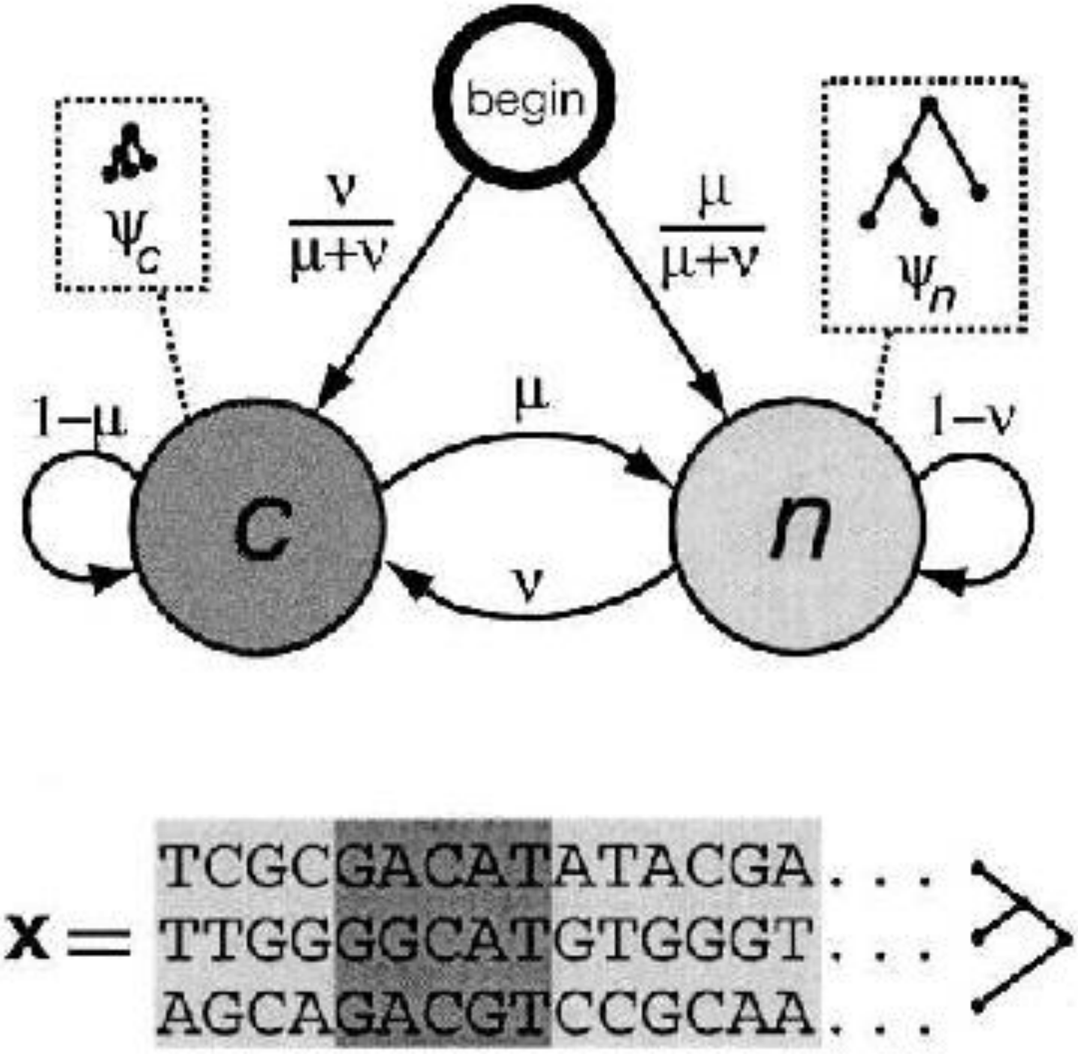
Lecture 16

- PhastCons

PhastCons PhyloHMM



TCGCGACATATACGA . . .
TTGGGGCATGTGGGT . . .
AGCAGACGTCCGCAA . . .

$\mu = a_{cn}$

$\nu = a_{nc}$

from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.

Some general issues in applying probability models, in the PhyloHMM context

- Is the model computable?
- Is the model ‘reasonable’?
  - 2 states enough?
    - variability of mutation, selection within genome
    - changes in selected sites over time
    - but simplicity has its advantages!
      - interpretability
      - overfitting & parameter estimation less problematic
  - Markov condition on transition probabilities
  - treatment of gaps

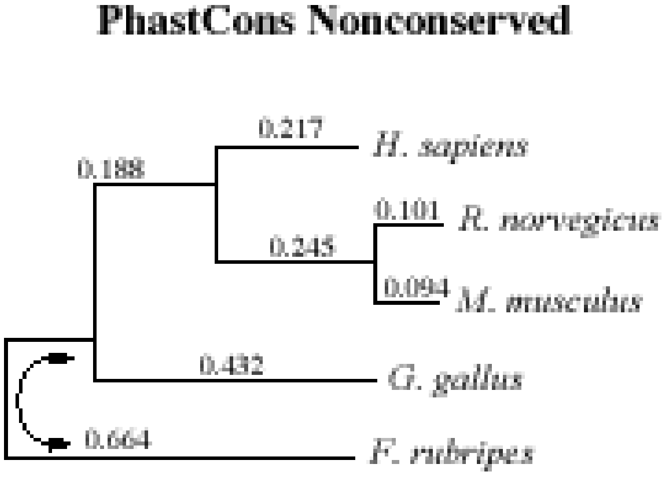
- How good is the input data?
  - – alignability of neutral sequence
  - – accuracy of genome sequence alignments

