Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes

A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hiller, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler

Genome Research **15** (2005) 1034–1050.

Speaker: Joseph Chuang-Chieh Lin

The Comparative & Evolutionary Genomics/Transcriptomics Lab.
Genomics Research Center, Academia Sinica
Taiwan

4 January 2012

Self introduction

Basic information

- Name: Joseph Chuang-Chieh Lin
- Birth: 5 Dec. 1979 at Tainan City
- Married since 2007 & one daughter
- Hobbies: running, reading, & eating

⋆ Homepage:
http://idv.sinica.edu.tw/josephcclin

⋆ Email addresses:
josephcclin@gate.sinica.edu.tw

Education background

- B.S.: Mathematics,
National Cheng Kung University,
1998 – 2002
- M.S.: CSIE,
National Chi Nan University
2002 – 2004,
Thesis supervisor:
Prof. Richard Chia-Tung Lee
- Ph.D.: CSIE,
National Chung Cheng University
2004 – 2011,
Dissertation supervisors:
Prof. Maw-Shang Chang &
Prof. Peter Rossmanith

Self introduction (contd.)

Research interests

- Fixed-parameter algorithms
- Randomized algorithms (property testing)
- Graph theory & algorithms
- Bioinformatics

Outline

Introduction

- Despite tremendous progress in vertebrate genomics, it is not exactly clear which regions are functional.
- One of the best strategies known for finding functional sequences is to *look for sequences that are conserved across species*.
- The primary reason for cross-species sequence conservation is believed to be negative (purifying) selection.

Introduction

- Despite tremendous progress in vertebrate genomics, it is not exactly clear which regions are functional.
- One of the best strategies known for finding functional sequences is to *look for sequences that are conserved across species*.
- The primary reason for cross-species sequence conservation is believed to be negative (purifying) selection.

Introduction

- Despite tremendous progress in vertebrate genomics, it is not exactly clear which regions are functional.
- One of the best strategies known for finding functional sequences is to *look for sequences that are conserved across species*.
- The primary reason for cross-species sequence conservation is believed to be negative (purifying) selection.

Introduction (contd.)

- Thanks to
  - a recent explosion in the number of sequenced genomes &
  - the development of multiple sequence alignment tools,

  it is possible to conduct large-scale searches for conserved sequences.

Introduction (contd.)

Mouse Genome Sequencing Consortium 2002; Chiaromonte *et al.* 2003, Roskin *et al.* 2003, Cooper *et al.* 2004, Rat Genome Sequencing Project Consortium 2004, etc.:

- $\approx$ 5% more bases in mammalian genomes are under purifying selection.
  - Protein-coding genes are believed to account for only $\approx$ 1.5%.
  - $\approx$ 3.5% of bases are thought to be functional but not to code for proteins.

Bergman *et al.*, Kellis *et al.*, Stein *et al.*:

- While there are less non-coding regions in the genomes of insects, worms, and yeasts, the functions of many conserved sequences in these genomes are not yet known.

Introduction (contd.)

Mouse Genome Sequencing Consortium 2002; Chiaromonte *et al.* 2003, Roskin *et al.* 2003, Cooper *et al.* 2004, Rat Genome Sequencing Project Consortium 2004, etc.:

- $\approx$ 5% more bases in mammalian genomes are under purifying selection.
  - Protein-coding genes are believed to account for only $\approx$ 1.5%.
  - $\approx$ 3.5% of bases are thought to be functional but not to code for proteins.

Bergman *et al.*, Kellis *et al.*, Stein *et al.*:

- While there are less non-coding regions in the genomes of insects, worms, and yeasts, the functions of many conserved sequences in these genomes are not yet known.

Drawbacks of previous methods

- Pairwise alignments + simple percent identity-based method (e.g., Dermitzakis *et al.* 2002, Nobrega *et al.* 2003).
  - It becomes essential to use "multiple" alignments.
  - Phylogeny information & branch lengths?
- Phylogenetic shadowing method (Boffelli *et al.* 2003).
  - Using a sliding window of fixed size can be a limitation.
    - small $\Rightarrow$ difficult to tell effectively between conserved and nonconserved regions.
    - large $\Rightarrow$ small conserved regions may be missed.

In this study: phastCons

phastCons: HMM + space + time (phylogeny) + conservation analysis.

- Identify conserved elements.
- Based on a phylogenetic hidden Markov model.
- Considering nucleotide substitutions which occur at each site in a genome and how this process changes from one site to the next.
- Do NOT require a sliding window of fixed size.
- Nearly all parameters can be estimated from the data by maximum likelihood.

Data sets

Table S1: Summary of genomes and assemblies

Species	Group	UCSC assembly	Reference
H. sapiens	vertebrate	hg17	International Human Genome Sequencing Consortium, 2001
M. musculus	vertebrate	mm5	Mouse Genome Sequencing Consortium, 2002
R. norvegicus	vertebrate	rn3	Rat Genome Sequencing Project Consortium, 2004
G. gallus	vertebrate	galGal2	International Chicken Genome Sequencing Consortium, 2004
F. rubripes	vertebrate	fr1	Aparicio et al., 2002
D. melanogaster	insect	dm1	Adams et al., 2000
D. yakuba	insect	droYak1	(In prep.)
D. pseudoobscura	insect	dp2	Richards et al., 2005
A. gambiae	insect	anoGam1	Holt et al., 2002
C. elegans	worm	ce2	C. elegans Sequencing Consortium, 1998
C. briggsae	worm	cb1	Stein et al., 2003
S. cerevisiae	yeast	sacCer1	http://www.yeastgenome.org
S. castelli	yeast	–	Cliften et al., 2003
S. kluyveri	yeast	–	Cliften et al., 2003
S. kudriavzevii	yeast	–	Cliften et al., 2003
S. mikatae	yeast	–	Kellis et al., 2003
S. bayanus	yeast	–	Kellis et al., 2003
S. paradoxus	yeast	–	Kellis et al., 2003

- Annotations for only the reference genomes.

Data sets (contd.)

Table S2: Summary of multiple alignments

Group	n^a	Reference Genome	Aligned Genomes (Coverage)[b]	Tot. Cov.[c]
vertebrate	5	*H. sapiens*	*M. musculus* (35.4%), *R. norvegicus* (34.0%), *G. gallus* (3.5%), *F. rubripes* (1.6%)	40.0%
insect	4	*D. melanogaster*	*D. yakuba* (85.1%), *D. pseudoobscura* (58.1%), *A. gambiae* (14.6%)	86.9%
worm	2	*C. elegans*	*C. briggsae* (43.8%)	43.8%
yeast	7	*S. cerevisiae*	*S. paradoxus* (96.6%), *S. mikatae* (93.0%), *S. kudriavzevii* (89.7%), *S. bayanus* (89.5%), *S. castelli* (63.4%), *S. kluyveri* (56.1%)	96.6%

[a]Number of species.
[b]Genomes aligned to reference genome and fraction of bases in reference genome covered by local alignments with each one.
[c]Fraction of bases in reference genome covered by alignments with at least one other genome.

- Using MULTIZ ver.10.
  - A phylogenetic tree directed multiple alignment.
  - NOT a real aligner, as it uses BLASTZ.

Methods

Hidden Markov model

(Refer to the slides for the textbook *An Introduction to Bioinformatics Algorithms*, by Jones & Pevzner 2004)

- Can be viewed as an abstract machine with (finite) hidden states that emits symbols from an alphabet Σ.
- Each state has its own probability distribution, and the machine switches between states according to this probability distribution (stationary distribution; equilibrium).
- While in a certain state, the machine makes two decisions:
  - What state should I move to next?
  - What symbol from Σ should I emit?

Hidden Markov model

(Refer to the slides for the textbook *An Introduction to Bioinformatics Algorithms*, by Jones & Pevzner 2004)

- Can be viewed as an abstract machine with (finite) hidden states that emits symbols from an alphabet Σ.
- Each state has its own probability distribution, and the machine switches between states according to this probability distribution (stationary distribution; equilibrium).
- While in a certain state, the machine makes two decisions:
  - What state should I move to next?
  - What symbol from Σ should I emit?

Why "hidden"

- Observers can see the emitted symbols of an HMM but have no ability to know which state the HMM is currently in.
- Thus, the goal is to infer the most likely hidden states of an HMM based on the given sequence of emitted symbols.

