Functional Importance and Selective Constraints in Human non-coding DNA

Shamil Sunyaev

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

Most of the Genome is Non-coding

… and probably is an evolutionary junkyard

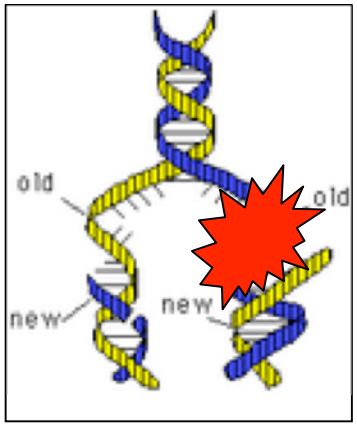
However, many genomic regions are highly conserved!

acgtcttcccttaggatc

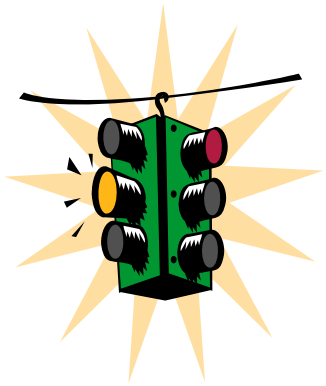
gcatcttcccttaggcgc

Definition:

Conservation \Con`ser*va"tion\, n. [L. conservatio: cf. F. conservation.] The preservation of a genetic sequence over time due to natural selection.



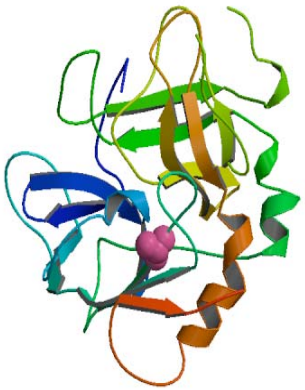
Natural selection



Conservation as a functional genomics assay

Phenotype

Effect on molecular function



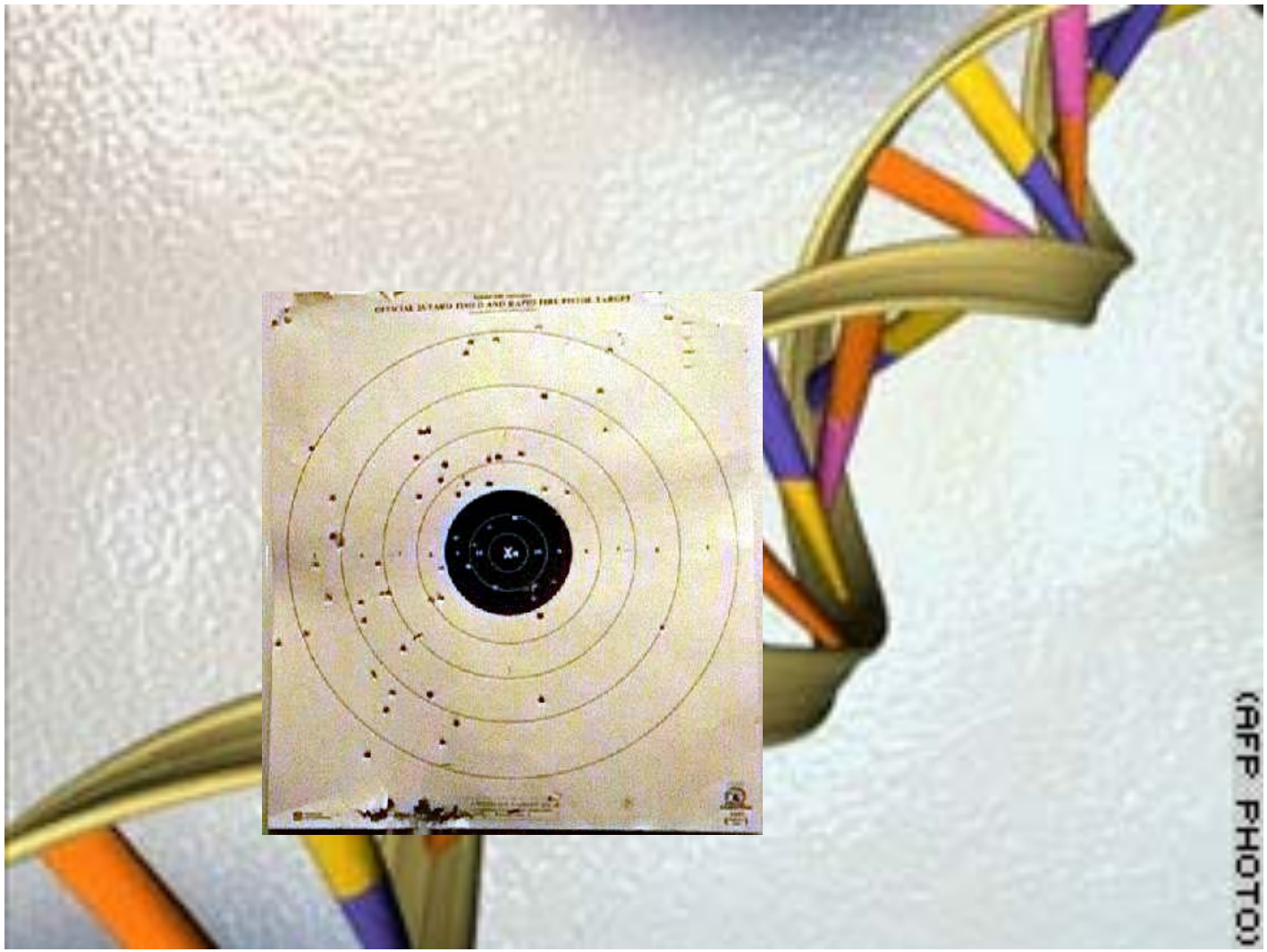
Practical aspect of conservation

- Prediction of new non-coding functional elements
- Search for allelic variation underlying human phenotypes (even if we do not know function)

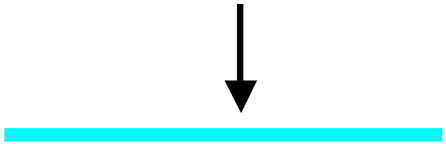
Theoretical aspect of conservation

- Estimation of genomic rate of deleterious mutations (*U > 1* already implies synergistic epitasis; estimates including non-coding sequence are much larger).
- Genetics of common phenotypes (Why do they exist in spite of purifying selection?).

Individual human genome is a target for deleterious mutations



Common disease / Common variant



Trade off (antagonistic pleiotropy)
Balancing selection
Recent positive selection
Reverse in direction of selection

Examples

APOE	**Alzheimer's disease**
AGT	**Hypertension**
CYP3A	**Hypertension**

Multiple mostly rare variants

Many deleterious alleles in mutation-selection balance

Examples

Plasma level of HDL-C
Plasma level of LDL-C
Colorectal adenomas

Questions:

- How many nucleotides in the human genome are selectively constrained?
- How are selectively constrained nucleotides distributed along the genome?
- How strong is selective pressure in non-coding regions?
- Do comparative and functional genomics data agree?

Selection coefficient

Selection coefficient - is a measure of reproductive disadvantage (or advantage) associated with a mutation.

Wild type

$N_1 = 4$

New mutation

$N_2 = 3$

$$\frac{N_2}{N_1} = 1 - s$$

Selection coefficient

Divergence (K)

Every new mutation eventually will be either fixed or lost

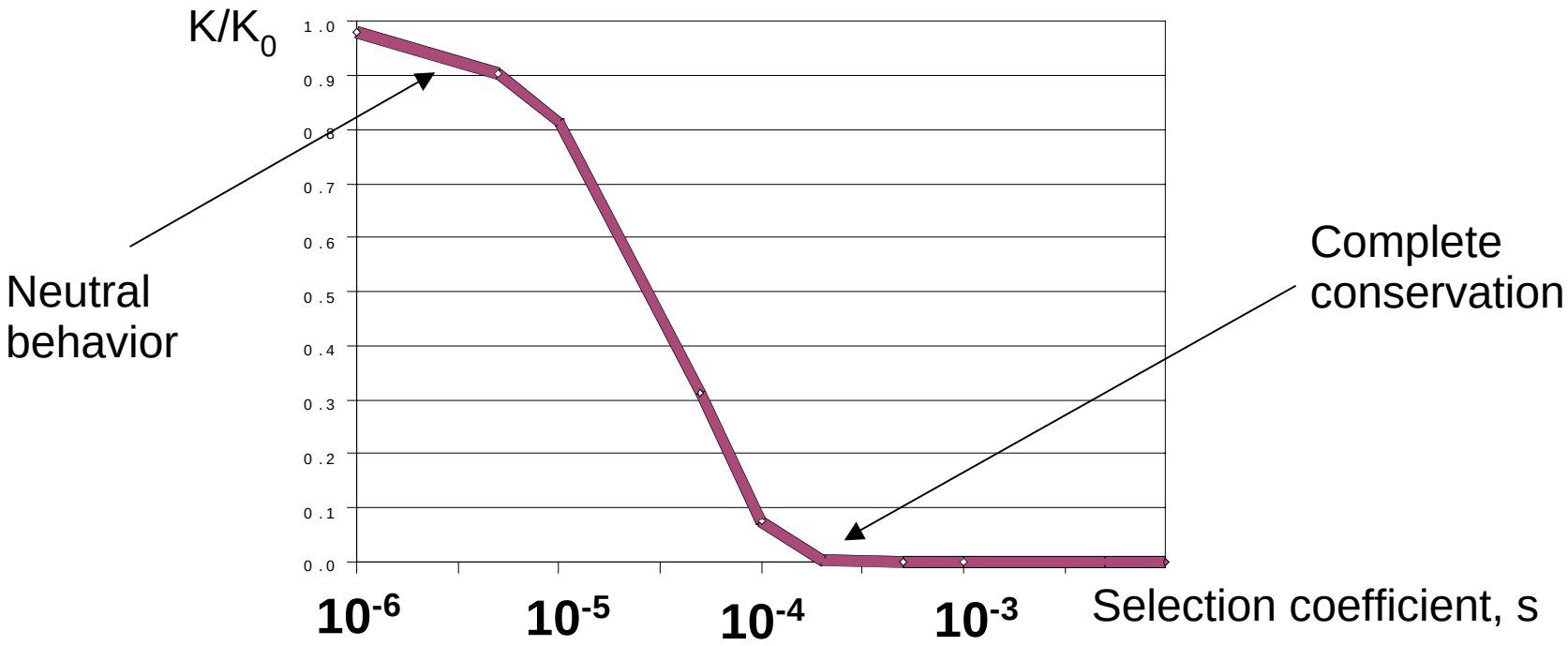
$$K = K_0 2N_e \frac{(1-e^{-2s})}{(1-e^{-4N_e s})}$$