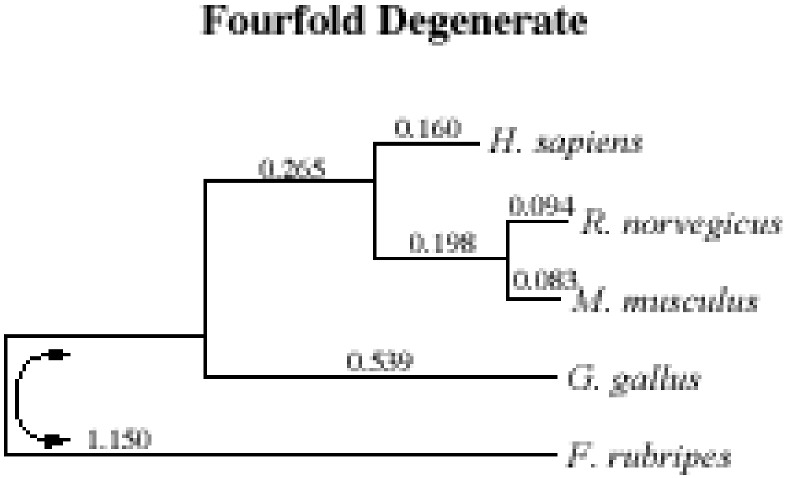
- Are results reliable?
  - – no true ‘test set’ – instead, putative false positive rate, and ‘biological plausibility’ of findings

Alignment issues

- Multiz: progressive pairwise alignments
- accurate multiple genome alignment *not* a solved problem!
  - statistical assessment: Prakash & Tompa (2005, 2007, 2009)
  - ENCODE region alignment analyses: Margulies EH *et al.* 2007
  - major issues:
    - accurate gap placement (even for close species!!)
    - discrimination among paralogous sequences (e.g. repeats, duplications)
    - short ‘junk’ alignment segments
  - *in principle*, more sequences should give more accurate alignments
- inaccurate alignments can cause
  - neutral rate to be *overestimated*
  - conserved segments to be *overidentified*
    - because more slowly mutating (or better aligned) neutral segments may be called conserved

- for distantly related species, neutrally evolving regions no longer alignable
  - – analyze 4D sites in coding sequences to estimate neutral rates
    - CDS alignments much more reliable, but
    - synonymous sites somewhat atypical (some selection; composition & mutation patterns)


PhastCons Nonconserved


Fourfold Degenerate

The Genetic Code

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

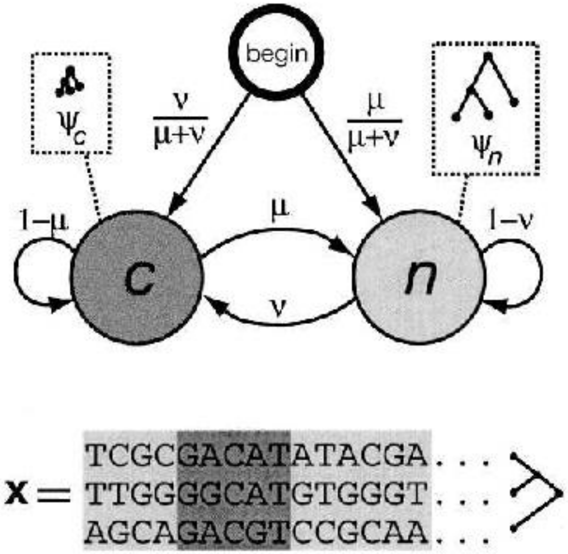
Notation

- $\mu = a_{cn}$, $\omega = 1/\mu$ (expected length of conserved elt)
- $\nu = a_{nc}$
- expected ‘coverage’ γ (frac of genome that is conserved):

$$= \text{Elen (cons seg)} / (\text{Elen(cons seg)} + (\text{Elen(neut seg)})$$

$$= (1/\mu) / (1/\mu + 1/\nu)$$

$$= \nu / (\mu + \nu)$$



- transition probs imply *a priori* length dist'ns for conserved & non-conserved segments
  - prob(cons seg has length *n*) is

