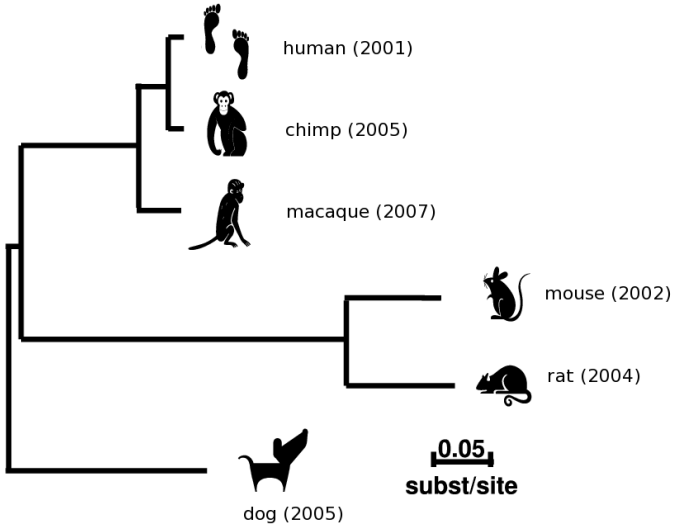
Announcements

- Homework 1 is due next Tuesday, November 9 22:00
  submit in Moodle, guests by email to brejovadcs.fmph.uniba.sk
  discussion regarding questions in MS Teams
- Work on the journal club
  (read the paper, plan the meeting no later than Nov. 23)
- Next week Bratislava in the red zone
  we will try to keep the possibility of in-person classes

Comparative Genomics

Tomáš Vinař

November 4, 2021



Comparative genomics (komparatívna genomika)

- Genome evolution:
  - Single point mutations (this lecture)
  - Short insertions and deletions
  - Large-scale events: rearrangements and duplications
- Mutations according to their effect:
  - Neutral
  - Deleterious (škodlivé)
    $\Rightarrow$ **purifying selection (purifikačný výber)**
  - Advantageous (prospešné)
    $\Rightarrow$ **positive selection (pozitívny výber)**
- By comparing several genomes,
  find regions that evolve in an unusual way
  (e.g. conserving an important function, evolving a new function)

Comparative genomics

- Start with multiple alignment of several genomes
  (aligned sites should have originated from the same ancestral sequence)

```
          Human AGTGGCTGCCAGGCTG---GGATGCTGAGGCCTTGTTTGCAGGGAGGT
         Rhesus AGTGGCTGCCAGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGAGGT
          Mouse GGTGGCTGCCGGGCTG---GGTGGCTGAGGCCTTGTTGGTGGGGTGGT
            Dog AGTGGCTGCCCGGCTG---GGTGGCTGAGGCCTTATTTGCAGGGAGGT
          Horse GATGGCTGCCGGGCTG---GGCTGCCGAGGCCTTGTTCGTGGGGAGGT
      Armadillo AGTGGCTGCCGGGCTG---GGAGGCCAAGGCCTTGTTCGCGGGCAGGT
        Chicken AGTGGCTGCCAGTCTGCGCCGTGGCCGACGTCTTGCTCGGGGGAAGGT
X. tropicalis AATGGCTTCCATTTTGTGCCGCTGCTGAGGTCTTGTTCTGGGGAAGAT
```

- **Methods:** Combine techniques for sequence annotation (HMMs) and evolutionary models

