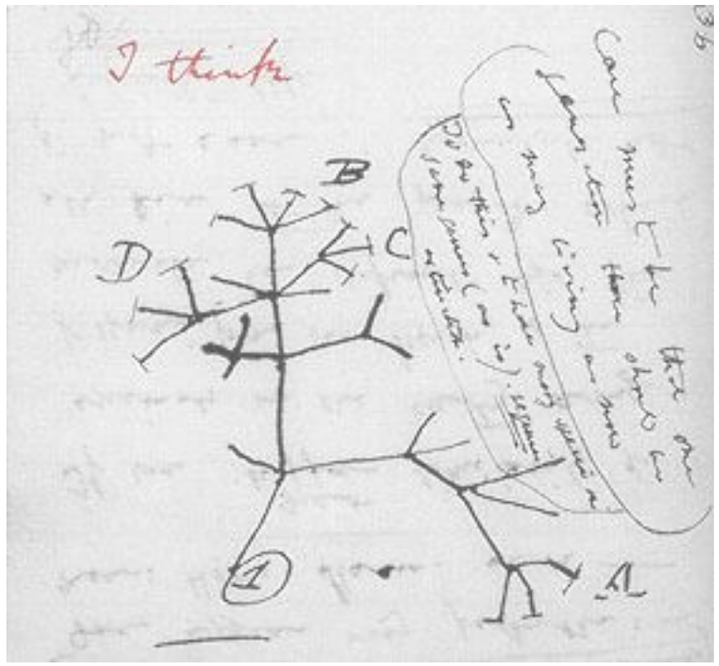
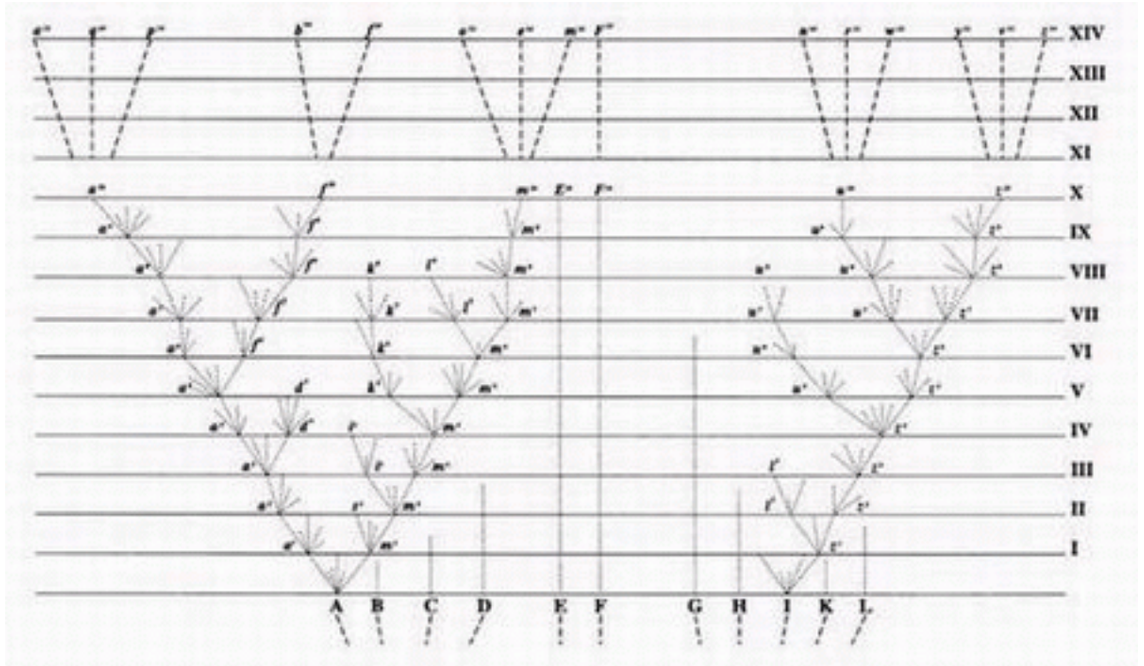
Evolutionary Ideas

Ryan Hernandez
Department Bioengineering and Therapeutic Sciences
Institute for Quantitative Biosciences (QB3)
Institute for Human Genetics
UCSF



Darwin's notebook, 1837



Darwin, *On the Origin of Species*, 1859

Is a species' genome size correlated with its "complexity"?

Species	Genome size (Mbp)
E. coli (bacteria)	5.4
S. cerevisiae (yeast)	12.1
A. thaliana (mustard weed)	115
D. melanogaster (fruit fly)	133
D. rerio (zebrafish)	1,688
H. sapiens (human)	3,272
P. aethiopicus (lungfish)	140,000
A. dubia (amoeba)	670,000

Frog genome sizes

- The ornate burrowing frog, Limnodynastes ornatus, has a genome several times **smaller** than the human genome (0.9Gb).

- The European fire-bellied toad, Bombina bombina, has a genome several times **larger** than the human genome (8Gb).

Number of genes across species

- **Hypothesis**: More "complex" organisms will have more genes in their genome

Species	# genes
E. coli (bacteria)	~4,200
S. cerevisiae (yeast)	~6,600
D. melanogaster (fruit fly)	~13,500
C. elegans (roundworm)	~20,000
A. thaliana (mustard weed)	~24,000
H. sapiens (human)	???

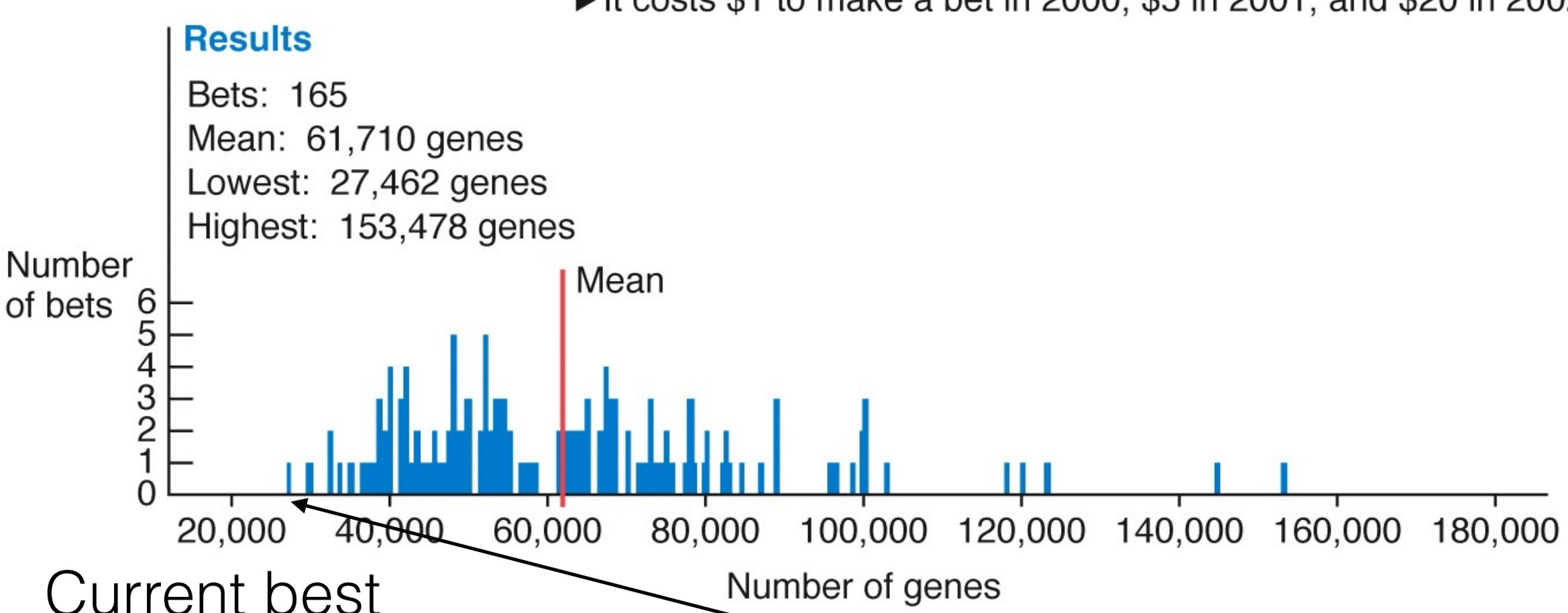
How many genes are there in the human genome?

Home▶ Genesweep http://www.ensembl.org/Genesweep/

Gene Sweepstakes

The Gene Sweepstakes will run between 2000 and 2003.

▶ It costs $1 to make a bet in 2000, $5 in 2001, and $20 in 2002.



Current best guess: ~22,000

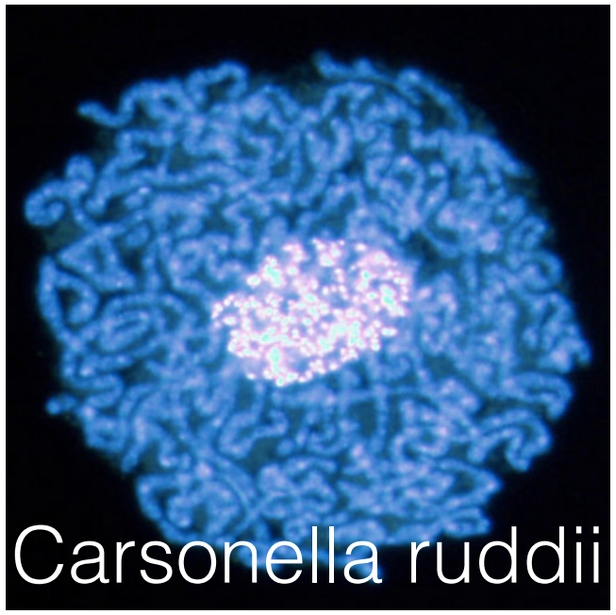
Minimum guess: 25,947

Which animal has the most genes?

- near-microscopic freshwater crustacean
- ~31,000 genes
- More than one-third of Daphnia's genes are undocumented in any other organism
- Genome size: ~200Mb.

Colbourne, *et al.*, The Ecoresponsive Genome of Daphnia pulex. *Science* (2011).

Which species has the fewest genes?



- Tiny bacteria
- 182 genes
- Genome size: ~160kb.

Nakabachi, *et al.*, The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science* (2006).

The Human Genome

- What does the 3.3 billion base pair human genome look like?
  - **Coding sequences** – 22,000 genes make up ~1.2% of the total sequence
  - **Regulatory sequences** – Make up < 5% of the total sequence
  - Much of our genome consists of DNA with **no known function!**
    - But don't call it "junk". Let's just say its complicated!

Human Protein Coding Exome

- **Exome**: ~22,000 genes
- **Gene**: average ~10 exons
- **Exon**: average ~165bp
- **Intron**: average ~2,700bp
- **Total genomic region**: ~50kb!

Phylogenetics

1. Select a sequence of interest (gene, regulatory region, transposable element, or even a whole genome).

2. Identify **homologs**:

   - Objects that derive from a common ancestor.

   - **Orthologs**: thru speciation; **paralogs** thru duplication.

3. Align sequences.

4. Calculate phylogeny.

} Not necessarily independent!

5. Determine confidence

CREATE PHYLOGENY

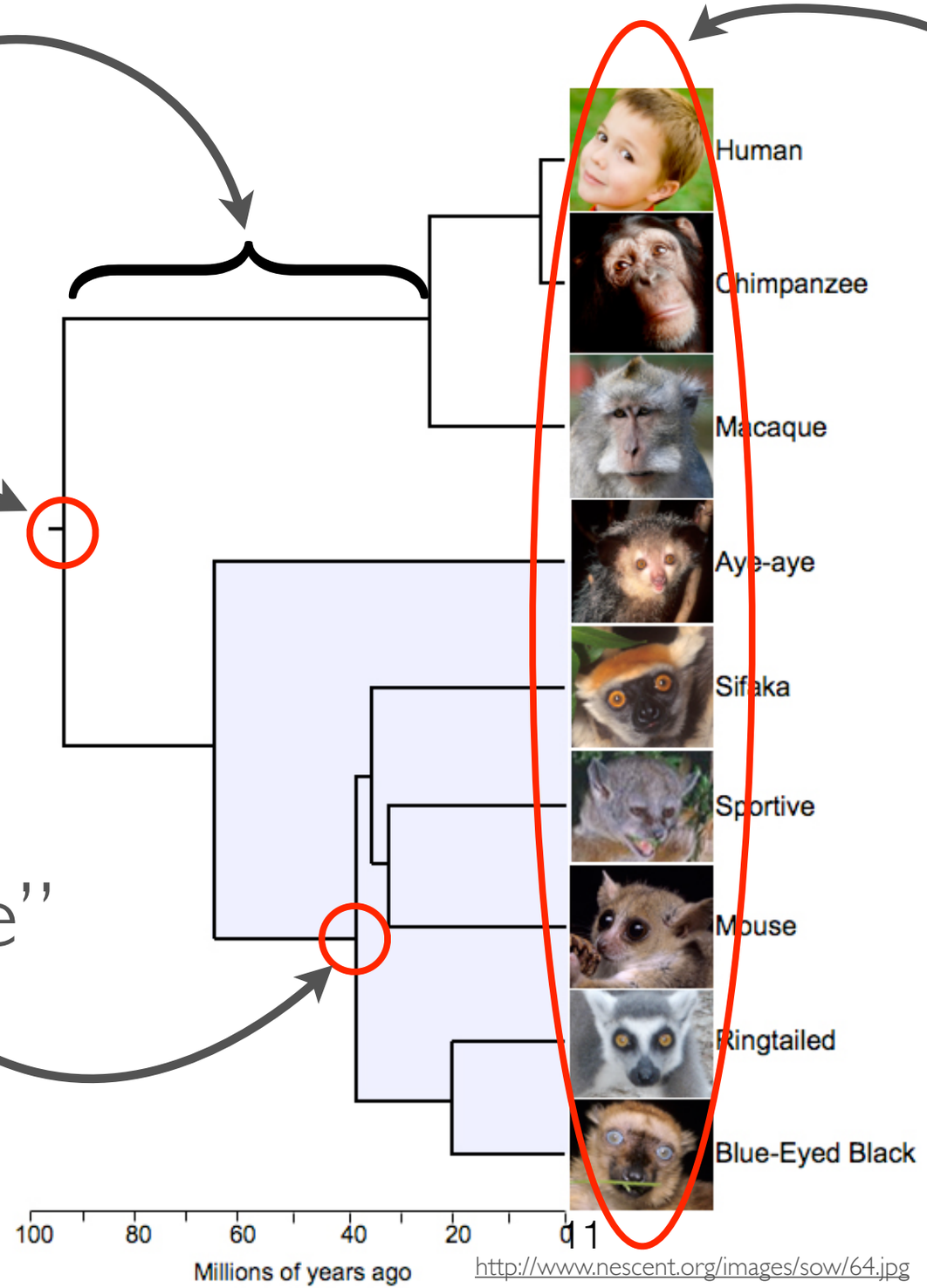
We will focus on bifurcating trees.