$$(a_{cc})^{n\text{-}1}a_{cn} = (a_{cc})^{n\text{-}1}(1 – a_{cc})$$

  - geometric distribution
  - expected length (Elen) ω of conserved segment is

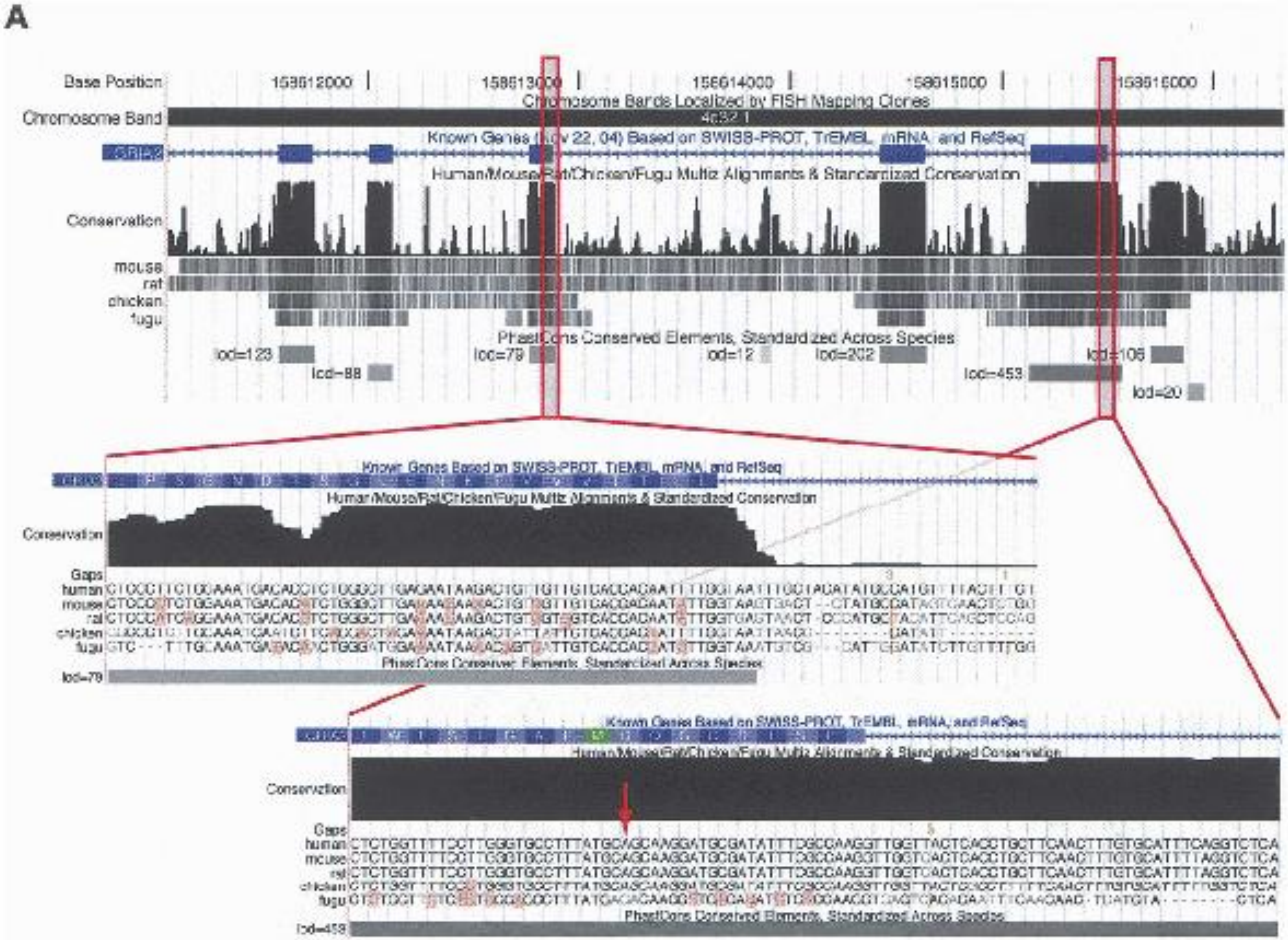
$$1.0 / (1 – a_{cc}) = 1.0 / a_{cn}$$

special case: $a_{cc} = .5 = a_{nn} \Rightarrow$ positions are independent

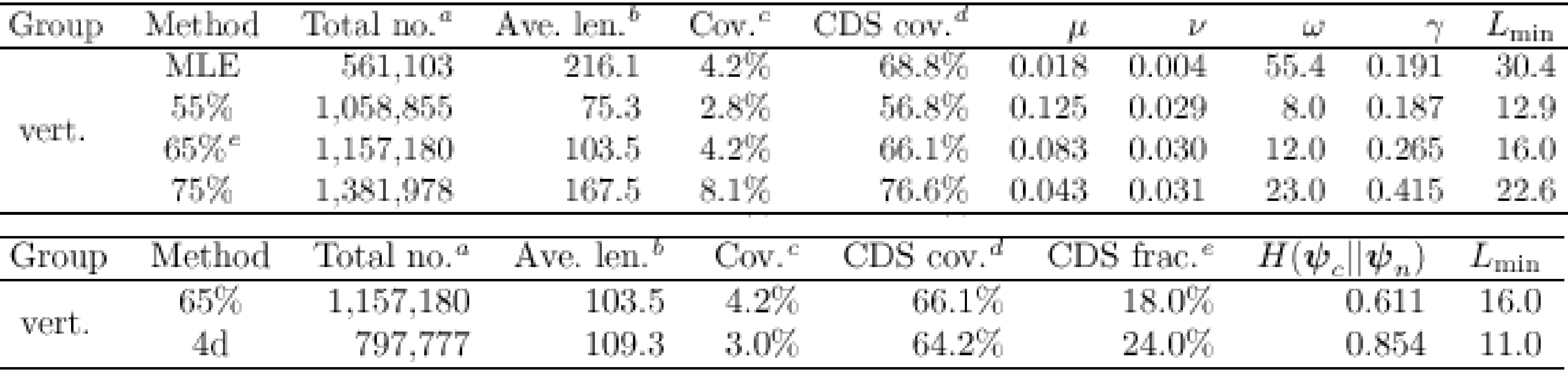
PhastCons Parameter Estimation

- parameters estimated separately in 1 Mb windows using EM algorithm
  - full maximum likelihood analysis, or
  - constraining some parameters

