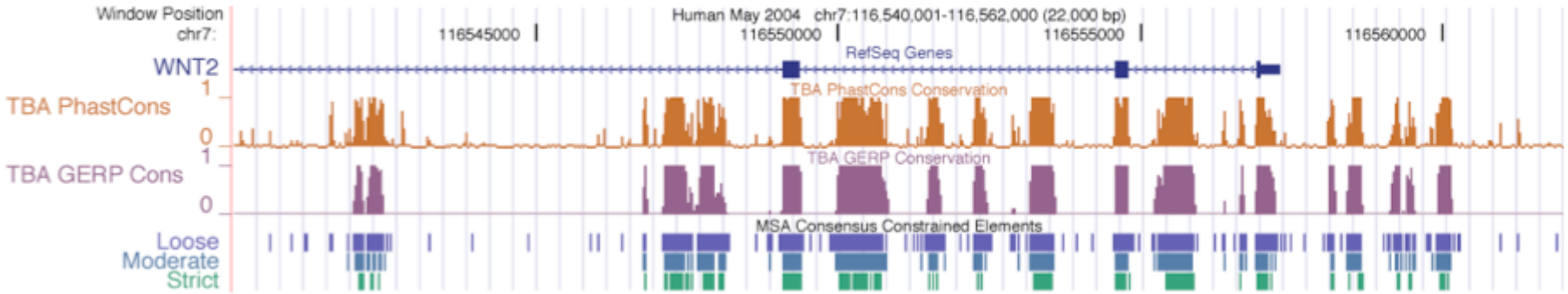
- 5% under constraint
  - Lower bound estimate of the portion of the human genome that is functional
    - DNA sequences not included in this estimate:
      - Those diverged for new functions in different lineages
      - Those acquired new function recently through adaptive evolution
  - 1.2% protein-coding
  - 0.7% UTR of mature mRNA
  - Remaining 3% (larger than ~2% for mRNA!)
    - Noncoding RNAs
    - Regulatory sequences

PhastCons

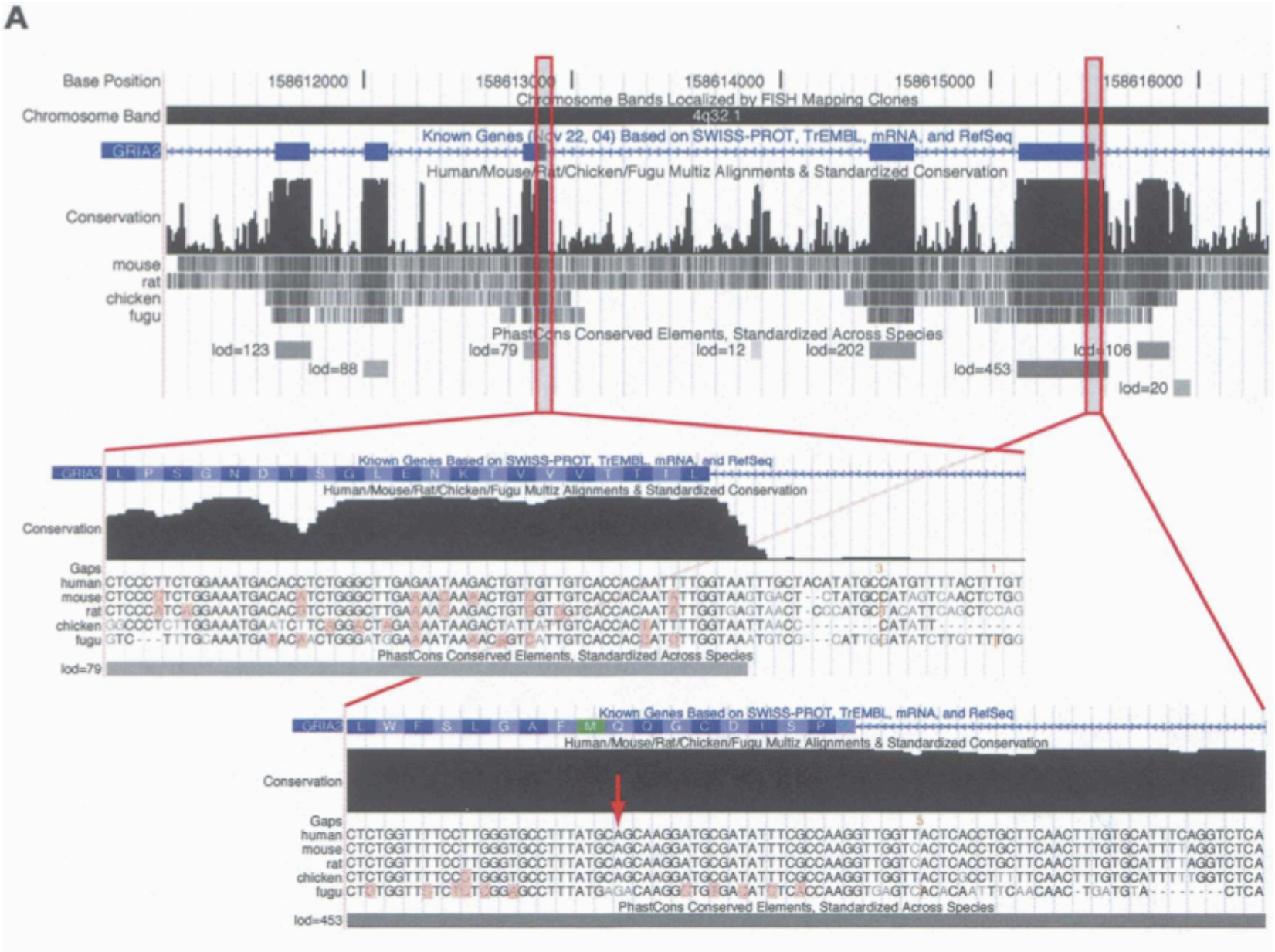
- Before: most groups have used pairwise alignments and simple percent identity-based methods
  - E.g,. At least 100bp with >70% identity
  - E.g., VISTA, zPicture
- Multiple genomes: use multiple alignments and use the phylogeny of the species

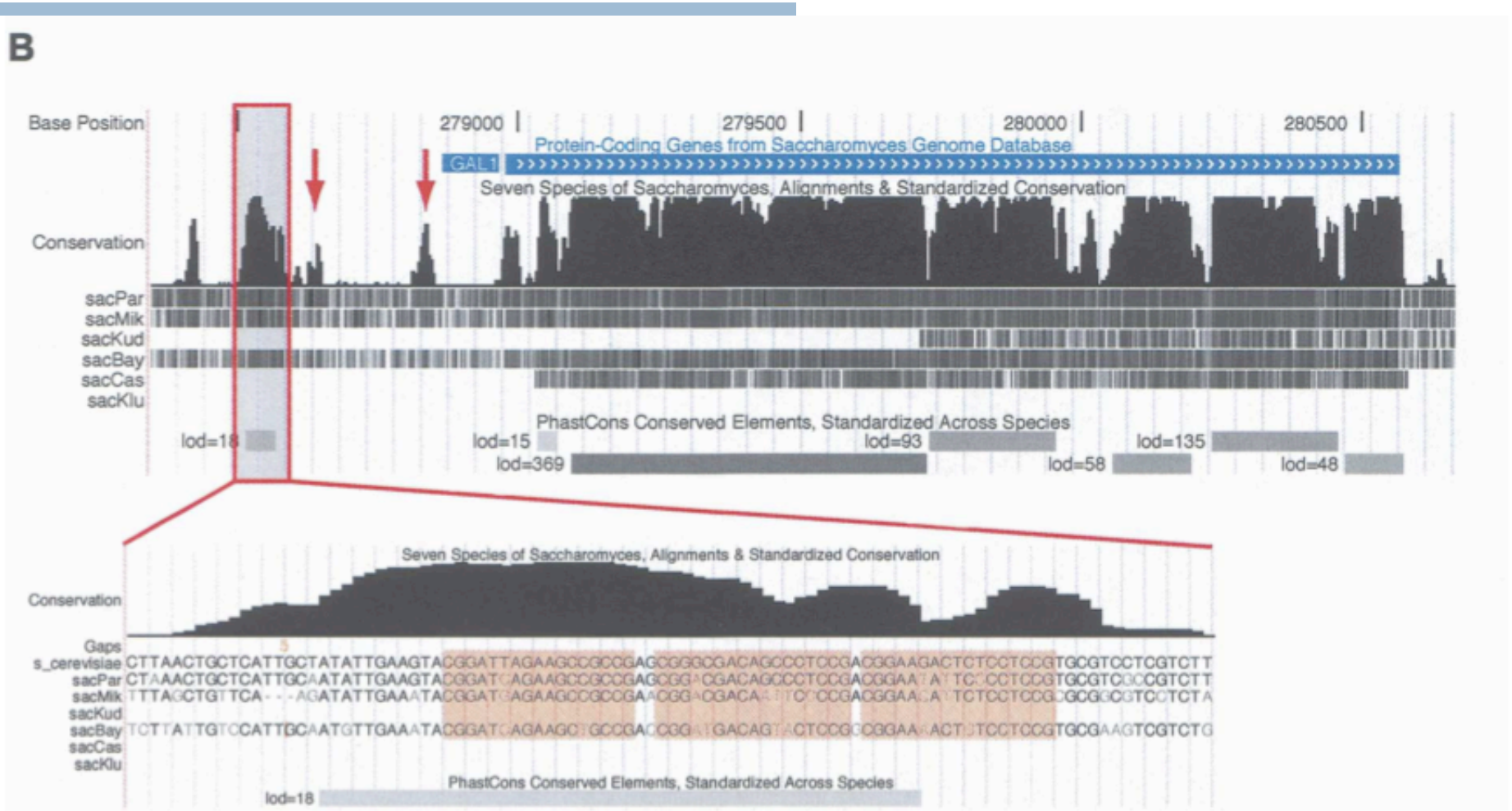
Identifying specific sequences under constraint

- phastCons
  - Phylogenetic Hidden Markov Model
  - Two states of conservation: one neutral and one constrained
  - Posterior probability that any aligned position came from the constrained state