s – selection coefficient

N_e - effective population size

For humans estimated to be ~ 10 000

For mice estimated to be ~ 85 000



Divergence depends on…

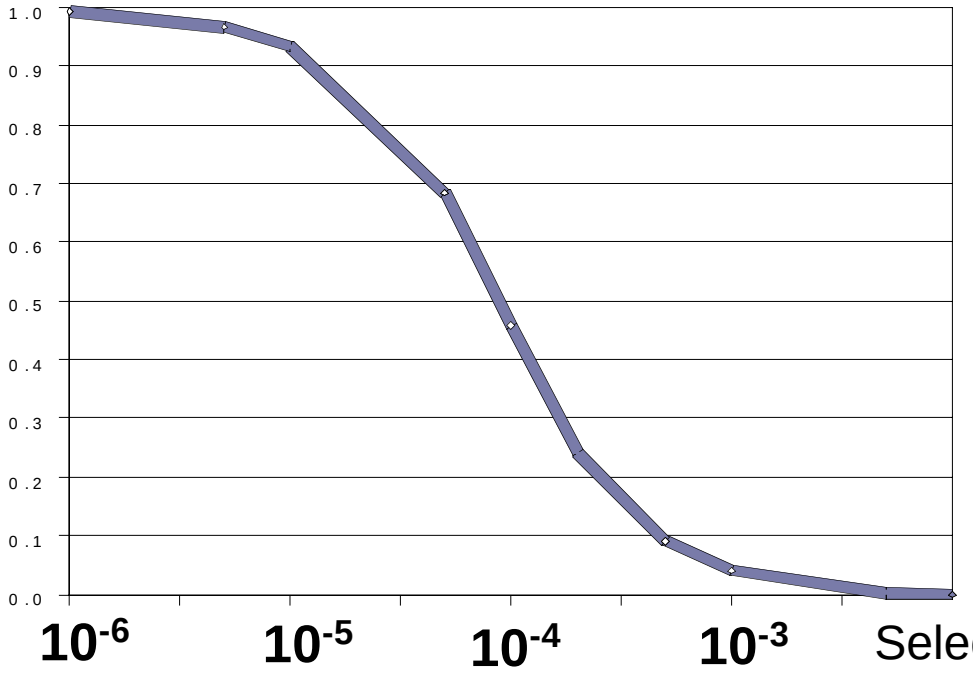
- Mutation rate
- Phylogenetic distance (mostly unknown)
- Selection (negative or positive)

Divergence (conservation) is not informative about strength of selection.

Nucleotide diversity (π)

Database SNP density is proportional to π

$$\pi = \pi_0 \frac{2N_e s + e^{-2N_e s} - 1}{N_e s\left(1 - e^{-2N_e s}\right)}$$

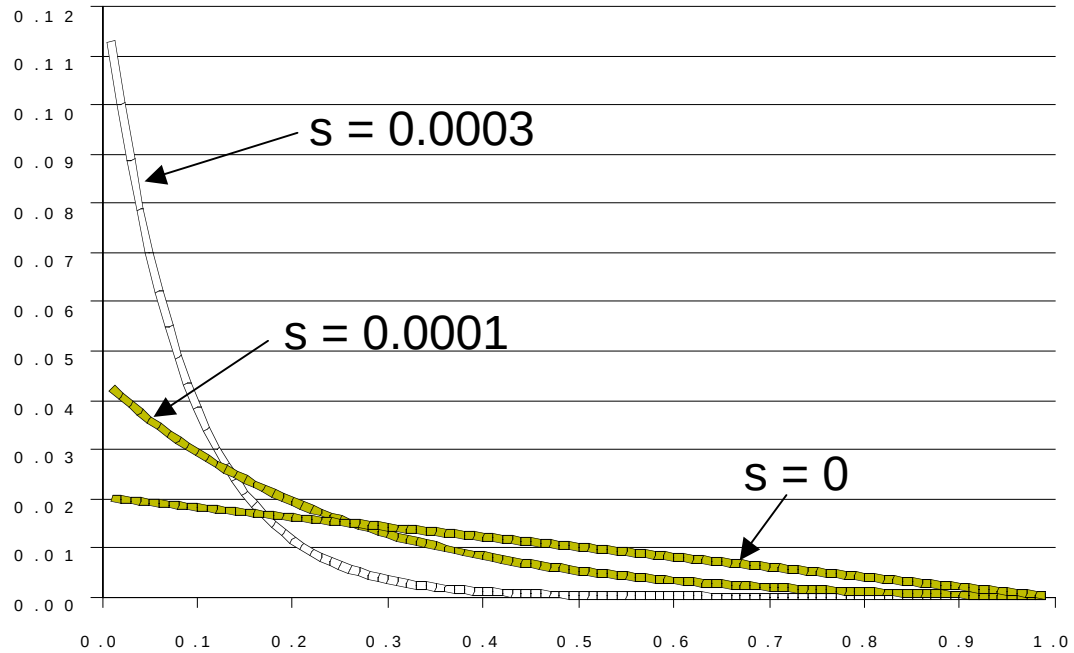


Nucleotide diversity depends on…

- Mutation rate
- Coalescent variance
- Background selection
- Hitchhikings
- Selection (negative or positive)

Allele frequency spectrum

Proportion of new alleles with particular frequency



New allele frequency

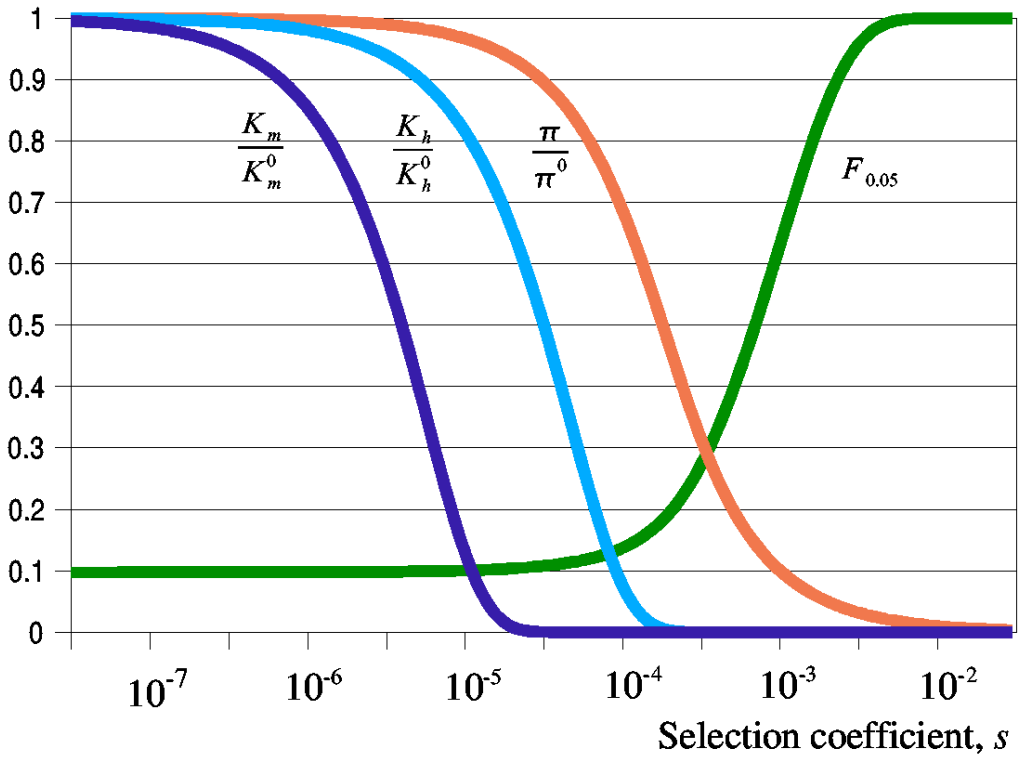
The higher selection coefficient *s* (stronger selection) – the higher proportion of low frequency alleles

$F_{0.05}$ – the proportion of new alleles with frequency below 5%

Allele frequency spectrum depends on…

- Demographic history
- Biased gene conversion
- Background selection (if associated with inefficient negative selection)
- Hitchhikings
- Selection (negative or positive)

Relationships between K, π, $F_{0.05}$ and s



$K_m(s) / K_m(s=0)$

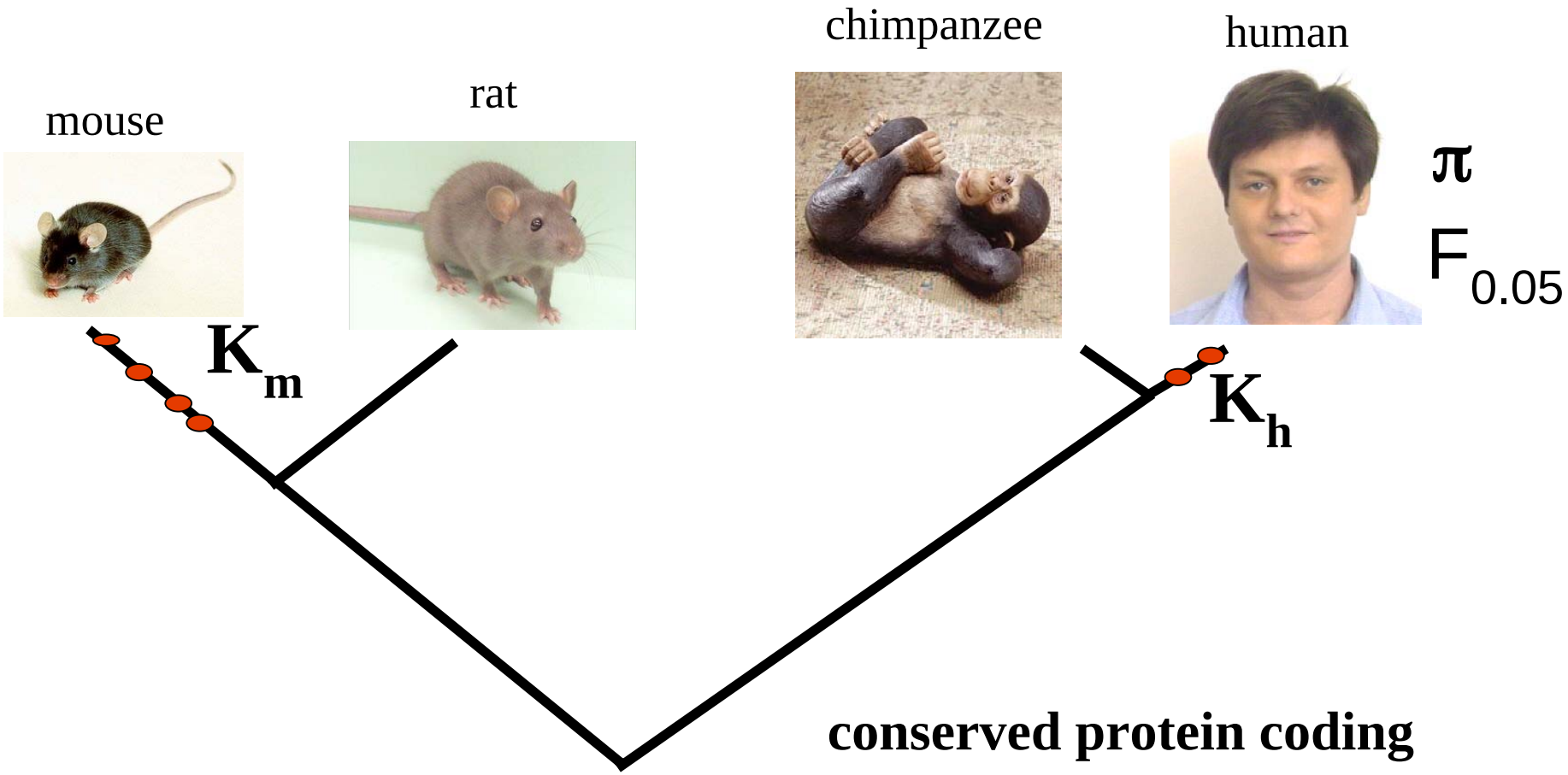
$K_h(s) / K_h(s=0)$

$\pi(s) / \pi(s=0)$

$F_{0.05}(s)$

- Conservation (very low fixation rate of new mutations) does not necessary mean high selection coefficients
- Conservation is not enough to estimate the level of selective pressure