  & averaged over genome
- full MLE results don’t match biologists’ intuition -- too much ‘smoothing’:
  - fewer, & larger, conserved elements
  - long, apparently non-conserved regions within conserved elements
  - attributed to fact that (prior) geometric length dist’n inappropriate


A
Base Position
156612000
156613000
156614000
156615000
156616000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
4q32.1
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
GRIA2
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
mouse
rat
chicken
fugu
PhastCons Conserved Elements, Standardized Across Species
lod=123
lod=88
lod=79
lod=12
lod=202
lod=453
lod=105
lod=20
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
human
mouse
rat
chicken
fugu
PhastCons Conserved Elements, Standardized Across Species
lod=79
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
human
mouse
rat
chicken
fugu
PhastCons Conserved Elements, Standardized Across Species
lod=453

from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.



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

Group	Method	Total no.[a]	Ave. len.[b]	Cov.[c]	CDS cov.[d]	CDS frac.[e]	$H(\psi_c\|\|\psi_n)$	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

Instead: -- impose constraints

- coverage constraint:
  - 65% of coding bases covered by conserved elts
  - (target value based on earlier mouse/human analysis)
- smoothness constraint:
  - PIT (≡ expected min. amt of phylogenetic info required to predict a conserved element)

    = 9.8 bits

    - (forced to be same for all species groups)

- constraints met by ‘tuning’ γ and ω (or equivalently transit probs)
  - choose γ and ω,
  - get ML estimates of other parameters by EM algorithm
  - see whether get desired coverage & PIT
  - if not, adjust γ and ω & redo

- $L_{\min}$: expected min length of a conserved segment that could appear in a Viterbi path
- at $L_{\min}$,

  expected loglike of staying in state n

  = expected loglike of switching to c & back again, so

$$(L_{\min}+1)\log(1-\nu)+L_{\min}\sum_{x} P(x|\psi_c)\log P(x|\psi_n)$$

$$= \log\nu + \log\mu + (L_{\min}-1)\log(1-\mu) + L_{\min}\sum_{x} P(x|\psi_c)\log P(x|\psi_c)$$

- $L_{\min} = \dfrac{\log\nu + \log\mu - \log(1-\nu) - \log(1-\mu)}{\log(1-\nu) - \log(1-\mu) - H\,(\psi_c||\psi_n)}$

- where

$$H(\psi_c||\psi_n) = \sum_x P(x|\psi_c) \log \frac{P(x|\psi_c)}{P(x|\psi_n)}$$

= rel entropy of *c*-state emission prob dist'n w.r.t.

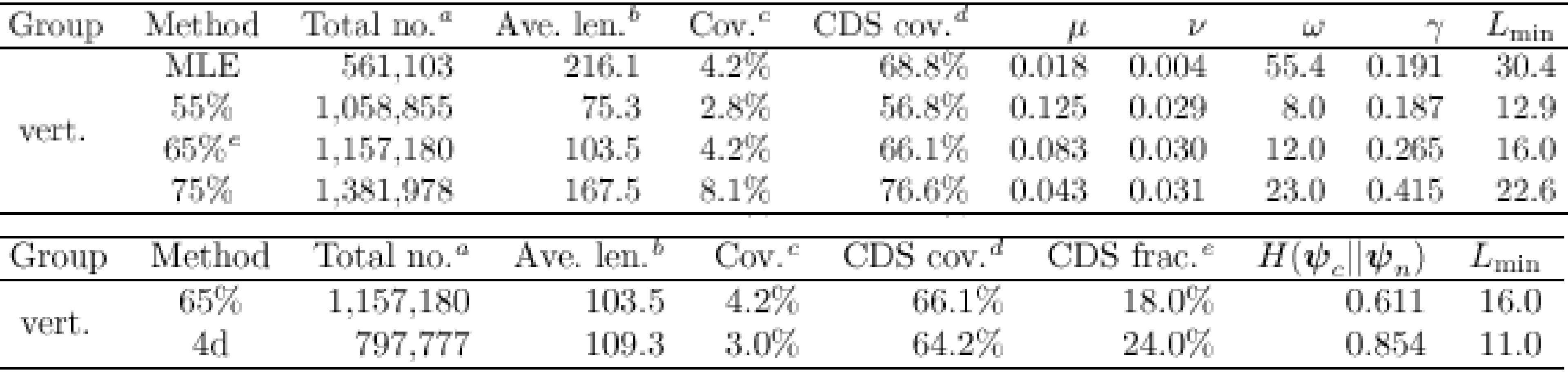
n-state dist'n

- PIT (phylogenetic information threshold)

= $L_{\min} H(\psi_c||\psi_n)$.