Application 1: Finding functional elements of the genomes

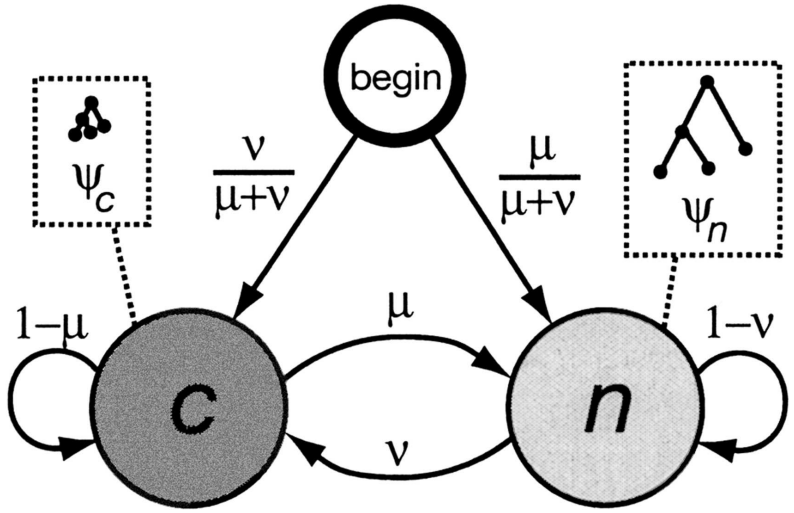
Consequences of purifying selection

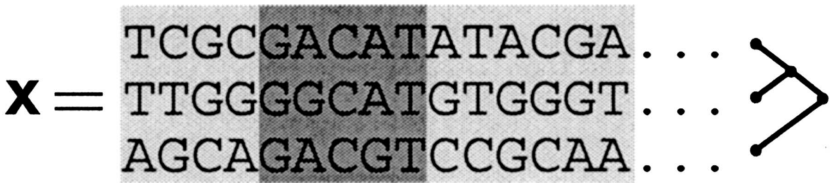
- Important functional sequences are likely to be conserved: they appear to evolve slower
- Non-functional sequences evolve faster
- **Example:** protein coding genes in humans and mouse
  - coding regions: 85% identity (98% of their total length aligned)
  - introns: 69% identity (48% of their total length aligned)
- **Task:** find **well-conserved sequences** between organisms
- Majority of conserved sequences will correspond to known functional elements (coding genes, regulation sequences, etc.)
- Conserved sequences that do not overlap known functional elements: interesting objects for further research

PhastCons: detection of conserved sequences

Phylogenetic HMM:

combination of an HMM and a phylogenetic tree



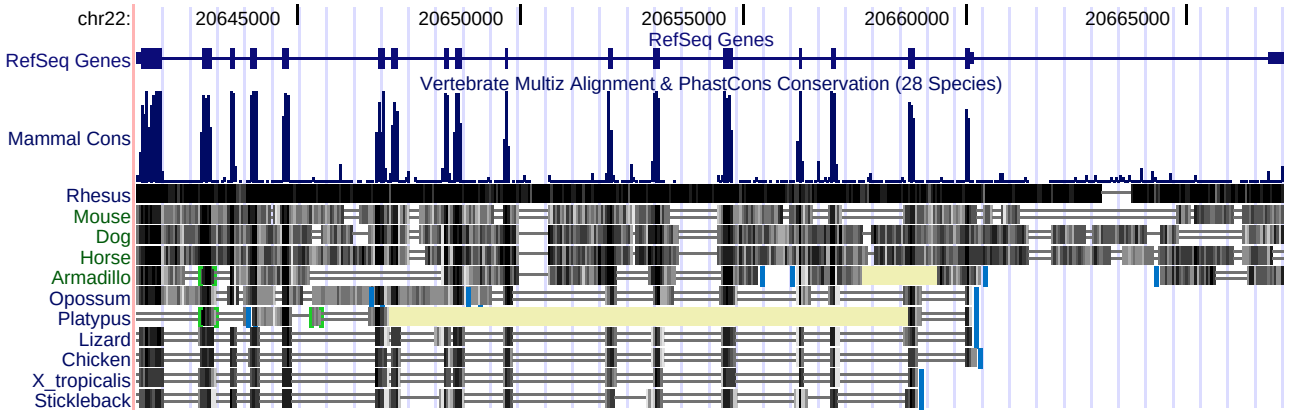


Source: [Siepel et al., 2005]

- Two states: conserved and neutral
- Each state emits a whole column of a sequence alignment
- Conserved sequences have shorter tree branches, causing less sequence divergence