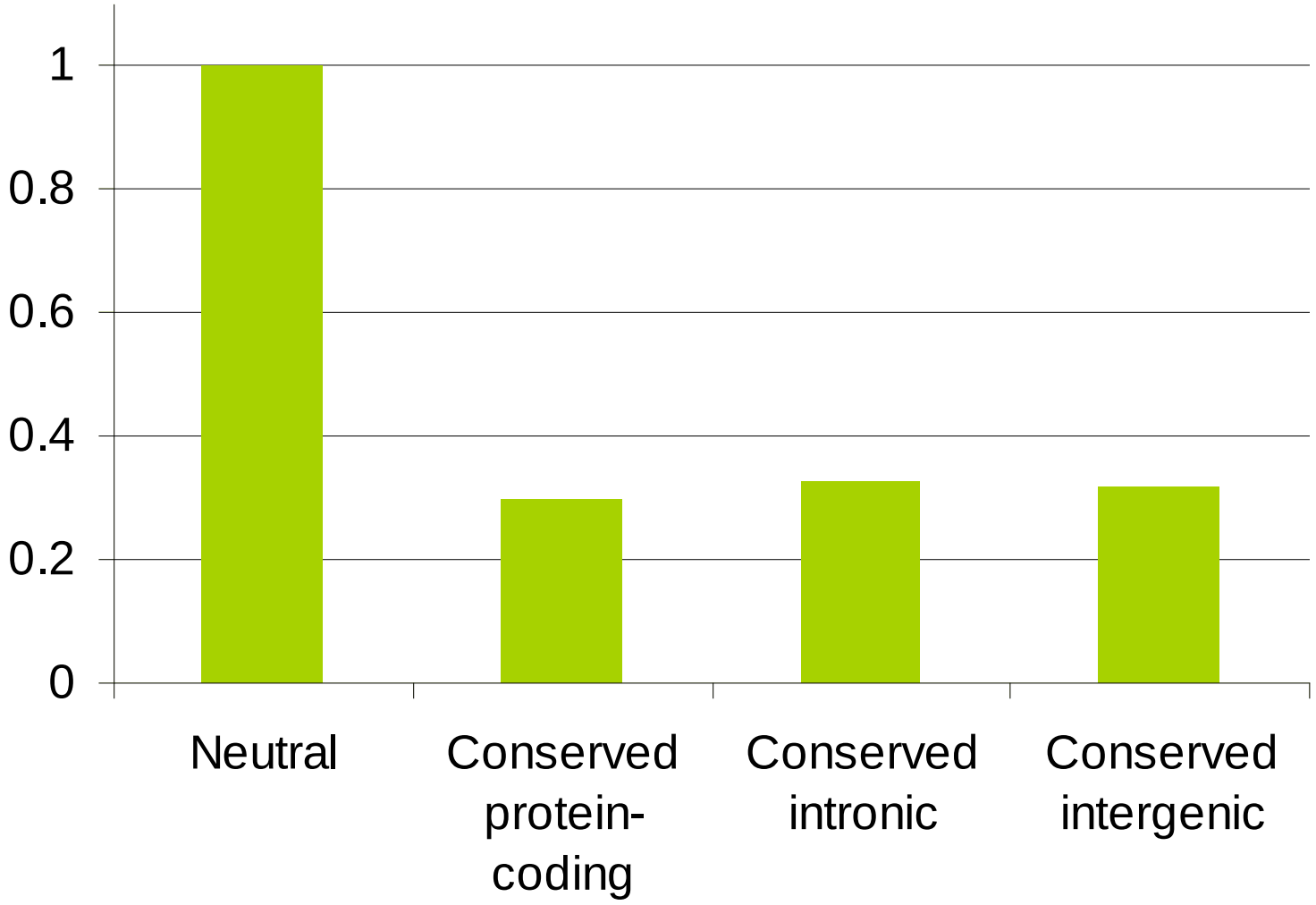
Genome analysis



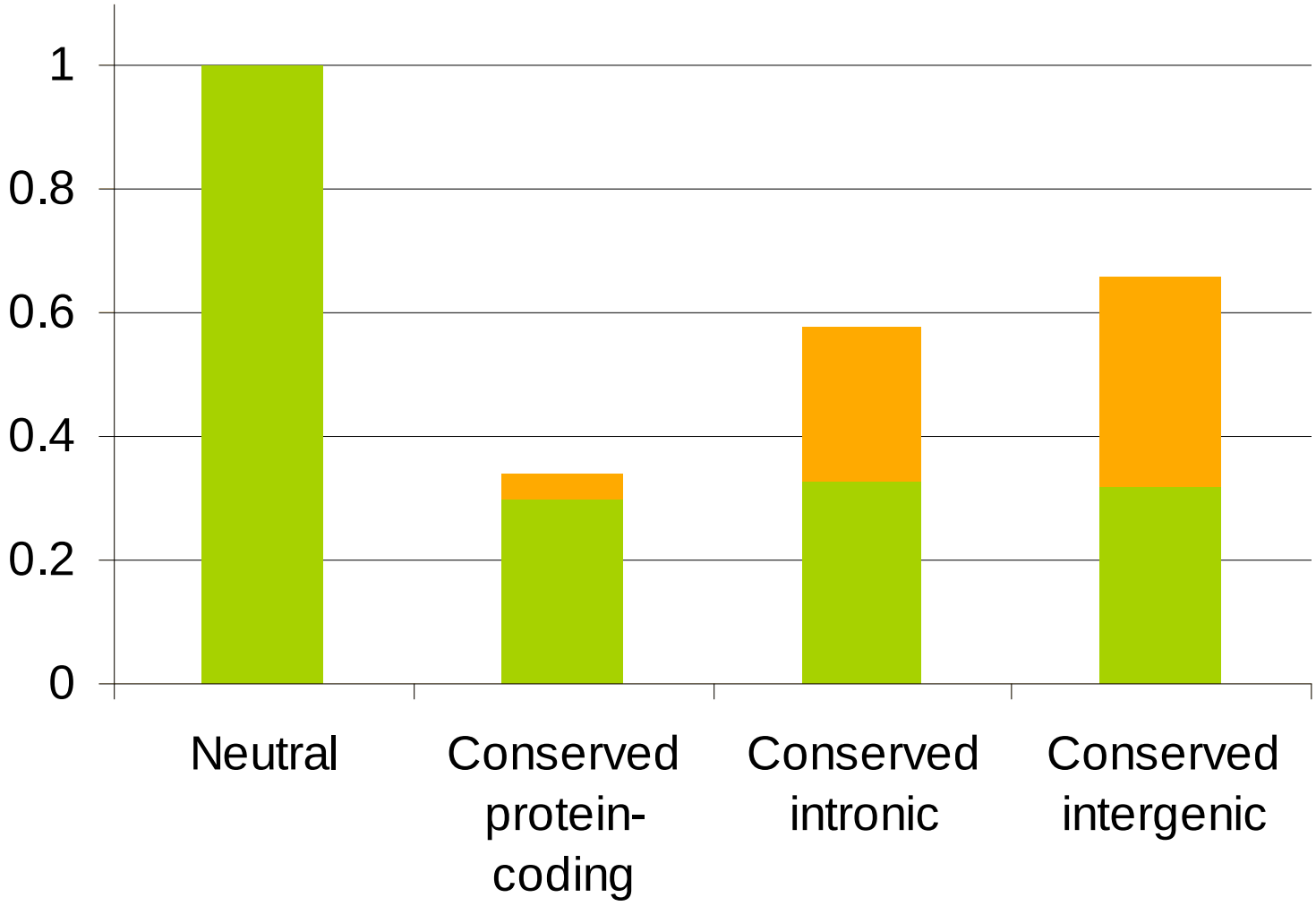
conserved protein coding
conserved intronic
conserved intergenic region

>90% in 50 nt window

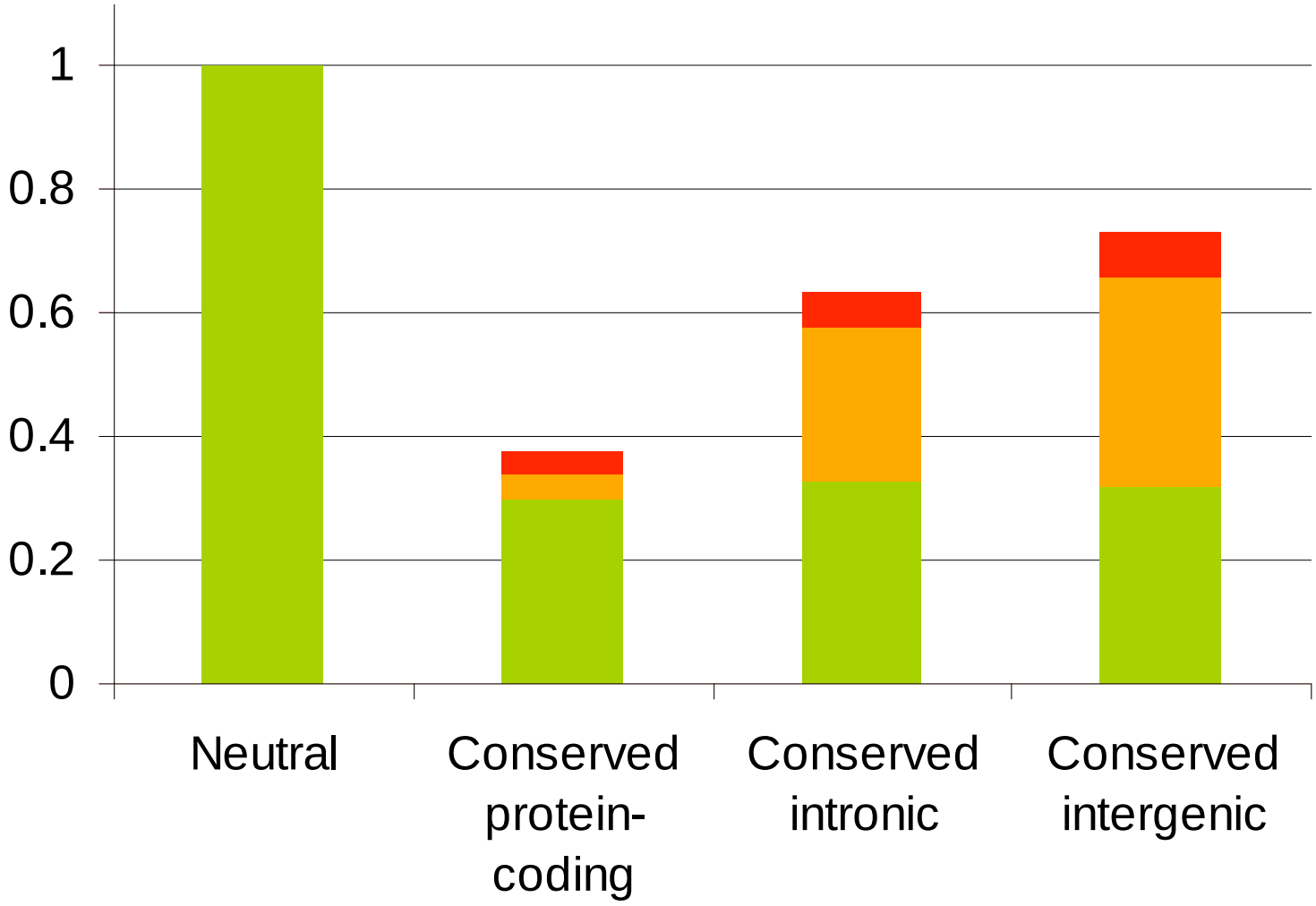
Divergence in the mouse lineage (K_m) in conserved genomic regions