A simple example

HMM for Fair Bet Casino

- The *Fair Bet Casino* in *HMM* terms:

  $\Sigma = \{0, 1\}$ (0 for ***T***ails and 1 ***H***eads)

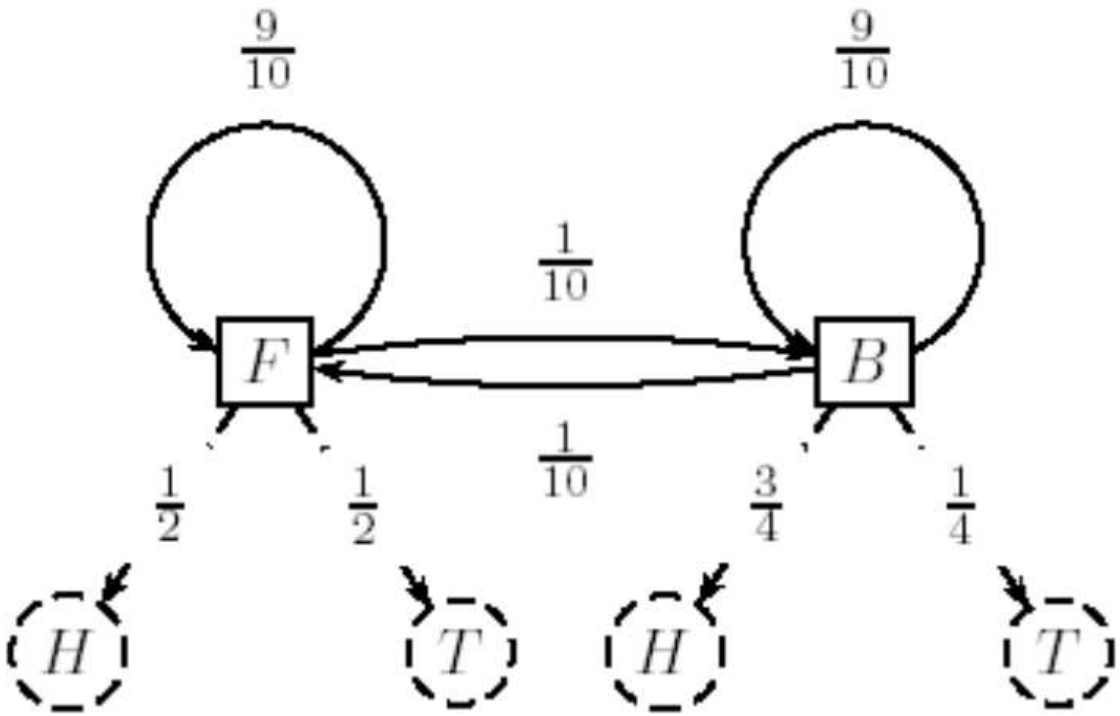
  $Q = \{F,B\}$ – F for Fair & B for Biased coin.

- Transition Probabilities A *** Emission Probabilities E

	Fair	Biased
Fair	$a_{FF} = 0.9$	$a_{FB} = 0.1$
Biased	$a_{BF} = 0.1$	$a_{BB} = 0.9$

	Tails(0)	Heads(1)
Fair	$e_F(0) = \frac{1}{2}$	$e_F(1) = \frac{1}{2}$
Biased	$e_B(0) = \frac{1}{4}$	$e_B(1) = \frac{3}{4}$

A simple example (contd.)



Hidden paths

- A path $\mathbf{z} = z_1, z_2, \ldots, z_n$ in the HMM: a *sequence of states*.
- Consider a path $\mathbf{z} =$ F F F B B B B B F F F and the emitted sequence $\mathbf{x} = 0\ 1\ 0\ 1\ 1\ 1\ 0\ 1\ 0\ 0\ 1$.

$$\begin{array}{l} \mathbf{x} \\ \mathbf{z} \\ \mathbf{Pr}(x_i \mid z_i) \\ \mathbf{Pr}(z_i \mid z_{i-1}) \end{array} = \left\{ \begin{array}{ccccccccccc} 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ F & F & F & B & B & B & B & B & F & F & F \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{3}{4} & \frac{3}{4} & \frac{3}{4} & \frac{1}{4} & \frac{3}{4} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{9}{10} & \frac{9}{10} & \frac{1}{10} & \frac{9}{10} & \frac{9}{10} & \frac{9}{10} & \frac{9}{10} & \frac{1}{10} & \frac{9}{10} & \frac{9}{10} \end{array} \right\}$$

- $\mathbf{Pr}(\mathbf{x} \mid \mathbf{z})$: probability that sequence $\mathbf{x}$ was generated by the path $\mathbf{z}$:

$$\mathbf{Pr}(\mathbf{x} \mid \mathbf{z}) = a_{\pi_0, z_1} \cdot \prod_{i=1}^{n} e_{z_i}(x_i) \cdot a_{z_i, z_{i+1}}.$$

Hidden paths

- A path $\mathbf{z} = z_1, z_2, \ldots, z_n$ in the HMM: a *sequence of states*.
- Consider a path $\mathbf{z} =$ F F F B B B B B F F F and the emitted sequence $\mathbf{x} = 0\ 1\ 0\ 1\ 1\ 1\ 0\ 1\ 0\ 0\ 1$.

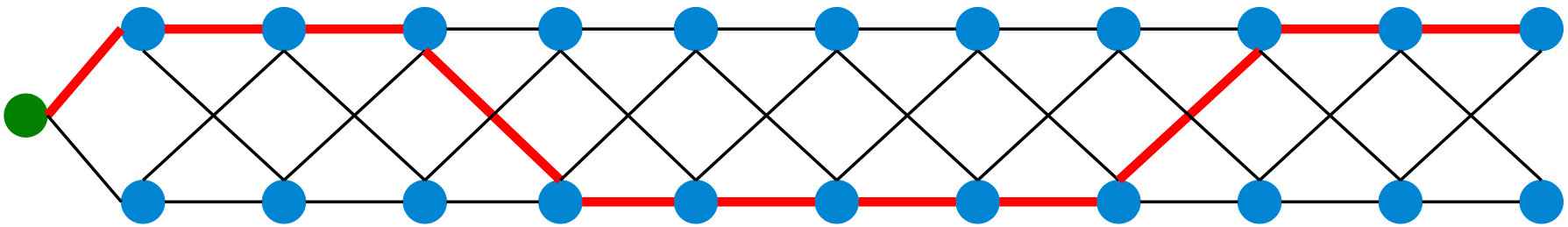
$$\begin{array}{l} \mathbf{x} \\ \mathbf{z} \\ \mathbf{Pr}(x_i \mid z_i) \\ \mathbf{Pr}(z_i \mid z_{i-1}) \end{array} = \left\{ \begin{array}{ccccccccccc} 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ F & F & F & B & B & B & B & B & F & F & F \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{3}{4} & \frac{3}{4} & \frac{3}{4} & \frac{1}{4} & \frac{3}{4} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{9}{10} & \frac{9}{10} & \frac{1}{10} & \frac{9}{10} & \frac{9}{10} & \frac{9}{10} & \frac{9}{10} & \frac{1}{10} & \frac{9}{10} & \frac{9}{10} \end{array} \right\}$$

- $\mathbf{Pr}(\mathbf{x} \mid \mathbf{z})$: probability that sequence $\mathbf{x}$ was generated by the path $\mathbf{z}$:

$$\mathbf{Pr}(\mathbf{x} \mid \mathbf{z}) = a_{\pi_0, z_1} \cdot \prod_{i=1}^{n} e_{z_i}(x_i) \cdot a_{z_i, z_{i+1}}.$$

Hidden paths (contd.)

Fair



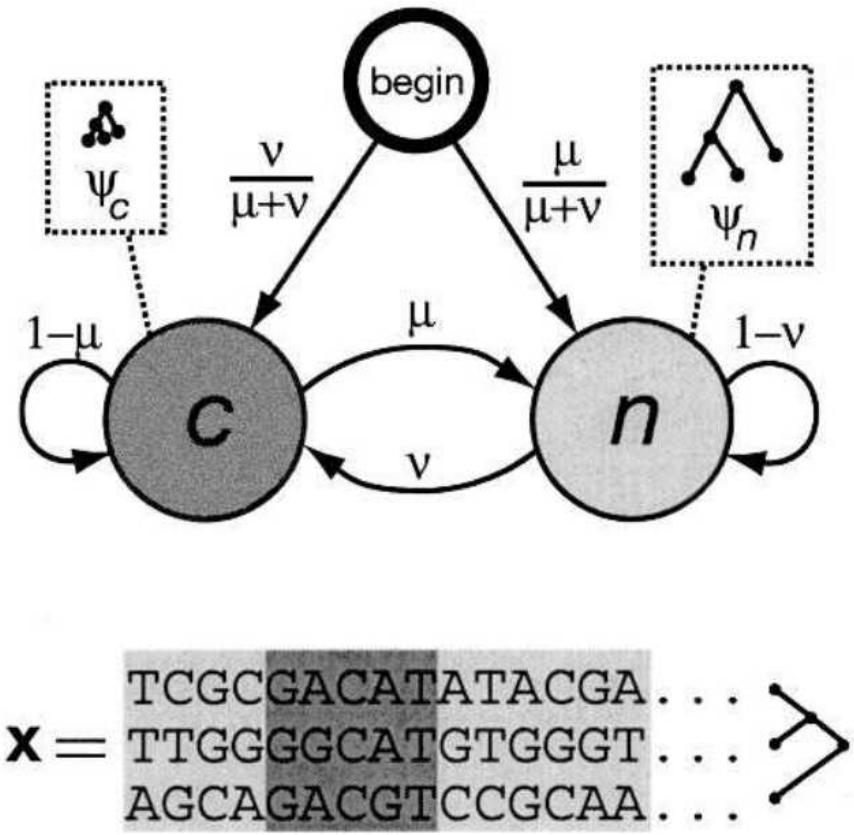
Biased

phylo-HMM

phastCons: HMM + space + time (phylogeny) + conservation analysis.