"Lineage"
"branch"

"External nodes"
"Leaf nodes"
"Taxa"

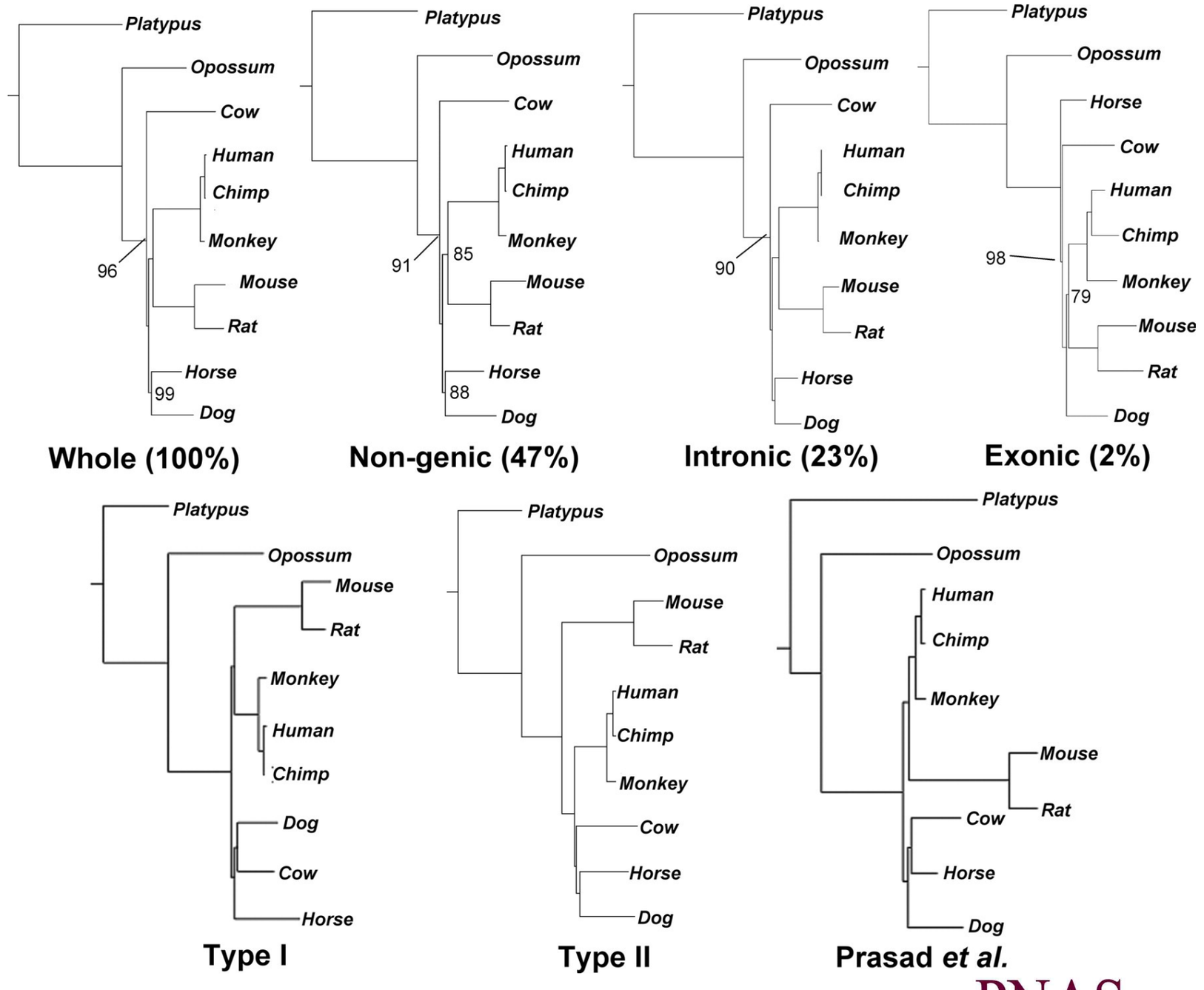
"Root"

"Internal node"



http://www.nescent.org/images/sow/64.jpg

Phylogenetic trees from different parts of the genome



Sims G E et al. PNAS 2009;106:17077-17082

The Neutral Theory

- Forty years ago, Kimura (1968) and King and Jukes (1969) proposed that most new mutations are neutral (or lethal) and that most genetic variation is of no functional relevance.

- Though highly controversial at the time, the neutral theory is now regarded as a good approximation of the truth for most species.

Molecular Clock

- At sites unaffected by natural selection, divergence accumulates at a roughly constant rate.

- We can use orthologous sequence data across different species to estimate the time when the species split from each other.

- This is the basis for the field of phylogenetics, which seeks to understand the historical and evolutionary relationships of all species.

Phylogenetics

- There are several widely used methods for constructing phylogenetic trees:
  - Parsimony-based methods
  - Heuristic methods (e.g., Neighbor-Joining)
  - Maximum-likelihood based methods
  - Bayesian Methods

Phylogenetics complications

- Some methods cannot handle the large-scale data sets that are now commonplace
- Mutation rates do change over time
- Multiple mutations at the same nucleotide site can obscure evolutionary relationships
- Analyses of different parts of the genome can lead to different phylogenetic trees
- Horizontal gene transfer (in bacteria) violate the basic assumptions of phylogenetics

Natural Selection

- Genomic approaches to looking for natural selection:
  - Codon based models (comparison of orthologous sites across many species)
  - Identification of function through conservation

The Effect of "Positive Selection"

Adaptive

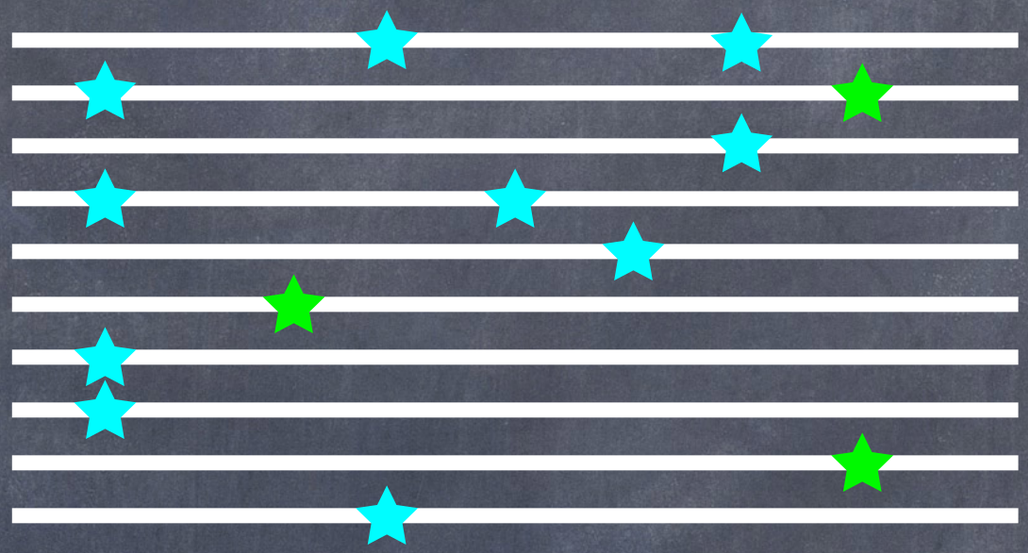
Neutral

Nearly Neutral

Mildly Deleterious

Fairly Deleterious

Strongly Deleterious



The Effect of "Positive Selection"