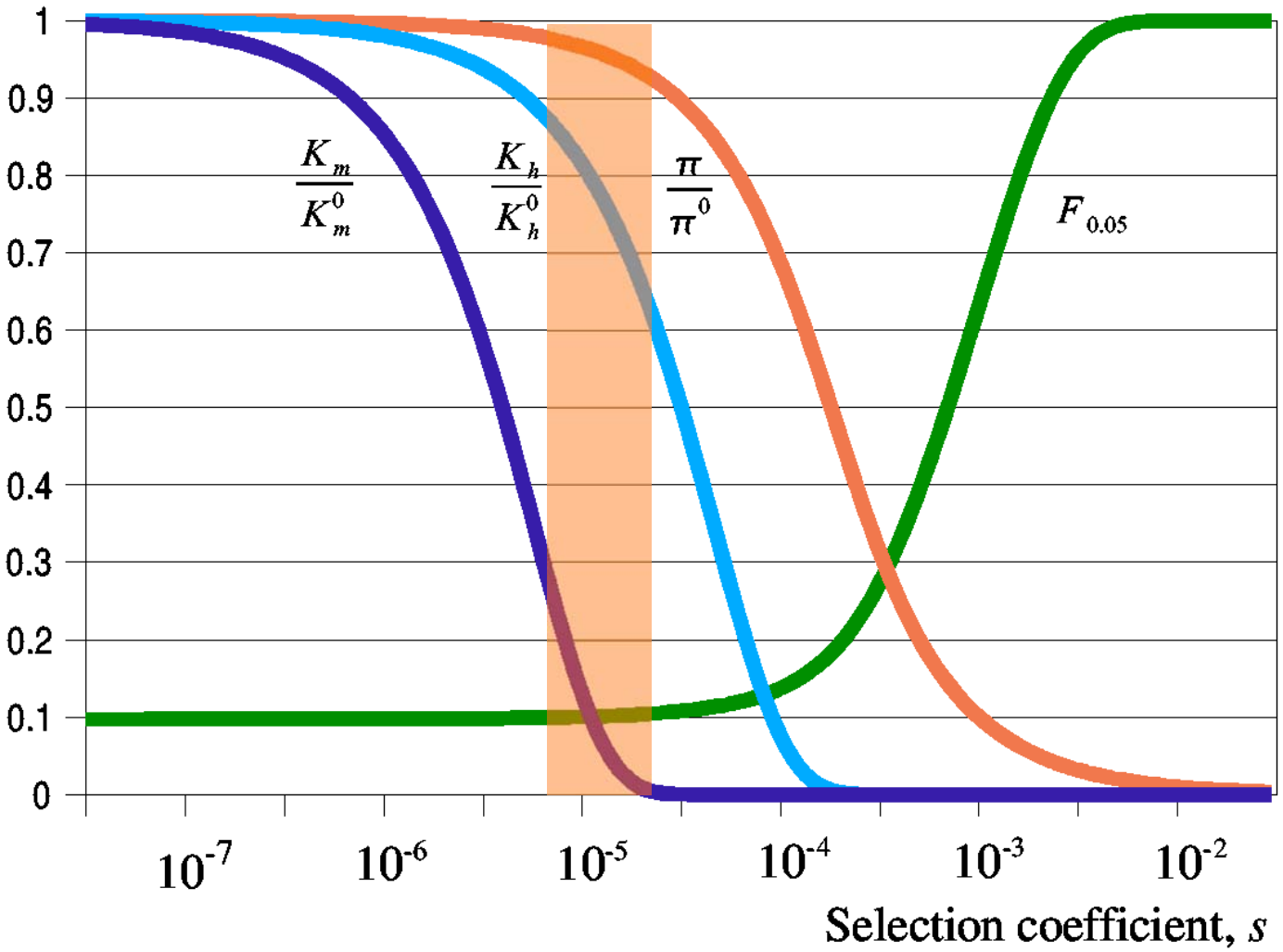
Divergence in the human lineage (K_h) in conserved genomic regions



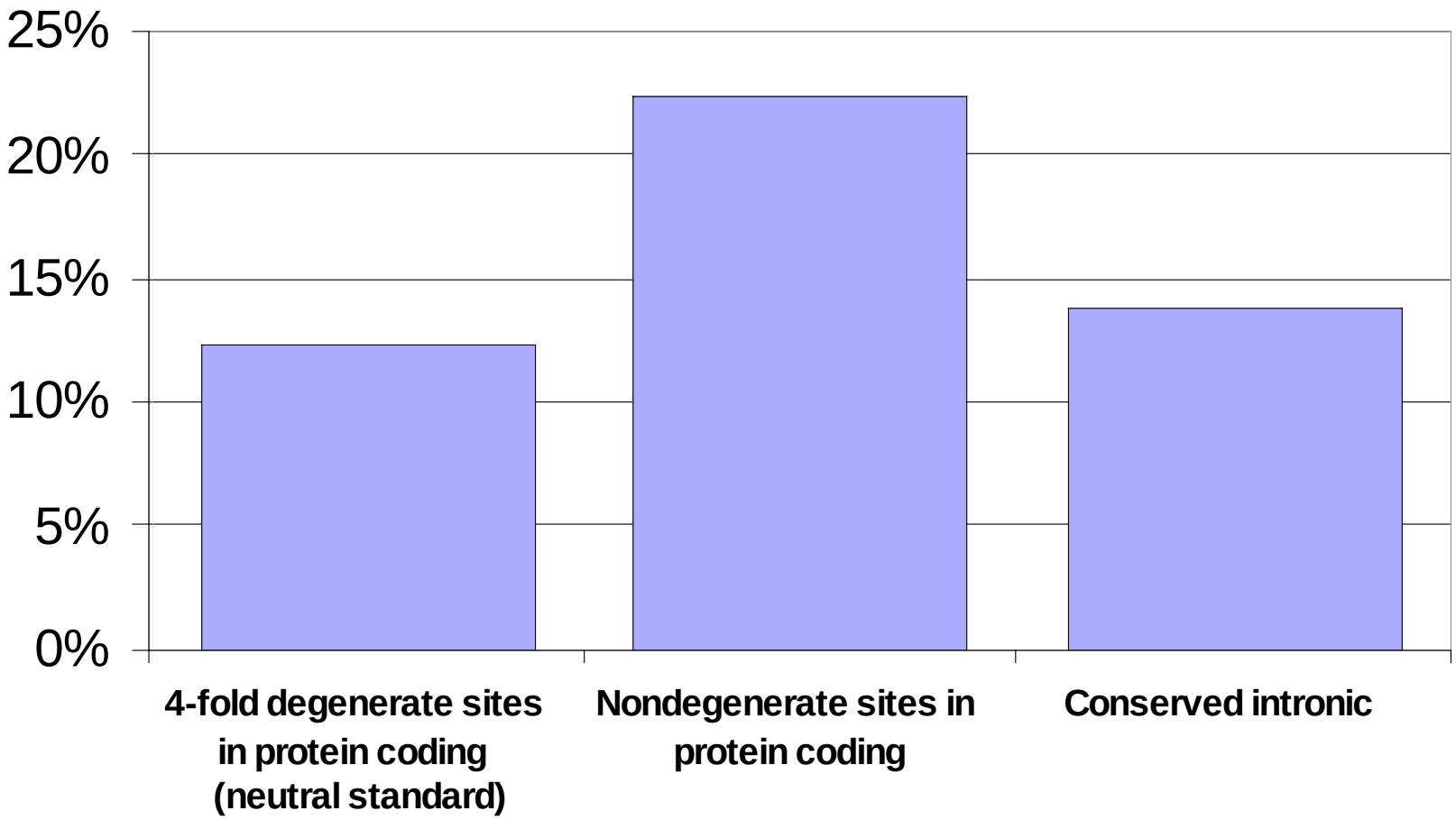
Nucleotide diversity (π) in conserved genomic regions



Selective pressure in conserved genomic regions: theory



Fraction of rare new alleles ($F_{0.05}$) in conserved genomic regions

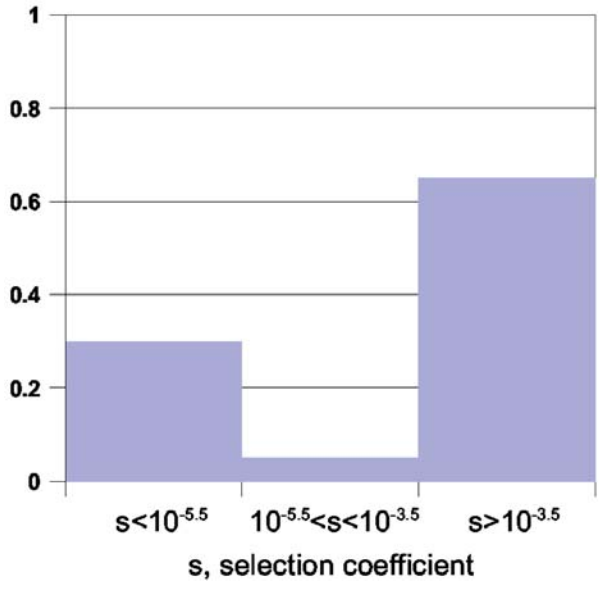


Selective constraints in conserved genomic regions

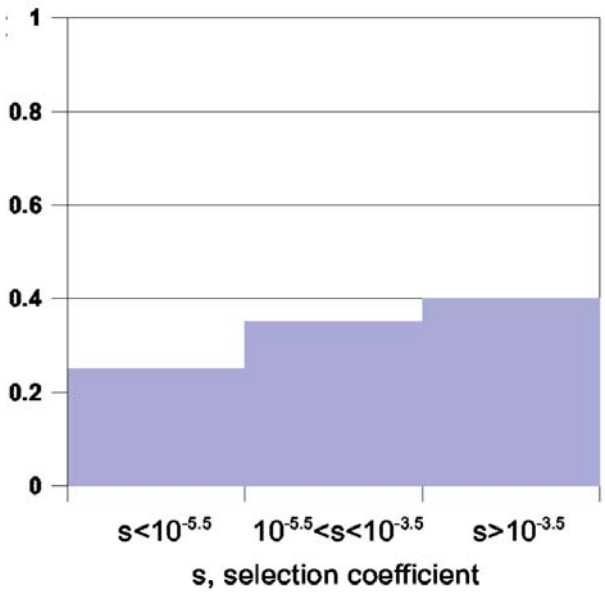
- A genome-wide relaxation of selective constraint in the primate lineage
- This relaxation most likely resulted from a smaller human effective population size
- Relaxation is much more profound in conserved non-coding regions than in protein-coding regions
- Mutations at a large proportion of sites in conserved non-coding regions are associated with very small fitness effect