- $\mathbf{x} = (x_1, \ldots, x_L) \Rightarrow$ a multiple alignment.
  - $x_i \Rightarrow i$th column in the alignment $\mathbf{x}$.
- nonconserved phylogenetic model: $\psi_n = (Q, \pi, \tau, \beta)$;
  conserved phylogenetic model: $\psi_c = (Q, \pi, \tau, \rho\beta)$.
  - Q: 4×4 (nucleotide) substitution rate matrix.
  - π: a vector of background probabilities for A,G,C,T.
  - τ: the phylogeny (binary tree; assumed to be known).
  - β: a vector of branch lengths of τ (expected substitutions per site).
  - ρ: the scaling parameter for ψ_c.
- The free parameters: $\boldsymbol{\theta} = (Q, \beta, \rho, \mu, \nu) \Rightarrow$ estimated from the data.
  - μ and ν: the transition probs. and initial-state probs. of the Markov chain.

phylo-HMM (contd.)



An example of Q

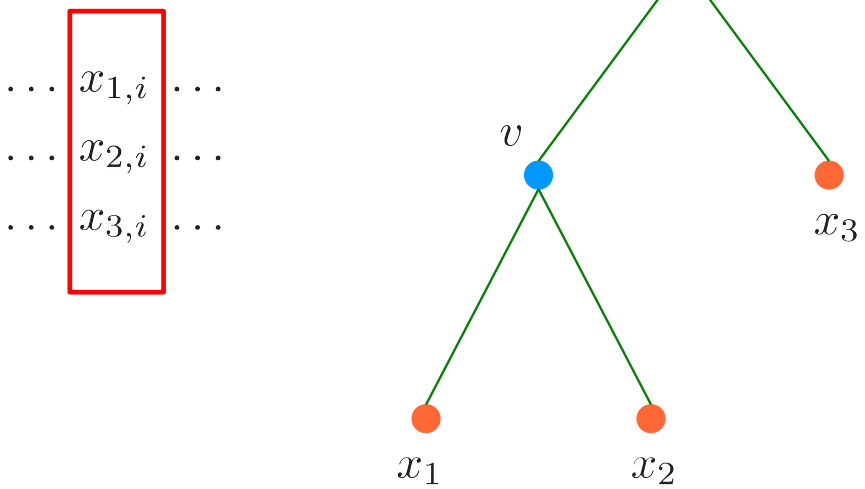
$$Q_{REV} = \{q_{i,i}\} = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & - \end{pmatrix}$$

$$q_{i,i} = -\sum_{j;j\neq i} q_{i,j}.$$

- $\mathbf{P}(t) = \mathbf{S}e^{\mathbf{\Lambda} t}\mathbf{S}^{-1}$, where $Q = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$.
  - the matrix of substitution probabilities for length t.
- Q is by convention *scaled* so that the branch lengths are equal to the expected substitutions per site.

The site probability in an alignment conditioned on θ

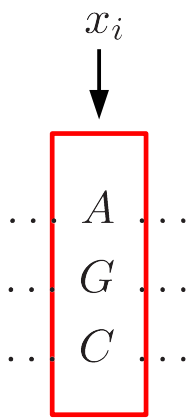
- $\mathbf{Pr}(b|a,t)$: the probability of base b replacing base a over a branch of length t.
- $\mathbf{Pr}(L_u|a)$: the probability of all the leaves below u given that the base assigned to u is an a (conditioned on θ).

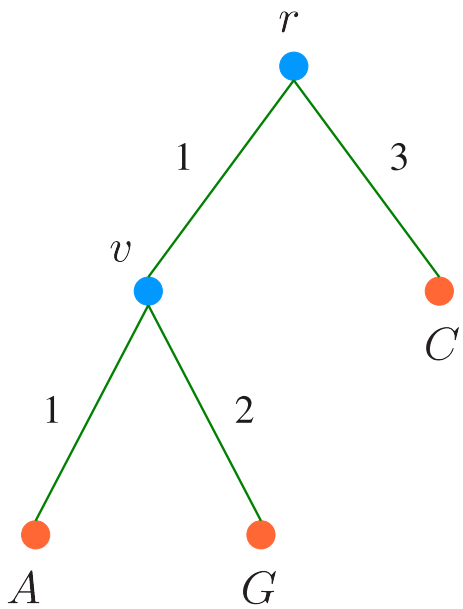


- $\mathbf{Pr}(x_i \mid \theta) = \sum_a \pi_a \cdot \mathbf{Pr}(L_r|a)$ (r is the tree-root).

The site probability in an alignment conditioned on θ (contd.)

$$\mathbf{Pr}(L_v|A) = \mathbf{Pr}(A|A,1) \cdot \mathbf{Pr}(G|A,2)$$
$$\mathbf{Pr}(L_v|G) = \mathbf{Pr}(A|G,1) \cdot \mathbf{Pr}(G|G,2)$$
$$\mathbf{Pr}(L_v|C) = \mathbf{Pr}(A|C,1) \cdot \mathbf{Pr}(G|C,2)$$
$$\mathbf{Pr}(L_v|T) = \mathbf{Pr}(A|T,1) \cdot \mathbf{Pr}(G|T,2)$$
$$\mathbf{Pr}(L_r|A) = \sum_{b \in \Sigma} \mathbf{Pr}(b|A,1) \cdot \mathbf{Pr}(C|A,3)$$
$$\mathbf{Pr}(L_r|G) = \sum_{b \in \Sigma} \mathbf{Pr}(b|G,1) \cdot \mathbf{Pr}(C|G,3)$$
$$\mathbf{Pr}(L_r|C) = \sum_{b \in \Sigma} \mathbf{Pr}(b|C,1) \cdot \mathbf{Pr}(C|C,3)$$
$$\mathbf{Pr}(L_r|T) = \sum_{b \in \Sigma} \mathbf{Pr}(b|T,1) \cdot \mathbf{Pr}(C|T,3)$$





- $\mathbf{Pr}(x_i \mid \theta) = \sum_a \pi_a \cdot \mathbf{Pr}(L_r|a)$.

The likelihood function

$$L(\theta \mid \mathbf{x}) = \mathbf{Pr}(\mathbf{x} \mid \theta) = \sum_{\mathbf{z}} \prod_{i=1}^{L} \mathbf{Pr}(x_i \mid z_i, \theta) \mathbf{Pr}(z_i \mid z_{i-1}, \theta).$$

- The complete log likelihood for an alignment $\mathbf{x}$ and a specific path $\mathbf{z}$:

$$l(\theta \mid \mathbf{x}, \mathbf{z}) = \log \mathbf{Pr}(\mathbf{x}, \mathbf{z} \mid \theta) = \log \prod_{i=1}^{L} \mathbf{Pr}(x_i \mid z_i, \theta) \mathbf{Pr}(z_i \mid z_{i-1}, \theta).$$

Constraints on coverage and PIT threshold

- ⋆ The fraction of conservative sites in coding regions: $\approx 65\%$.
  - An assumption: coding regions *evolve in fundamentally similar way across species groups*.
- ⋆ PIT $:= L_{min} \cdot H(\psi_c||\psi_n) > 9.8$ (phylogenetic information threshold).
  - $L_{min} = \frac{\log \nu + \log \mu - \log(1-\nu) - \log(1-\mu)}{\log(1-\nu) - \log(1-\mu) - H(\psi_c||\psi_n)}$.
    - Expected min. length of a sequence of conserved sites required for a conserved element to be predicted.
  - $H(\psi_c||\psi_n) = \sum_i \mathbf{Pr}(x_i \mid \psi_c) \log \frac{\mathbf{Pr}(x_i|\psi_c)}{\mathbf{Pr}(x_i|\psi_n)}$
    - Relative entropy of distribution associated with ψ_c w.r.t. that associated with ψ_n.
  - PIT is set the same for all species groups.