Adaptive

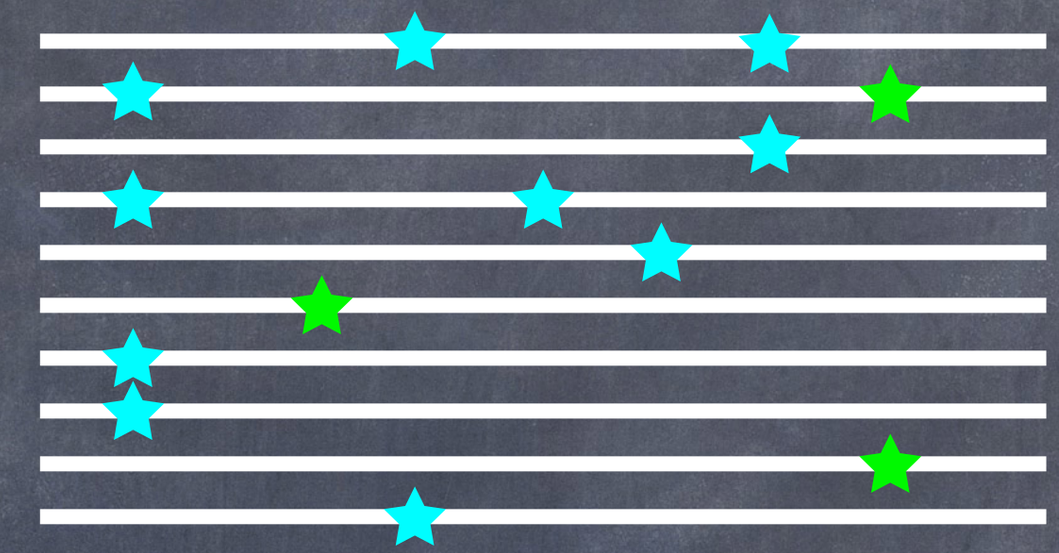
Neutral

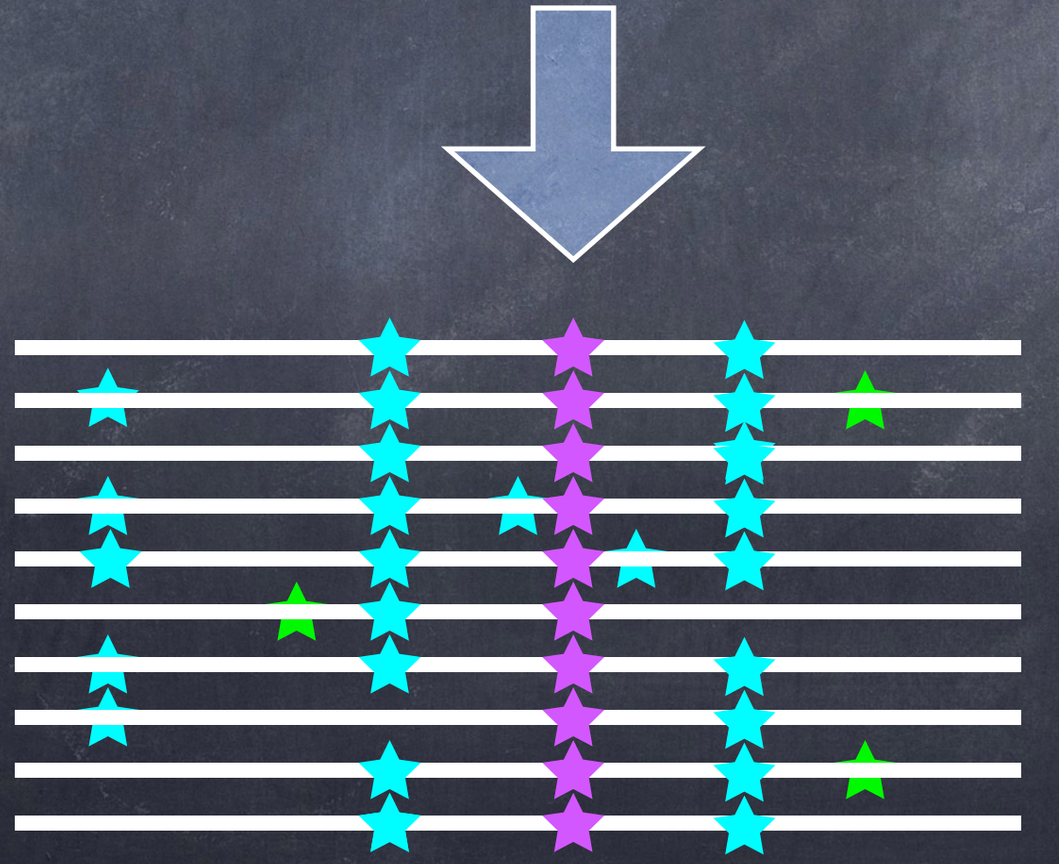
Nearly Neutral

Mildly Deleterious

Fairly Deleterious

Strongly Deleterious

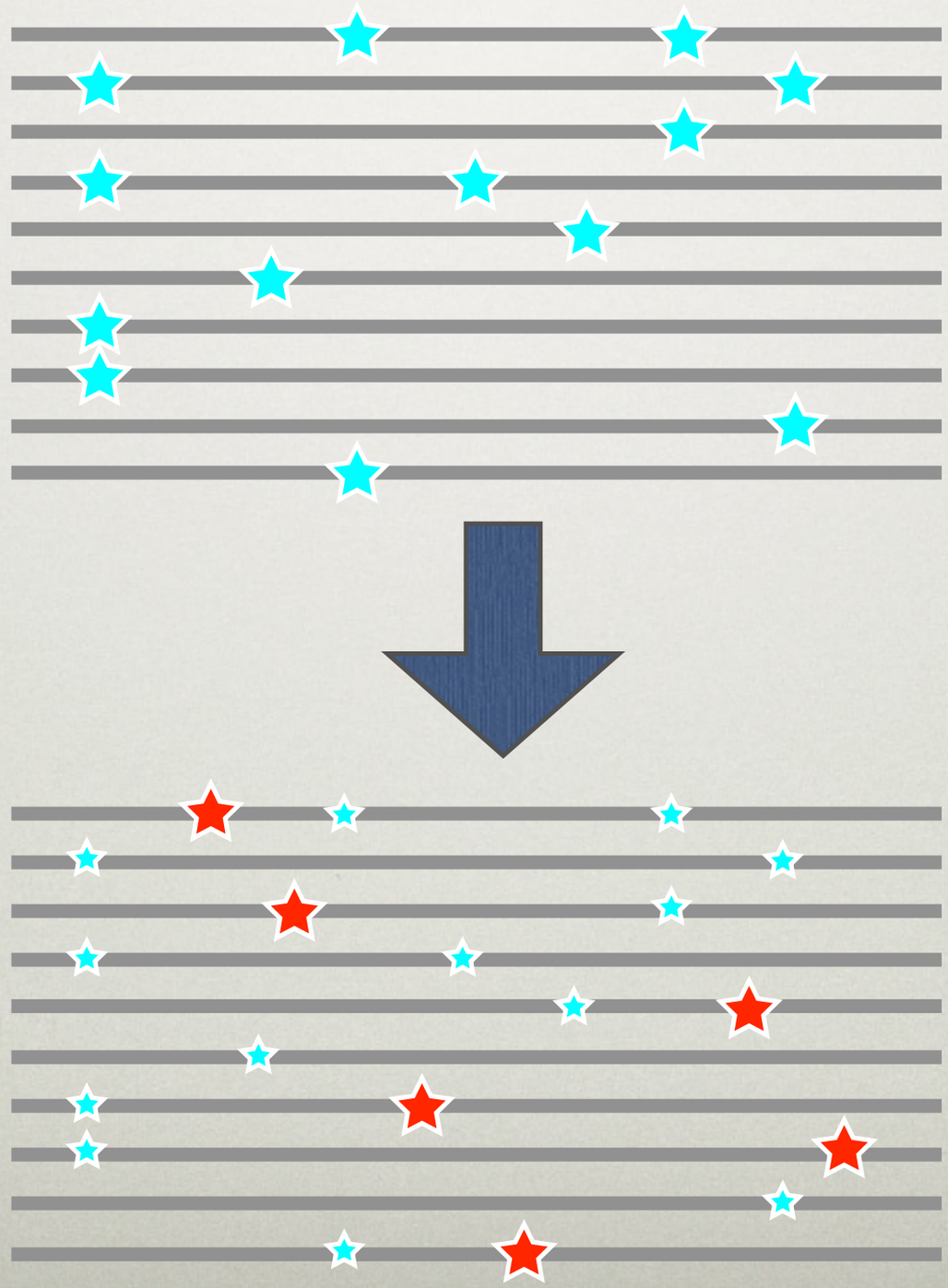




"Selective Sweep"

The Effect of Negative Selection

Chromosomes in a population with standing variation



Deleterious mutations will arise in the next generation

Negative selection: the action of natural selection purging deleterious mutations.

Codon-based models

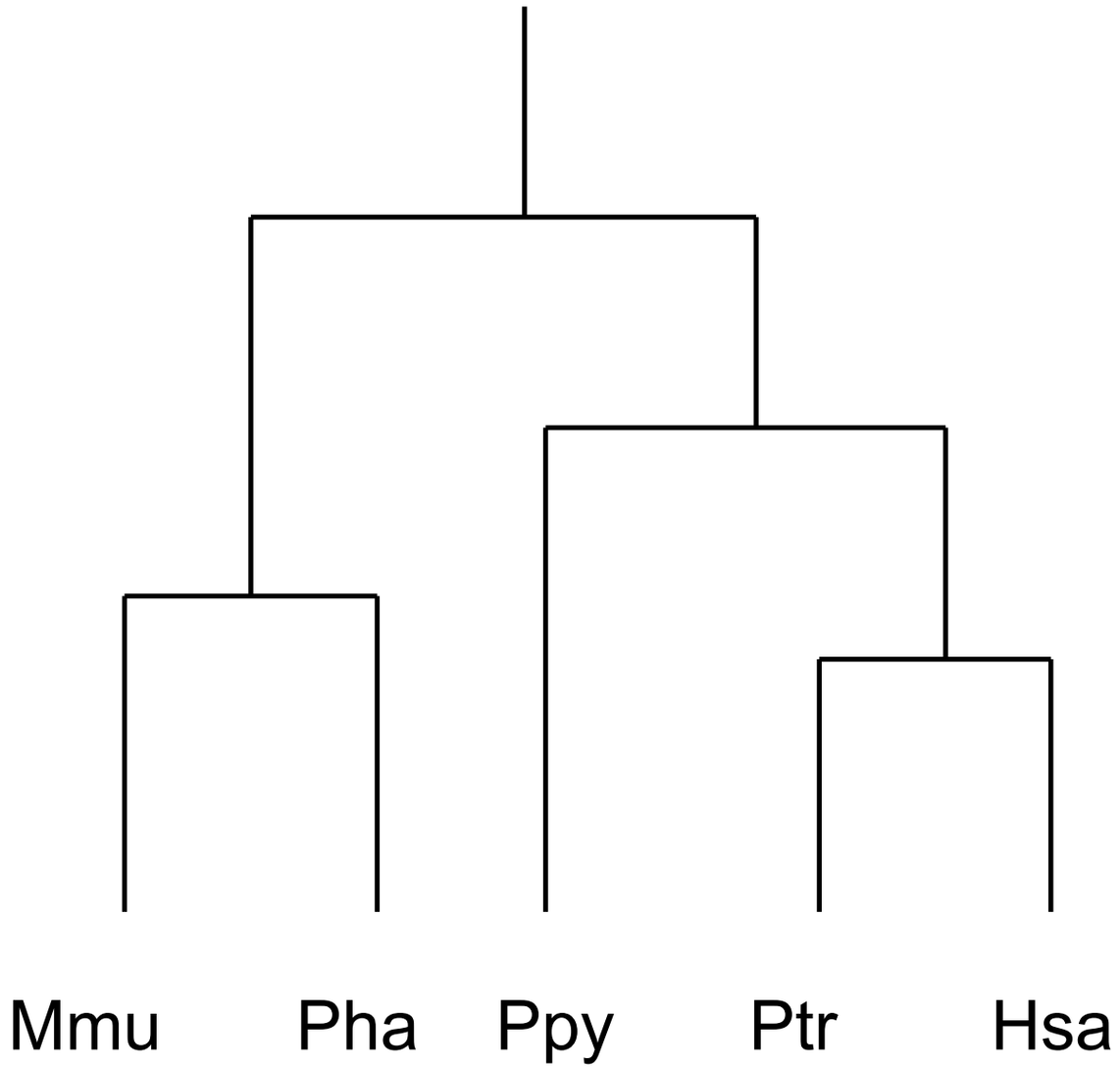
(e.g., Goldman & Yang 1994; Nielsen & Yang 1998)

- Suppose one has sequence data from multiple (>> 2) species from a single locus.

- How can one use these data to infer the specific sites that have been subject to natural selection?

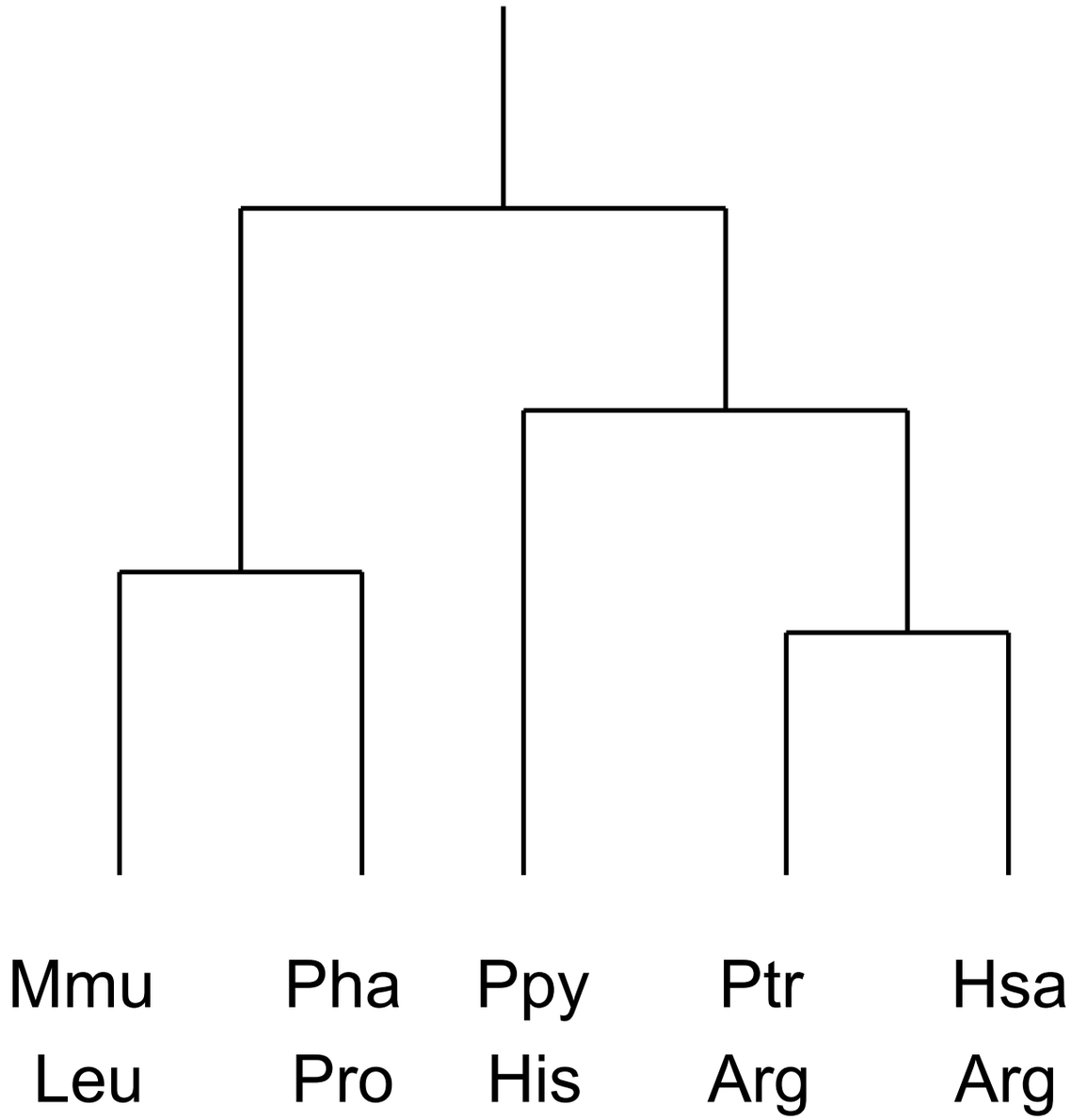
- Repeated fixation of functional mutations in coding regions over evolutionary timescales can lead to a disproportional number of amino acid substitutions relative to silent substitutions (synonymous).