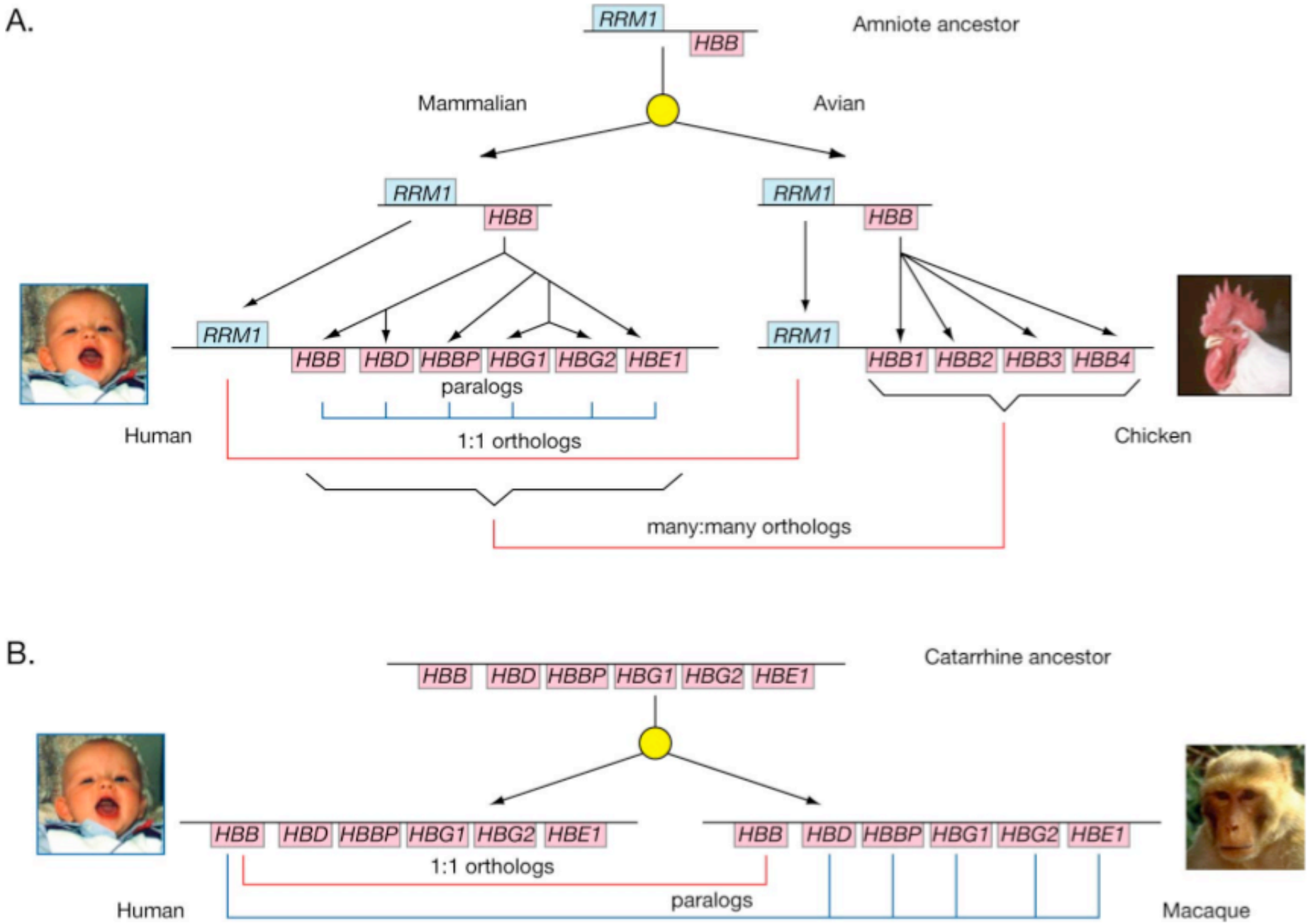
PhastCons examples



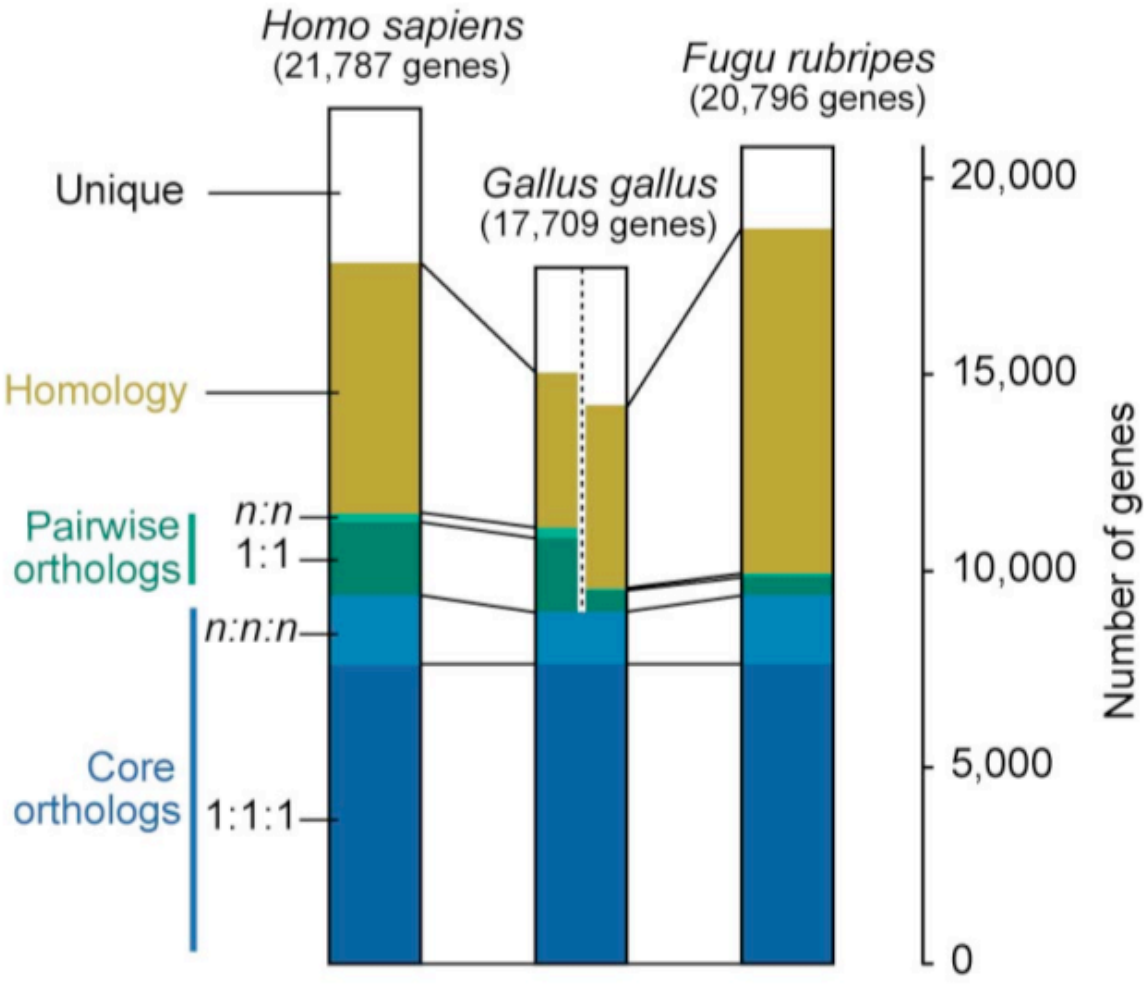

B
Base Position
279000
279500
280000
280500
Protein-Coding Genes from Saccharomyces Genome Database
GAL1
Seven Species of Saccharomyces, Alignments & Standardized Conservation
Conservation
sacPar
sacMik
sacKud
sacBay
sacCas
sacKlu
PhastCons Conserved Elements, Standardized Across Species
lod=18
lod=15
lod=93
lod=135
lod=369
lod=58
lod=48
Seven Species of Saccharomyces, Alignments & Standardized Conservation
Conservation
Gaps
s_cerevisiae
sacPar
sacMik
sacKud
sacBay
sacCas
sacKlu
PhastCons Conserved Elements, Standardized Across Species
lod=18

Evolution within protein-coding genes

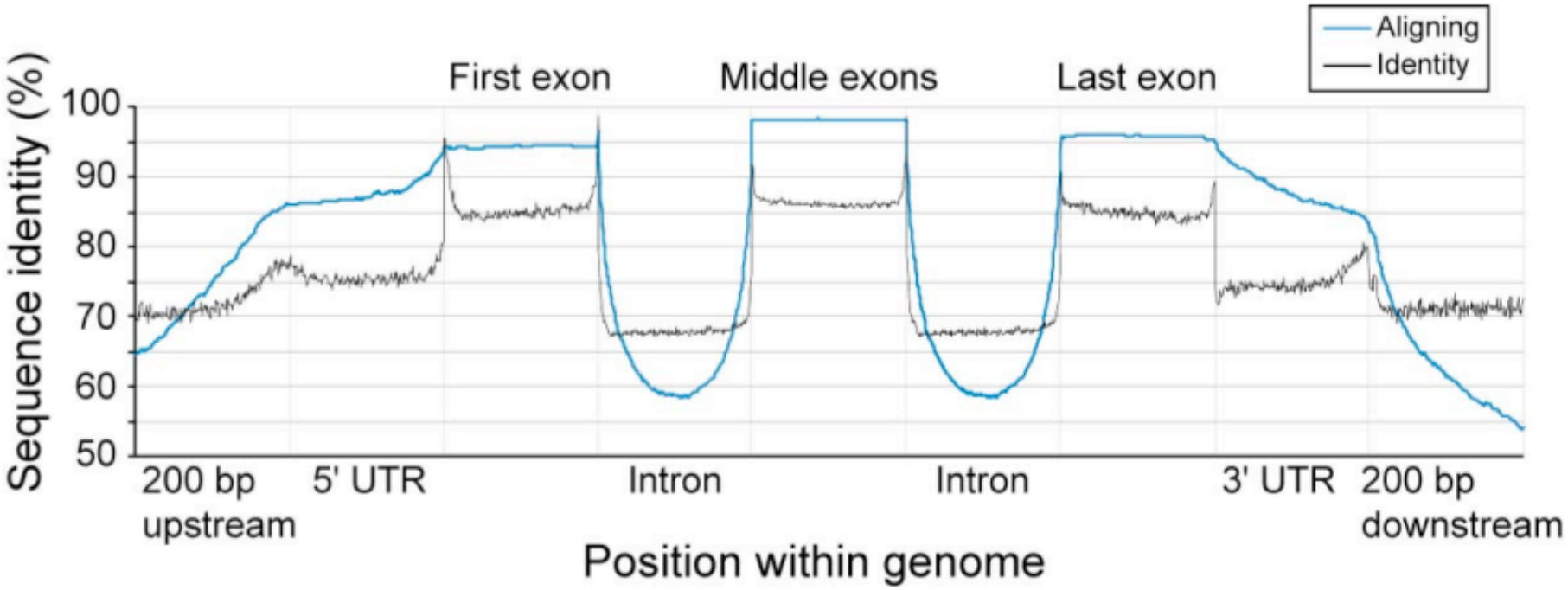
- First define a set of genes in each species
  - Evidence-based
    - ESTs, mRNAs, proteins
  - *Ab initio* (Hidden Markov Models)
    - *Genscan, genmark*
  - Adding alignments of other species
    - *Twinscan, exoniPhy*
- Then define a set of orthologous genes
  - Homologous genes that separated because of a speciation event
    - RRM1 1:1 relationship
      - Any changes can be interpreted as changes since the time of divergence of the species
    - Beta-like globin genes: members of multigene families: distinguish genes that separated as a result of gene duplication (*paralogous* genes) from the orthologous genes (separated by speciation event)