Distribution of selective coefficients

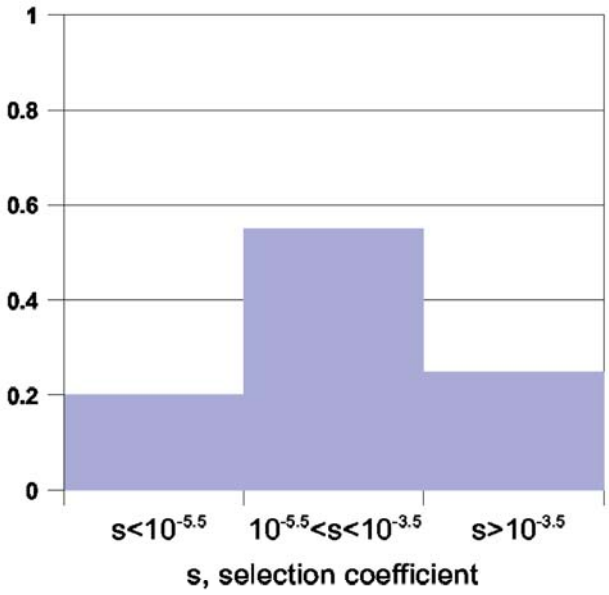
Protein coding



Introns



Intergenic



Evolution of non-coding regions

? ? ? ? ?

What can explain this staggering enrichment in sites at the borderline of neutrality in conserved non-coding regions?

Evolution of non-coding regions

? ? ? ? ?

What can explain this staggering enrichment in sites at the borderline of neutrality in conserved non-coding regions?

...synergistic epistasis!

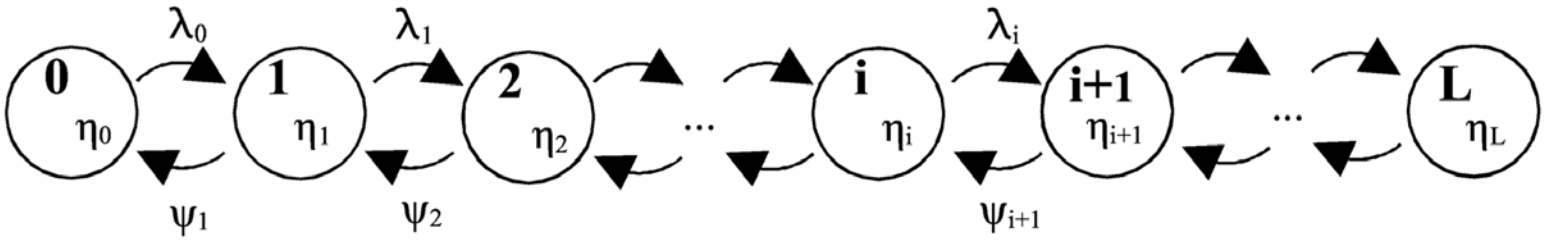
Model of non-coding regions evolution

SPECULATIONS!

- Non-coding region “is trying” to maintain an overall similarity to some optimal sequence
- All nucleotides are of equal importance
- Mutations on average tend to “move away” region from the optimal sequence
- With each new deleterious mutation the remaining nucleotides become more important

Non-coding regions evolution

Sequence evolution in the framework of birth-death processes



$$\eta_{i+1} = \frac{\lambda_i}{\Psi_{i+1}} \eta_i$$

The recursive formula for an equilibrium distribution

Transition probabilities from state to state

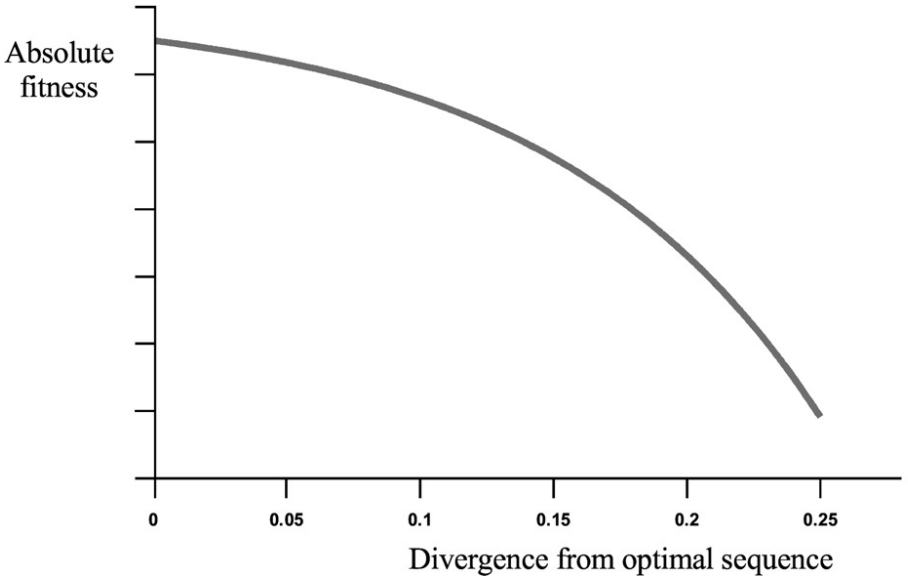
$$\lambda_i = \mu (L - i) \frac{(1 - e^{2s})}{(1 - e^{4Ns})}$$

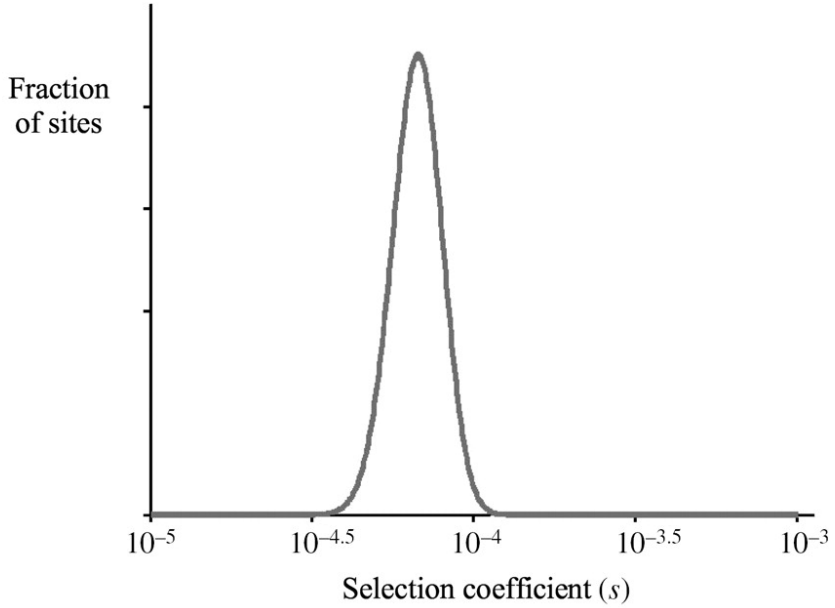
$$\Psi_{i+1} = \mu \frac{1}{3} (i + 1) \frac{(1 - e^{-2s})}{(1 - e^{-4Ns})}$$

$$F_i = \frac{1}{1 + 10^{-4} e^{10 \frac{i}{L}}}$$

fitness dependence on number of deleterious mutations

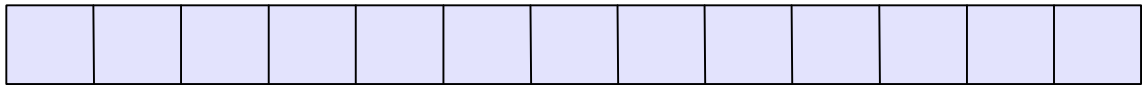
Sequence evolution in the framework of birth-death processes