Hypothetical example



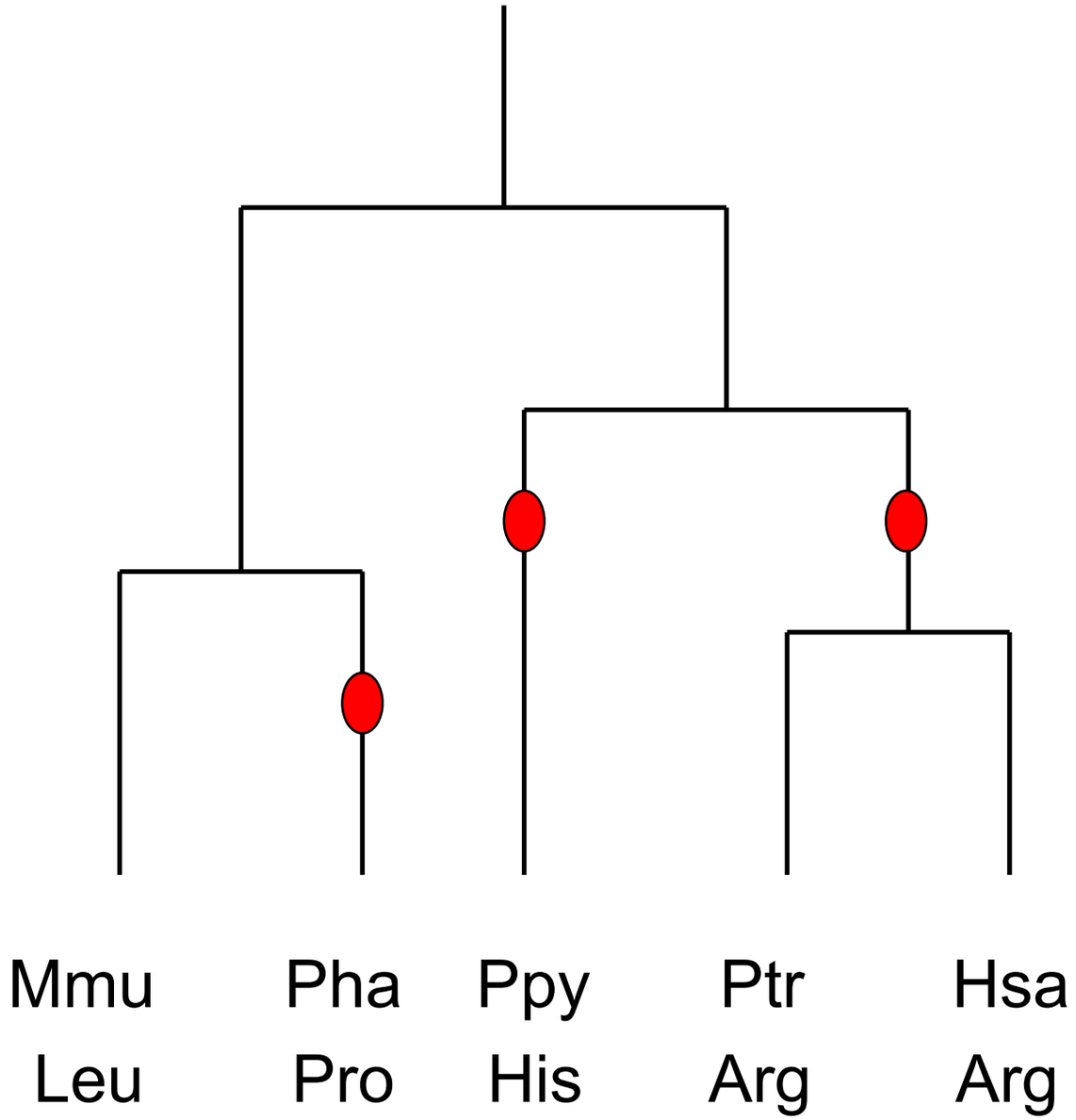
- The exons of a single gene are sequenced in 5 species:
  - macaque, baboon, orang, chimp and human.
- Between each pair of species, there is at most one non-synonymous change per site.

Hypothetical example



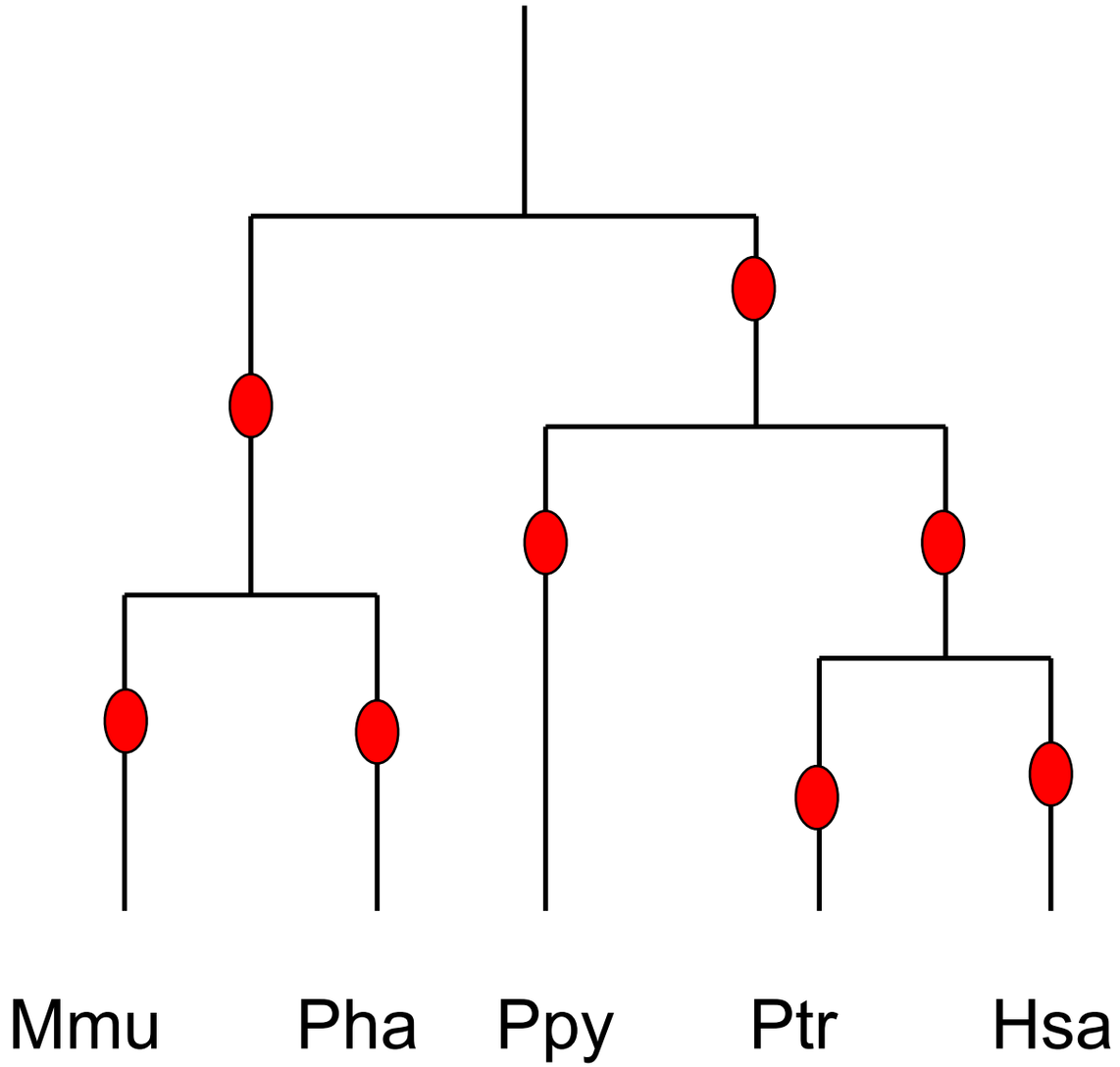
- Suppose at one codon, we observe the following amino acids.

Hypothetical example



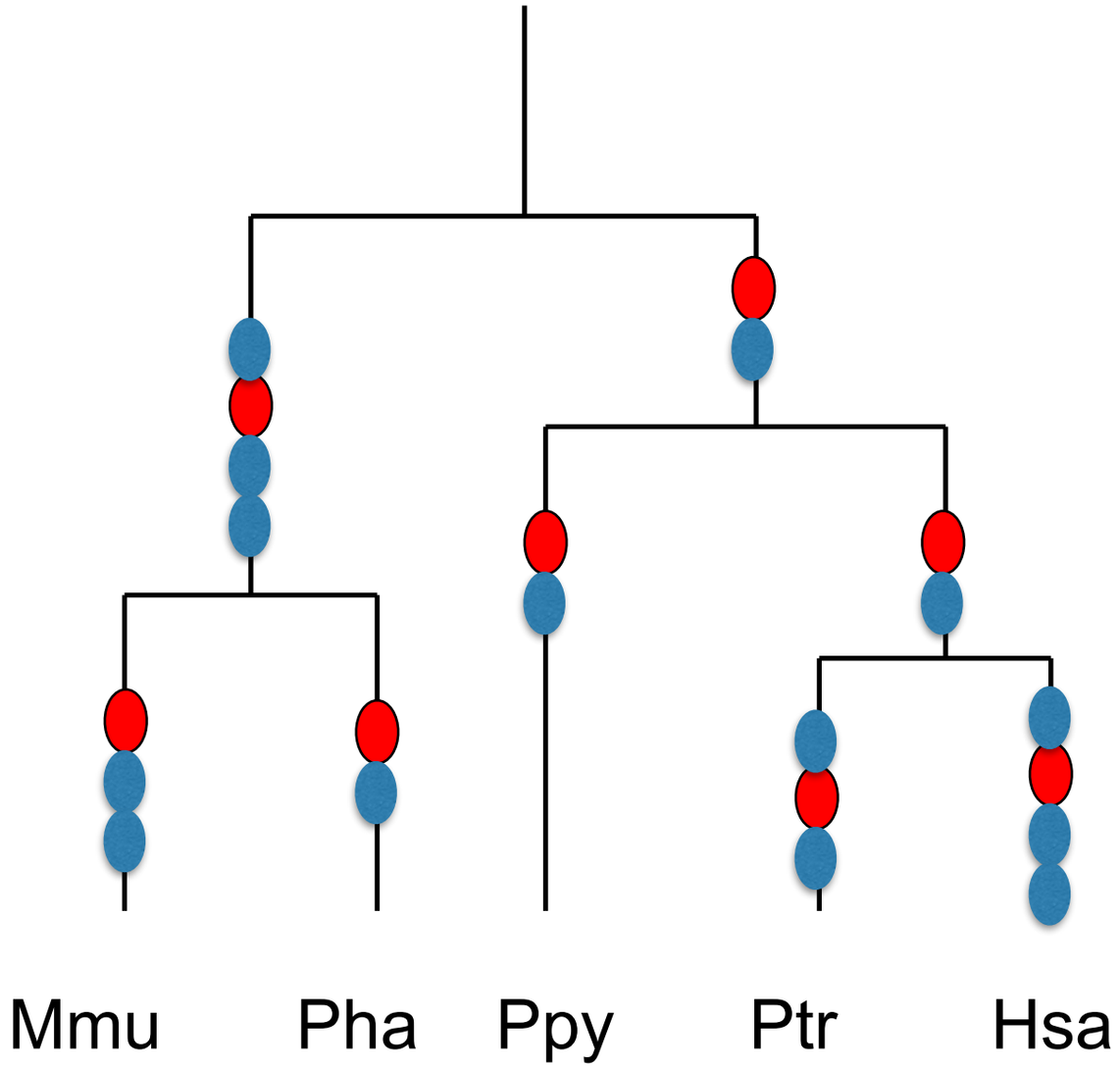
- Parsimony: 3 changes happen at this position, leading to four different amino-acid residues.
- Three (or more) non-synonymous changes at the same codon may be unlikely to have happened by chance.

Hypothetical example



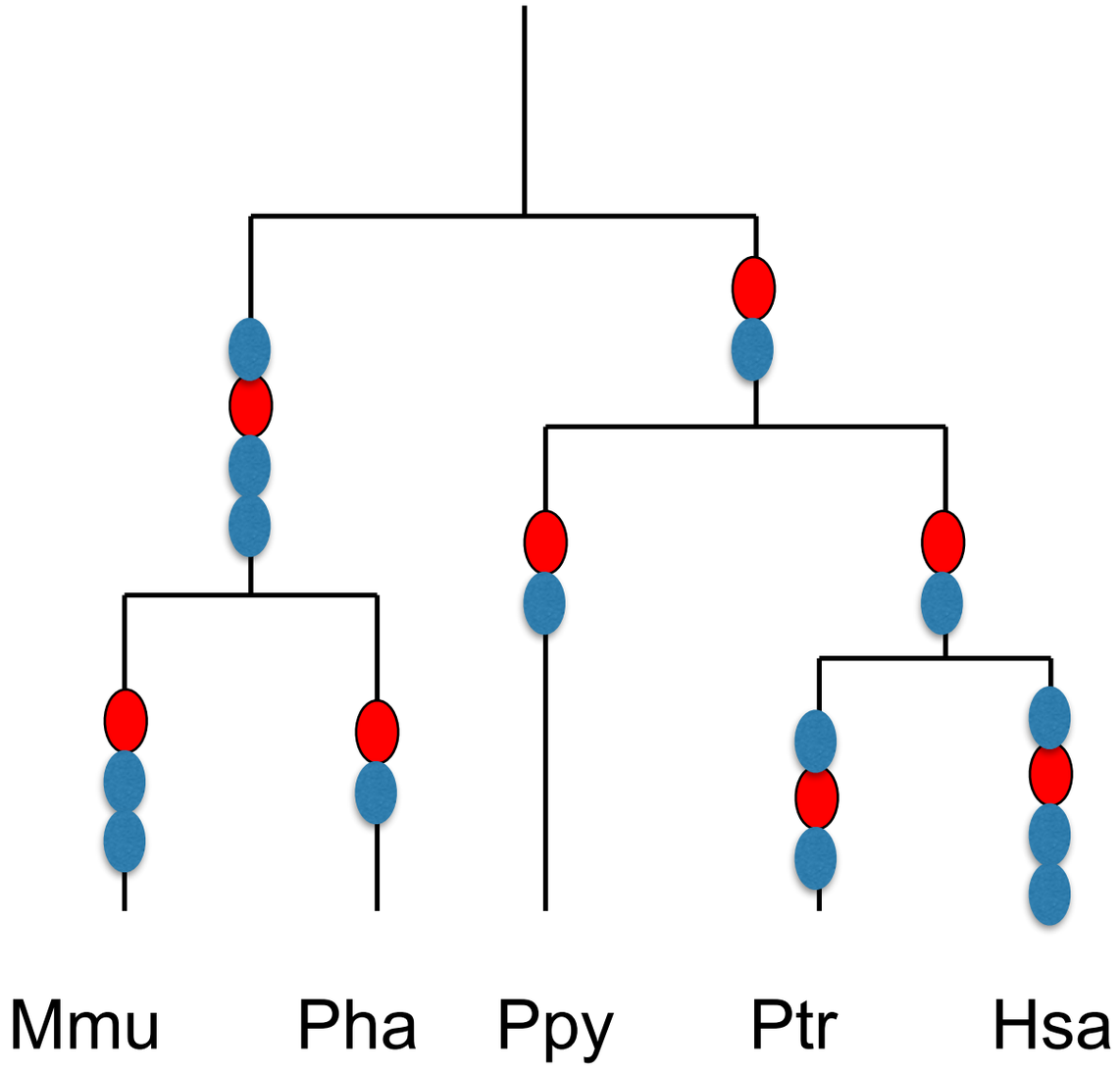
- Overall sites, suppose we observe 8 amino acid substitutions.
- How unlikely is this observation?