A.
RRM1
HBB
Amniote ancestor
Mammalian
Avian
RRM1
HBB
RRM1
HBB
RRM1
HBB
HBD
HBBP
HBG1
HBG2
HBE1
RRM1
HBB1
HBB2
HBB3
HBB4
paralogs
Human
1:1 orthologs
Chicken
many:many orthologs
B.
HBB
HBD
HBBP
HBG1
HBG2
HBE1
Catarrhine ancestor
HBB
HBD
HBBP
HBG1
HBG2
HBE1
HBB
HBD
HBBP
HBG1
HBG2
HBE1
1:1 orthologs
paralogs
Human
Macaque

- 1:1 orthologs: reciprocal highest similarity is good guide



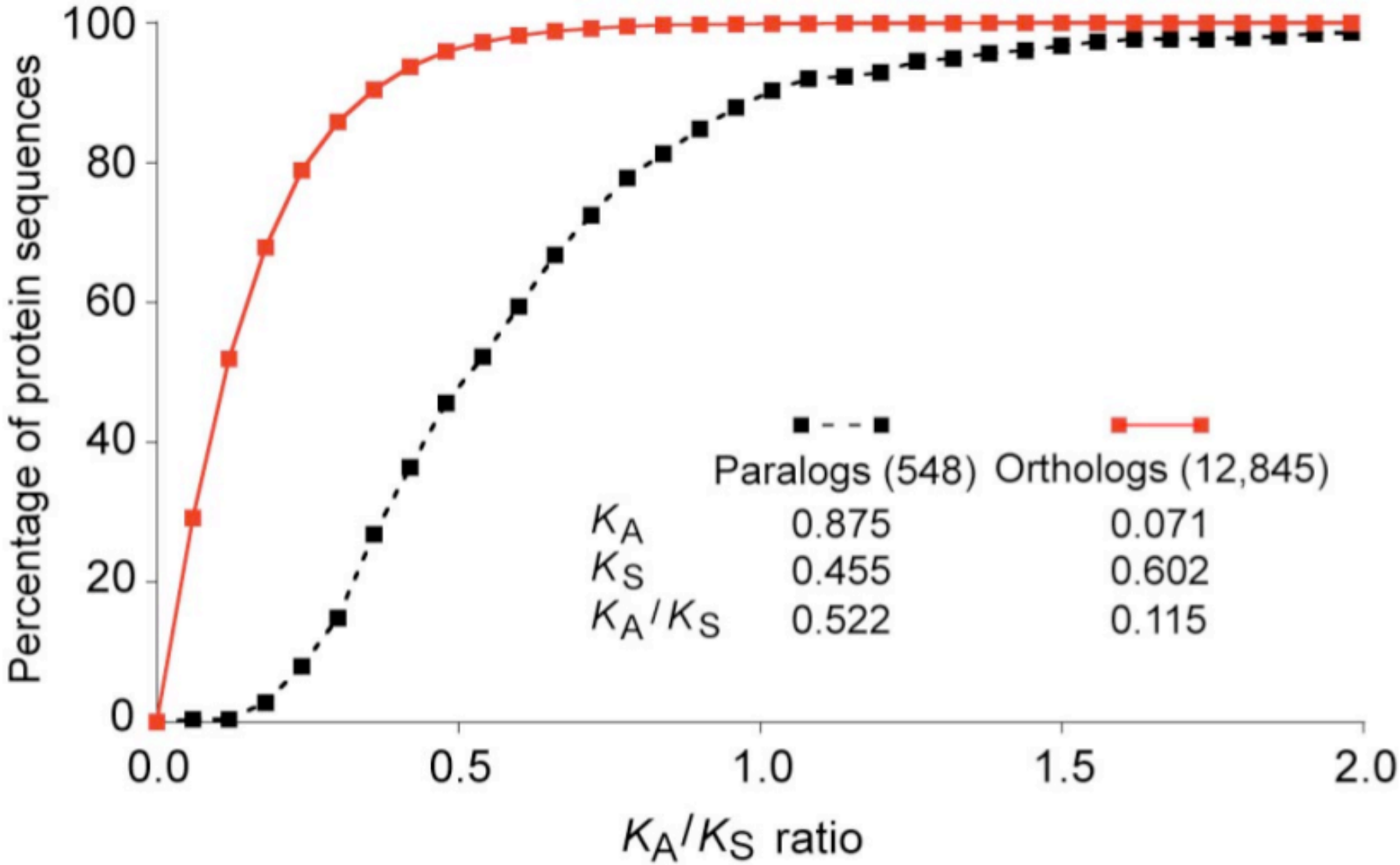
Different parts of genes



Protein-coding exons

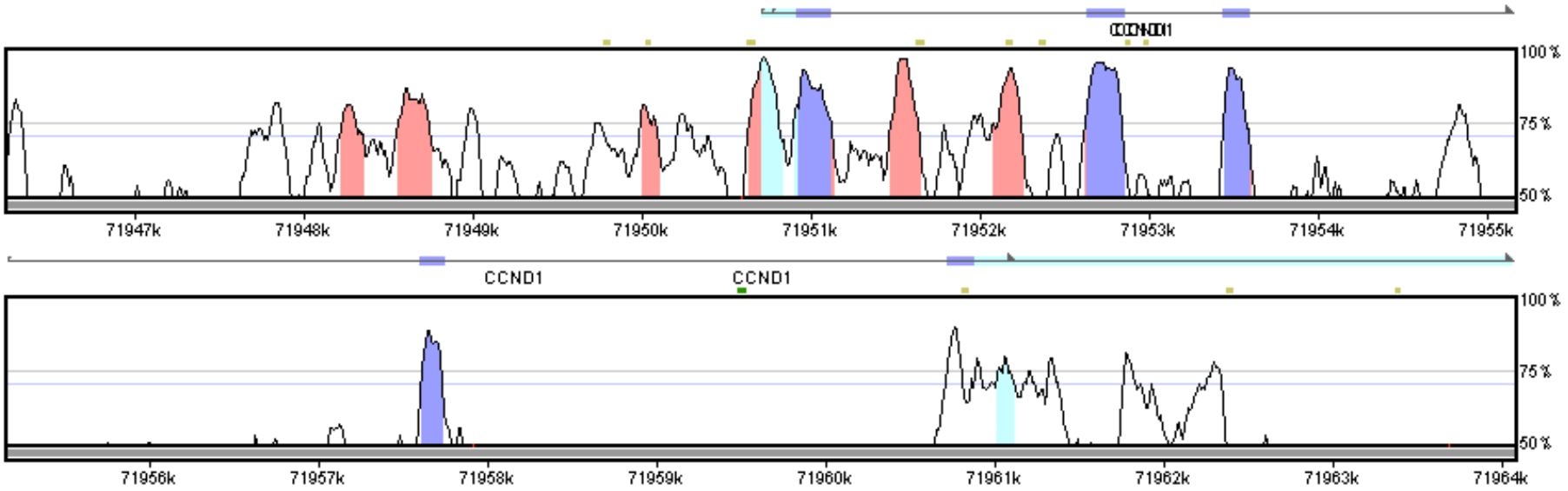
- Histone H4: only 1 AA change between peas and cows
- Fibrinopeptides (segments of fibrinogen cleaved off by thrombin during blood clotting): little or no selective pressure (AA sequence seems not important for its function)

- Ks = # synonymous substitutions per synonymous site (estimate of neutral rate)
- Ka = # non-synonymous substitutions per non-synonymous site (estimate of stringency of the purifying selection or strength of adaptive evolution)
- Ka/Ks ~ 0.2 : indicative of constraint
- Ka/Ks > 1 : adaptive evolution


Percentage of protein sequences
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
K_A/K_S ratio
Paralogs (548)
Orthologs (12,845)
K_A 0.875 0.071
K_S 0.455 0.602
K_A/K_S 0.522 0.115

Evolution in non-protein-coding regions

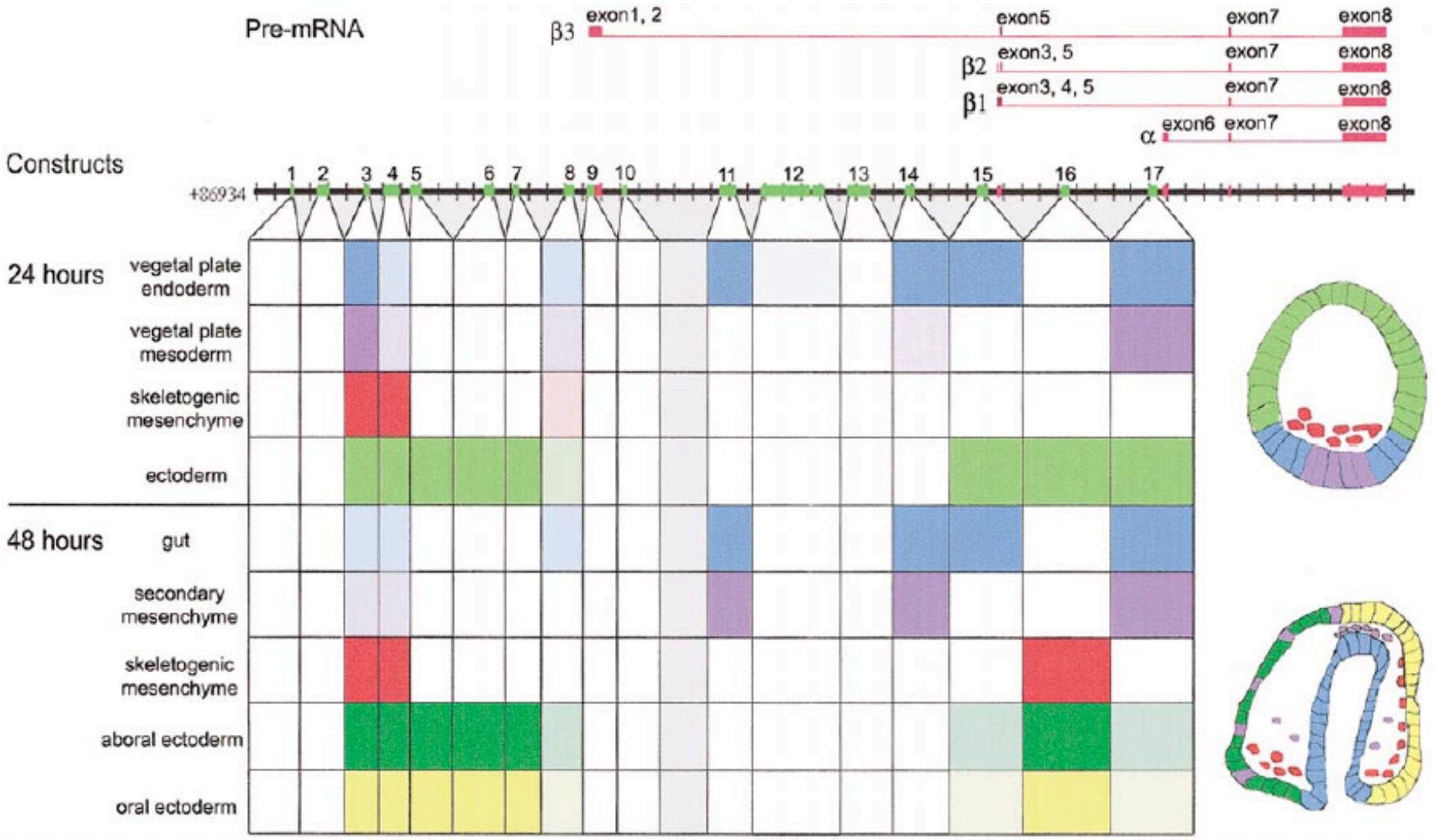
Conserved non-coding sequences (CNS)