= 'expected min amt of phylogenetic info required to predict conserved element'

- Final param estimates (for vertebrates):
  - $\gamma = 0.265$
  - $\omega = 12.0$ bp
  - $H(\psi_c \| \psi_n) = .608$ bits / site
  - $L_{\min} = 16.1$ bp
  - PIT $= L_{\min} H(\psi_c \| \psi_n) = 9.8$ bits

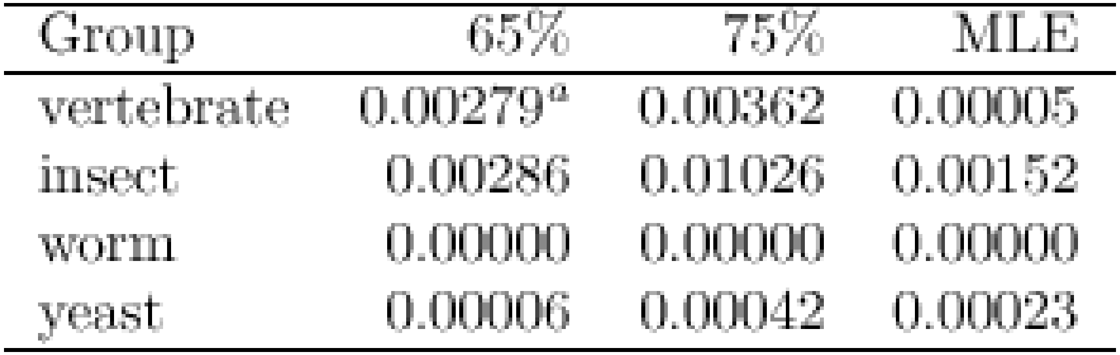


Group	Method	Total no.[a]	Ave. len.[b]	Cov.[c]	CDS cov.[d]	μ	ν	ω	γ	$L_{\min}$
vert.	MLE	561,103	216.1	4.2%	68.8%	0.018	0.004	55.4	0.191	30.4
	55%	1,058,855	75.3	2.8%	56.8%	0.125	0.029	8.0	0.187	12.9
	65%[e]	1,157,180	103.5	4.2%	66.1%	0.083	0.030	12.0	0.265	16.0
	75%	1,381,978	167.5	8.1%	76.6%	0.043	0.031	23.0	0.415	22.6

Group	Method	Total no.[a]	Ave. len.[b]	Cov.[c]	CDS cov.[d]	CDS frac.[e]	$H(\psi_c \vert\vert \psi_n)$	$L_{\min}$
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

Estimating false positive rates

- simulate 1 Mb alignment
  - by sampling 4D sites (with replacement) from aligned CDSs
  - caveat: these not typical of all neutral sites!
- predict cons elts (using prev param estimates)
- frac of bases in cons elts:



Group	65%	75%	MLE
vertebrate	0.00279[a]	0.00362	0.00005
insect	0.00286	0.01026	0.00152
worm	0.00000	0.00000	0.00000
yeast	0.00006	0.00042	0.00023

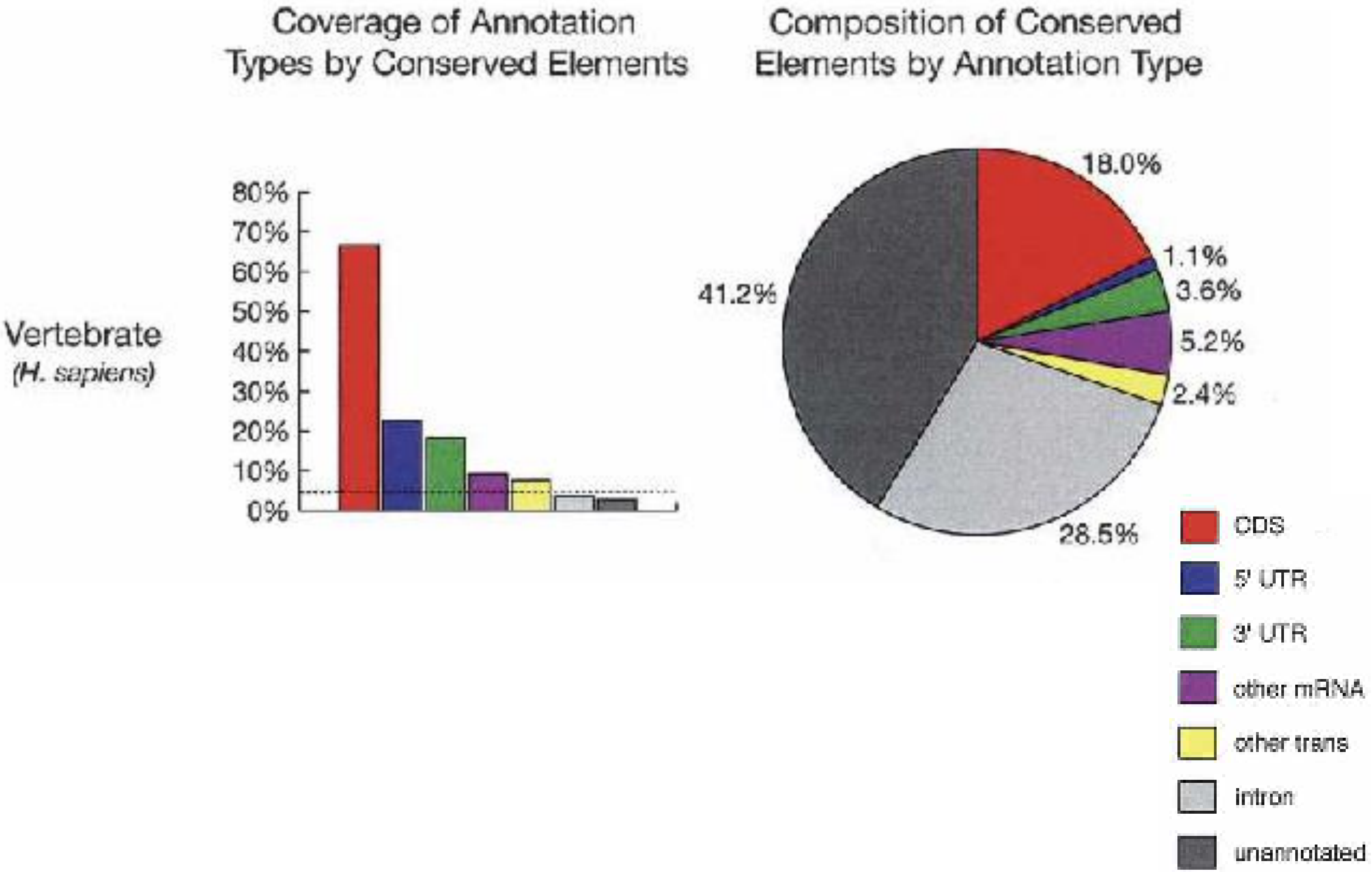
- does not address (important) issue of rate of false positive bases within, or flanking, true conserved elements
- also: genes more G+C rich than genome average, & have somewhat higher mutation rate (due in part to more frequent CpGs)