Hypothetical example



- What if we observed 14 synonymous substitutions?

Hypothetical example



	NS	SYN
Subst	8	14
Sites	300	100

dN=0.027 **dS=0.14**

- ω=dN/dS
  = 0.191
- Fisher Exact Test
  p=0.00019

Definitions

- Define ω as the ratio of the non-synonymous and the synonymous substitution rates: ω=dN/dS.

- Then:
  - ω = 0 → complete constraint
  - ω < 1 → selective constraint
  - ω = 1 → neutrality
  - ω > 1 → selectively advantageous

A Caveat

- Anisimova et al. (2003) looked at the effect of recombination on codon-based likelihood ratio tests. Recombination causes different codons to have different topologies and branch lengths (especially if closely related species are studied).

- They found that with high recombination rates the type I error rate can be as high as 90 %.

Incomplete Lineage Sorting

- Hobolth, et al., PLoS Gen (2007):
    - The genealogical relationship of human, chimpanzee, and gorilla varies along the genome.

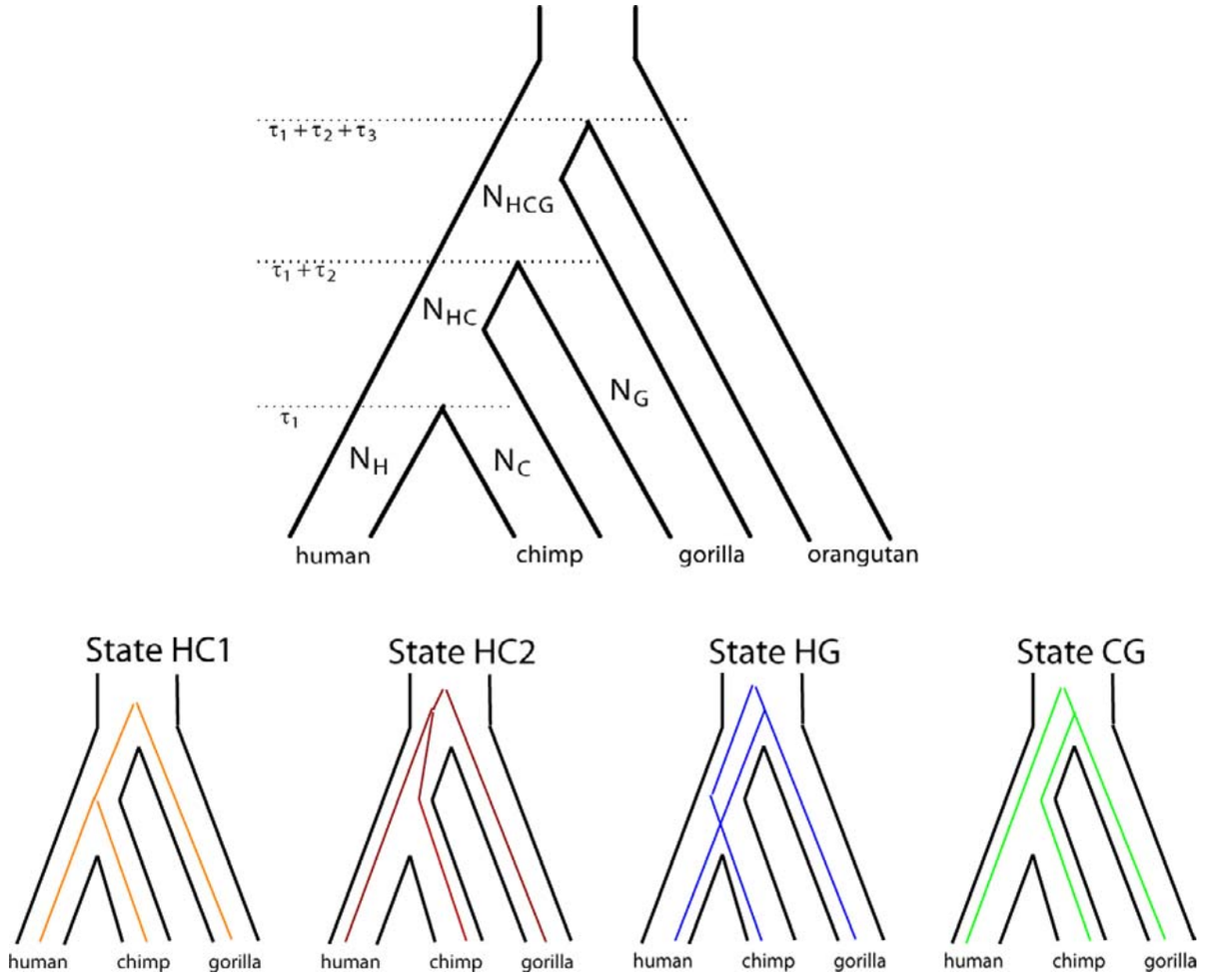
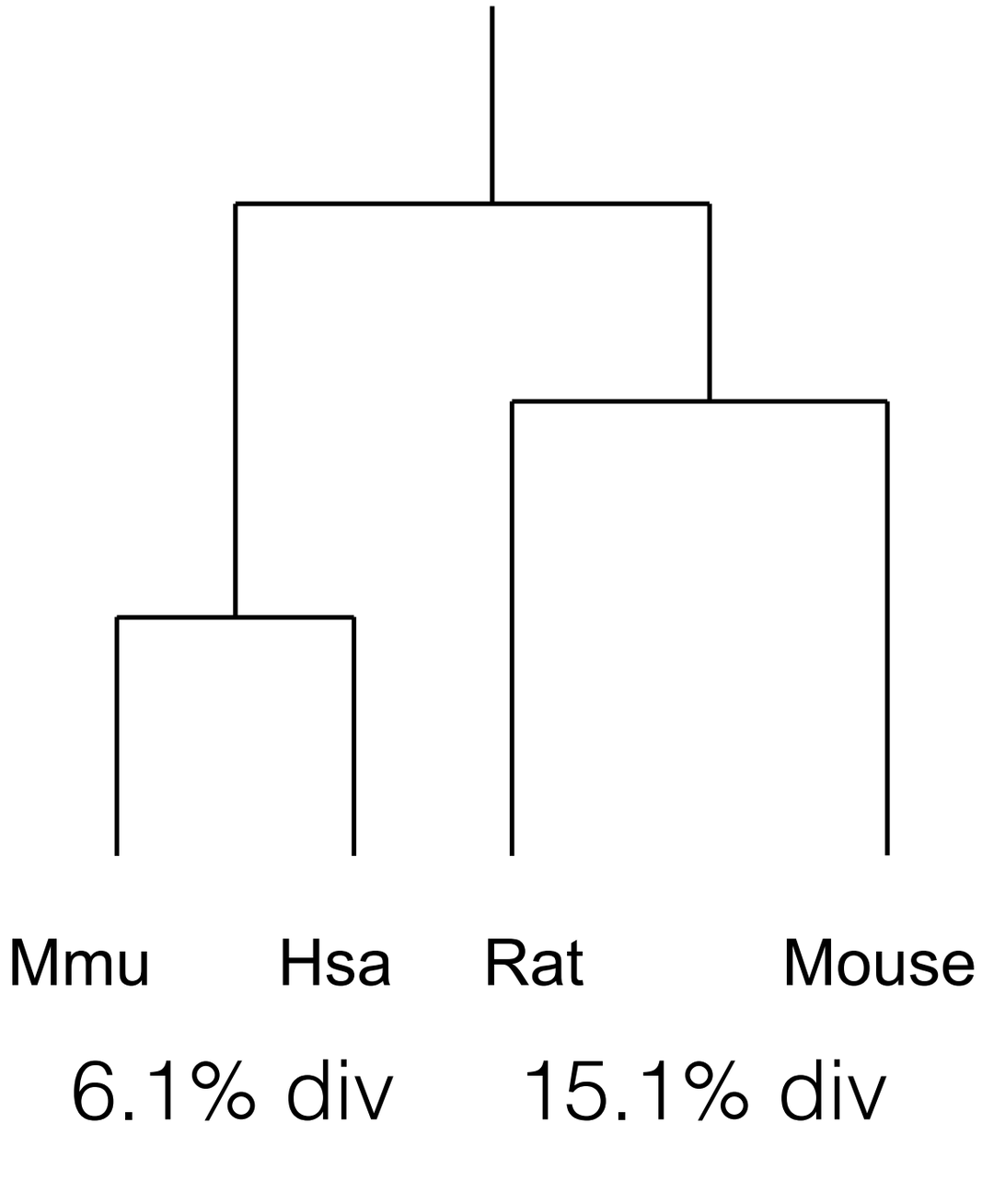
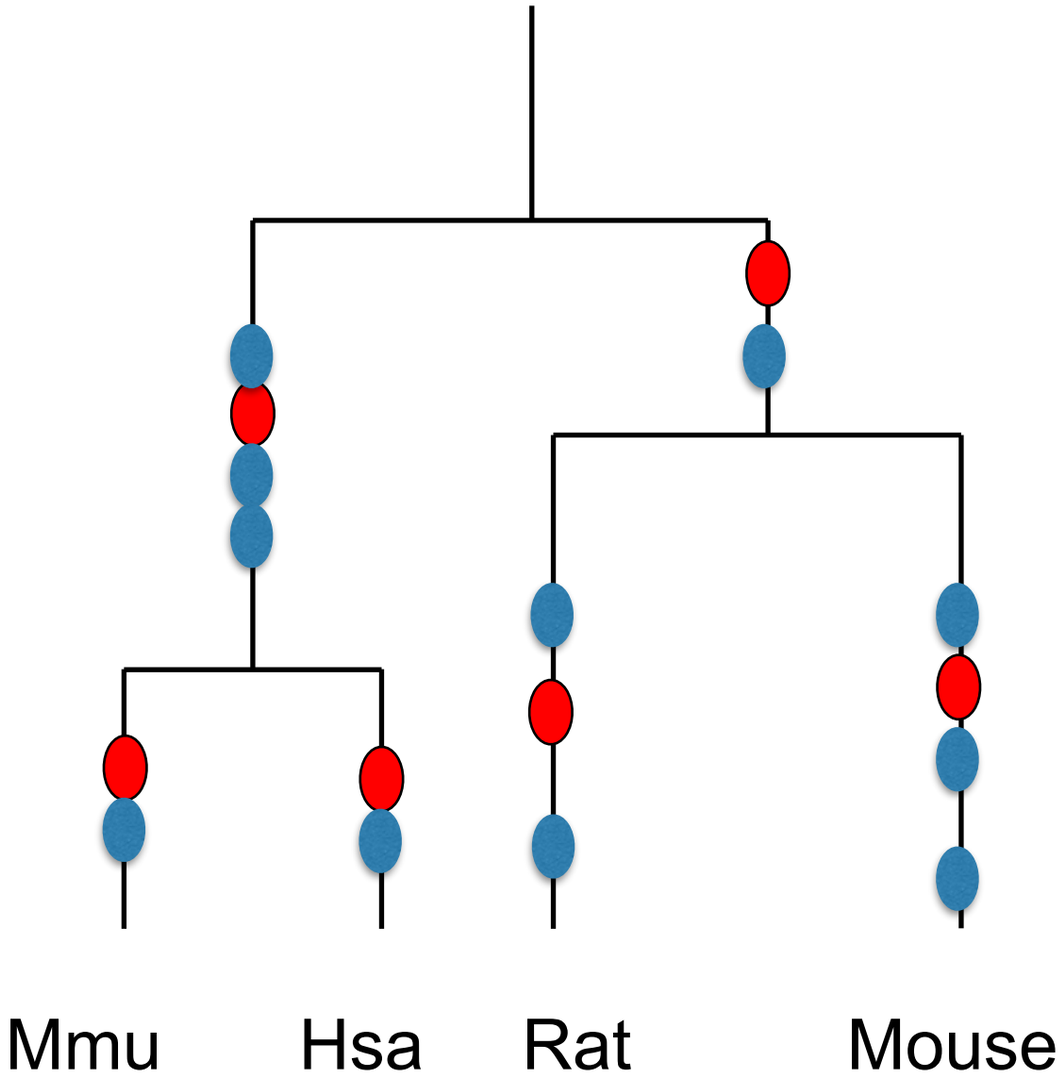