AVID, LAGAN, BLASTZ

Conserved non-coding sequences

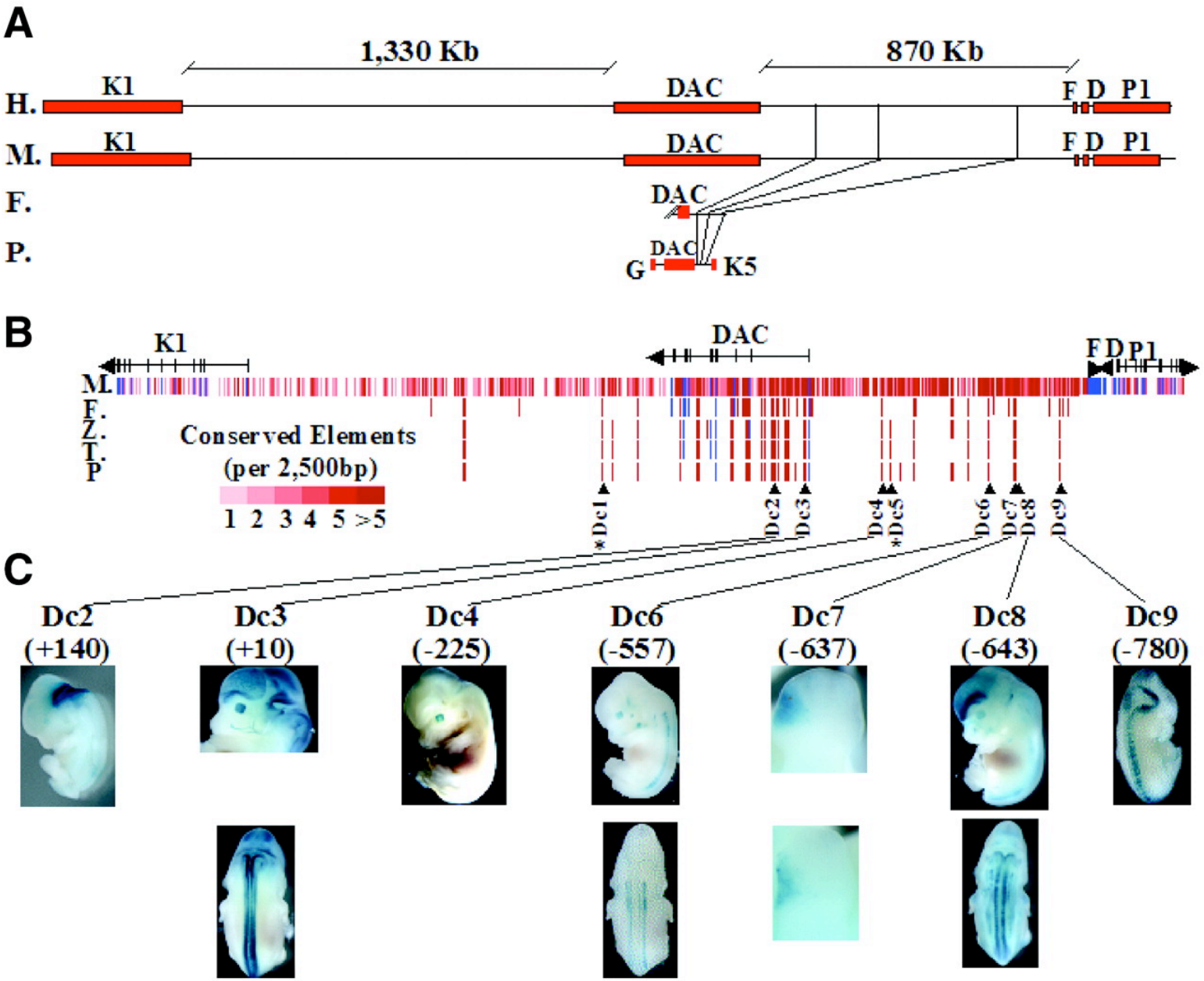
Strongylocentrotus purpuratus - Lytechinus variegatus: otx gene Yuh et al. (2002) *Dev Biol* 246:1

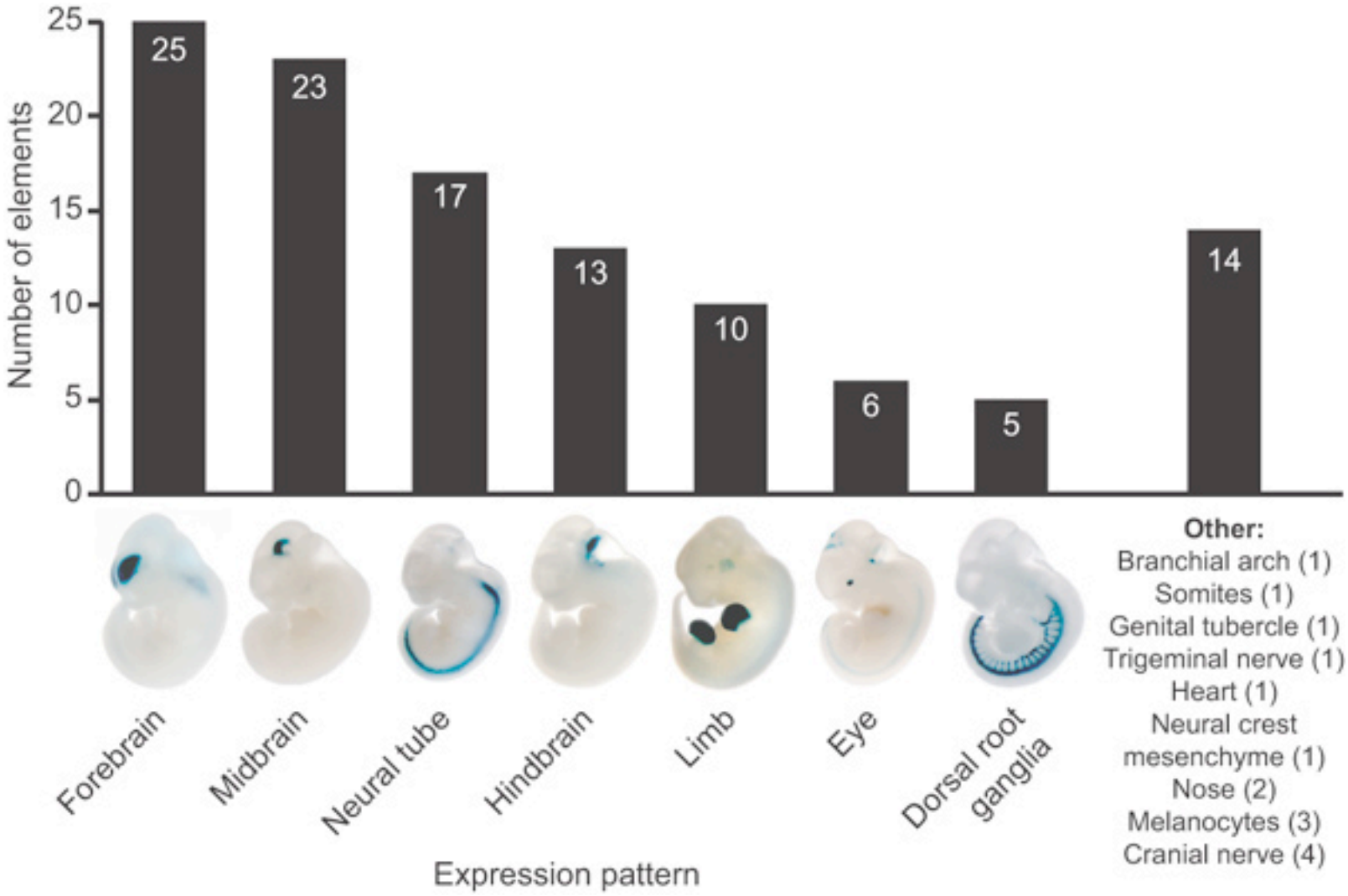


Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJ, Davidson EH. Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. Dev Biol. 2002 Jun 1;246(1):148-61.

Conserved non-coding sequences

Nobrega et al. (2003) *Science* 302 (5644)





76 (45%) of 167 tested enhancer show expression

The majority in developing nervous system

Pennachio et al., Nature 2006

VISTA Enhancer Browser

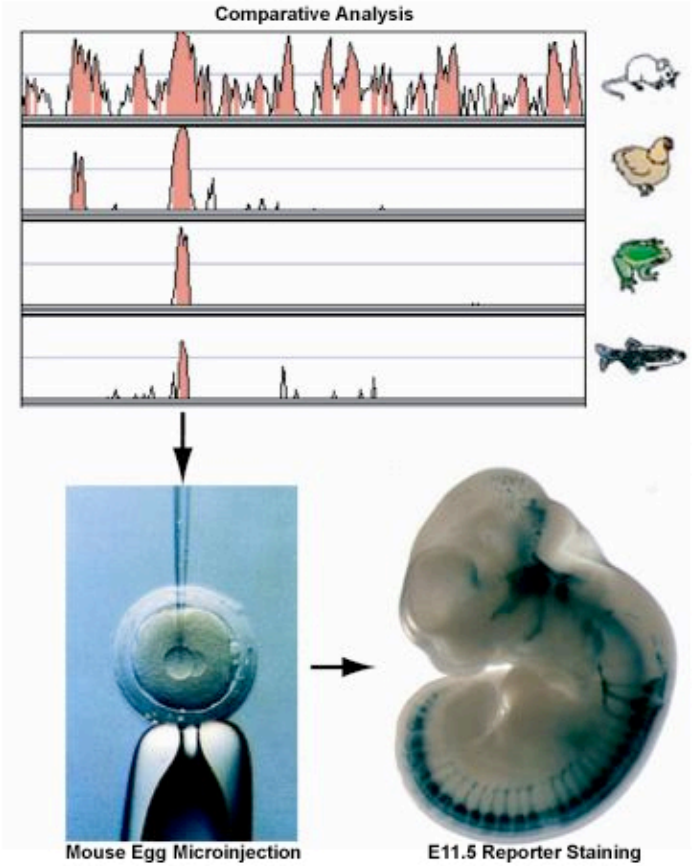
whole genome enhancer browser

The **VISTA Enhancer Browser** is a central resource for experimentally validated human noncoding fragments with gene enhancer activity as assessed in transgenic mice. Most of these noncoding elements were selected for testing based on their extreme conservation with other vertebrates. The results of this enhancer screen are provided through this publicly available website.