Parameters set by the user

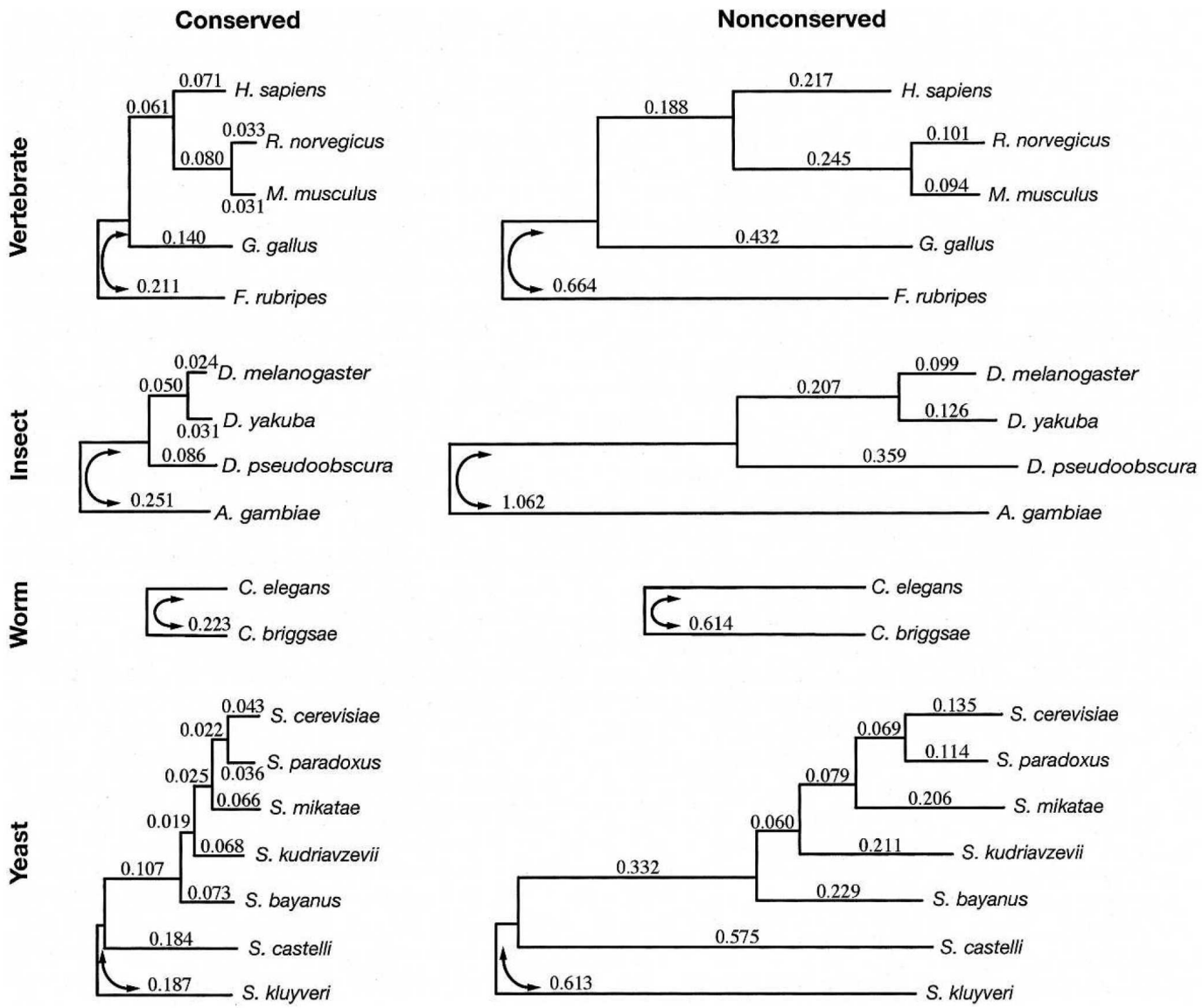
- γ: expected coverage by conserved elements.
  - $\gamma = \frac{\nu}{\mu+\nu}$.
  - ▷ larger $\gamma \Rightarrow$ higher coverage by conserved elements and larger conservation scores.
- ω: expected length of a conserved element.
  - $\omega = \frac{1}{\mu}$.
  - ▷ larger $\omega \Rightarrow$ more similar scores at adjacent sites.

The scores computed by phastCons

- The log-odds score of an element predicted from position i to position j in an alignment of length L:

$$s_{ij} = \log \frac{\mathbf{Pr}[x_i, \ldots, x_j \mid \psi_c]}{\mathbf{Pr}[x_i, \ldots, x_j \mid \psi_n]} = \sum_{k=i}^{j} [\log \mathbf{Pr}(x_k \mid \psi_c) - \log \mathbf{Pr}(x_k \mid \psi_n)].$$

phylo-HMM (assumed tree topologies with estimated branch lengths)



- The estimated nonconserved branch lengths were fairly consistent with the other results (neutral evolving DNA in mammals; by Cooper *et al.*, 2004), yet NOT accurate in all respects.
- In particular, the branches to "chicken" and "Fugu" were significantly underestimated.
  - Similar effects were observed with other species groups.
- Nevertheless, such inaccuracies do not strongly influence the results.

Results

Some terminologies

- "element": a continuous sequence of nucleotide bases.
- "HCE (highly conserved elements)"?
  - top 5000 elements for vertebrate and insect sets.
  - top 1000 elements for worm and yeast sets.
- "ultraconserved": a region of human DNA of length 200 nucleotides or greater, which is entirely identical in both rats and mice.

Synteny filtering

- Predicted elements in the vertebrate data set were discarded if they did not fall on the *syntenic net* between human and mouse.
  - syntenic net: a subset of chained local alignments.
- The human and mouse genomes were selected because of their evolutionary distance and assembly quality.
- A similar filter was applied to the insect predictions.
  - A predictions were discarded if they did not fall on either the D. melanogaster/D. yakuba syntenic net or the D. melanogaster/D. pseudoobscura syntenic net.

Table S5: Predicted Conserved Elements and Estimated Parameters Under Four Different Estimation Methods

Group	Method	Total no.[a]	Ave. len.[b]	Cov.[c]	CDS cov.[d]	μ	ν	ω	γ	$L_{\min}$
vert.	MLE	561,103	216.1	4.2%	68.8%	0.018	0.004	55.4	0.191	30.4
	55%	1,058,855	75.3	2.8%	56.8%	0.125	0.029	8.0	0.187	12.9
	65%[e]	1,157,180	103.5	4.2%	66.1%	0.083	0.030	12.0	0.265	16.0
	75%	1,381,978	167.5	8.1%	76.6%	0.043	0.031	23.0	0.415	22.6
insect	MLE	352,620	173.1	48.3%	73.3%	0.018	0.019	56.1	0.522	17.1
	55%	502,974	92.9	36.9%	57.0%	0.050	0.029	20.0	0.370	13.5
	65%	467,232	120.6	44.5%	68.3%	0.036	0.031	28.1	0.468	13.5
	75%	427,815	156.9	53.1%	78.4%	0.025	0.039	40.0	0.610	12.7
worm	MLE	71,419	258.8	18.4%	43.1%	0.017	0.014	58.6	0.455	83.7
	55%	108,588	181.2	19.6%	45.8%	0.037	0.033	27.0	0.470	56.4
	65%	98,415	268.9	26.4%	56.4%	0.019	0.031	53.0	0.620	60.6
	75%	87,228	357.5	31.1%	62.5%	0.010	0.030	100.0	0.750	65.6
yeast	MLE	57,610	134.1	63.6%	77.8%	0.021	0.034	47.1	0.615	11.5
	55%	71,726	78.8	46.5%	58.0%	0.067	0.023	15.0	0.260	12.5
	65%	62,640	107.9	55.6%	68.9%	0.043	0.029	23.0	0.400	11.7
	75%	62,754	124.1	64.0%	78.1%	0.025	0.041	40.0	0.620	10.6

[a] Total number of predicted conserved elements; here and below, only elements passing synteny filters are considered (where applicable)

[b] Average length of a predicted conserved element (bp)

[c] Genome-wide coverage by predicted conserved elements

[d] Coverage of coding regions by predicted conserved elements. These numbers differ slightly from the coverage targets (55%, 65%, and 75%) because of the adjustment for alignment coverage in coding regions and because of the removal of nonsyntenic predictions.

[e] Slight differences with the numbers reported in the main text result from the use of only a portion of the genome in parameter estimation (a random sample of 100 1Mb intervals).