  ⇒ *underestimating* false pos rate
- also: randomization procedure destroys underlying mutation rate variation

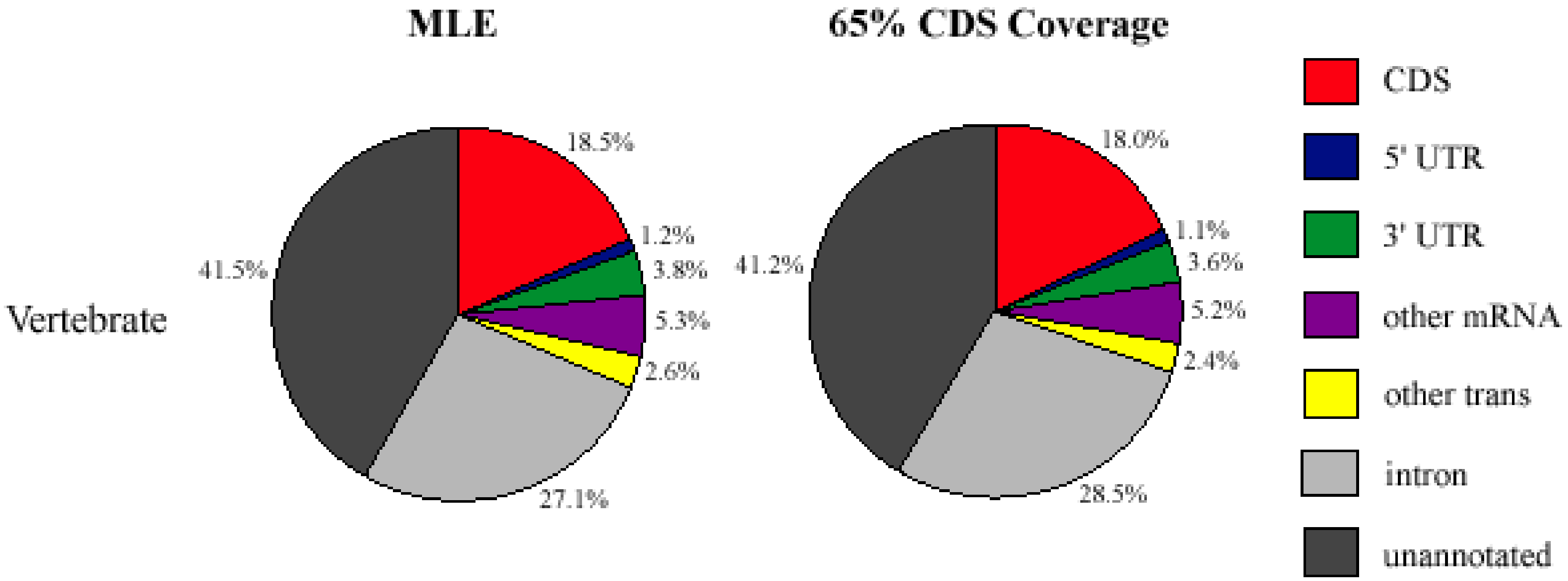
  ⇒ *underestimating* false pos rate

Characteristics of phastCons predicted conserved elements

- 1.18 million elements
- constitute 4.3% of human sequence
  - 66% of coding bases
    - 88% of coding exons overlap predicted elt
  - 23% of 5'UTR bases
    - 63% of exons
  - 18% of 3'UTR bases
    - 64% of exons
  - 42% of RNA gene bases
    - 56% of genes
  - 3.6% of intronic bases
  - 2.7% of intergenic bases
  - < 1% of mammalian 'ancestral repeats' (ARs)