This program is located at Lawrence Berkeley National Laboratory. See Handbook for additional details on this work or visit the Experimental Results to view data.

The browser also features relevant results by external contributors and a large collection of additional genome-wide conserved noncoding elements which are candidate enhancer sequences. We invite external groups to submit computational predictions of developmental enhancers.

As of **10/27/2008** the database contains information on **867** in vivo tested elements - **353** elements with enhancer activity.



Keyword Search

Search

Examples: gene, accession number, locus link, genomic position

Expression Pattern Search

Advanced Search

Regulatory regions ~ conserved non-coding sequences

- Pennachio etc.
  - 2 genomes, pairwise, sliding window, min % identity
- Now: rather using multiple-alignment based constraint estimates (phastCons, GERP, binCons)

Exercise 1

- What is the mouse orthologous gene of human albumin preproprotein (ALB)?

Exercise 2

- Retrieve all conserved elements
  - PhastConsElements28way (‘most conserved’)
  - That don’t overlap with mRNA
    - Human mRNAs (all_mrna)
  - That don’t overlap with repeats
    - RepeatMasker (rmsk)
  - That have a score >= 300
- Look at ALB and AFP (liver function)
  - Make custom track of known regulatory elements
    - ftp://ftp.esat.kuleuven.be/sista/aerts/teaching/ComparativeGenomics/albumin_afp_elements_and_high_scoring.bed

Exercise 3

- Use Galaxy this time
  - Overlap phastcons regions (with score >= 300) with core promoters
  - Limit to chrom1
  - Use TSS => 200 bp upstream
    - Compare with 200bp downstream of TSS
    - Get BED from UCSC/Galaxy or get TSS BED from Mysql (flybase_gene, ucsc_gene, …) and work on TSS locations locally (-200 for + strand, +200 for - strand)
    - Do for vertebrates, for yeast, for flies, for worms

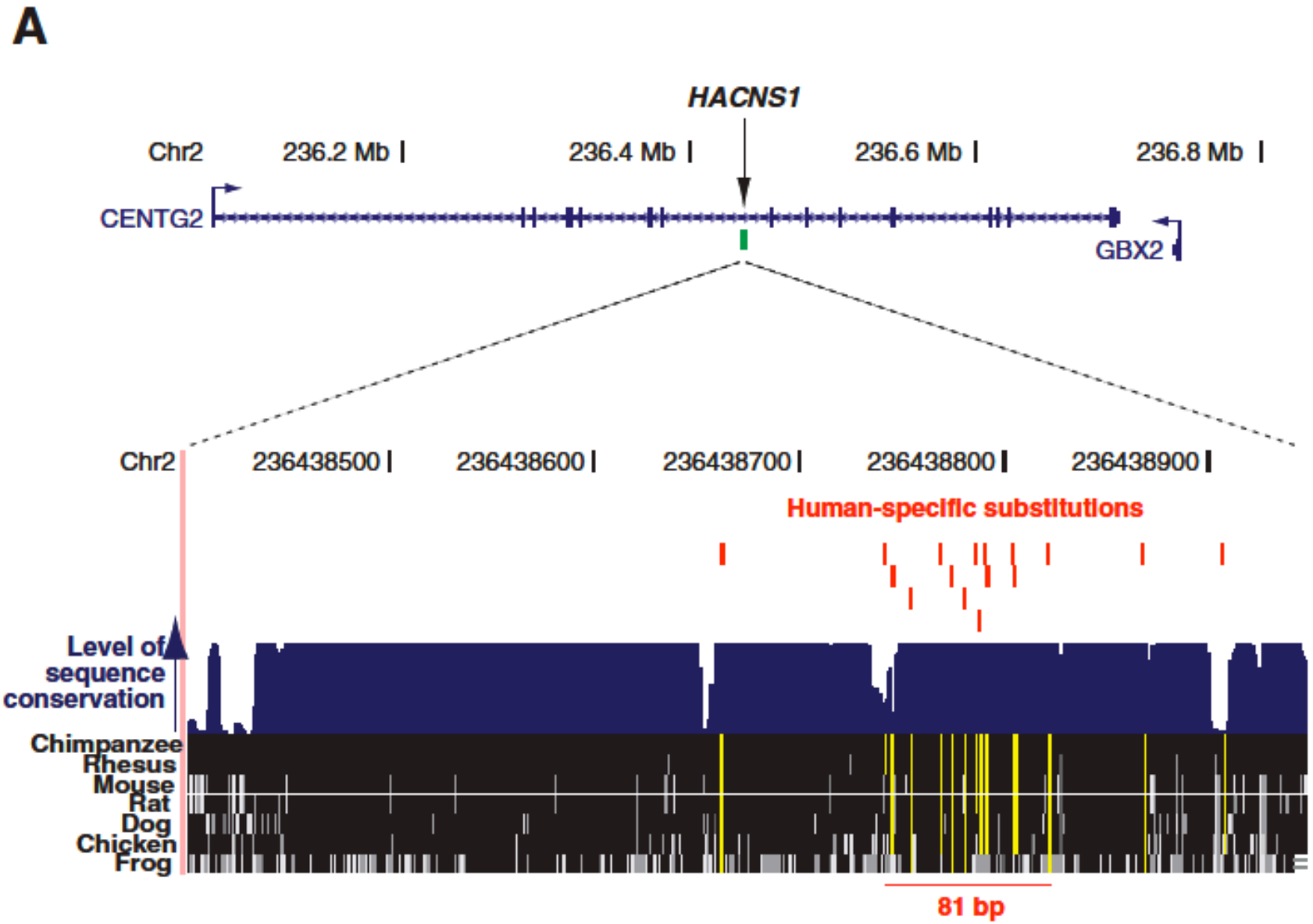
Ultraconserved elements (UCEs)

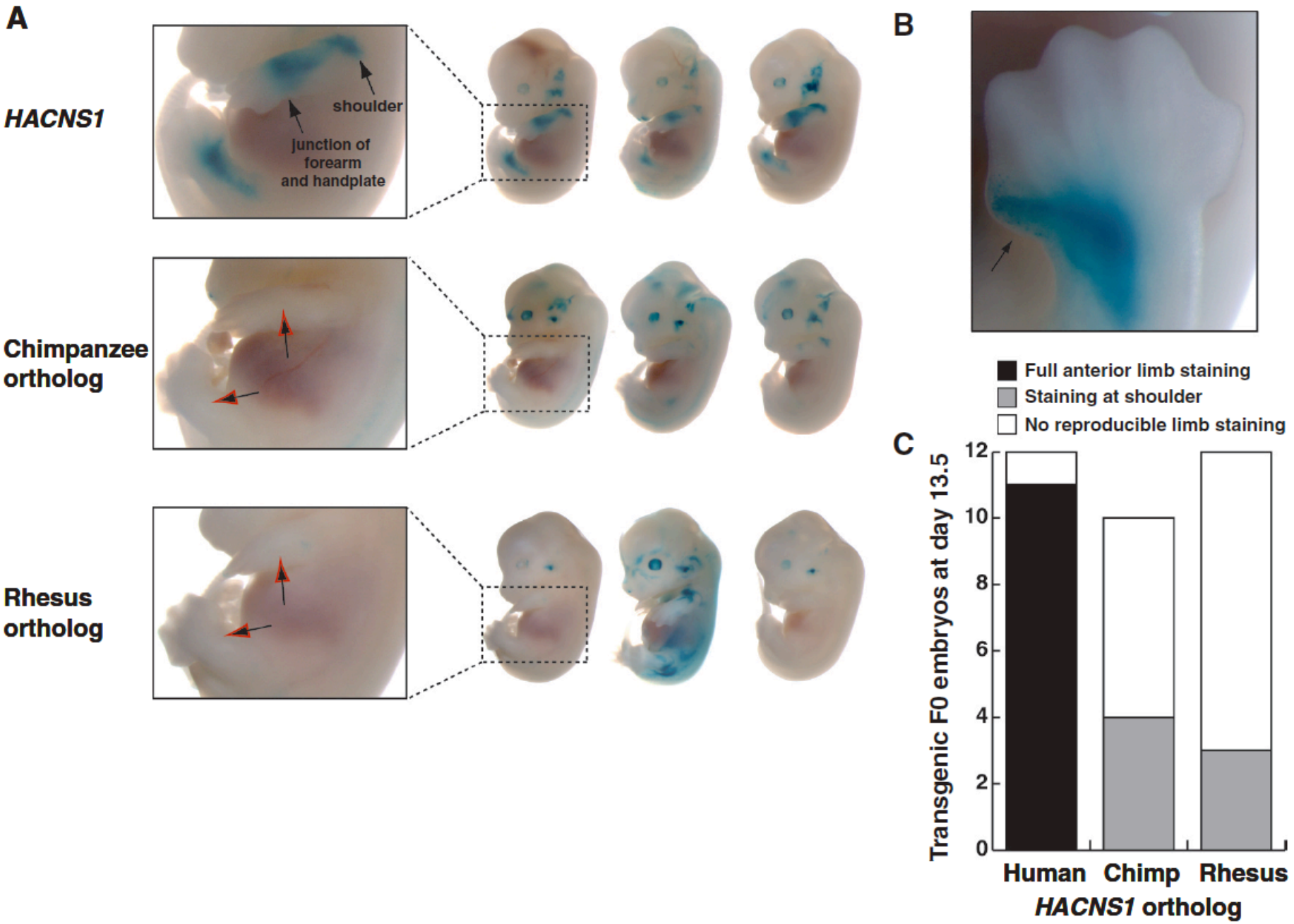
- 481 segments identical between human and mouse for at least 200bp
- ~ 1/2 of those tested serve as tissue-specific enhancers in transgenic mouse embryos
- UCEs: restricted to vertebrates
- While protein sequences are often conserved between vertebrates and invertebrates