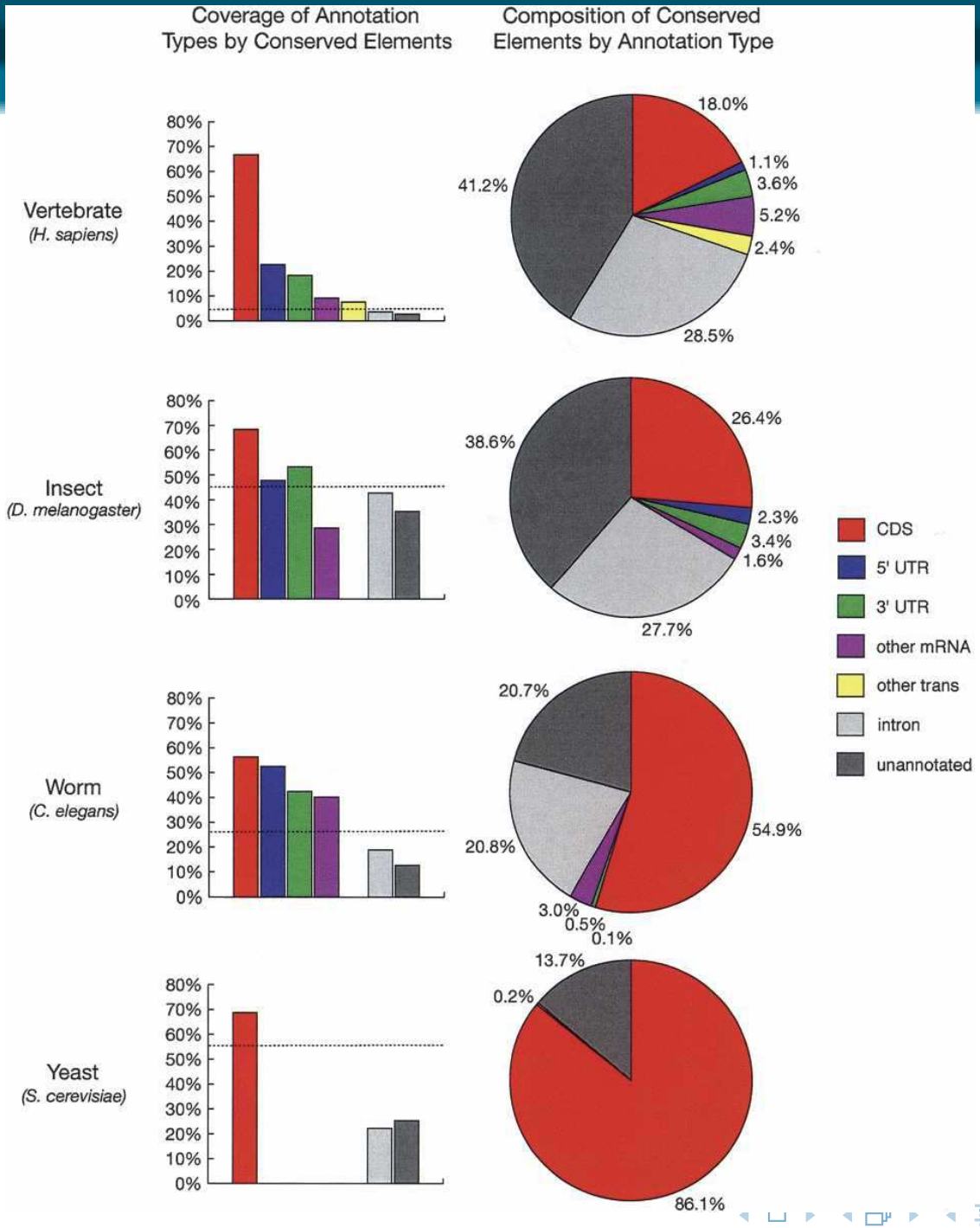
Prediction based on 4d sites

- Fourfold degenerate sites (4d sites)?
  - e.g., the 3rd position of the codens (i.e., GGA, GGG, GGC, GGU) for the amino acid Glycine.
- Replacing the nonconserved model with a model estimated from 4d sites.

Table S6: Conserved Elements Based on PhastCons Nonconserved Model (65% Coverage Target) vs. Elements Based on 4d Model

Group	Method	Total no.[a]	Ave. len.[b]	Cov.[c]	CDS cov.[d]	CDS frac.[e]	$H(\boldsymbol{\psi}_c \| \boldsymbol{\psi}_n)$	$L_{\min}$
vert.	65%	1,157,180	103.5	4.2%	66.1%	18.0%	0.611	16.0
	4d	797,777	109.3	3.0%	64.2%	24.0%	0.854	11.0
insect	65%	467,232	120.6	44.5%	68.3%	26.4%	0.725	13.5
	4d	554,823	110.0	48.3%	75.0%	26.8%	1.032	9.5
worm	65%	98,415	268.9	26.4%	56.4%	54.9%	0.159	60.6
	4d	195,062	188.0	36.6%	69.3%	48.7%	0.403	25.4
yeast	65%	62,640	107.9	55.6%	68.9%	86.1%	0.836	11.7
	4d	94,615	86.8	67.6%	81.8%	84.1%	1.914	5.3

[a]Total number of predicted conserved elements; here and below, only elements passing synteny filters are considered (where applicable)
[b]Average length of a predicted conserved element (bp)
[c]Genome-wide coverage by predicted conserved elements
[d]Coverage of coding regions by predicted conserved elements.
[e]Fraction of conserved elements in coding regions, at the base level

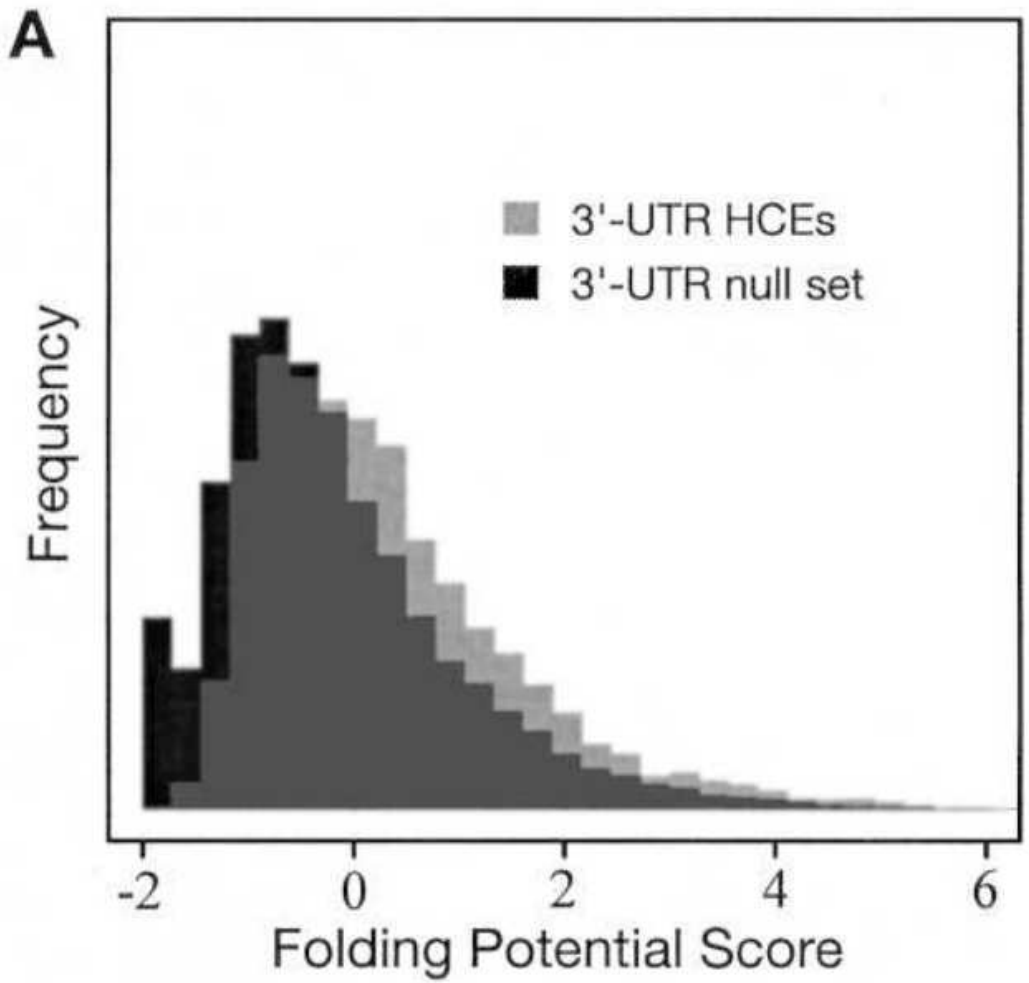

Coverage of Annotation Types by Conserved Elements
Composition of Conserved Elements by Annotation Type
Vertebrate
(H. sapiens)
Insect
(D. melanogaster)
Worm
(C. elegans)
Yeast
(S. cerevisiae)
80%
70%
60%
50%
40%
30%
20%
10%
0%
18.0%
1.1%
3.6%
5.2%
2.4%
41.2%
28.5%
26.4%
38.6%
2.3%
3.4%
1.6%
27.7%
20.7%
54.9%
20.8%
3.0%
0.5%
0.1%
13.7%
0.2%
86.1%
CDS
5' UTR
3' UTR
other mRNA
other trans
intron
unannotated