Figure 1. Genetic and Species Relationships May Differ

LYG1: an anti-bacterial enzyme



LYG1: an anti-bacterial enzyme



	NS	SYN
Primates	16	5
Rodents	0	25

- Fisher Exact Test $p=2.05\times10^{-8}$

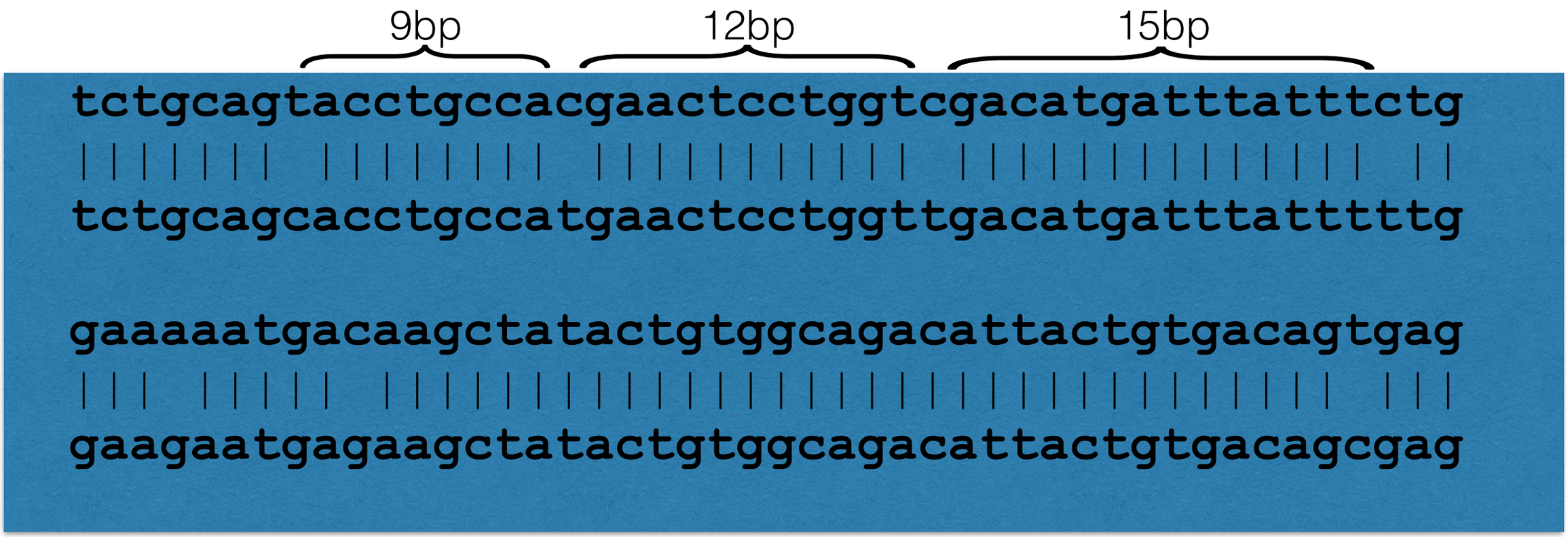
Natural Selection Revisited

- Genomic approaches to looking for natural selection:
  - Codon based models (comparison of orthologous sites across many species)
  - Identification of function through conservation

Selective Constraint

- Comparison of the genomes of evolutionarily distant species has helped identify:
  - Novel genes
  - Intron/exon boundaries
  - cis- and trans-acting regulatory elements
  - Conserved sequences with unknown function

Random pig-human alignment



- Signs that the aligned sequence is from an exon:
    - Overall level of sequence identity is higher than average
    - Distances between fixed differences is often a multiple of 3

Evolutionary Conservation as a Tool for Interpretation

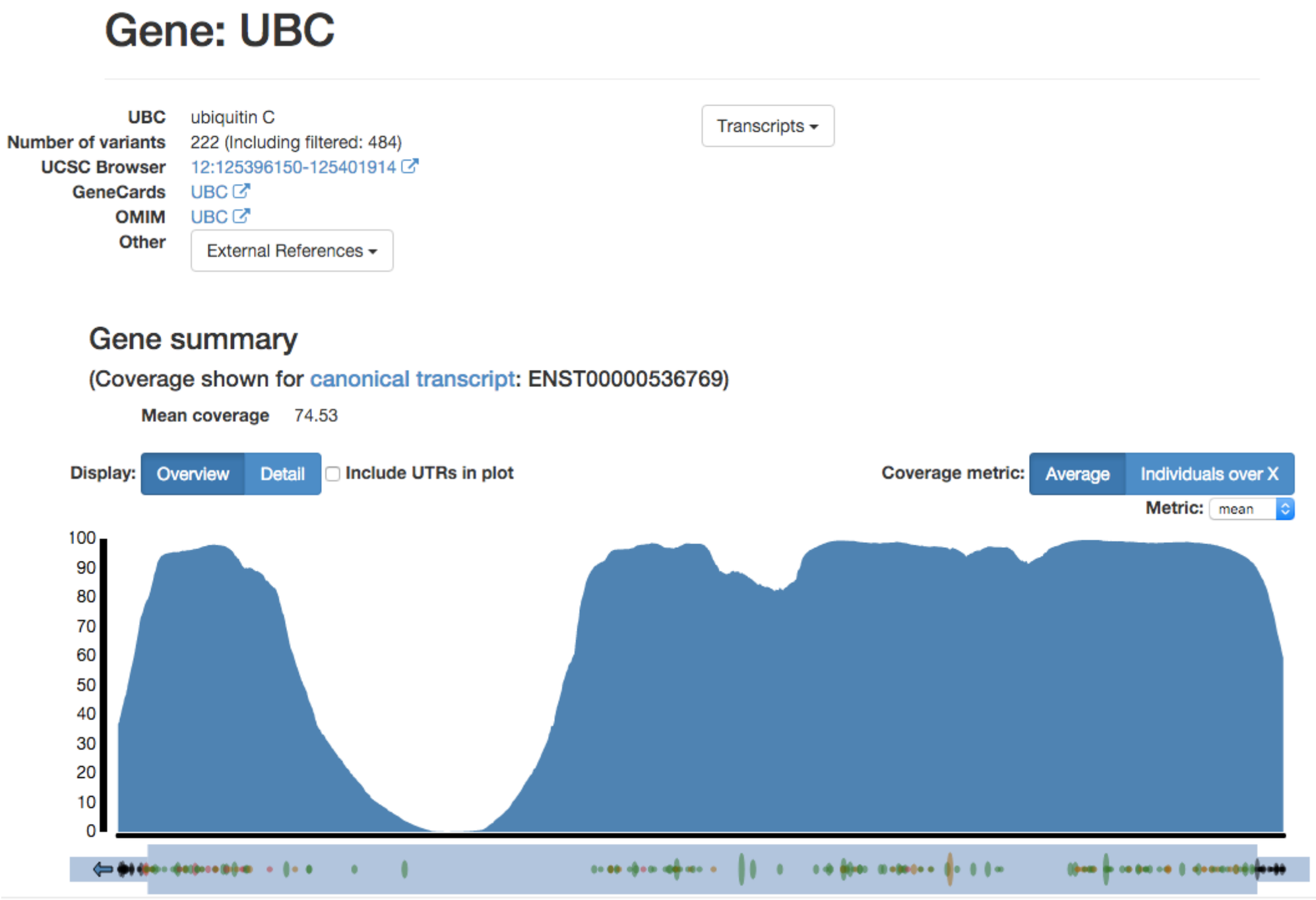
- Exome Aggregation Consortium released variant lists and frequencies from **60,706 exomes**!

Analysis of protein-coding genetic variation in 60,706 humans

Exome Aggregation Consortium[#], Monkol Lek[1,2,3,4], Konrad J Karczewski[1,2*], Eric V Minikel[1,2,5*], Kaitlin E Samocha[1,2,6,5*], Eric Banks[2], Timothy Fennell[2], Anne H O'Donnell-Luria[1,2,7], James S Ware[2,8,9,10,11], Andrew J Hill[1,2,12], Beryl B Cummings[1,2,5], Taru Tukiainen[1,2], Daniel P Birnbaum[2], Jack A Kosmicki[1,2,6,13], Laramie Duncan[1,2], Karol Estrada[1,2], Fengmei Zhao[1,2], James Zou[2], Emma Pierce-Hoffman[1,2], David N Cooper[14], Mark DePristo[15], Ron Do[16,17,18,19], Jason Flannick[2,20], Menachem Fromer[1,6,21,16,17], Laura Gauthier[15], Jackie Goldstein[1,2,6], Namrata Gupta[2], Daniel Howrigan[1,2,6], Adam Kiezun[15], Mitja I Kurki[2,22], Ami Levy Moonshine[15], Pradeep Natarajan[2,23,24,25], Lorena Orozco[26], Gina M Peloso[2,24,25], Ryan Poplin[15], Manuel A Rivas[2], Valentin Ruano-Rubio[15], Douglas M Ruderfer[21,16,17], Khalid Shakir[15], Peter D Stenson[14], Christine Stevens[2], Brett P Thomas[1,2], Grace Tiao[15], Maria T Tusie-Luna[27], Ben Weisburd[2], Hong-Hee Won[2,23,24,25], Dongmei Yu[22,28], David M Altshuler[2,29], Diego Ardissino[30], Michael Boehnke[31], John Danesh[32], Roberto Elosua[33], Jose C Florez[2,23,24], Stacey B Gabriel[2], Gad Getz[15,23,34], Christina M Hultman[35], Sekar Kathiresan[2,23,24,25], Markku Laakso[36], Steven McCarroll[6,8], Mark I McCarthy[37,38,39], Dermot McGovern[40], Ruth McPherson[41], Benjamin M Neale[1,2,6], Aarno Palotie[42], Shaun M Purcell[21,16,17], Danish Saleheen[43,44,45], Jeremiah Scharf[22,28], Pamela Sklar[21,16,17,46,47], Patrick F Sullivan[48,49], Jaakko Tuomilehto[50], Hugh C Watkins[51], James G Wilson[52], Mark J Daly[1,2,6], Daniel G MacArthur[1,2†]

Evolutionary Conservation as a Tool

- Exome Aggregation Consortium released variant lists and frequencies from **60,706 exomes**!

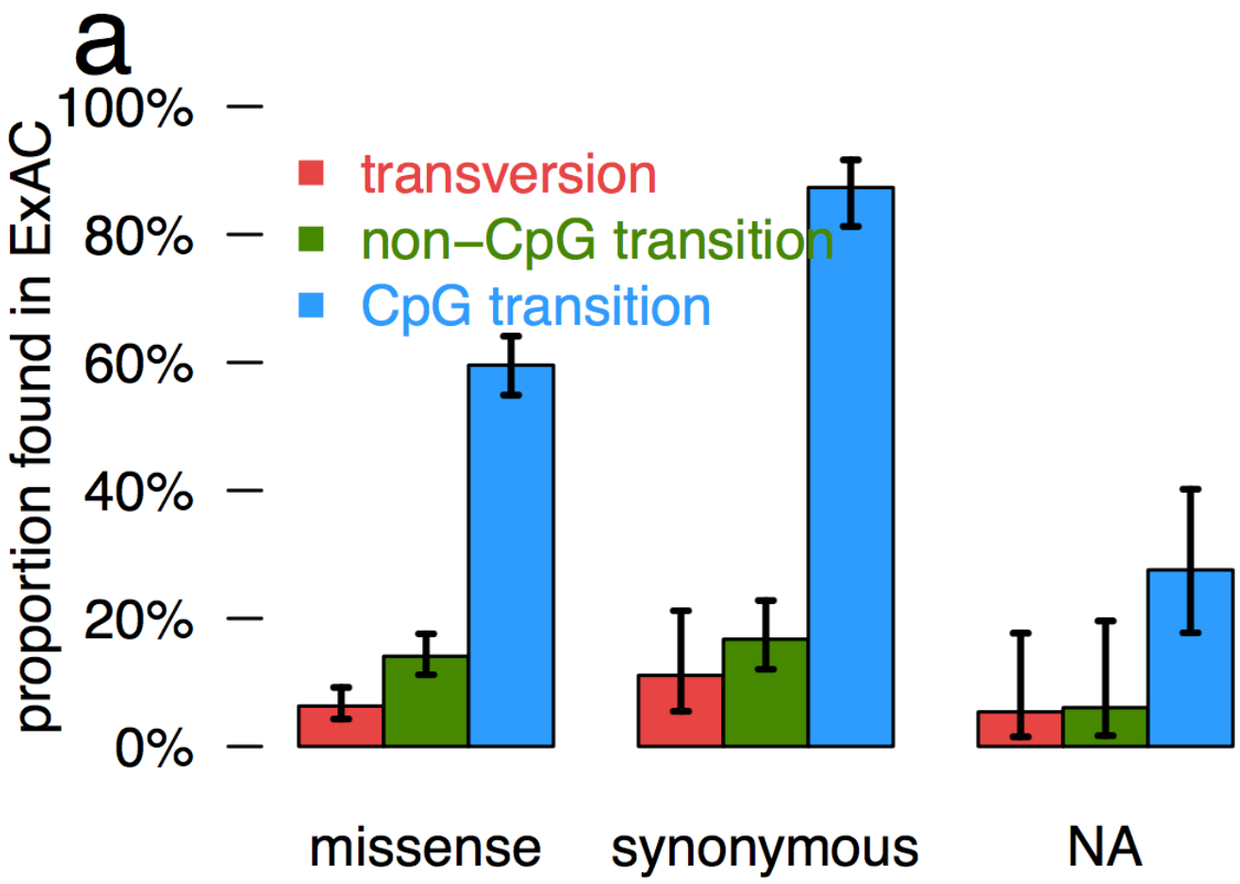


Evolutionary Conservation as a Tool

- Exome Aggregation Consortium released variant lists and frequencies from **60,706 exomes**!