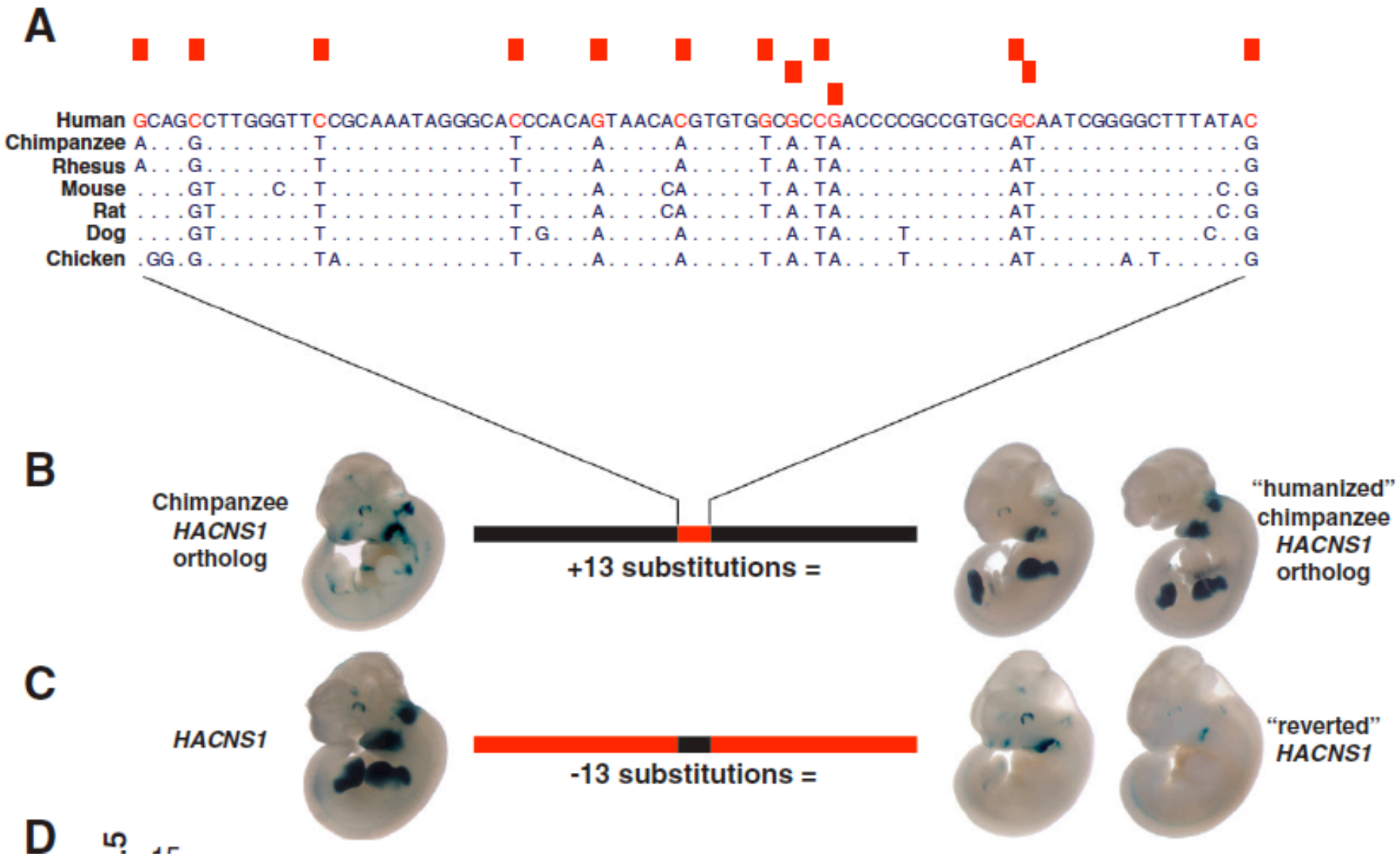
- Slowly changing regions across mammals with rapid change along the human lineage: *human accelerated regions*

Example of human accelerated conserved non-coding sequence

Prabhakar et al Science Sep 2008




A
HACNS1
shoulder
junction of forearm and handplate
Chimpanzee ortholog
Rhesus ortholog
B
C
Full anterior limb staining
Staining at shoulder
No reproducible limb staining
Transgenic F0 embryos at day 13.5
0
2
4
6
8
10
12
Human
Chimp
Rhesus
HACNS1 ortholog


A
Human GCAGCCTTGGGTTCCGCAAATAGGGCACCCACAGTAACACGTGTGGCGCCGACCCCGCCGTGCGCAATCGGGGCTTTATAC
Chimpanzee A...G.....T.....T.....A.....A.....T.A.TA.....AT.....G
Rhesus A...G.....T.....T.....A.....A.....T.A.TA.....AT.....G
Mouse ....GT....C..T.....T.....A....CA.....T.A.TA.....AT.....C.G
Rat ....GT.....T.....T.....A....CA.....T.A.TA.....AT.....C.G
Dog ....GT.....T.....T.G...A.....A.....A.TA....T.....AT.....C..G
Chicken .GG.G.....TA.....T.....A.....A.....T.A.TA....T.....AT.....A.T.....G
B
Chimpanzee HACNS1 ortholog
+13 substitutions =
"humanized" chimpanzee HACNS1 ortholog
C
HACNS1
-13 substitutions =
"reverted" HACNS1
D

Discovery of short regulatory motifs by comparative genomics

articles

Sequencing and comparison of yeast species to identify genes and regulatory elements

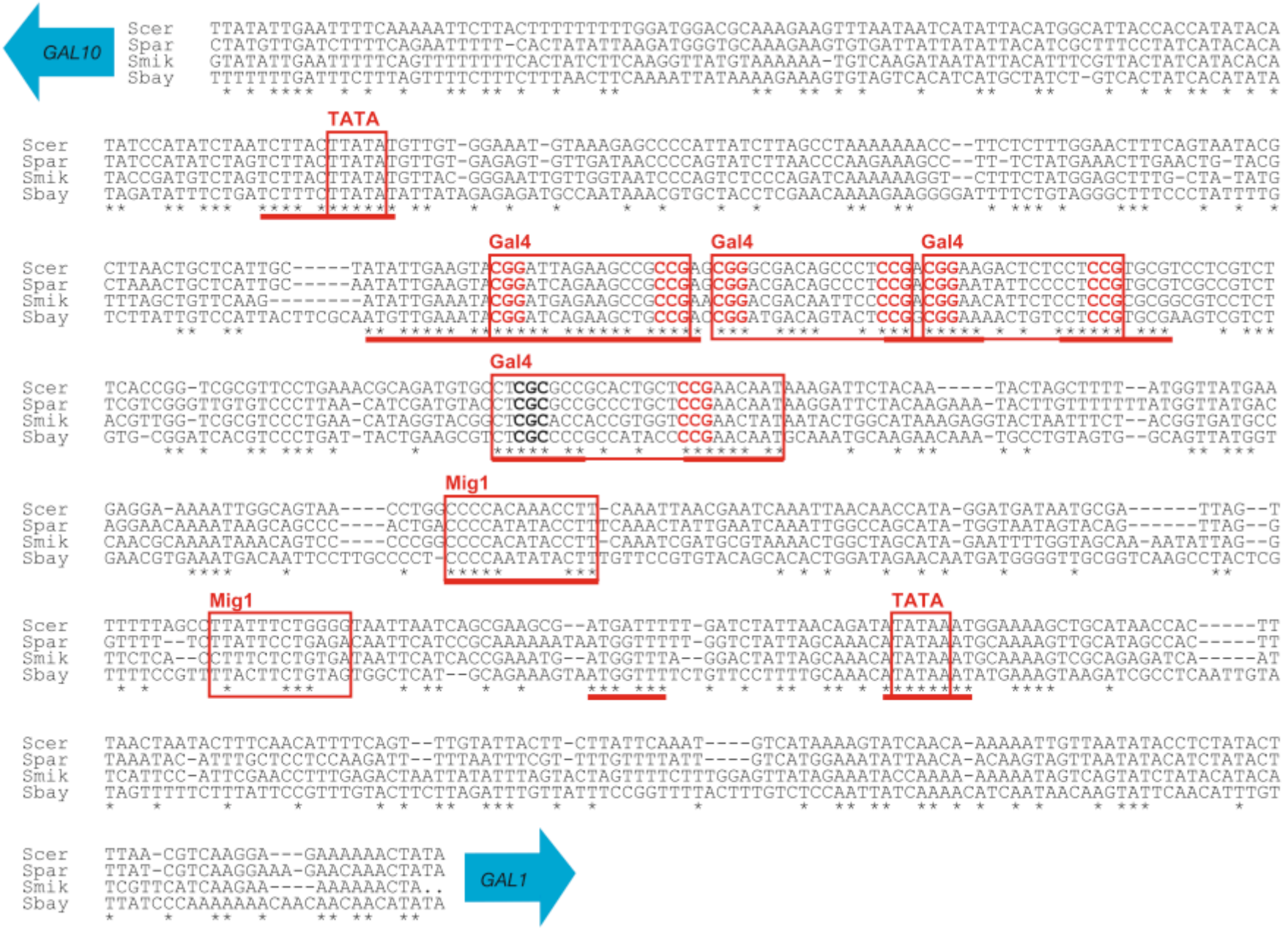
Manolis Kellis*†, Nick Patterson*, Matthew Endrizzi*, Bruce Birren* & Eric S. Lander*‡

** Whitehead/MIT Center for Genome Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA*
† Department of Computer Science and ‡ Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Identifying the functional elements encoded in a genome is one of the principal challenges in modern biology. Comparative genomics should offer a powerful, general approach. Here, we present a comparative analysis of the yeast *Saccharomyces cerevisiae* based on high-quality draft sequences of three related species (*S. paradoxus*, *S. mikatae* and *S. bayanus*). We first aligned the genomes and characterized their evolution, defining the regions and mechanisms of change. We then developed methods for direct identification of genes and regulatory motifs. The gene analysis yielded a major revision to the yeast gene catalogue, affecting approximately 15% of all genes and reducing the total count by about 500 genes. The motif analysis automatically identified 72 genome-wide elements, including most known regulatory motifs and numerous new motifs. We inferred a putative function for most of these motifs, and provided insights into their combinatorial interactions. The results have implications for genome analysis of diverse organisms, including the human.

Nature, 2003

Regulatory elements in 4 yeast species



Gal4

- 96 times in intergenic regions and 415 times in genic regions
- occurrences of the Gal4 motif in intergenic regions have a conservation rate (proportion conserved across all four species) that is about fivefold higher than for equivalent random motifs (12.5% compared with 2.4%)
- intergenic occurrences of the Gal4 motif are more frequently conserved than genic occurrences (12.5% compared with 3%). By contrast, random motifs are less frequently conserved in intergenic regions than in genic regions (3.1% compared with 7.0%), reflecting the lower overall level of conservation in intergenic regions.
- Third, the Gal4 motif shows a higher conservation rate in divergent compared with convergent intergenic regions (those that lie upstream compared with downstream of both flanking genes); no such preferences are seen for control motifs.

Known motifs

- 55 (from SCPD and YTFD)
- conservation score (MCS) on the basis of the conservation rate of the motif in intergenic regions: the MCS is measured in standard deviations above the rate for comparable control motifs
- Most of the known motifs show strong conservation, with 60% having MCS>4 (which is substantially higher than expected by chance). Some of the motifs, however, show relatively modest MCS. These motifs may be incorrect, suboptimal or not well conserved.