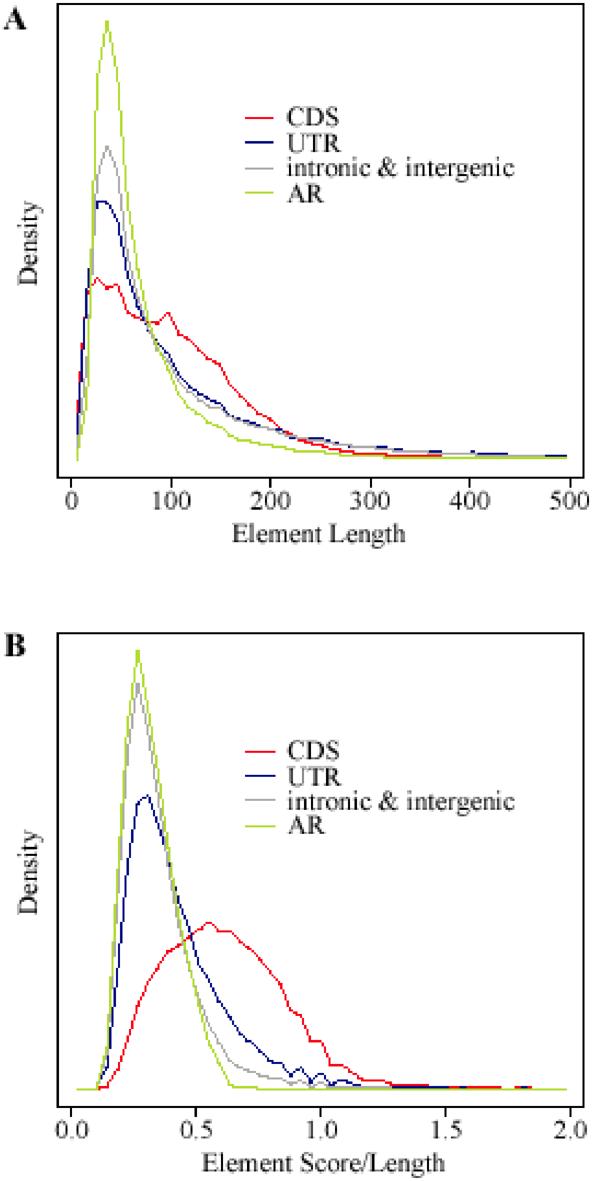
from Siepel A. *et al*. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 15:1034-50.



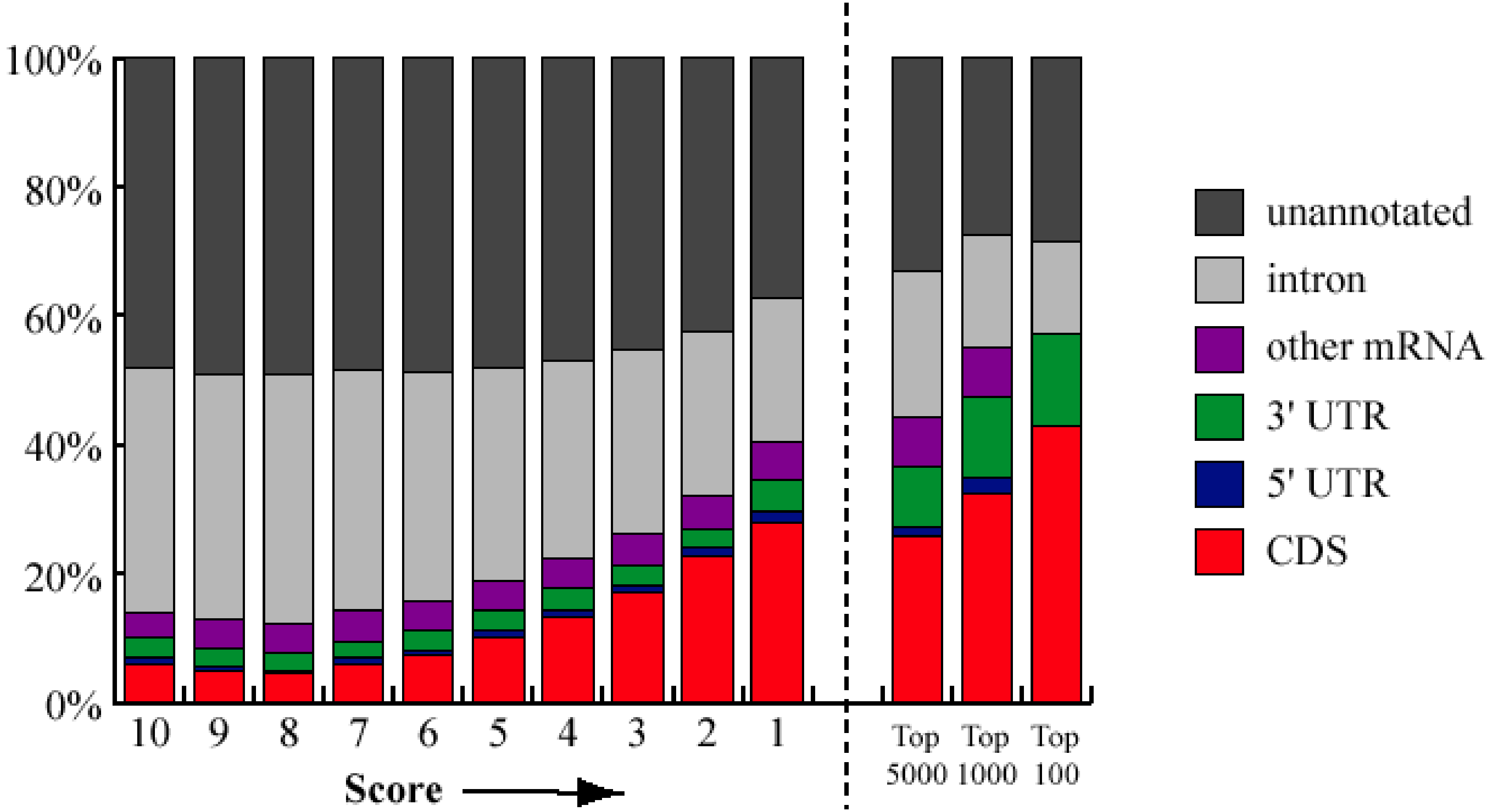
from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.