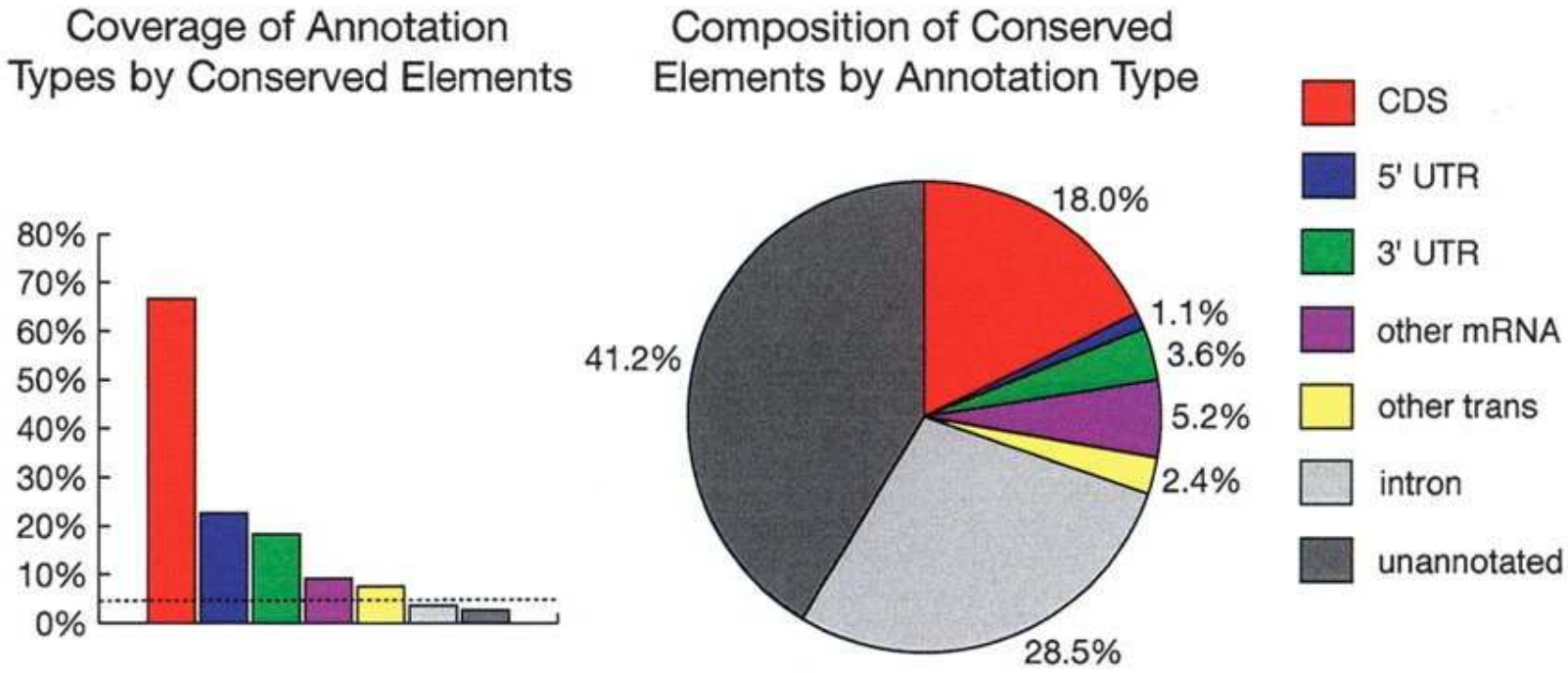
How to use phylogenetic HMMs

- The model gives a probability distribution over all possible alignments and annotations
  (here: annotation = markup of conserved / neutral regions)
- For a given alignment, we are looking for the annotation that would maximize this probability
- Can be done efficiently
  (combination of the Viterbi and Felsenstein algorithms)



Results of PhastCons application to four whole genomes

Alignment of human, mouse, chicken, fugu



Source: [Siepel et al., 2005]

Phylogenetic HMMs for gene finding

- Use states from a typical gene finder
- Each state has a separate evolutionary model (rate matrix, branch lengths)
- Mutation frequencies in coding regions are three-periodic; this helps to find genes

How much we can improve on gene finding results?

	Exons		Genes	
Program	sn	sp	sn	sp
AUGUSTUS (1 genome)	52%	63%	24%	17%
NSCAN (alignment)	68%	82%	35%	37%

Guigo et al 2006, 1% of the human genome

Genetic code

Ala / A	GCT, GCC, GCA, GCG	Leu / L	TTA, TTG, CTT, CTC, CTA, CTG
Arg / R	CGT, CGC, CGA, CGG, AGA, AGG	Lys / K	AAA, AAG
Asn / N	AAT, AAC	Met / M	ATG
Asp / D	GAT, GAC	Phe / F	TTT, TTC
Cys / C	TGT, TGC	Pro / P	CCT, CCC, CCA, CCG
Gln / Q	CAA, CAG	Ser / S	TCT, TCC, TCA, TCG, AGT, AGC
Glu / E	GAA, GAG	Thr / T	ACT, ACC, ACA, ACG
Gly / G	GGT, GGC, GGA, GGG	Trp / W	TGG
His / H	CAT, CAC	Tyr / Y	TAT, TAC
Ile / I	ATT, ATC, ATA	Val / V	GTT, GTC, GTA, GTG
START	ATG	STOP	TAA, TGA, TAG

Application 2: Detecting positive selection in protein coding genes

- **Positive selection:** process that helps to fix **advantageous mutations** in a genome
- Unusually high number of mutations that can lead to change of function
- Mutations in protein coding genes:
  - Synonymous: do not change encoded amino acid
    e.g. ACA (Thr) $\rightarrow$ ACT (Thr)
  - Nonsynonymous: change the amino acid
    e.g. ACA (Thr) $\rightarrow$ AAA (Lys)
- We create a probabilistic model of evolution distinguishing synonymous and nonsynonymous mutations $\Rightarrow$ identification of sequences with unusually high fraction of nonsynonymous mutations