- 1 variant ~8bp!

- Variant observed at 60-90% of all CpG sites in exome!



Evolutionary Conservation as a Tool for Interpretation

- Exome Aggregation Consortium released variant lists and frequencies from **60,706 exomes**!

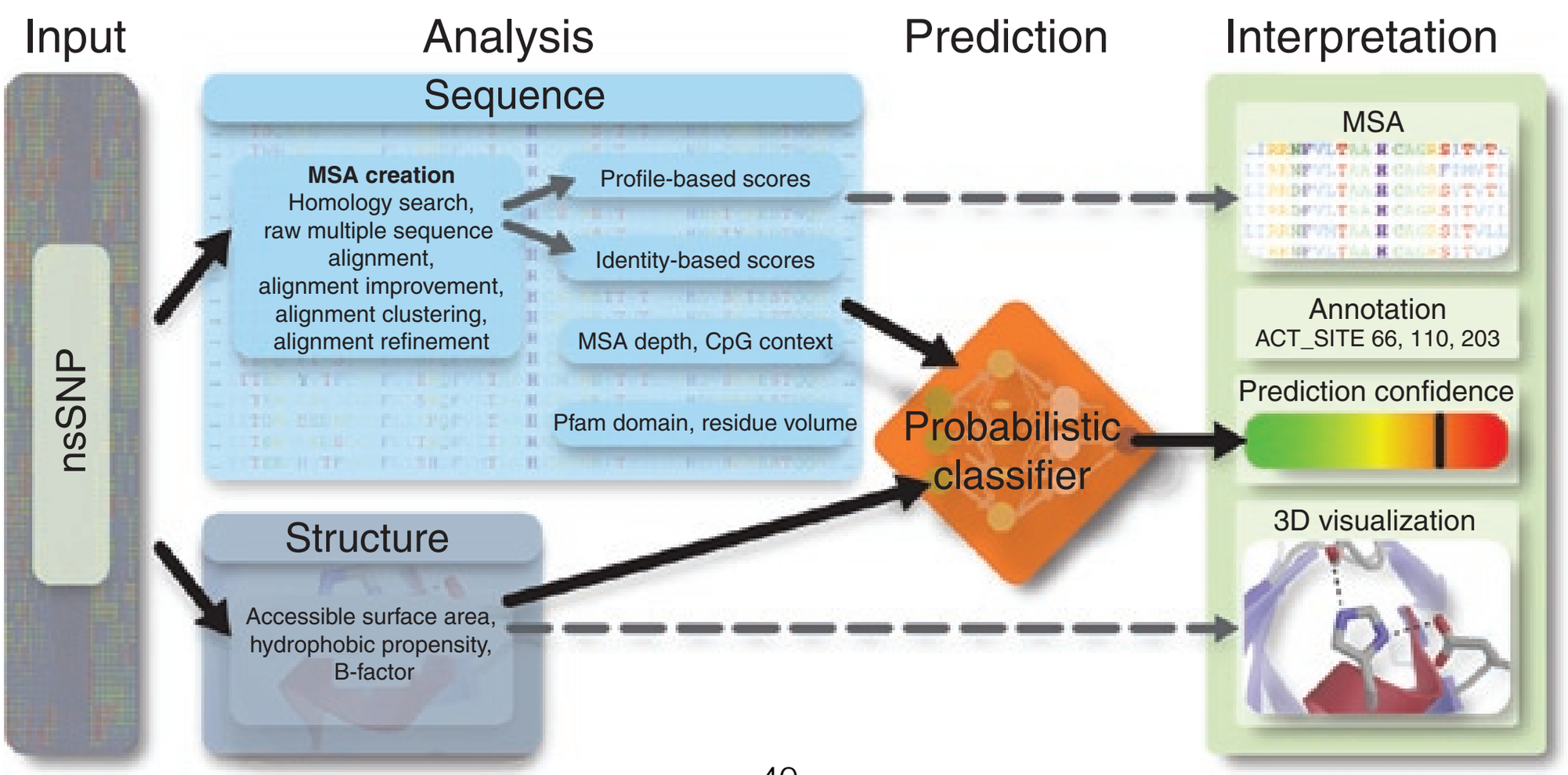
- Evolutionary conservation is one way that putatively “functional” variants will be identified.

- Often joint with structural information when available.

PolyPhen-2

- One of the most popular tools for predicting damaging effects of missense mutations.

- Uses 8 sequence-based and 3 structure-based predictive features (chosen from initial set of 32).



What about the 99%?

- Outside the exome, few tools existed until very recently (like within the last few months).

- Previously, we relied on **PhastCons**
    - HMM-based method for identifying evolutionarily conserved blocks in the genome.

- and **PhyloP**
    - A Likelihood Ratio Test method for identifying sites within a genome that are conserved or accelerated (compared to neutral background).

- Both are integrated into the UCSC Genome Browser.

Ultraconserved Regions

- Comparison of the human, mouse and rat genomes identified 481 segments of >200 bp (and more than 5000 segments of >100bp) that are completely conserved (100% identity) across the three species.

- Over half of these ‘ultraconserved’ segments do not occur in exons, and presumably have some sort of regulatory or structural function.

Bejerano et al. (2004)

Ultraconserved Regions

- There are at least two possible explanations for this remarkable degree of conservation:
    - Low mutation rate
    - Purifying selection
- How can we distinguish between these two possibilities?

Predictions for within-species variation

Low mutation rate

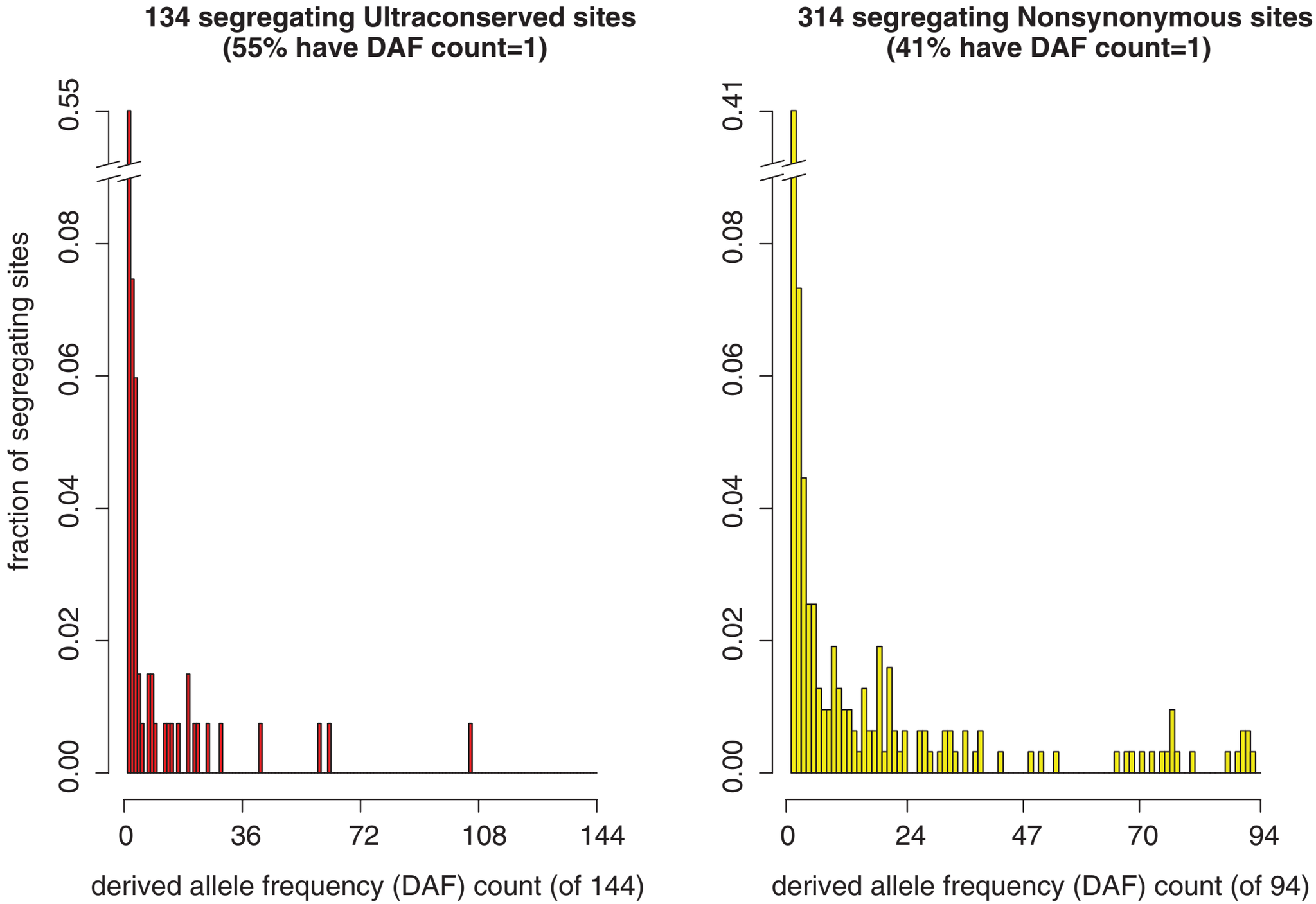
- Very few (if any) polymorphisms
- Frequency of any observed mutations should look like genome-wide background.

Purifying selection

- Some polymorphisms
- Polymorphisms will be at low frequency

Frequency spectrum

Ultraconserved regions vs. non-synonymous sites



Human Genome Ultraconserved Elements Are Ultraselected.
Katzman, et al. 2007.

Ultraconserved Regions

- How critical are these ultraconserved regions for the viability and fertility of an individual?

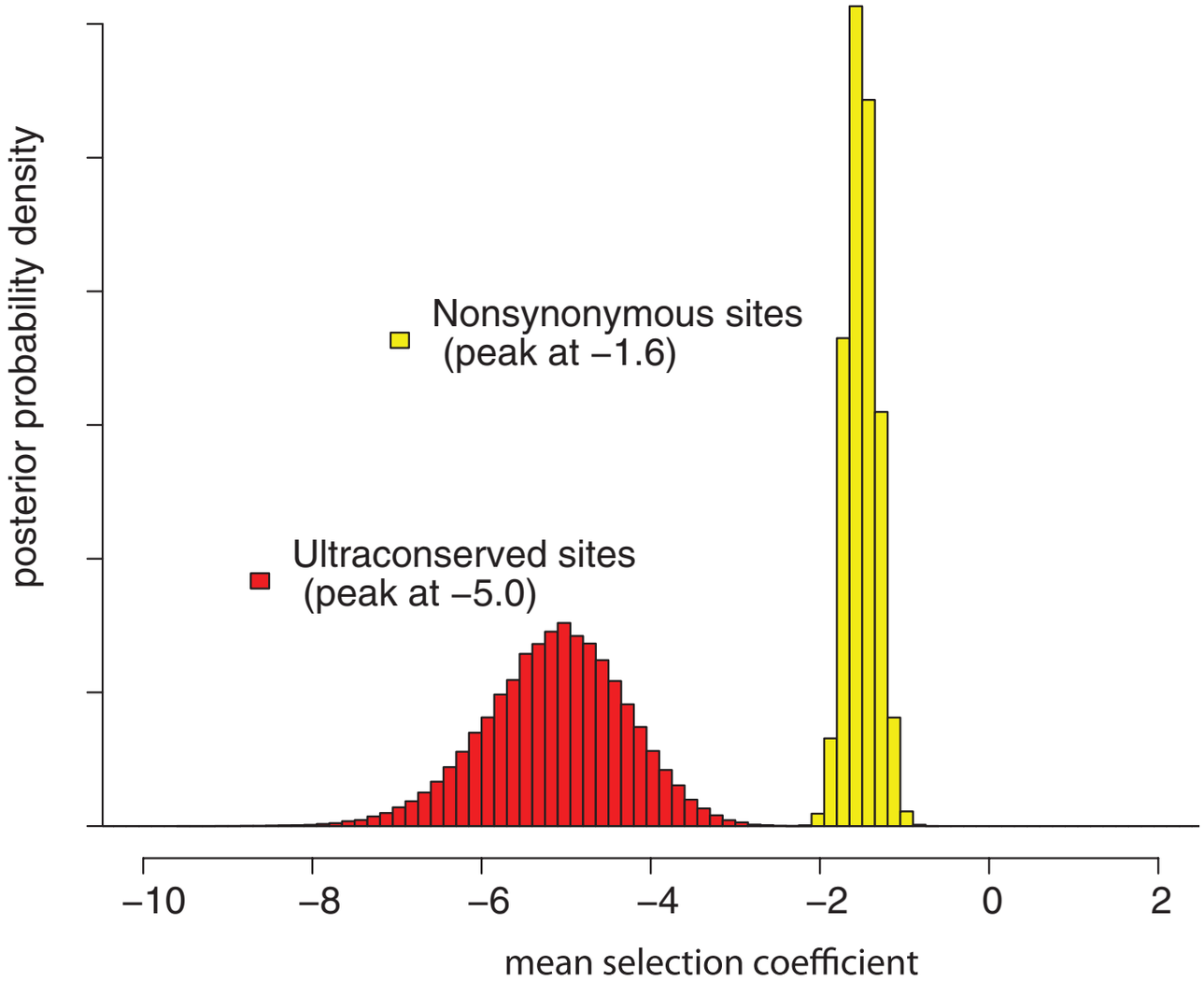
- Ahituv et al. (2007) created mouse knockouts that deleted four of the ultraconserved regions.