Length dist’ns of conserved elements

- lengths approx. geometrically distributed, avg 104 bp
- length dist’n depends on annotation category



from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.



from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.

Highly conserved elements (HCEs)

- top 5000 in score; cover 0.14% of human genome
  - mean length 781 bp (range 318-4922)
- probably a more sensible category to study than ‘ultraconserved elements’
- non-randomly distributed with respect to genes
  - overrepresented in or near regulatory (DNA-, RNA-binding) genes, some other classes (e.g. ion channels)
  - overrepresented in 3’ UTRs – some associated with miRNA binding sites
  - also enriched in ‘stable gene deserts’
- enriched for RNA-folding potential
- why long highly conserved regions? clusters of binding sites?

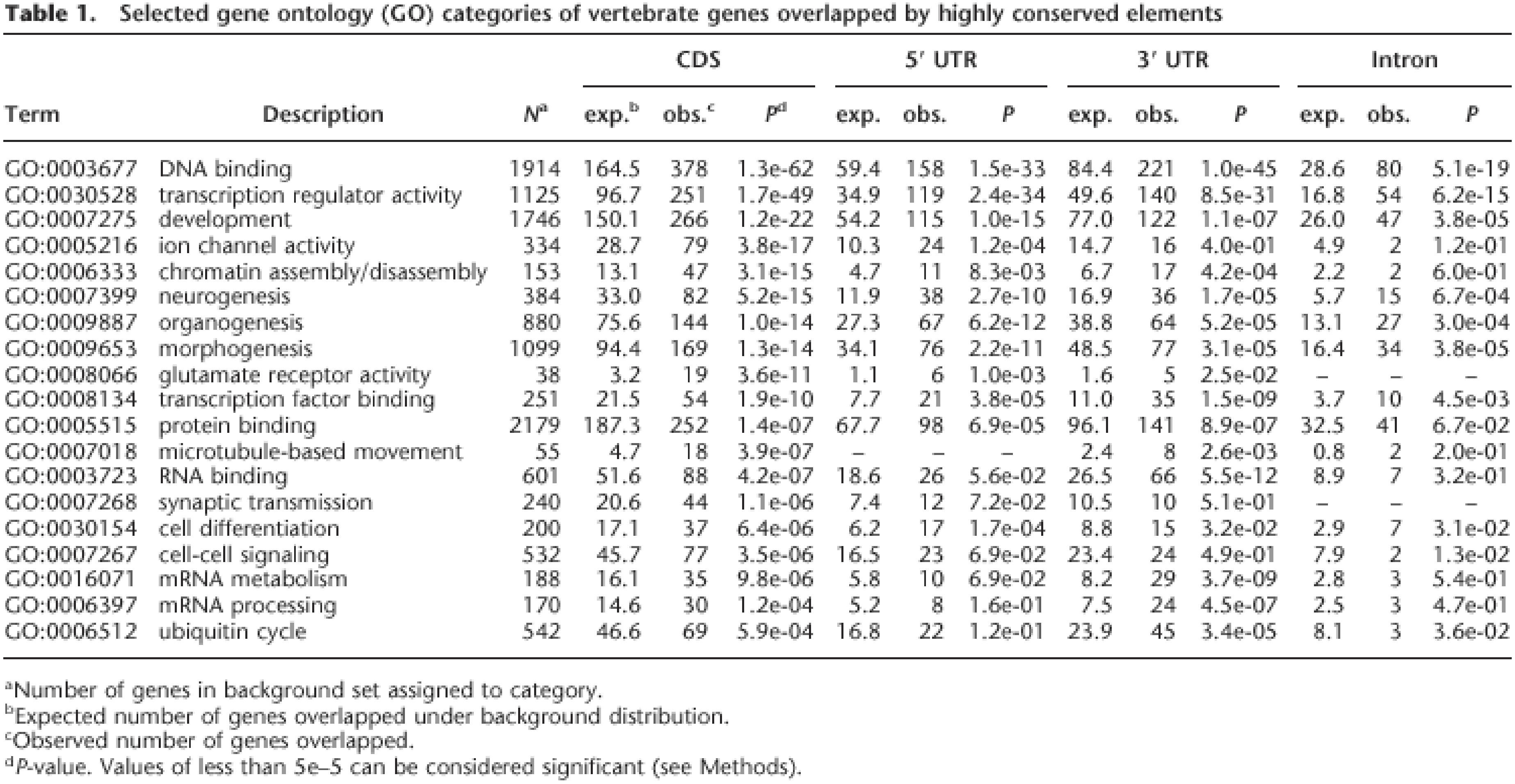


Table 1. Selected gene ontology (GO) categories of vertebrate genes overlapped by highly conserved elements

			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).

from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.