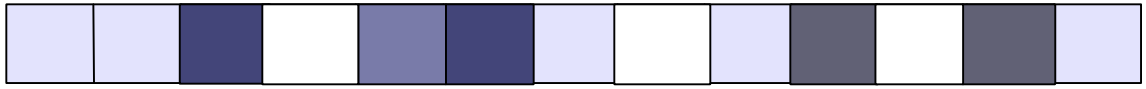
Model of non-coding regions evolution

Conserved non-coding region



All nucleotides have small selection coefficient

Conserved coding

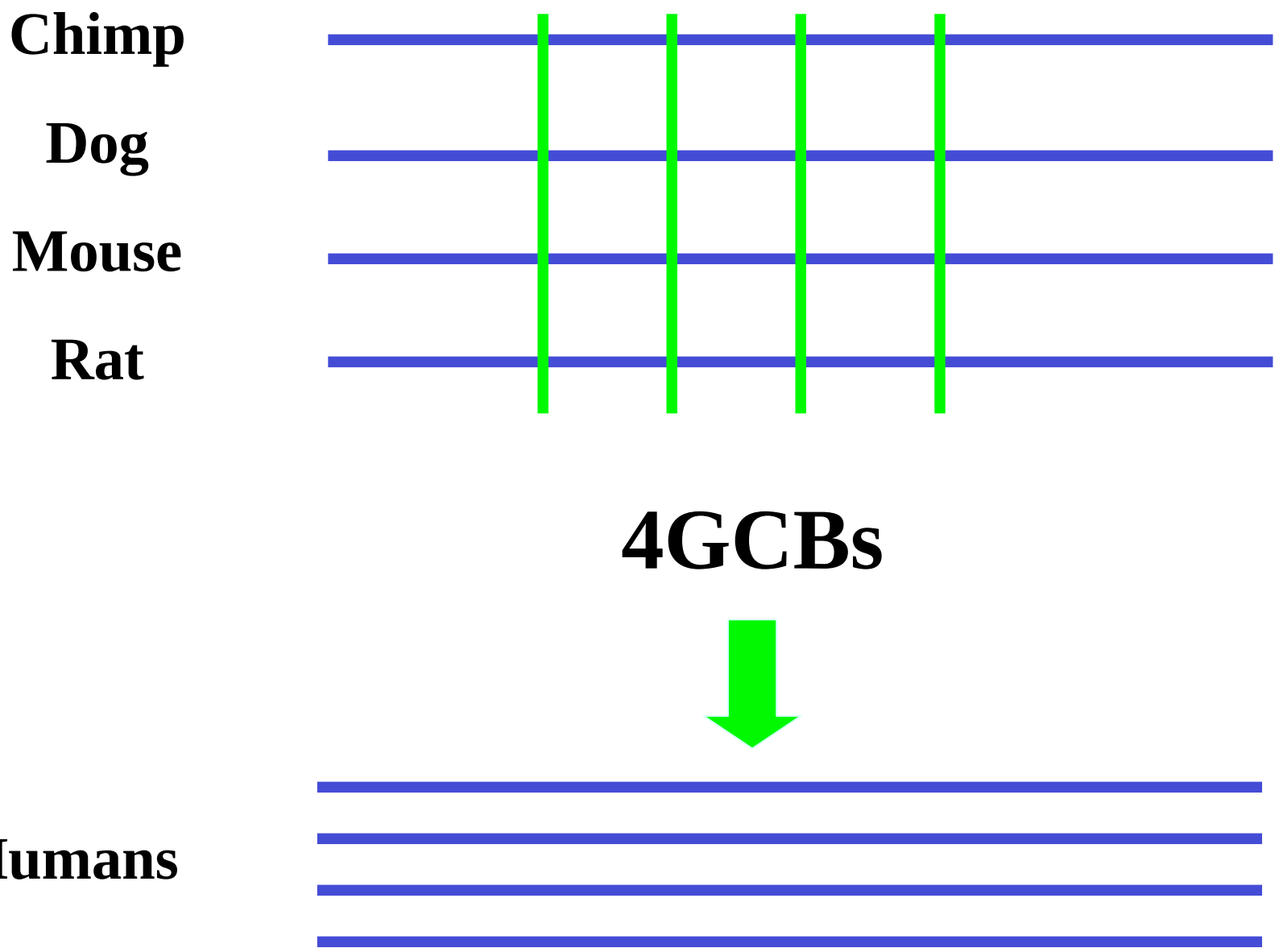


Some nucleotides have very high selection coefficient

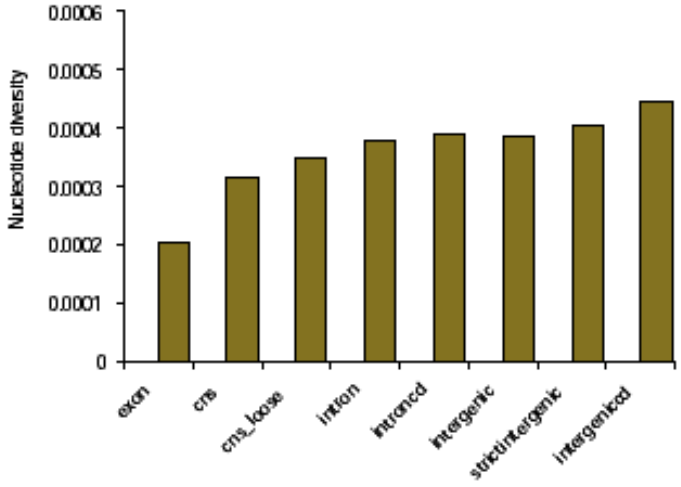
This model predicts mutations of mostly small effect even in very conserved non-coding regions

Questions:

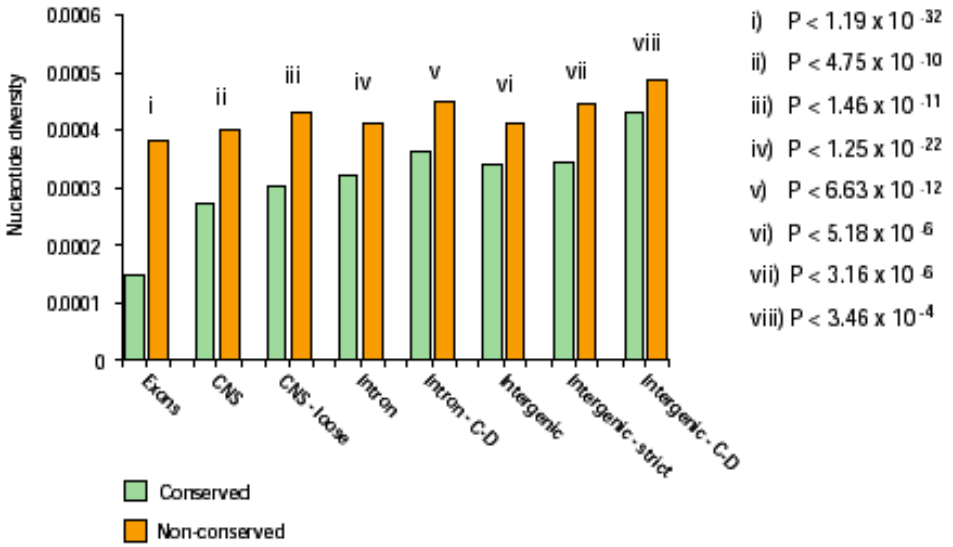
- How many nucleotides in the human genome are selectively constrained?
- How are selectively constrained nucleotides distributed along the genome?
- How strong is selective pressure in non-coding regions?
- Do comparative and functional genomics data agree?


Chimp
Dog
Mouse
Rat
4GCBs
Humans

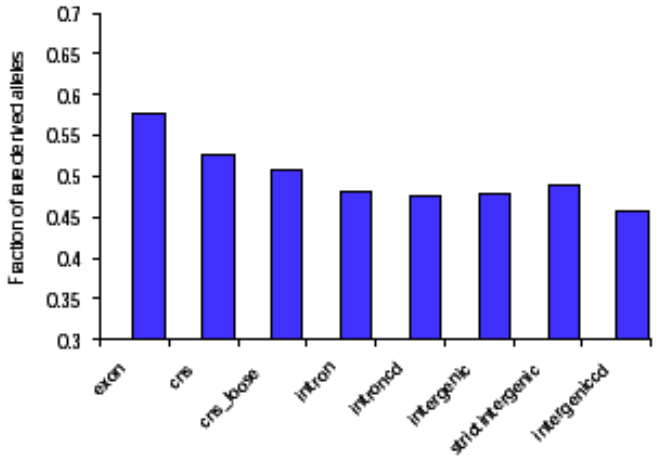
(a)


Nucleotide diversity in different genomic sequence features
Nucleotide diversity
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
exon
cns
cns_loose
intron
intronCd
intergenic
strictIntergenic
intergenicCd

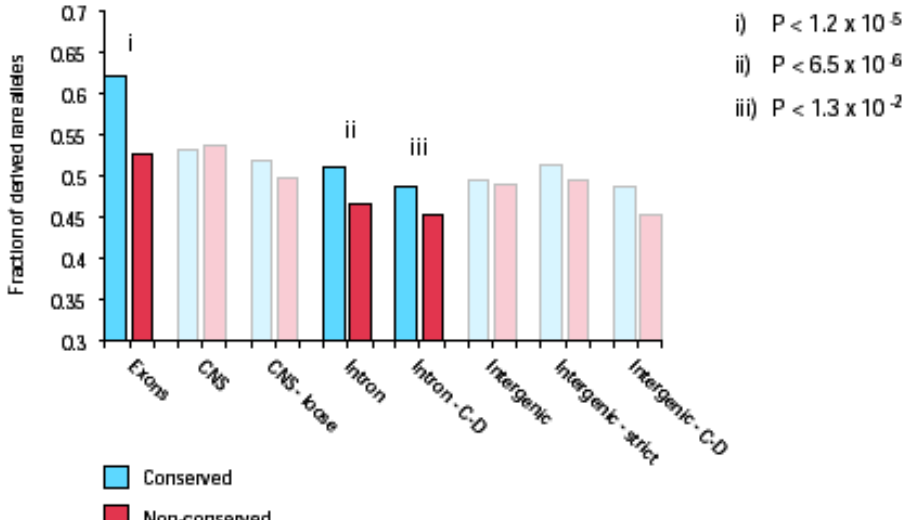
(b)


Nucleotide diversity in 4GCBs and non-4GCBs
Nucleotide diversity
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
i
ii
iii
iv
v
vi
vii
viii
Exons
CNS
CNS - loose
Intron
Intron - C-D
Intergenic
Intergenic - strict
Intergenic - C-D
Conserved
Non-conserved
i) P < 1.19 x 10 -32
ii) P < 4.75 x 10 -10
iii) P < 1.46 x 10 -11
iv) P < 1.25 x 10 -22
v) P < 6.63 x 10 -12
vi) P < 5.18 x 10 -6
vii) P < 3.16 x 10 -6
viii) P < 3.46 x 10 -4

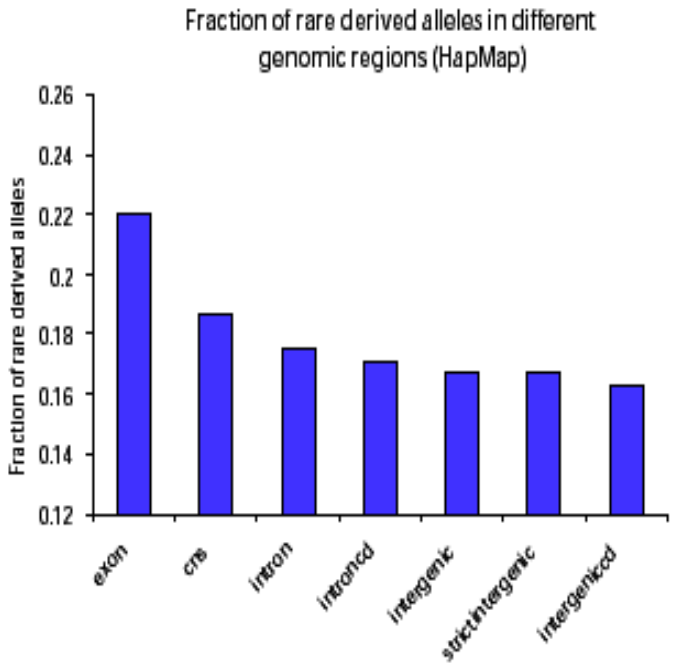
(c)


Fraction of rare derived alleles in different genomic sequence features
Fraction of rare derived alleles
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
exon
cns
cns_loose
intron
intronCd
intergenic
strictIntergenic
intergenicCd