From Jukes-Cantor to more general substitution models

- Jukes-Cantor assumes all mutations are equally probable
- In general μ_{xy} is the substitution rate from base x to base y
- **Substitution rate matrix** (matica rýchlostí)

$$\begin{pmatrix} -\mu_A & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & -\mu_C & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & -\mu_G & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & -\mu_T \end{pmatrix}$$

$$\begin{pmatrix} -\mu_A & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & -\mu_C & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & -\mu_G & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & -\mu_T \end{pmatrix}$$

For given time interval t, we can compute probability of each possible substitution (**transition probabilities**):

$$\Pr(X = C \,|\, Y = A, t)$$

Decreasing the number of parameters — HKY model

Hasegawa, Kishino and Yano [Hasegawa et al., 1985]

$$\begin{pmatrix} -\mu_A & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & -\mu_C & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & -\mu_G & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & -\mu_T \end{pmatrix} \qquad \mu_{x,y} = \begin{cases} \alpha\pi_y & \text{if } x \Leftrightarrow y \text{ is transition} \\ \beta\pi_y & \text{if } x \Leftrightarrow y \text{ is transversion} \end{cases}$$

- frequencies $\pi_A, \pi_C, \pi_G, \pi_T$
  (equilibrium, do not change over time)
- **transition rate (rýchlosť tranzícií)** α: $C \Leftrightarrow T, A \Leftrightarrow G$
- **transversion rate (rýchlosť tranzverzií)** β: $\{C, T\} \Leftrightarrow \{A, G\}$
- Only four parameters: $\pi_A, \pi_C, \pi_G, \kappa = \alpha/\beta$

Codon substitution models

Rate matrices on **codons** rather than single nucleotides

Rate of substitution from codon i to codon j:

$$\mu_{i,j} = \begin{cases} 0, & \text{if } i,j \text{ differ at } > 1 \text{ positions,} \\ \alpha\pi_j, & \text{synonymous transitions,} \\ \beta\pi_j, & \text{synonymous transversions,} \\ \omega\alpha\pi_j, & \text{nonsynonymous transitions,} \\ \omega\beta\pi_j, & \text{nonsynonymous transversions.} \end{cases}$$

Example: $\mu_{AAC,GGC} = 0$, $\mu_{CTA,CTT} = \beta\pi_{CTT}$, $\mu_{CTA,CCA} = \omega\alpha\pi_{CCA}$

Parameters: Codon frequencies π_j, ω, $\kappa = \alpha/\beta$

Selection: neutral evolution $\omega = 1$, positive selection $\omega > 1$, purifying selection $\omega < 1$

Application of codon substitution models

```
           F   V   I   H   D   S   E   G   D   G   E   C   M   Q   E
human    TTT GTG ATC CAC GAC TCC GAG GGG GAC GGC GAG TGC ATG CAG GAG
marmoset TTT GTG ATC CAC GAG AAC AAC AAG GAC GGC GAG TGC ATG CAG GAT
           F   V   I   H   E   N   N   K   D   G   E   C   M   Q   D
```

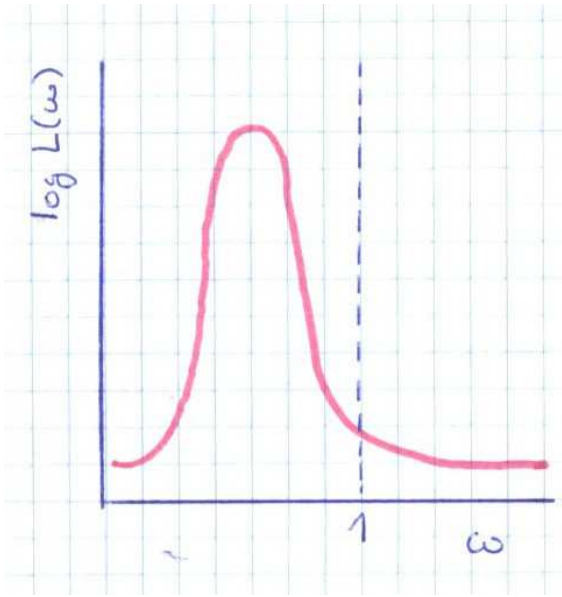
- Using whole genomes, estimate basic model parameters