- They chose regions thought to have a regulatory role on genes of known function.

- Surprisingly, they found that the mouse knockouts were completely viable and fertile, with no observable phenotypic abnormalities!

Estimated selection coefficients

(Katzman et al. 2007)



- It only takes weak selection (Ns ~ 5) to produce both a skew in the frequency spectrum and a lack of fixed differences across species.

- However, this corresponds to selection coefficients ~0.05% (N=10,000), so it is understandable if there are no obvious phenotypic effects on knockout mice.

What about the 99%?

- Two recent tools have been developed for genome-wide functional prediction:
  - CADD
  - fitCons

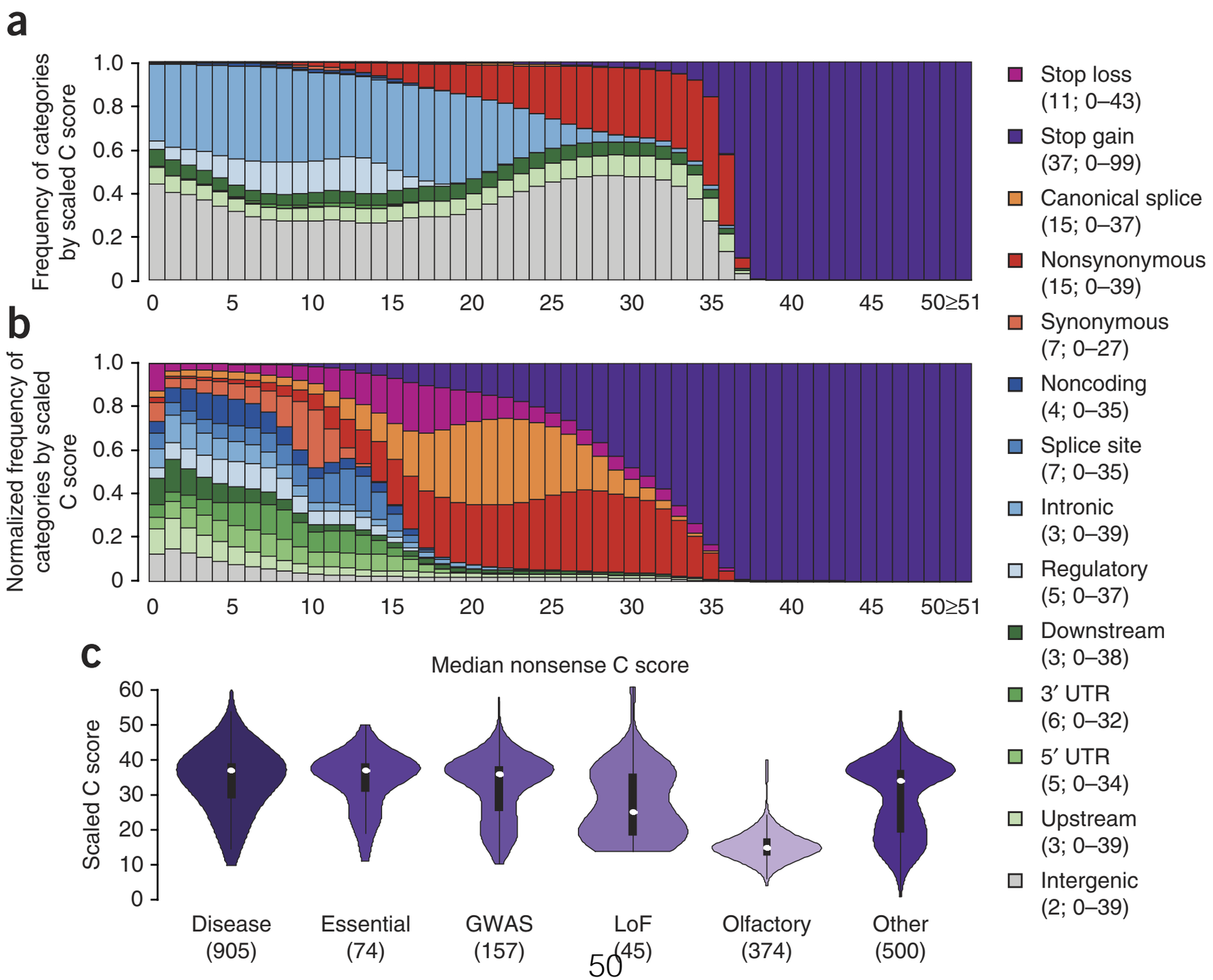
A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]

- Combined Annotation–Dependent Depletion (CADD)
- A method for objectively integrating **88** diverse annotations into a single measure (C score) for each possible variant at every position in the genome.
- CADD is the result of using a **support vector machine** trained to differentiate 14.7 million high-frequency human-derived alleles from 14.7 million simulated variants.

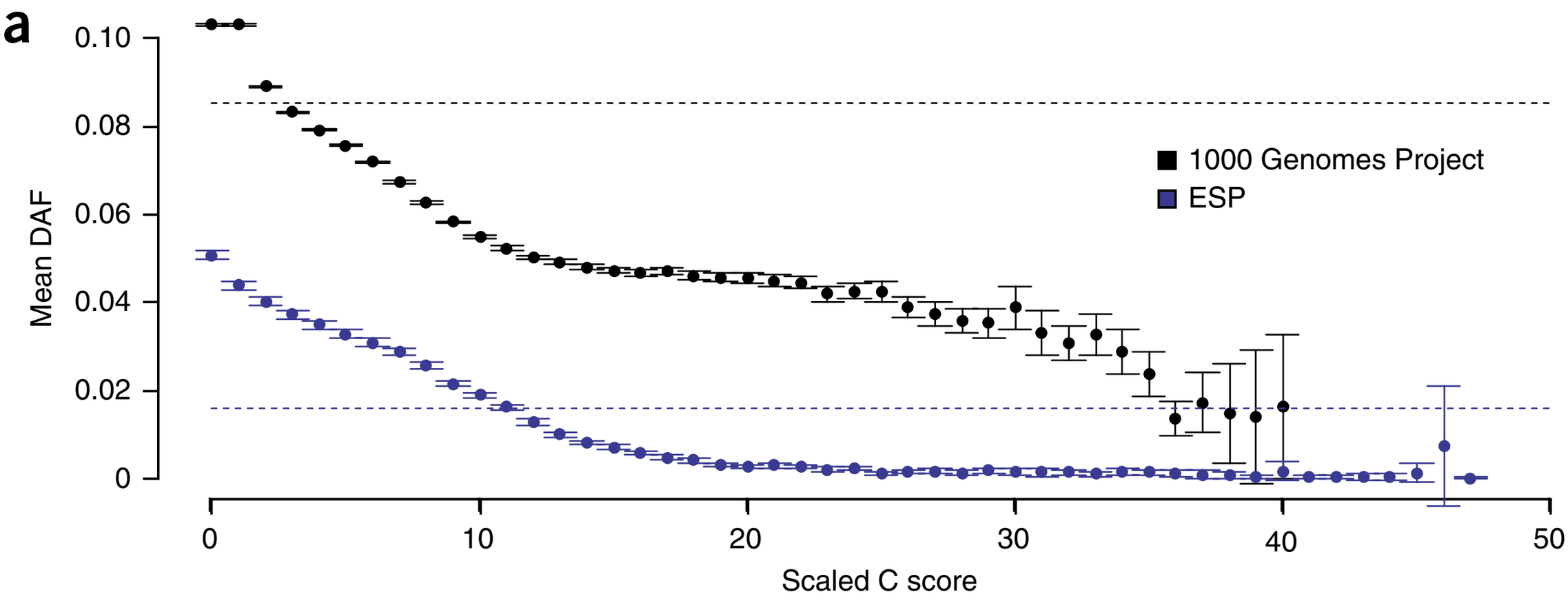
A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]



A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]



TECHNICAL REPORTS

A method for calculating probabilities of fitness consequences for point mutations across the human genome

Brad Gulko[1], Melissa J Hubisz[2], Ilan Gronau[2,3] & Adam Siepel[1–3]

- A computational approach for estimating the probability that a point mutation at each nucleotide position in a genome will have a fitness consequence (**fitCons**).

- Scores can be interpreted as an evolution-based measure of potential genomic function.

- fitCons scores for three human cell types based on publicly available genomic data and made them available as UCSC Genome Browser tracks.

A method for calculating probabilities of fitness consequences for point mutations across the human genome

Brad Gulko[1], Melissa J Hubisz[2], Ilan Gronau[2,3] & Adam Siepel[1–3]

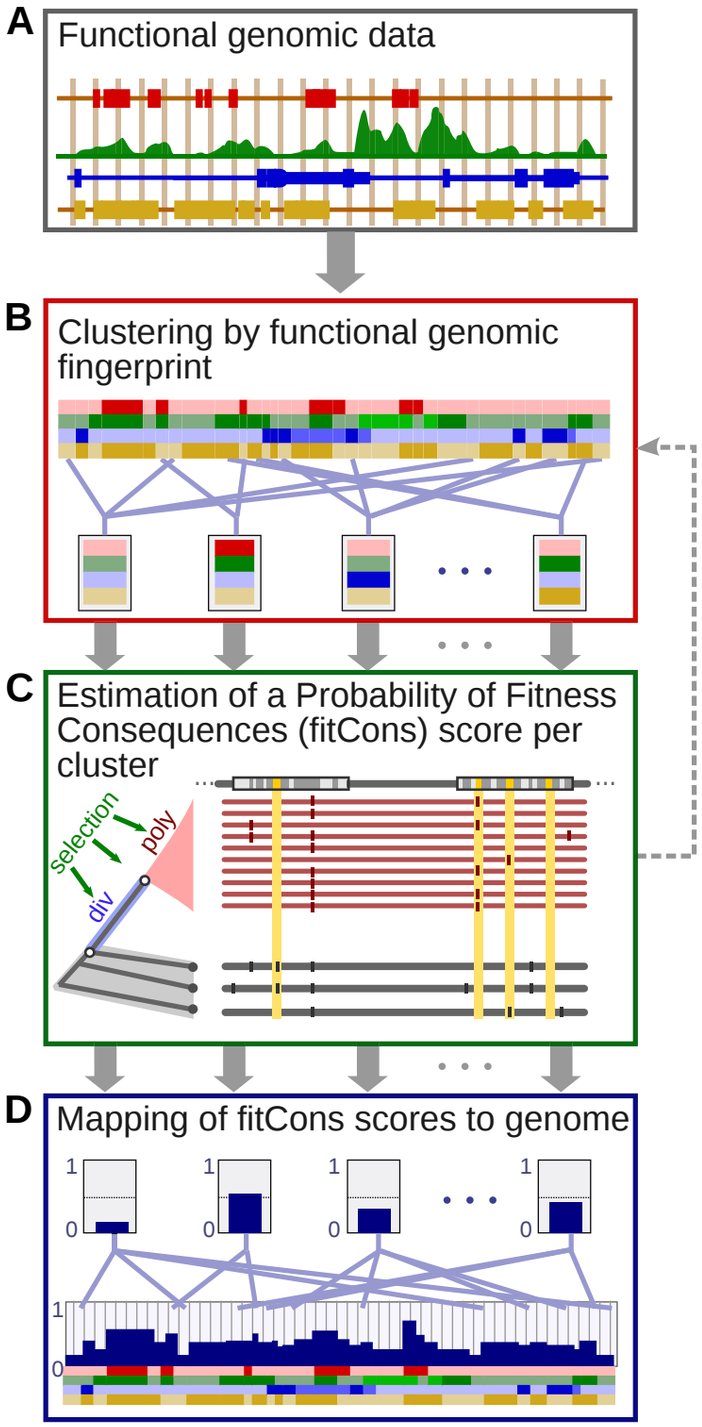
- Like conventional evolutionary conservation scores, fitCons scores are clearly elevated in known coding and noncoding functional elements, but they show considerably better sensitivity than conservation scores for many noncoding elements.

- They perform exceptionally well in distinguishing ChIP-seq-supported transcription factor binding sites, expression quantitative trait loci, and predicted enhancers from putatively nonfunctional sequences.

- The fitCons scores indicate that **4.2-7.5%** of nucleotide positions in the human genome have influenced fitness since the human-chimpanzee divergence.

A method for calculating probabilities of fitness consequences for point mutations across the human genome

Brad Gulko[1], Melissa J Hubisz[2], Ilan Gronau[2,3] & Adam Siepel[1–3]



- fitCons will likely be one of the most popular tools for genome-wide functional predictions.
  - In humans…

A method for calculating probabilities of fitness consequences for point mutations across the human genome

Brad Gulko[1], Melissa J Hubisz[2], Ilan Gronau[2,3] & Adam Siepel[1–3]