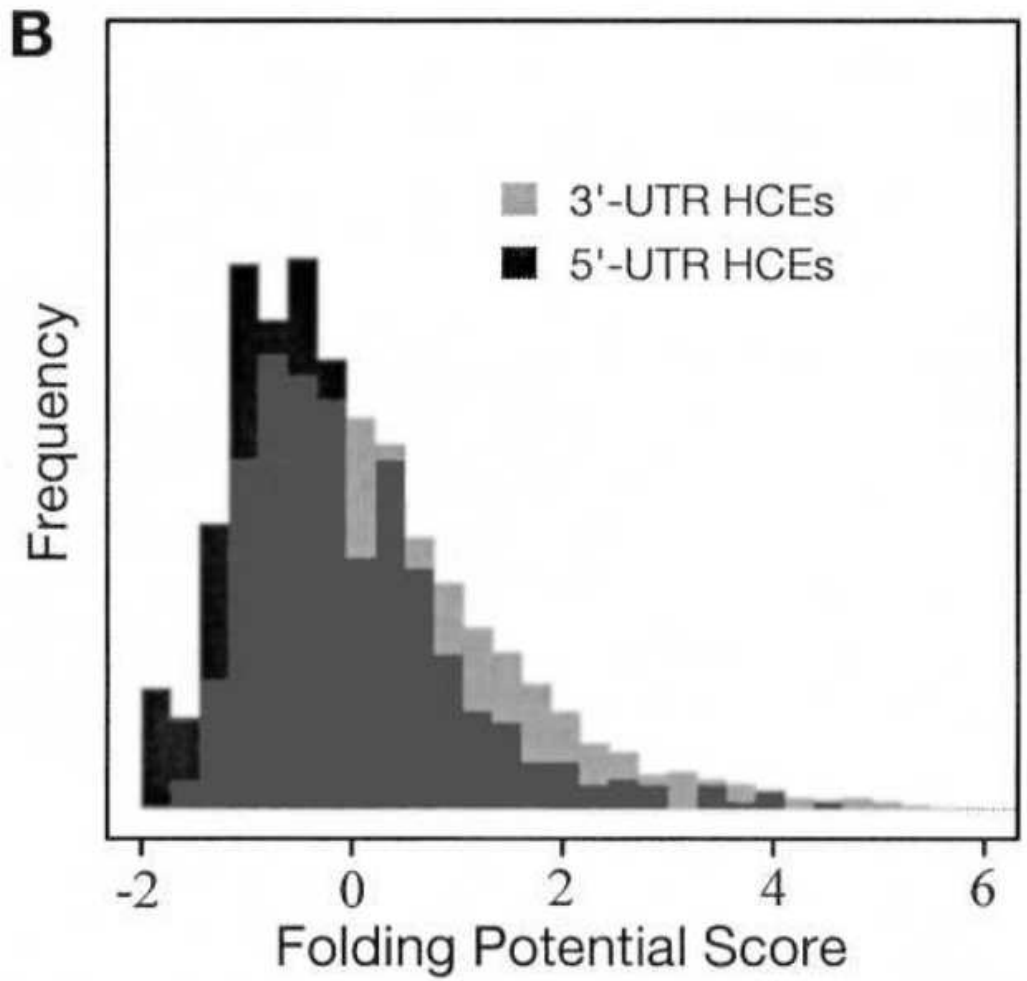
Some more highlights

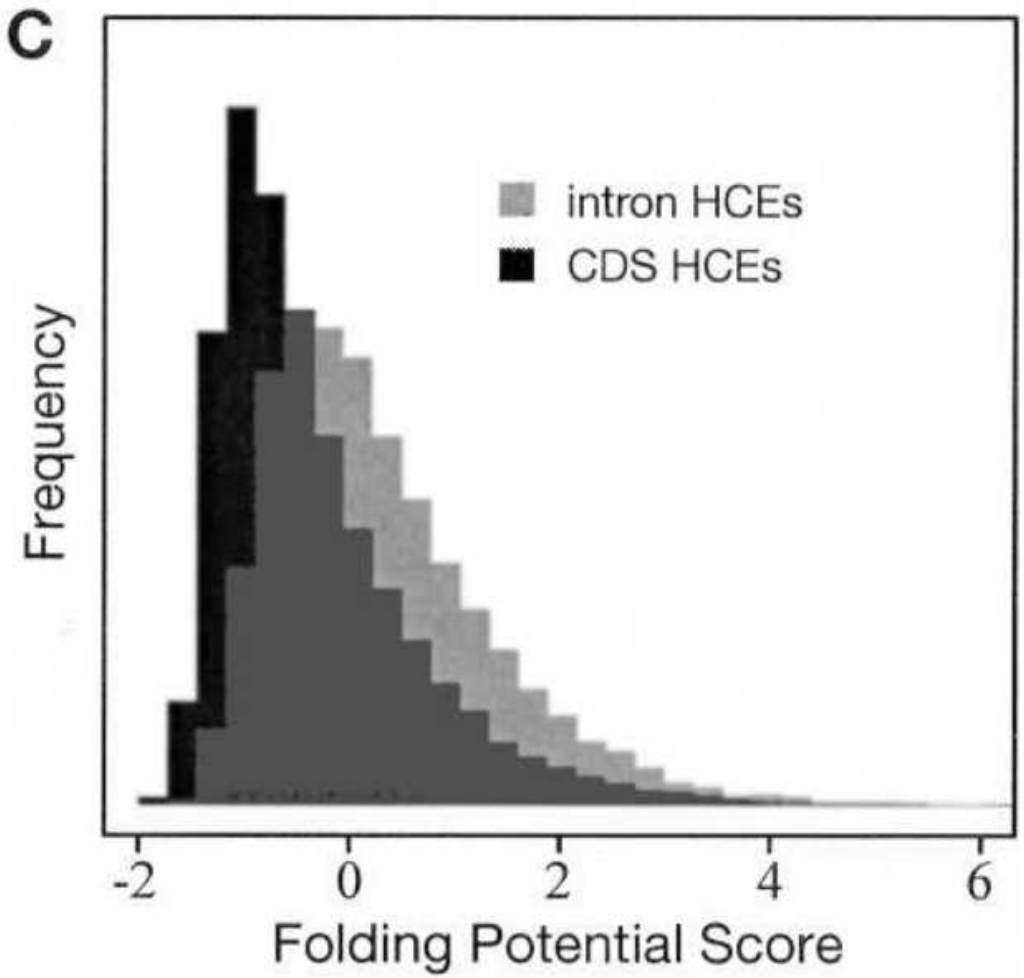
- Only $\approx 42\%$ of the vertebrate HCE overlap exons of known protein coding genes ($> 93\%$ for the other species groups).
- Some of the most extreme conservation in vertebrates is seen in 3'UTRs, particularly of genes that regulate other genes.
  - Less pronounced in insects and was not observed in worms; data for yeast was not available.
- HCE in vertebrate 3'UTRs show strong statistical evidence of an enrichment for local RNA secondary structure.
  - Consistent with the hypothesis of a role in post-transcriptional regulation.

Some more highlights (contd.)

- Testing HCEs in UTRs for statistical evidence of secondary structure using phylo-SCFG which is similar to phylo-HMM.
  - SCFG: stochastic *context-free grammar*.
  - Using folding potential score (FPS).
- It is worth noting that the non-HCE 5'UTRs had significantly higher FPSs than the non-HCE 5'UTRs.
  - This suggests that there is also widespread secondary structure in 3'UTRs outside of HCE.