Table 2 **Known motifs and related discovered motifs**

	Known motif	
Factor	Motif	MCS†
ABF1	**RTCRYnnnnnACG**	50.0
UME6	**TCGGCGGCTA**	20.9
CBF1	**RTCACRTG**	19.0
NDT80	**TCGGCGGCTDW**	18.6
REB1	**TTACCCGG**	17.8
MCM1a	TT**WCC**Cn**WWW**R**GGAAA**	16.5
SWI6	**ACGCGT**	16.4
PHO4	**CACGTG**	16.1
MBP1	**ACGCGTnA**	14.8
SWI4	**TTTTCGCG**	12.4
DAL81	GATAAG	12.1
RPN4	**TTTTGCCACC**	11.5
MSN2	**CCCCT**	11.3
MSN4	**CCCCT**	11.3
PDR1	**CCGCGG**	9.3
ESR2	**AAAAWTTTT**	8.9
MIG1	**CCCCRS**WWWW	8.7
MIG1b	**CCCCGC**	8.4
BAS1	**TGACTC**	8.3
GCN4	**ATGACTCAT**	8.2
GAL4	**CGGnnnnnnnnnnnCCG**	8.0
HSF1b	**TTCTAGAA**	7.8
ESR1	**GATGAG**	7.7
MET31	AAA**CTGTGGC**	6.8
AFT1	Y**RCACCC**R	6.8
TEA1	CGGnCGG	6.8
PUT3	**CGG**nnnnnnnnnn**CCG**	6.2
HAP2	**TGATTGG**C	5.7
RAP1	**ACACCCATACAT**TT	5.2
LEU3	**CCGGnnCCGG**	4.9
MCM1b	Y**TTCCTAATT**W**G**nnCn	4.8
INO4	**CATGTGAA**AT	4.1
INO2	**CATGTG**AAAT	4.1
GLN3	GATAAK	3.8
ADR1	GGAGA	3.7
FKH2	**TTGTTTAC**ST	3.6
FKH1	**TTGTTTAC**ST	3.6
RLM1	**CTAWW**W**WT**AG	3.6
SWI5	**KGCTGR**	3.4
HAP1	C**G**Gnnn**TAnCGG**	2.5
XBP1	M**CTCGARRR**nR	2.5
MAC1	TT**TGCTCA**	2.3
TBF1	**TTAGGG**	2.3
MSE	**TTTTGTG**	1.4
STE12	R**TGAAACA**	0.7
DIG1	R**TGAAACA**	0.7
MET4	**TGGCAAA**TG	0.7
HAP4	**TnRTTGGT**	0.5
SMP1	ACTACTAWWWWTAG	0.4
ACE2	**GCTGGT**	−0.6
YAP1	TTACTAA	−1.1
CIN5	TTACTAA	−1.1
RME1	GAACCTCAA	−1.4
HAC1	CAGCGTG	−1.4
GCR1	**GGAAG**	−18.5

New motifs

- Mini-motifs
  - XYZn(0–21)UVW
    - TAGGAT, ATAnnGGC (45,760 motifs in total)
  - (1) intergenic conservation (CC1), the mini-motif shows a significantly high conservation rate in intergenic regions;
  - (2) intergenic-genic conservation (CC2), the mini-motif shows significantly higher conservation in intergenic regions than in genic regions;
  - (3) upstream-downstream conservation (CC3), the mini-motif shows significantly different conservation rates when it occurs upstream compared with downstream of a gene.
- 2,400 mini-motifs: high scores by one or more criteria

Full motifs

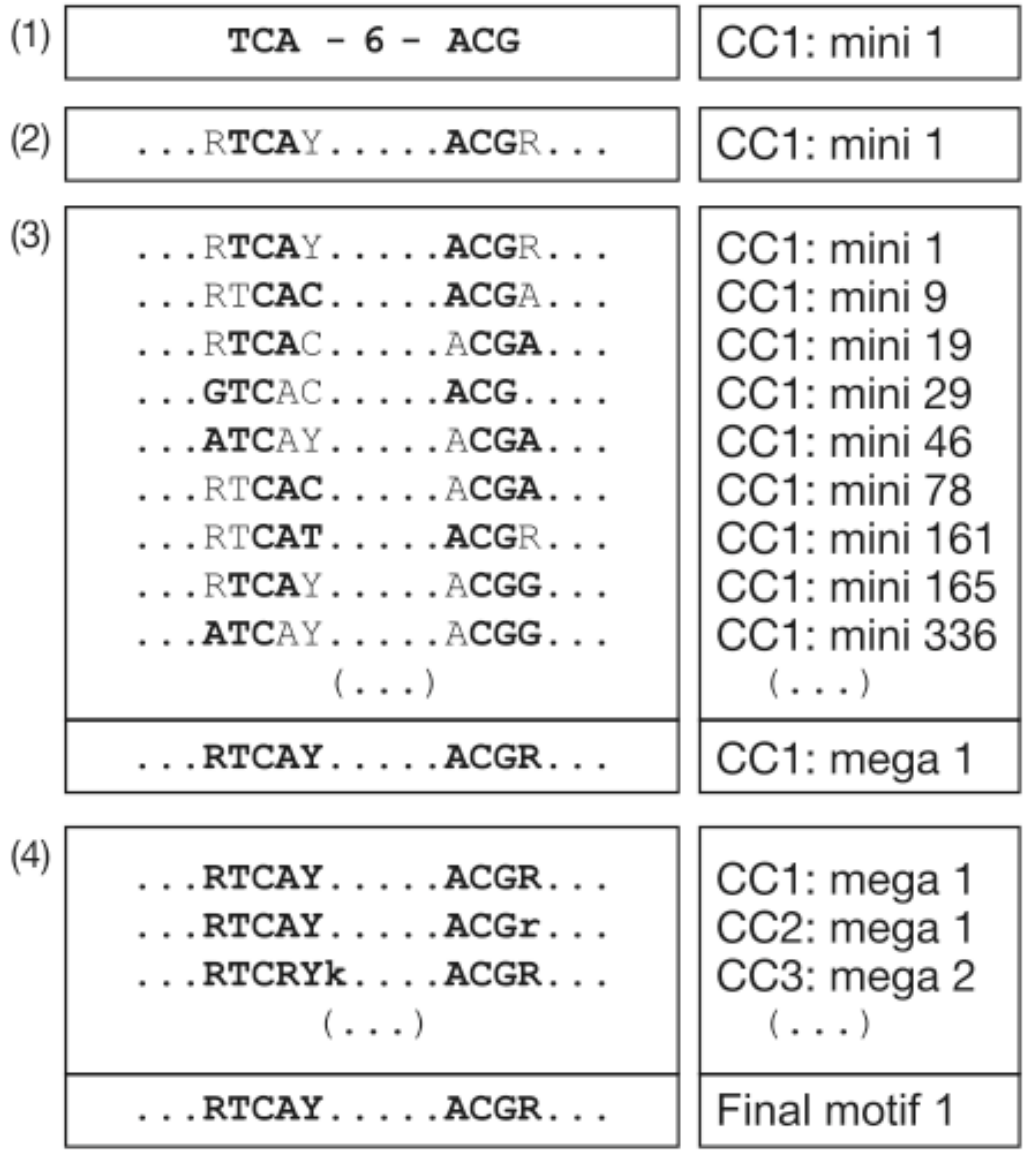

(1)
TCA - 6 - ACG
CC1: mini 1
(2)
...RTCAY.....ACGR...
CC1: mini 1
(3)
...RTCAY.....ACGR...
...RTCAC.....ACGA...
...RTCAC.....ACGA...
...GTCAC.....ACG....
...ATCAY.....ACGA...
...RTCAC.....ACGA...
...RTCAT.....ACGR...
...RTCAY.....ACGG...
...ATCAY.....ACGG...
(...)
CC1: mini 1
CC1: mini 9
CC1: mini 19
CC1: mini 29
CC1: mini 46
CC1: mini 78
CC1: mini 161
CC1: mini 165
CC1: mini 336
(...)
...RTCAY.....ACGR...
CC1: mega 1
(4)
...RTCAY.....ACGR...
...RTCAY.....ACGr...
...RTCRYk....ACGR...
(...)
CC1: mega 1
CC2: mega 1
CC3: mega 2
(...)
...RTCAY.....ACGR...
Final motif 1

Table 3 **Discovered motifs**

No.	Discovered motif	Location*	MCS†	Best category‡	CCS§	Interpretation‖
1	YCGTnnnnnRYGAY	5′	36.2	ChIP: Abf1	90	Known: Abf1
2	RTTACCCGRM	5′	34.3	ChIP: Reb1	38	Known: Reb1
3	gcGATGAGmtgaraw	5′	24.7	Exp.: cluster 74	62	Known: Esr1 GATGAG
4	TSGGCGGCTAWW	5′	23.4	GO: meiosis	10	Known: Ume6/Ndt80
5	RTCACGTGV	5′	17.6	ChIP: Cbf1	27	Known: Cbf1/Pho4
6	WTATWTACADG	3′	17.4	Exp.: cluster 16 downstream	25	New: mitochondrial downstream
7	GRRAAAWTTTTCACT	5′	15.6	Exp.: cluster 74	37	Known: Esr2
8	TTCCnaAttnGGAAA	5′	13.8	ChIP: Mcm1	29	Known: Mcm1
9	CGTTTCTTTTTCY.	5′	13.5	GO: filamentation	7	New: filamentation
10	TYYTCGAGA.	5′	12.5	Exp.: cluster 86	5	Known: Xbp1 (Hsf1-co-ocuring)
11	TTTTCGCG	5′	12.0	ChIP: Swi4	21	Known: Swi4 fixed gap
11a	TTTT = CGCG¶	5′	12.0	ChIP: Swi4	–	New: Swi4 variable gap
12	TKACGCGTT	5′	12.0	ChIP: Mbp1	18	Known: Mbp1/Swi6
13	STGCGGnnnttTCTnnG	5′	11.8	GO: filamentation	11	New: filamentation
14	YCTATTGTT	5′	11.5	ChIP: Fkh2	6	New: Rlm1-like
15	TTTTGCCACCG	5′	11.0	GO: proteolysis	25	Known: Rpn4/Met4
16	tTTGTTTACnTTT	5′	10.8	ChIP: Fkh2	28	Known: Fkh1/2
17	RVACCCTD	5′	10.3	–	–	Known: Aft1
18	WCGCGTCGCGt	5′	10.2	ChIP: Mbp1	17	New: double Mbp1
19	GGGTnACCC	5′	10.0	ChIP: Reb1	8	New: Reb1 palindrome
20	GnnATGTGTGGGTGT	5′	9.9	ChIP: Fhl1	5	Known: Rap1
21	TTTTGTGTCRC	5′	9.9	ChIP: Sum1	14	Known: Mse
22	TTTCAnCGCGC	5′	9.8	–	–	New: no category
23	TATTAWTATTATtMtnatta	3′	9.5	–	–	New: no category
24	SCGnHGGS	5′	8.8	GO: filamentation	6	New: filamentation
25	ACAGCCGCRY	5′	8.6	Exp.: cluster 37	6	New: expression cluster 37
26	DCGCGGGGH	5′	8.1	Exp.: cluster 46	8	Known: Mig1b
27	SKGTGGSGc	5′	8.1	ChIP: Met31	5	Known: Met31
28	TTTTn(19)GCKCG	5′	7.8	–	–	Known: no category
29	HRCCCYTWDt	5′	7.8	Exp.: cluster 8	22	Known: Msn2/4
30	TKCCCnnnnGGG	5′	7.3	ChIP: Mcm1	15	Known: Mcm1 (hits tRNA)
31	GTGTCAGTAAt	5′	7.1	ChIP: Sum1	15	New: Sum1
32	RGTTTTTCCG	5′	7.1	ChIP: Rgt1	7	New: Rgt1
33	TTCTMGAAGA	5′	7.0	ChIP: Hsf1	10	Known: Hsf1
34	YCCGSGGS	5′	6.7	GO: filamentation	9	New: filamentation
35	CnCCTTTTATAC	5′	6.5	–	–	New: no category
36	CCSGTAnCGG	5′	6.5	ChIP: Leu3	8	Known: Leu3
37	SKTKCCTT	5′	6.4	GO: filamentation	7	New: filamentation
38	CTCCCCTTAT	5′	6.4	Exp.: cluster 8	11	Known: Msn2/4
39	GCCCGG	5′	6.3	GO: filamentation	10	New: filamentation
40	SGCGCGRB	5′	6.3	–	–	New: no category
41	CTCSGCS	5′	6.2	–	–	New: no category
42	TGnKAGCGCCG	5′	6.2	–	–	–
43	ATGACTCWT	5′	6.1	ChIP: Gcn4	44	Known: Gcn4/Bas1
44	CCGAnnnTCGG	5′	6.1	Exp.: cluster 46	6	New: facilitators palindrome
45	SCGMnnnnnnnKCG	5′	6.0	–	–	New: no category
46	CnCCGCGCnnTTTs	5′	6.0	–	–	New: no category
47	TTTTnnnnnnnnnnnnngGGGT	5′	5.8	–	–	New: no category
48	TGTRnCAW	3′	5.5	–	–	New: no category
49	YCSknnnnnnnnnKCGG	5′	5.4	Exp: cluster 46	6	Known: Put3
50	CGGnnnnnnnnnnnnnnKCGV	5′	5.4	–	–	New: no category