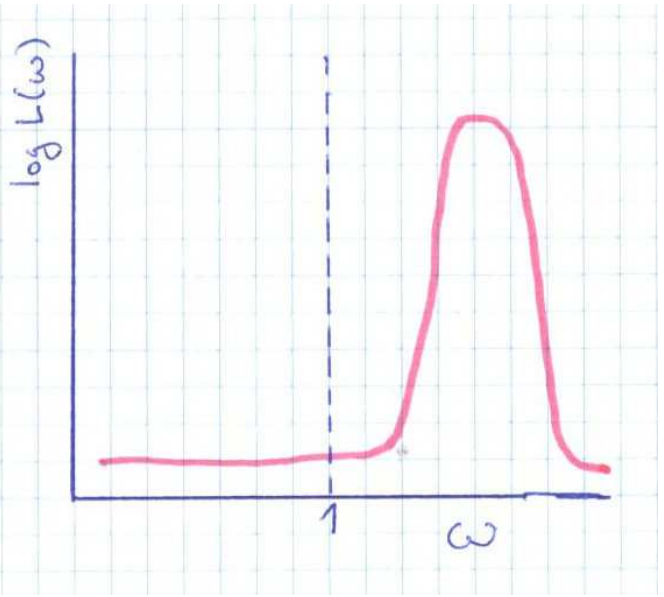
  $\pi_A, \pi_C, \pi_G, \pi_T, \kappa$

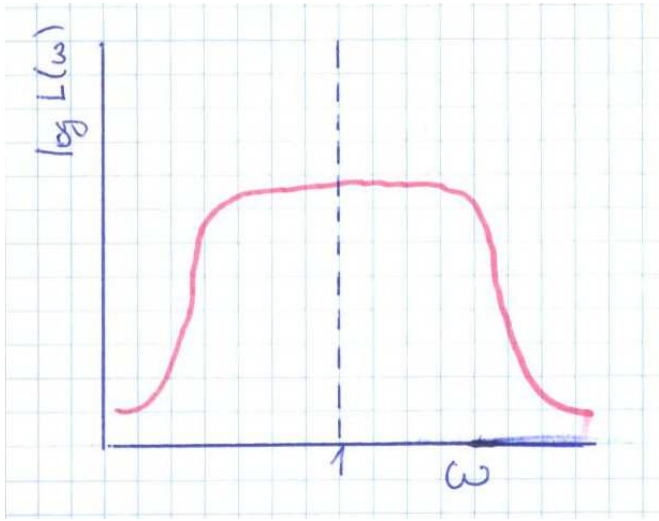
- For a given ω and t, we can compute likelihood (vierohodnosť)

$$L(\omega, t) = \Pr(H, M \,|\, \omega, t)$$

- We can observe how $L(\omega) = \max_t L(\omega, t)$ changes for different values of ω


log L(ω)
1
ω


log L(ω)
1
ω


log L(ω)
1
ω

Likelihood-ratio test (test pomerov vierohodností)

- Even if $L(\omega)$ achieves maximum for $\omega > 1$,
  this can be caused by a statistical variantion in the data
  $\Rightarrow$ we need a statistical test
- Compute likelihood $L_A = \max_{\omega<1} L(\omega)$
- Compute likelihood $L_B = \max_{\omega} L(\omega)$ (no restriction on ω)
- Always $L_B \geq L_A$
- If real $\omega < 1$, then $L_A \approx L_B$ (null hypothesis)
  we are interested in cases $L_B >> L_A$
  $\Rightarrow$ the gene is under positive selection (alt. hypothesis)

Assuming $\omega < 1$, we have $2\log(L_B/L_A) \approx \chi^2_1$
$\Rightarrow$ we can assign P-value to the null hypothesis $\omega < 1$

Detecting positive selection in protein coding genes (summary)

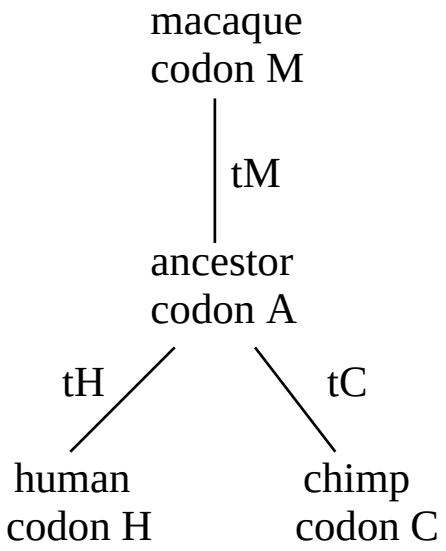
- Align sequences of the same gene from two species (at the codon level)
- Estimate basic parameters of the codon model using whole genome data
- Parameter ω models selection
- Compute likelihoods $L_A = \max_{\omega<1} L(\omega)$ and $L_B = \max_{\omega} L(\omega)$
- Using statistics $2\log(L_B/L_A)$, assign P-value to the null hypothesis $\omega < 1$
- Genes with small P-values are under the positive selection

"Simple" extension to multiple genomes

$$\Pr(A, H, C, M \,|\, \omega, t_H, t_C, t_M) =$$
$$\pi_A \cdot \Pr(H \,|\, A, t_H) \cdot \Pr(C \,|\, A, t_C) \cdot \Pr(M \,|\, A, t_M)$$