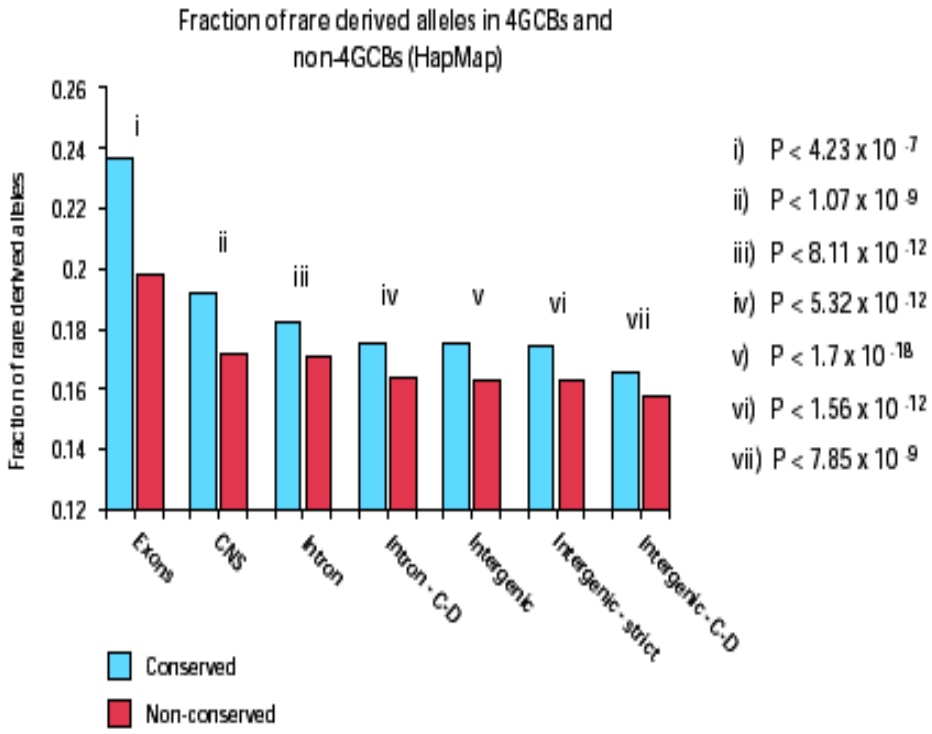
(d)


Fraction of derived rare alleles in 4GCBs and non-4GCBs
Fraction of derived rare alleles
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
i
ii
iii
Exons
CNS
CNS - loose
Intron
Intron - C-D
Intergenic
Intergenic - strict
Intergenic - C-D
Conserved
Non-conserved
i) P < 1.2 x 10 -5
ii) P < 6.5 x 10 -6
iii) P < 1.3 x 10 -2

(a)


Fraction of rare derived alleles in different genomic regions (HapMap)
Fraction of rare derived alleles
0.26
0.24
0.22
0.2
0.18
0.16
0.14
0.12
exon
cns
intron
intronCd
intergenic
strictintergenic
intergenicCd

(b)


Fraction of rare derived alleles in 4GCBs and non-4GCBs (HapMap)
Fraction of rare derived alleles
0.26
0.24
0.22
0.2
0.18
0.16
0.14
0.12
i
ii
iii
iv
v
vi
vii
Exons
CNS
Intron
Intron - C-D
Intergenic
Intergenic - strict
Intergenic - C-D
Conserved
Non-conserved
i) P < 4.23 x 10 -7
ii) P < 1.07 x 10 -9
iii) P < 8.11 x 10 -12
iv) P < 5.32 x 10 -12
v) P < 1.7 x 10 -18
vi) P < 1.56 x 10 -12
vii) P < 7.85 x 10 -9

Model

- All non-4GCBs are neutral (this is the most conservative assumption)
- 4GCBs are a mixture of neutral and functional sites
- All functional 4GCBs are associated with the same selection coefficient (this is the most conservative assumption)

Fraction of rare neutral alleles

$$F_{neutral}(1\%) = \frac{\int_0^1 \frac{\theta}{x} \cdot \left[mx(1-x)^{m-1} + \frac{m(m-1)}{2} x^2 (1-x)^{m-2} \right] \cdot dx}{\int_0^1 \frac{\theta}{x} \cdot \left(1 - x^m - (1-x)^m\right) dx}$$

$$F_{neutral}(1\%) = \frac{3}{2 \cdot \sum_{i=1}^{m-1} \frac{1}{i}}$$

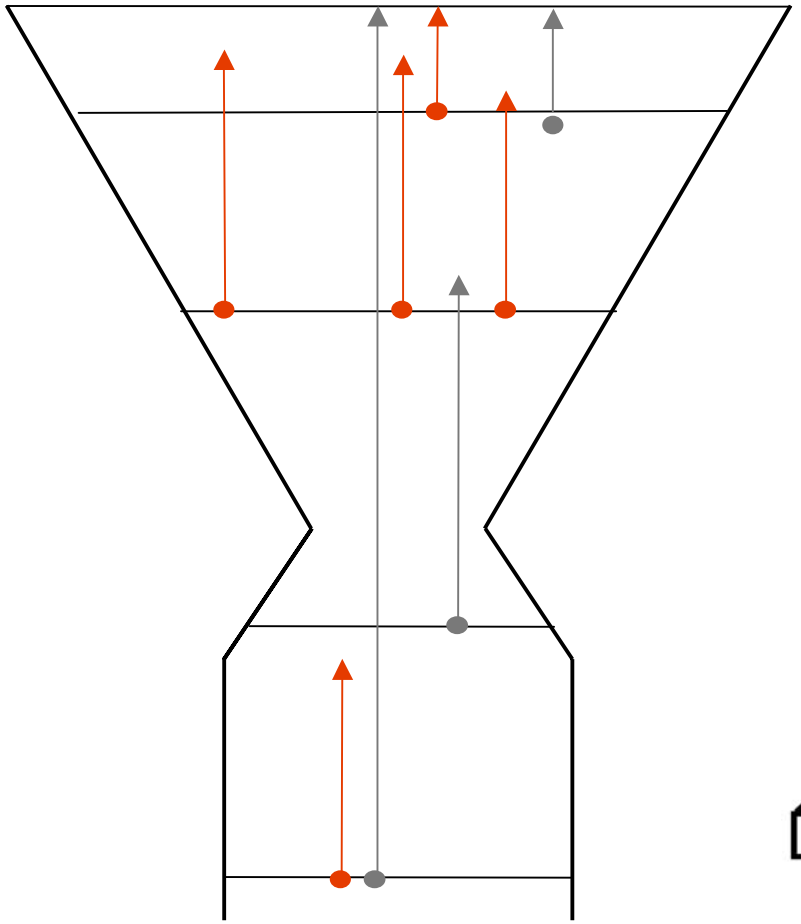
Mixture of neutral and functional sites

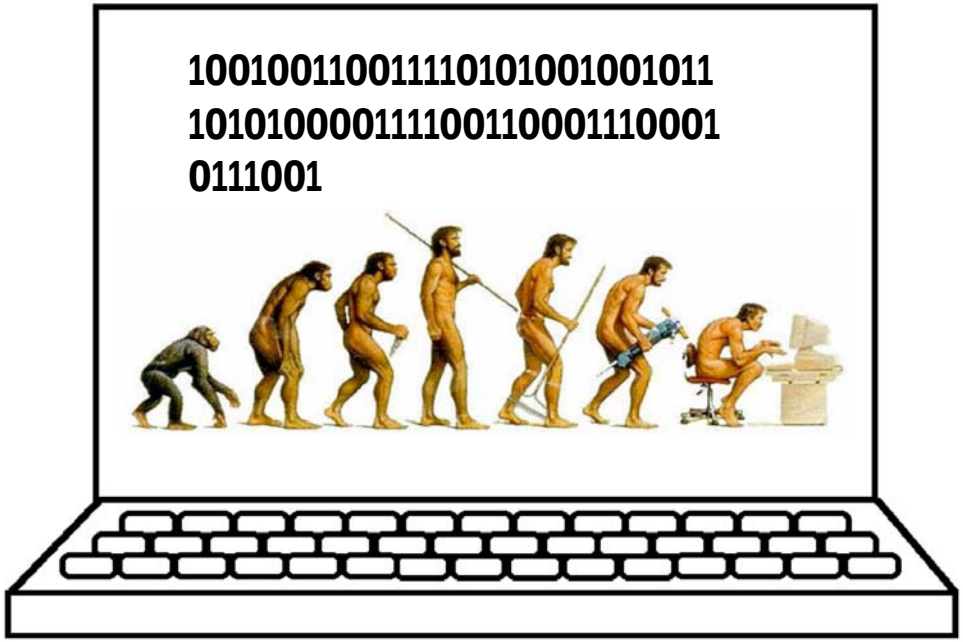
$$F_{mixture}(1\%) = \frac{\alpha \cdot n_{functional}(1\%) + \beta \cdot n_{neutral}(1\%)}{\alpha \cdot n_{functional} + \beta \cdot n_{neutral}}$$

$$n_{functional}(1\%) = \int_0^1 \frac{\theta\left(e^{-2N_e s(1-x)} - 1\right)}{x(1-x)\left(e^{-2N_e s} - 1\right)} \cdot \left[mx(1-x)^{m-1} + \frac{m(m-1)}{2} x^2 (1-x)^{m-2} \right] \cdot dx$$

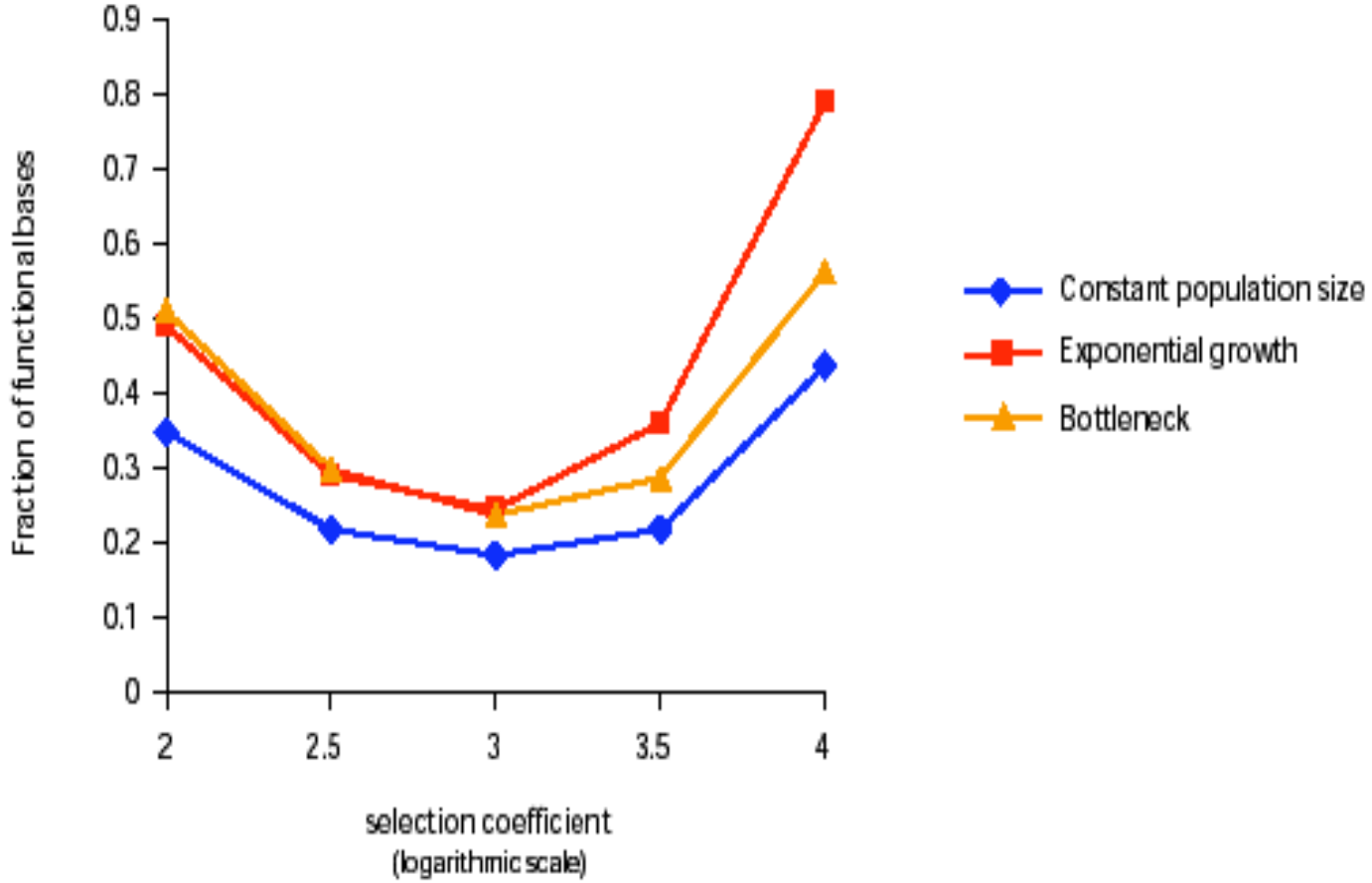
$$n_{functional} = \int_0^1 \frac{\theta\left(e^{-2N_e s(1-x)} - 1\right)}{x(1-x)\left(e^{-2N_e s} - 1\right)} \left(1 - x^m - (1-x)^m\right) \cdot dx$$

Direct simulation





How many functional sites are needed to produce observed allele frequency shift?