A
3'-UTR HCEs
3'-UTR null set
Frequency
-2
0
2
4
6
Folding Potential Score

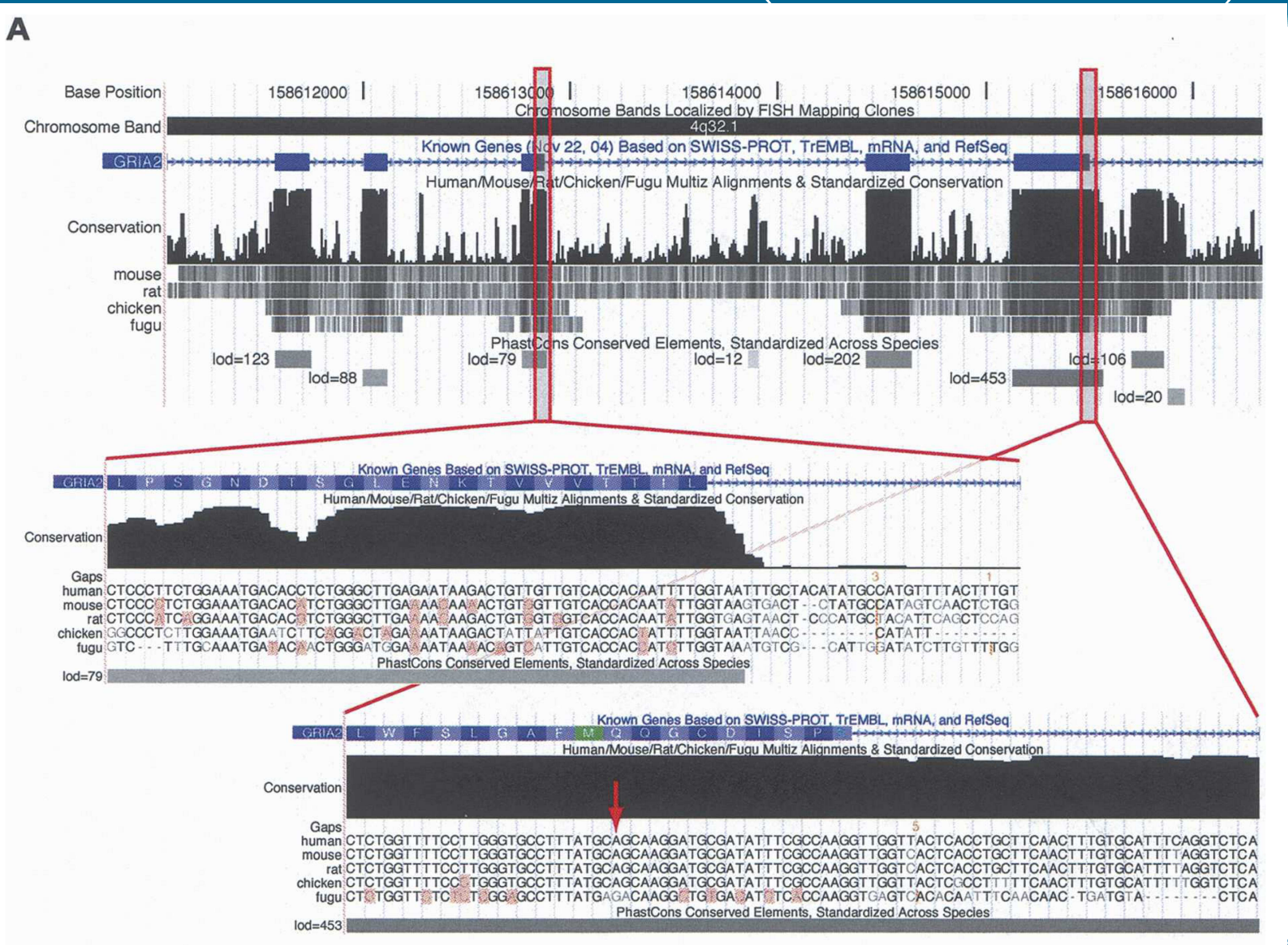

B
3'-UTR HCEs
5'-UTR HCEs
Frequency
-2
0
2
4
6
Folding Potential Score


C
intron HCEs
CDS HCEs
Frequency
-2
0
2
4
6
Folding Potential Score

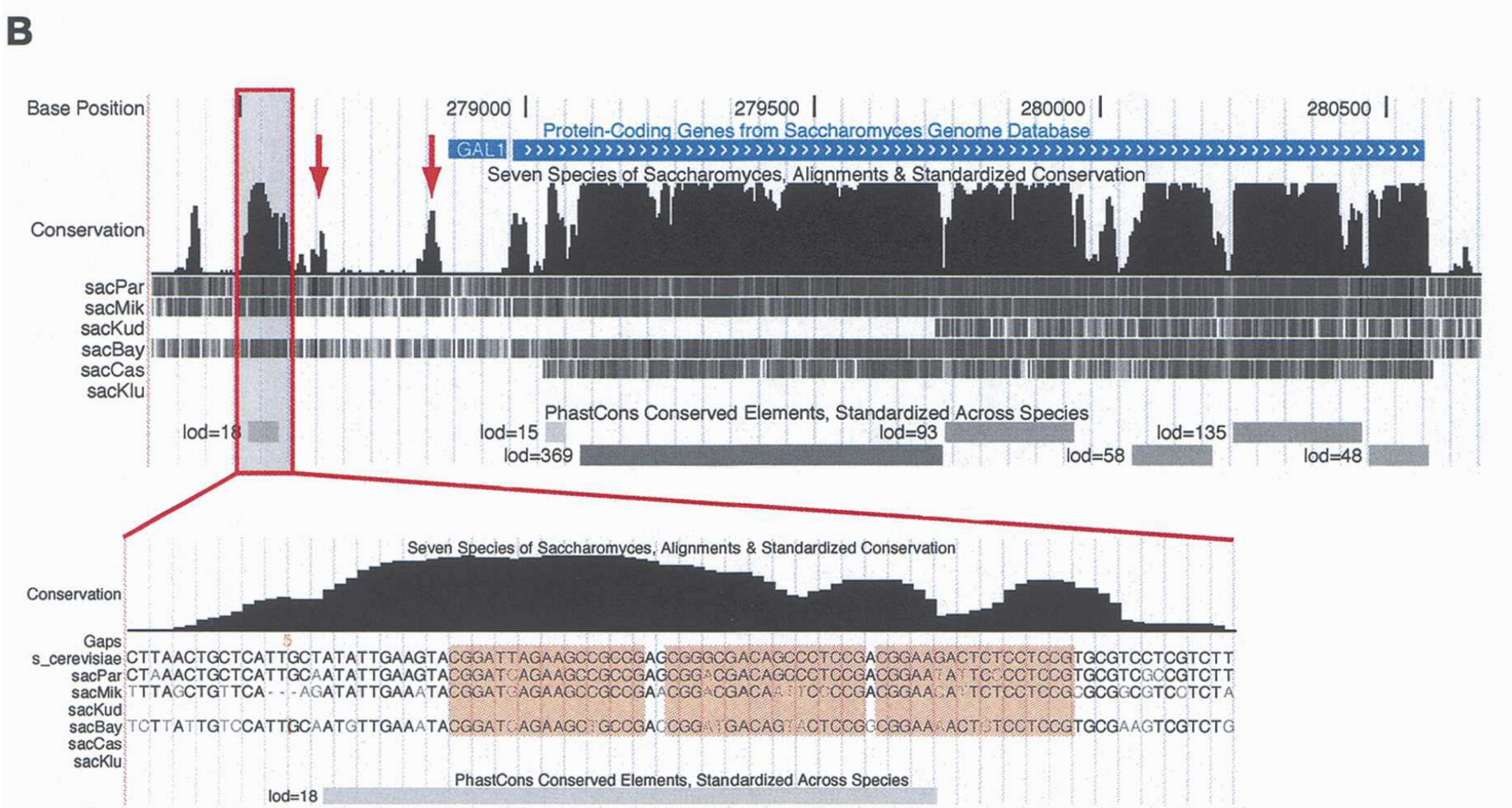
Some more highlights (contd.)

- Predicted conserved elements also include 42% of the bases in a set of 561 putative RNA genes, and 56% of these genes are overlapped by predicted conserved elements.
  - PhastCons are reasonably sensitive for detecting functional noncoding as well as protein-coding sequences.
- In vertebrates, **intergenic HCEs** are strongly enriched in stable gene deserts.
  - minimum length of 640kb & cover 25% of the human genome.
  - low G+C content, high SNP rates, ...
  - They may act as distal *cis*-regulatory elements (Ovcharenko *et al.* 2005).

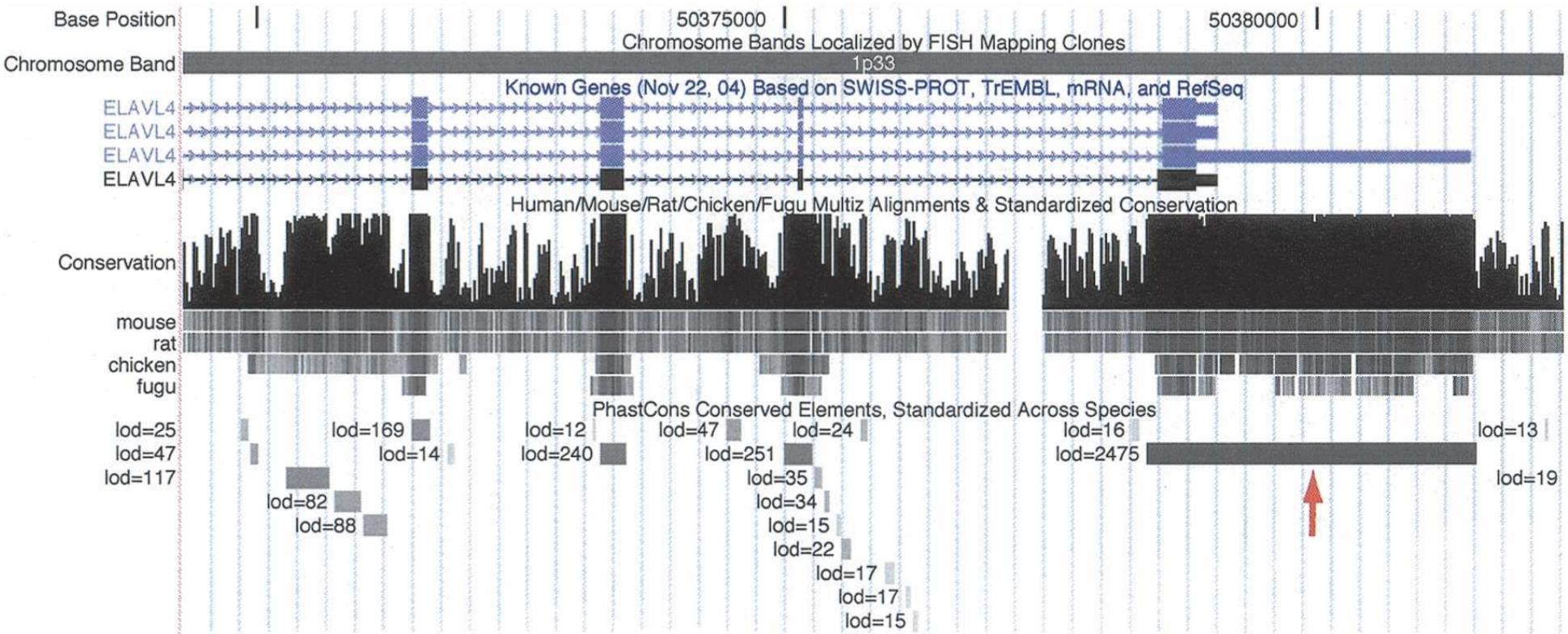
Screen shots of the conservation tracks (human & S. cerevisiae)



Screen shots of the conservation tracks (human & S. cerevisiae)



Screen shots of the conservation tracks (human & S. cerevisiae)



As to HCEs in the 3'UTRs of vertebrate genes

- It suggests that regulation in 3'UTRs plays a key role in critical regulatory networks.
- Post-transcriptional regulation by microRNA binding in 3'UTRs.
  - Human genes with predicted microRNA targets appear to be somewhat enriched for 3'UTR HCEs.
- Antisense transcription (Lipman 1997).