Ancestral sequences are not known:

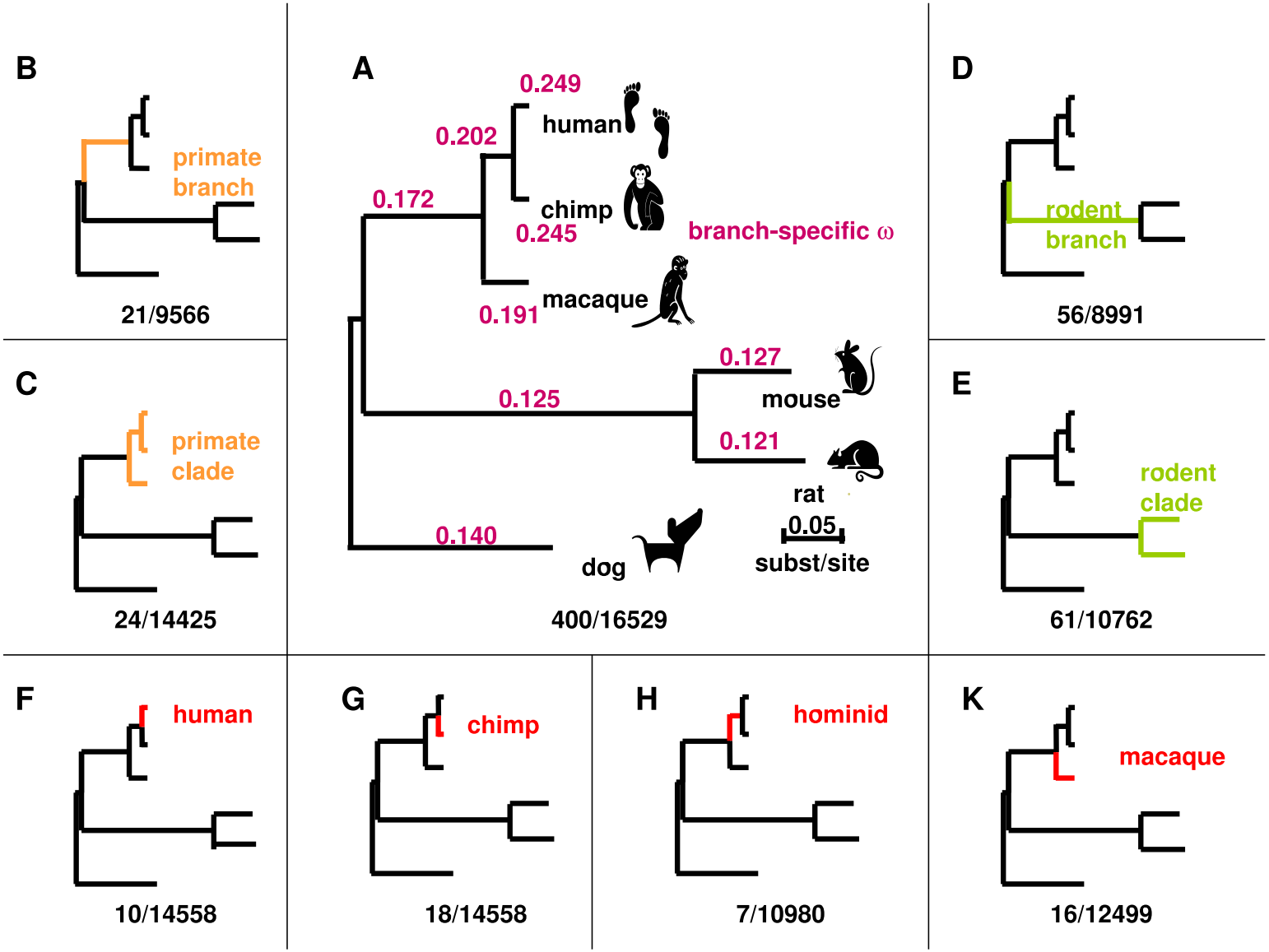
$$\Pr(H, C, M \,|\, \omega, t_H, t_C, t_M) =$$
$$\sum_A \Pr(A, H, C, M \,|\, \omega, t_H, t_C, t_M)$$



Likelihood ω:

$$L(\omega) = \max_{t_H, t_C, t_M} \Pr(H, C, M \,|\, \omega, t_H, t_C, t_M)$$

- This likelihood can be computed e.g. by PAML software
- There are also more complex models,
  e.g. with ω varying within a gene