- 72 full motifs with MCS > 4
  - Show strong overlap with 28/33 known motifs having MCS>4
  - Also 8/22 matches to known motifs with MCS<4
- Function?
  - Correlation with gene sets
    - 120 sets of genes with common GO annotation
    - 106 sets of genes: upstream region found to be bound by TF in ChIP/chip studies
    - 92 sets showing coordinate RNA regulation
  - Most of the known: OK
  - 25 of the 42 new motifs also enrichment: so assign a function

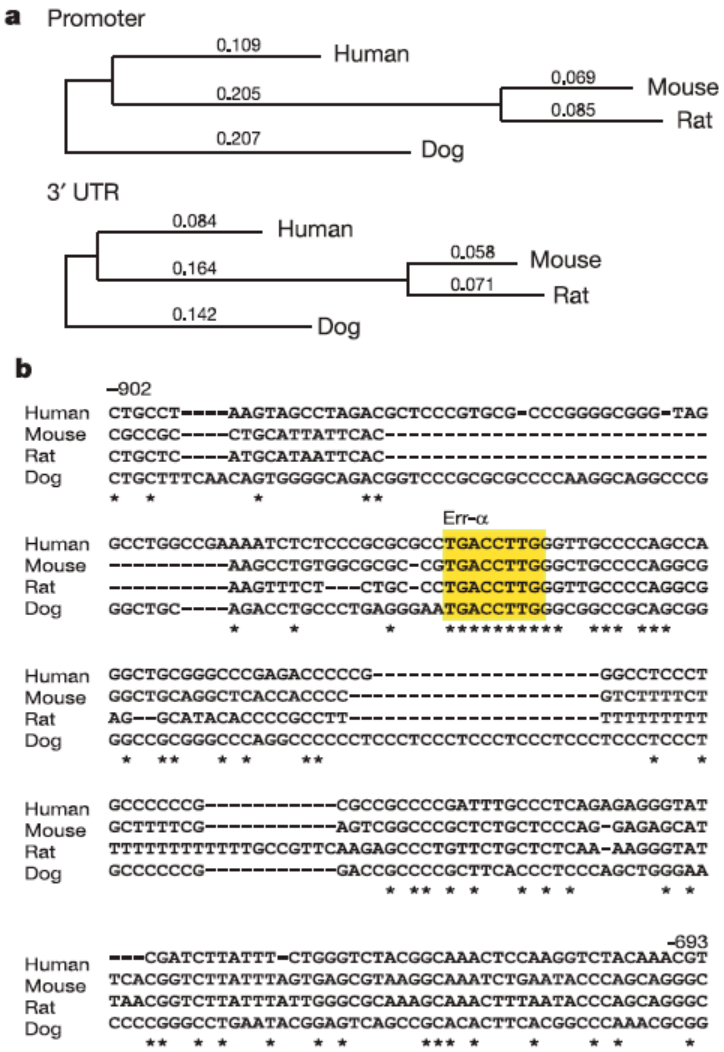
Motif combinations

- The Ste12 and Tec1 motifs show clear correlation, with about 20% of regions having a conserved occurrence of one also having a conserved occurrence of the other. This enrichment is not apparent when considering S. cerevisiae alone. We found that the genes that contain only the conserved Ste12 motif are enriched for those involved in mating and pheromone response, whereas those that contain conserved occurrences of both the Ste12 and Tec1 motifs are enriched for those involved in filamentous growth. These computational observations are consistent with recent work showing genome-wide evidence that Ste12 and Tec1 cooperate during starvation to induce filamentation

Genome-wide analysis of conserved motifs: human

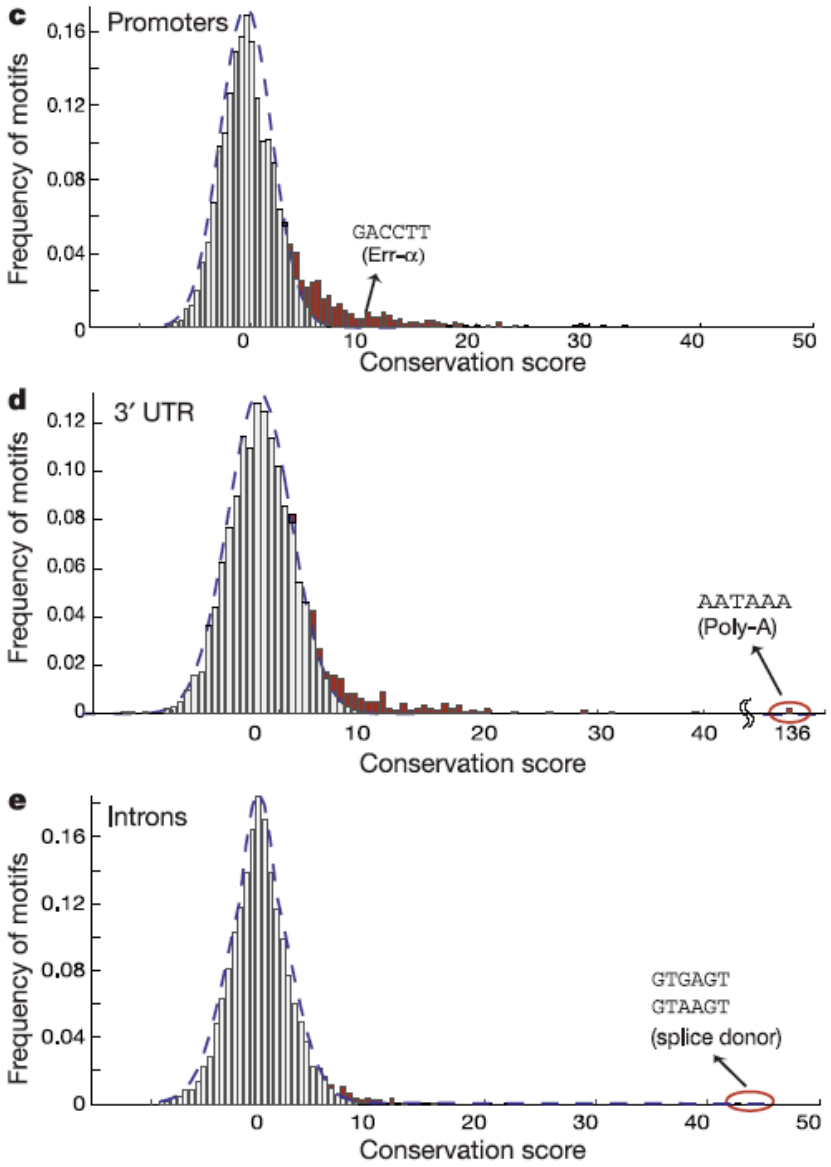
- Combination of
  - Proximal promoters
  - TFBS conservation
- Promoter alignments (4kb centered around annotated TSS)
- Example: Err-α
  - 434 occurrences in human promoters
    - 162 conserved across all 4 species
    - 37% conservation rate
  - Random 8-mer: 6.8% cons. rate
- Motif Conservation Score (MCS)
  - **binomial probability** of observing *k* conserved instances out of *n* given the probability p_0 of conservation for any one instance
  - Err-α: MCS = 25.2
- TRANSFAC motifs
  - 63% MCS>3
  - 50% MCS>5
- Random motifs:
  - 1.6% MCS>3
  - 0% MCS>5

Xie et al. (2005) *Nature* 434, 338-345



Genome-wide analysis of conserved motifs (ctd.)

- New motifs
  - 6-18 bp w/ allowed degeneracy
  - 174 motifs MCS>6
    - 59 w/ matches to known motifs
  - Validation:
    - correlation with tissue-specific gene expression
    - Positional bias of the motifs relative to the TSS



Regulatory elements in flies

- 12 drosophila papers in Nature, nov 2007
  - Evolution paper (12 Genomes Consortium)
    - How many known regulatory elements are under constraint?
  - Motif paper (Stark et al.)
    - Cfr previous slides, collection of 145 motifs

Constraint of regulatory sequences in fly

- mean selective constraint (C, the fraction of mutations removed by natural selection) relative to the ‘fastest evolving intron’ sites at the 5’ end of short introns, which represent putatively unconstrained neutral standards.
  - $C = 1 - O / E$
    - O = observed number of substitutions
    - E = expected number of substitutions (base pairs 8–30) of short introns (<80bp in length)
- REDfly known CRMs
  - Mean 0.643
- FlyReg footprints
  - 0.692
  - best match to position weight matrices within each foot- print; core motifs in transcription-factor-binding sites have a mean constraint of 0.773
  - The high level of constraint in cis-regulatory sequences also extends into flanking sequences, only declining to constraint levels typical of non-coding DNA 40 bp away
  - evidence that selective constraint is correlated with predicted transcription-factor-binding strength
- Compare to non-coding DNA overall: 0.555
- non-degenerate coding sites: 0.862

Conclusion regulatory elements

- Some regulatory sequences are under constraint
  - many conserved non-coding sequences human-mouse have regulatory function
  - Known elements analysed for their constraint (fly)
  - Known motifs have high MCS
- But still limited availability of known regulatory regions, especially in vertebrates => ENCODE: large scale experiments

End