Selected GO categories of **vertebrate** genes overlapped by HCE

Table 1. Selected gene ontology (GO) categories of vertebrate genes overlapped by highly conserved elements

			CDS			5′ UTR			3′ UTR			Intron		
Term	Description	N^a	exp.[b]	obs.[c]	P^d	exp.	obs.	P	exp.	obs.	P	exp.	obs.	P
GO:0003677	DNA binding	1914	164.5	378	1.3e-62	59.4	158	1.5e-33	84.4	221	1.0e-45	28.6	80	5.1e-19
GO:0030528	transcription regulator activity	1125	96.7	251	1.7e-49	34.9	119	2.4e-34	49.6	140	8.5e-31	16.8	54	6.2e-15
GO:0007275	development	1746	150.1	266	1.2e-22	54.2	115	1.0e-15	77.0	122	1.1e-07	26.0	47	3.8e-05
GO:0005216	ion channel activity	334	28.7	79	3.8e-17	10.3	24	1.2e-04	14.7	16	4.0e-01	4.9	2	1.2e-01
GO:0006333	chromatin assembly/disassembly	153	13.1	47	3.1e-15	4.7	11	8.3e-03	6.7	17	4.2e-04	2.2	2	6.0e-01
GO:0007399	neurogenesis	384	33.0	82	5.2e-15	11.9	38	2.7e-10	16.9	36	1.7e-05	5.7	15	6.7e-04
GO:0009887	organogenesis	880	75.6	144	1.0e-14	27.3	67	6.2e-12	38.8	64	5.2e-05	13.1	27	3.0e-04
GO:0009653	morphogenesis	1099	94.4	169	1.3e-14	34.1	76	2.2e-11	48.5	77	3.1e-05	16.4	34	3.8e-05
GO:0008066	glutamate receptor activity	38	3.2	19	3.6e-11	1.1	6	1.0e-03	1.6	5	2.5e-02	–	–	–
GO:0008134	transcription factor binding	251	21.5	54	1.9e-10	7.7	21	3.8e-05	11.0	35	1.5e-09	3.7	10	4.5e-03
GO:0005515	protein binding	2179	187.3	252	1.4e-07	67.7	98	6.9e-05	96.1	141	8.9e-07	32.5	41	6.7e-02
GO:0007018	microtubule-based movement	55	4.7	18	3.9e-07	–	–	–	2.4	8	2.6e-03	0.8	2	2.0e-01
GO:0003723	RNA binding	601	51.6	88	4.2e-07	18.6	26	5.6e-02	26.5	66	5.5e-12	8.9	7	3.2e-01
GO:0007268	synaptic transmission	240	20.6	44	1.1e-06	7.4	12	7.2e-02	10.5	10	5.1e-01	–	–	–
GO:0030154	cell differentiation	200	17.1	37	6.4e-06	6.2	17	1.7e-04	8.8	15	3.2e-02	2.9	7	3.1e-02
GO:0007267	cell-cell signaling	532	45.7	77	3.5e-06	16.5	23	6.9e-02	23.4	24	4.9e-01	7.9	2	1.3e-02
GO:0016071	mRNA metabolism	188	16.1	35	9.8e-06	5.8	10	6.9e-02	8.2	29	3.7e-09	2.8	3	5.4e-01
GO:0006397	mRNA processing	170	14.6	30	1.2e-04	5.2	8	1.6e-01	7.5	24	4.5e-07	2.5	3	4.7e-01
GO:0006512	ubiquitin cycle	542	46.6	69	5.9e-04	16.8	22	1.2e-01	23.9	45	3.4e-05	8.1	3	3.6e-02

[a]Number of genes in background set assigned to category.
[b]Expected number of genes overlapped under background distribution.
[c]Observed number of genes overlapped.
[d]*P*-value. Values of less than 5e–5 can be considered significant (see Methods).

Selected GO categories of **insect, worm, and yeast genes overlapped by HCE (CDS only)**

Group	Term	Description	N^a	exp.^{b}	obs.^{c}	P^d
insect	GO:0000166	nucleotide binding	762	234.0	423	2.1e-51
	GO:0007275	development	1214	372.8	575	3.6e-41
	GO:0009653	morphogenesis	702	215.5	377	8.1e-41
	GO:0009887	organogenesis	605	185.8	330	2.8e-37
	GO:0007552	metamorphosis	281	86.2	161	4.4e-21
	GO:0007399	neurogenesis	296	90.9	167	7.1e-21
	GO:0016462	pyrophosphatase activity	290	89.0	164	1.2e-20
	GO:0005515	protein binding	496	152.3	244	3.5e-19
	GO:0007267	cell-cell signaling	131	40.2	82	3.1e-14
	GO:0007268	synaptic transmission	119	36.5	76	5.9e-14
	GO:0030154	cell differentiation	237	72.7	125	6.0e-13
	GO:0004672	protein kinase activity	223	68.4	119	7.7e-13
	GO:0016310	phosphorylation	279	85.6	141	1.5e-12
	GO:0007017	microtubule-based process	90	27.6	58	3.4e-11
	GO:0048477	oogenesis	206	63.2	102	8.3e-09
	GO:0030528	transcription regulator activity	324	99.5	148	4.9e-09
	GO:0019953	sexual reproduction	300	92.1	133	2.7e-07
	GO:0004386	helicase activity	78	23.9	44	2.0e-06
	GO:0005244	voltage-gated ion channel activity	31	9.5	22	4.5e-06

Selected GO categories of **insect, worm, and yeast genes overlapped by HCE (CDS only)**

Group	Term	Description	N^a	exp.^{b}	obs.^{c}	P^d
worm	GO:0042302	structural constituent of cuticle	156	11.7	71	3.9e-39
	GO:0006811	ion transport	603	45.3	113	5.6e-21
	GO:0000166	nucleotide binding	1012	76.0	139	1.8e-13
	GO:0006520	amino acid metabolism	111	8.3	27	3.1e-08
	GO:0007155	cell adhesion	63	4.7	19	9.0e-08
	GO:0006412	protein biosynthesis	290	21.7	48	1.3e-07
	GO:0008158	hedgehog receptor activity	28	2.1	11	2.6e-06
	GO:0016787	hydrolase activity	1193	89.6	129	6.1e-06
	GO:0009451	RNA modification	37	2.7	11	5.6e-05
yeast	GO:0000166	nucleotide binding	619	98.1	251	9.7e-58
	GO:0003735	structural constituent of ribosome	201	31.8	89	1.4e-22
	GO:0004386	helicase activity	94	14.9	53	1.1e-19
	GO:0006520	amino acid metabolism	222	35.2	85	9.5e-17
	GO:0016787	hydrolase activity	765	121.3	196	2.4e-14
	GO:0006096	glycolysis	32	5.0	22	2.6e-11
	GO:0006006	glucose metabolism	75	11.8	32	2.3e-08
	GO:0016301	kinase activity	216	34.2	61	1.7e-06
	GO:0003723	RNA binding	335	53.1	85	2.3e-06
	GO:0000287	magnesium ion binding	86	13.6	31	3.3e-06
	GO:0009451	RNA modification	85	13.4	30	7.8e-06
	GO:0016310	phosphorylation	190	30.1	53	1.2e-05
	GO:0006333	chromatin assembly or disassembly	32	5.0	15	3.5e-05

Discussion

- Identified HCEs, on average, are longer and less perfectly conserved (perfect identity) in all four species groups.
- Some conserved elements spand hundreds or thousands bases.
  - HCEs may result from *overlapping constraints*.
    - e.g., overlapping binding sites,
    - binding sites overlapping RNA structural or protein-coding constraints (as in RNA editing sites within coding regions)
    - "hubs" of regulatory networks.
    - ...

Discussion (contd.)

- *Alternative splicing* might provide some useful clues about how unusual non-coding conservation arises.
  - fine tuned by evolution to promote splicing in certain tissue types or development stages.
  - the same sequence may bind > 1 factor or have roles both in protein binding and in determining secondary structure (overlapping constraints).
  - ⋆ However, an enrichment for HCEs has not been found in a set of ≈ 5000 alternatively retained cassette exons as compared with a background set of exons.

Discussion (contd.)

- Deficiencies of fixing coverage of coding regions by conserved elements:
  - Differences between groups in how coding regions evolve exists.
  - The sensitivity and specificity of methods for detecting conserved elements inevitably depend on;
    - the number of species;
    - their phylogeny;
    - the amount of missing data.
  - ...

Discussion (contd.)

- Alternative calibration methods have led to generally similar results. Certain basic conclusions appear to be fairly robust.
- Probably the weakest part of this paper: the worm data set.
  - the large degree of divergence between C. elegans and C. briggsae led to *low-alignment coverage*.
  - There might be alignment bias.
  - Only two species . . . ($\omega \uparrow$, short conserved elements tend to be missed)

Discussion (contd.)

- The phylo-HMM used is clearly not realistic (oversimplified).
- In general, the more parameter-rich and complex versions increase the computation burden of parameter estimation and prediction **without producing an appreciable improvement in the quality of program's output**.
- The quality of the whole-genome alignments produced by MULTIZ matters.
  - Remember to recompute predictions of conserved elements as multiple aligners improve.

Thanks for your attention.