B
primate branch
21/9566
A
0.249
human
0.202
chimp
0.172
0.245
branch-specific ω
macaque
0.191
0.127
mouse
0.125
0.121
rat
0.140
dog
0.05
subst/site
400/16529
D
rodent branch
56/8991
C
primate clade
24/14425
E
rodent clade
61/10762
F
human
10/14558
G
chimp
18/14558
H
hominid
7/10980
K
macaque
16/12499

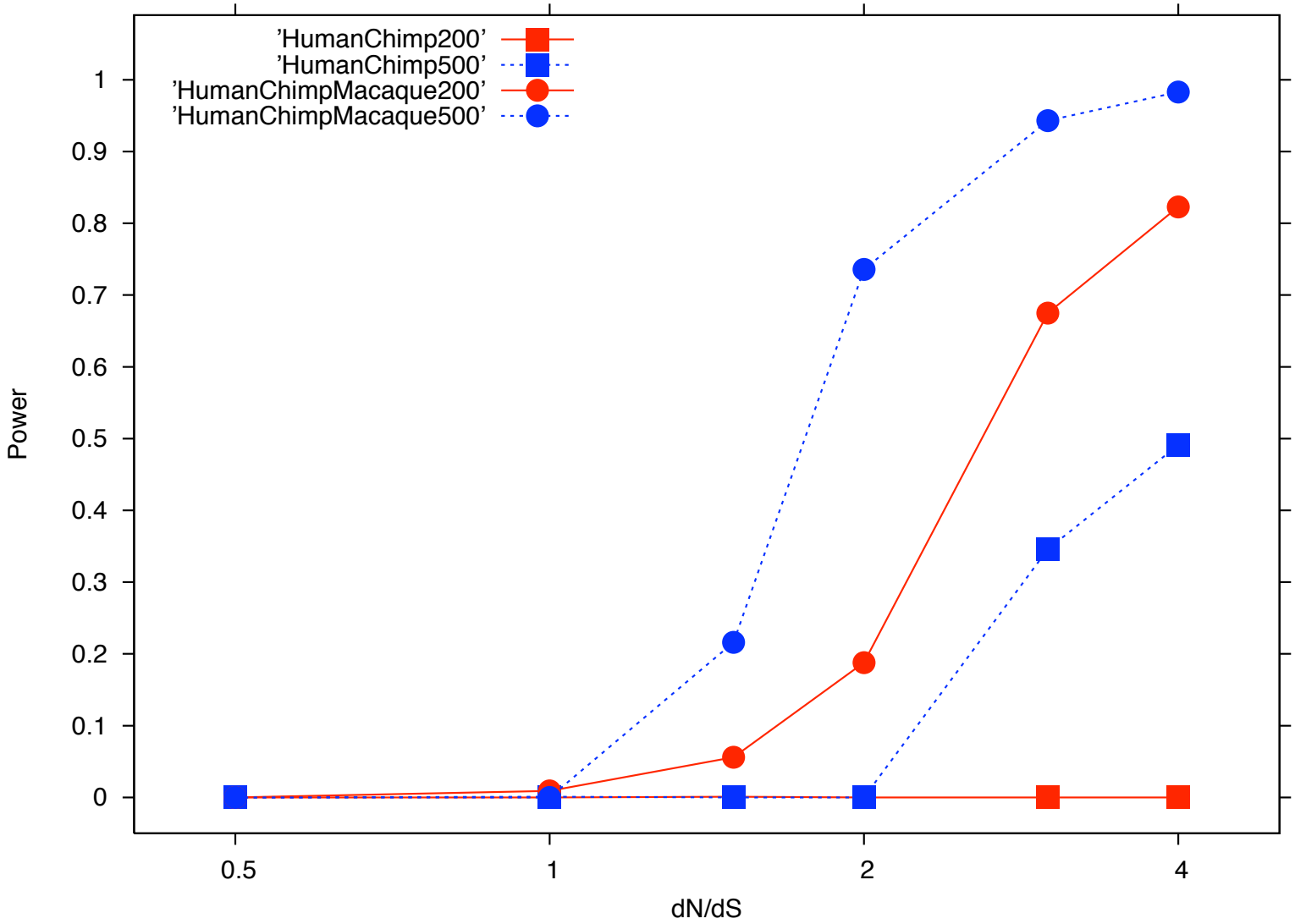
Functional categories enriched for positively selected genes

Defense: cellular defense response, antigen processing and presentation, response to virus, response to bacterium