Polymorphism to divergence ratio

$$\pi = \alpha \cdot 4 N_e \mu \frac{2N_e s + e^{-2N_e s} - 1}{N_e s\left(1 - e^{-2N_e s}\right)} + \beta \cdot 4 N_e \mu$$

$$D = \alpha \cdot 2 N_e \mu t \cdot \frac{1 - e^{-2N_e s}}{1 - e^{-4N_e s}} + \beta t \mu$$

$$R = \frac{\alpha \cdot \dfrac{2N_e s + e^{-2N_e s} - 1}{N_e s \cdot \left(1 - e^{-2N_e s}\right)} + \beta}{\alpha \cdot 2N_e \dfrac{1 - e^{-2N_e s}}{1 - e^{4N_e s}} + \beta}$$

Selective constraints in non-coding regions of the genome

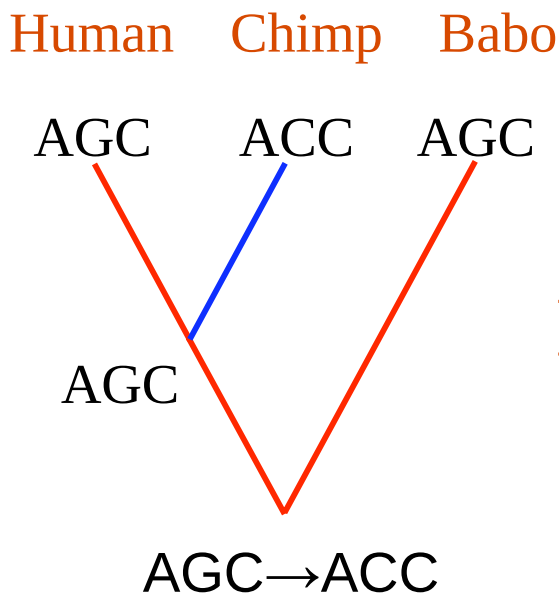
- Selectively constrained bases are diffusely distributed along the genome rather than condensed to highly conserved regions
- At least ~20% of 4GCBs are electively constrained (2% of the genome sequence)
- Probably additional constrained positions in non-alignable regions

Questions:

- How many nucleotides in the human genome are selectively constrained?
- How are selectively constrained nucleotides distributed along the genome?
- How strong is selective pressure in non-coding regions?
- Do comparative and functional genomics data agree?

Regions selected for the ENCODE project have 22 mammalian species sequenced

… and a lot of functional genomics data



Mutation rates are modeled as asymmetric and context specific.

The model incorporates insertions and deletions

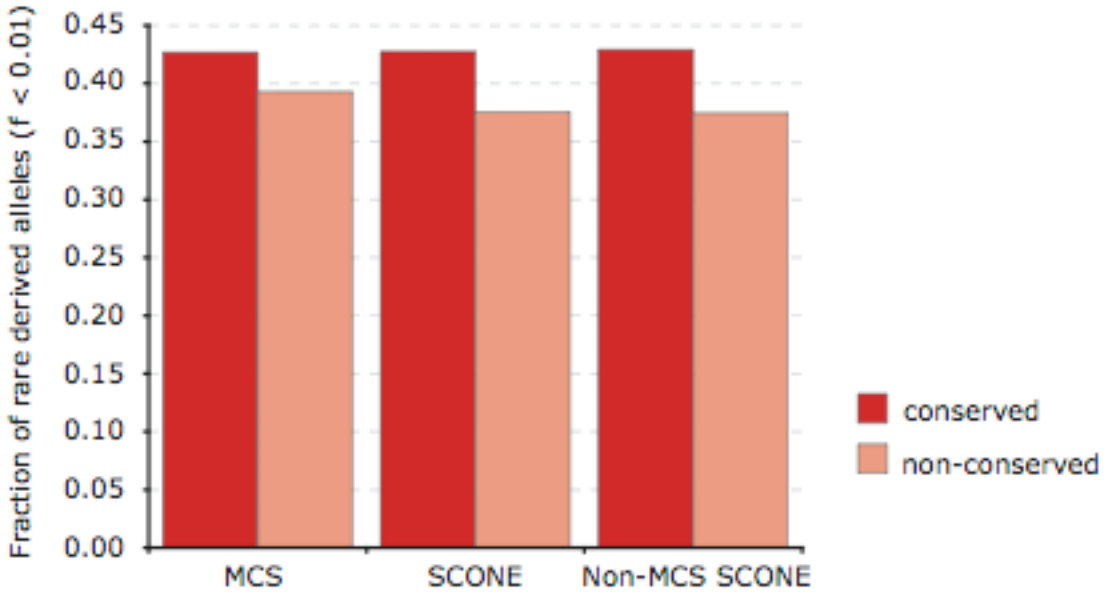
SCONE (Sequence CONservation Evaluation)

Instantaneous rate matrix of transitions Q

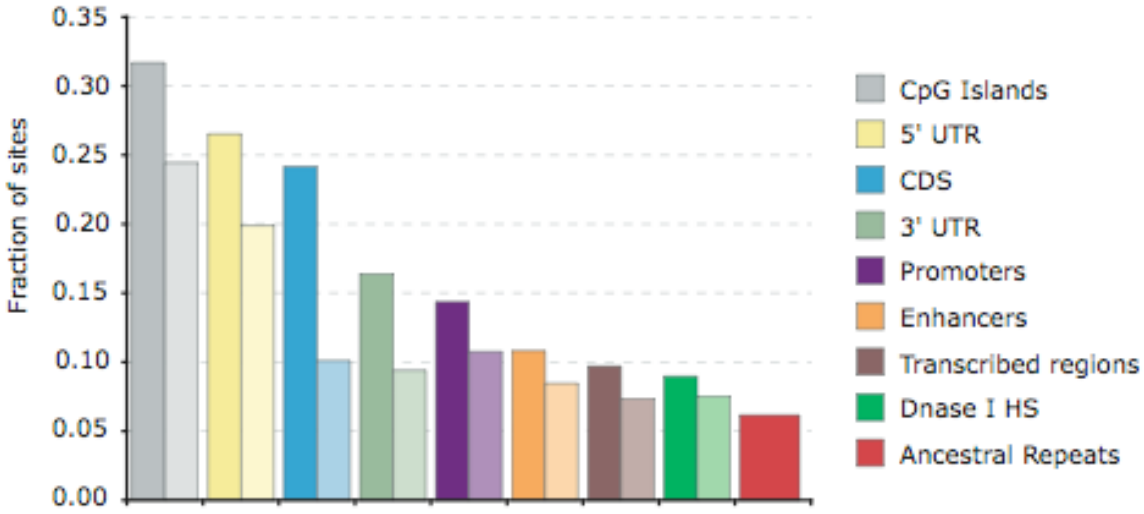
$$P(t) = e^{Qt}$$

- Ignores mutation rate heterogeneity along the genome
- Assumes uniformity between species
- Computes Bayesian estimate of evolutionary rate at the site
- Computes p-value via simulations

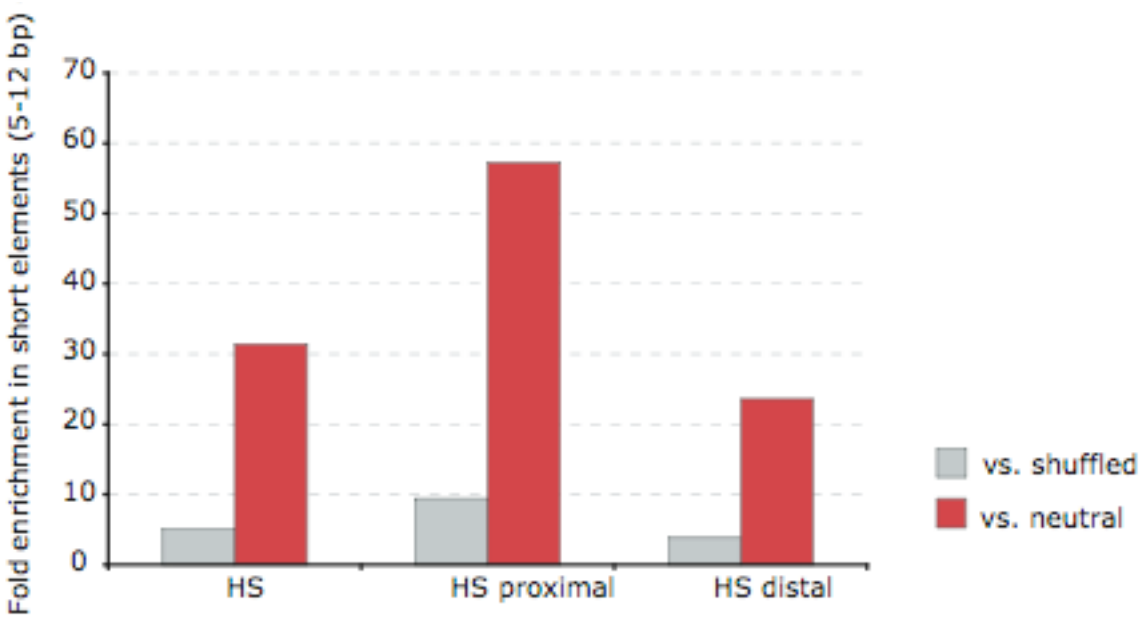
SCONE vs. ENCODE SNPs



Conservation of functional features



Clustering of conserved positions



Conservation of functional feature

- Conservation of most of ENCODE functional elements is due to a small number of positions.
- Most of conserved positions are outside of long conserved elements
- Conserved positions tend to cluster along the sequence

Acknowledgments

The lab: Saurabh Asthana,
Gregory Kryukov, Steffen Schmidt

University of Washington:
John Stamatoyannopoulos, William Noble