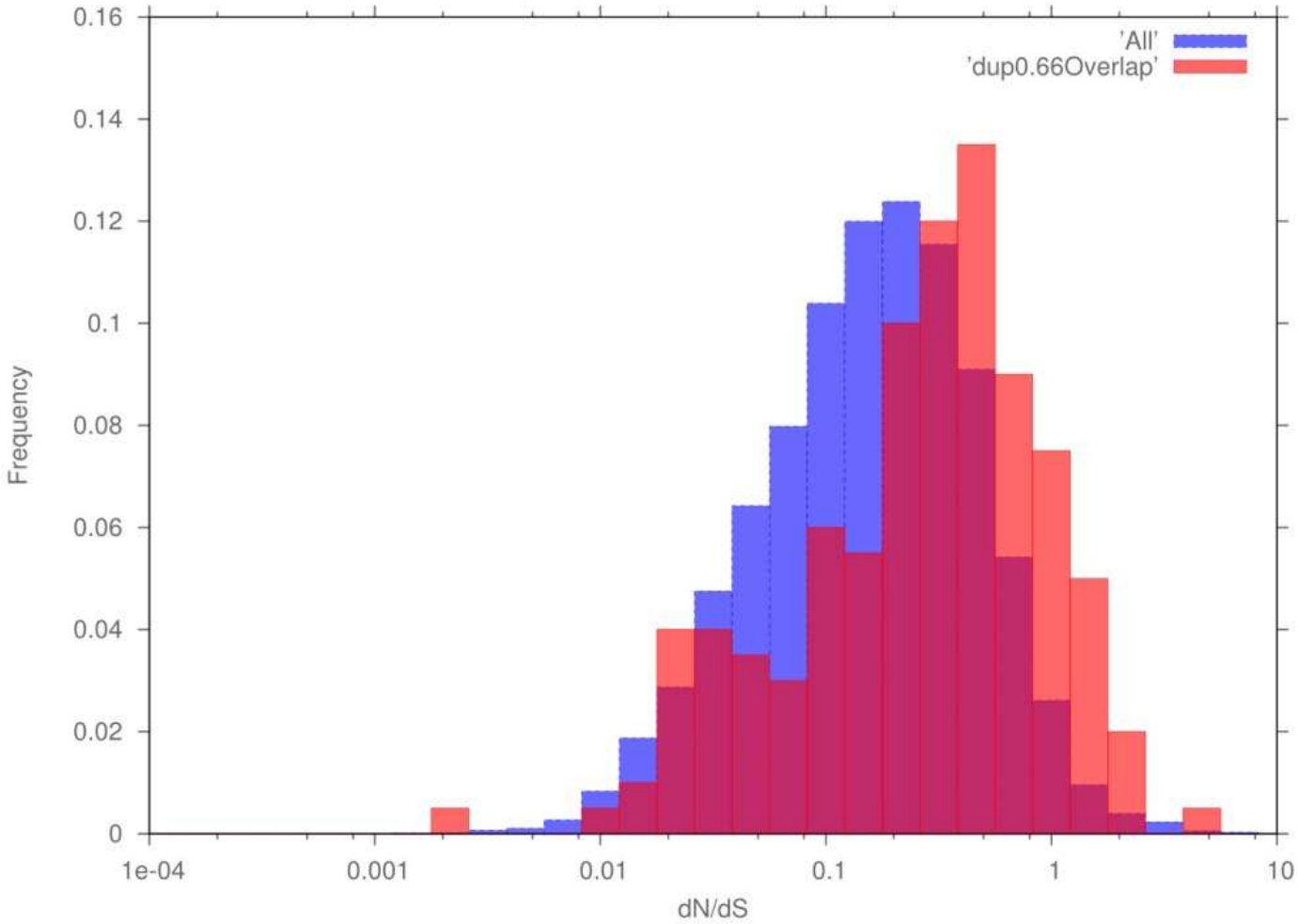
Immunity: adaptive immune response, adaptive immune response somatic recomb, lymphocyte mediated immunity, immunoglobulin mediated immune response, B cell mediated immunity, innate immune response, complement activation alternative pathway, regulation of immune system process, positive regulation of immune response, humoral immune response, complement activation classical pathway, humoral immune response circulating immunoglob, complement activation, activation of plasma proteins mute inflam resp, akute inflammatory response, response to wounding

Sensory perception: sensory perception of taste, G-protein coupled receptor protein signaling pathway, neurological process, sensory perception of chemical stimulus, sensory perception of smell

Adding genomes helps to improve power of the tests



Positive selection in duplicated genes



Summary

- Natural selection plays an important role in the evolution
- **Purifying selection:**
  - Conserved regions are likely to have some function
  - To find genes, we consider also typical codon mutations
- **Positive selection:**
  - Positive selection in genes causes high fraction of nonsynonymous changes (evolution at the protein level)
  - Duplicated genes are more often under positive selection
  - Hunt continues: we want to find genes causing human-specific features
- **Methods:** substitution models, phylogenetic HMMs, likelihood ratio tests