Exercises

1) When does the substitution rate depend on population size? When there is selection, otherwise K = 2ut

2) dN/dS = 0.34 for a gene. Is this most consistent with positive or negative selection? If 66% of nonsynonymous sites are under strong purifying selection, what is dN/dS for the remaining sites? 34% have dN/dS ~ 0 so remaining must be 1: 0.34 = 0*.66 + X*0.34

3) What can cause dN/dS > 1? positive selection on dN or negative selection on dS

4) Positive selection is acting on a gene, but dN/dS = 0.34 for the gene over the phylogenetic tree. What are three ways dN/dS could be used to detect selection? branch model, site model, branch*site model, sliding window

5) Why is mutation not a good explanation for dN != dS? interspersed

6) Why are codons with multiple substitutions on a single branch not included when detecting selection? multiple paths that differ in # syn and nonsyn changes

Exercises

1) Why is dN/dS high on trunk (influenza) compared to leaves (HIV) of a tree? Why is influenza evolution predictable? dN/dS is high on trunk to escape immune-resistant individuals whereas dN/dS is high on leaves to escape immune detecting within an individual rather. Influenza is predictable based on successful lineages in prior year

2) Why does recombination cause problems in phylogenetics? when lineages recombine there are multiple gene trees and assumptions of phylogenetics are violated

Schedule

Following week:

- 2/28 recitation, exercises, questions
- 3/2 in class exam 1:15 minutes
- Midterm exam
  - ~ 20 questions
  - Lectures 1-12 (end of next week)
  - Equations provided, calculator not needed
  - No phone, notes
- Lab06 is due Friday 3/4 midnight, not Tuesday

Today's objectives

- Comparative genomics overview
- Identifying conserved sequences
- Reality (lots of problems)
    - Mutation rate variation
    - Methylation and CpG sites
    - Bias gene conversion
    - Codon bias

Comparative genomics

1. Conservation (slower than neutral expectation)

- Annotation of genes, regulatory sequences and other functional elements
- Functional sequences will remain conserved across distantly related species whereas non-functional sequences will accumulate changes

2. Divergence

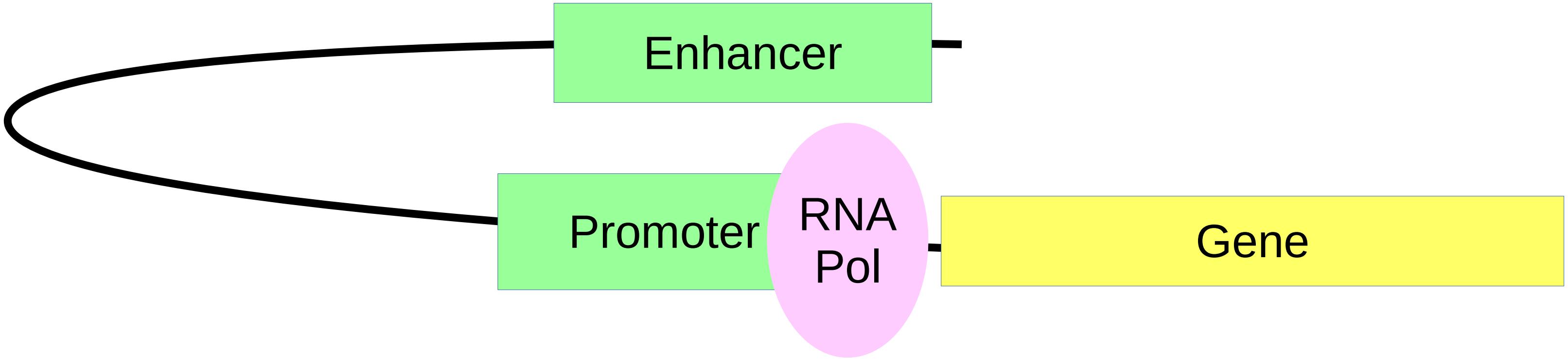
- Evolution (change) of genes, regulatory sequences and other functional elements, how they came to be
- The history of a species is written in its genome

Conservation: Identifying Functional Sequences

Protein coding genes: transcripts, gene finding algorithms, homology (works quite well)

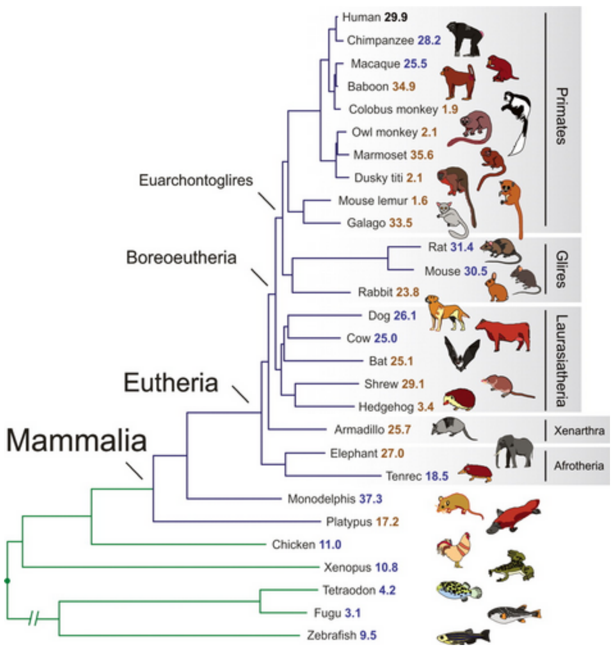
RNA genes: long (>200) noncoding RNA (lncRNA), small RNA, small nucleolar RNA (snoRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA (miRNA) (structure and conservation)

Regulatory sequences: enhancers, promoters (conservation and experimental assays)



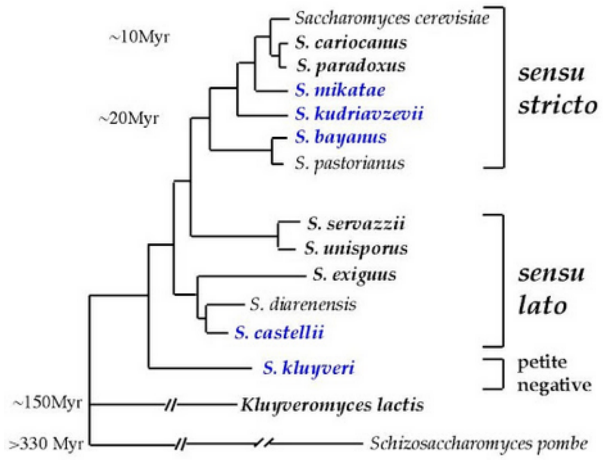
Comparative Genomics Sequencing (species related to model organisms)

Mammals (human and mouse)

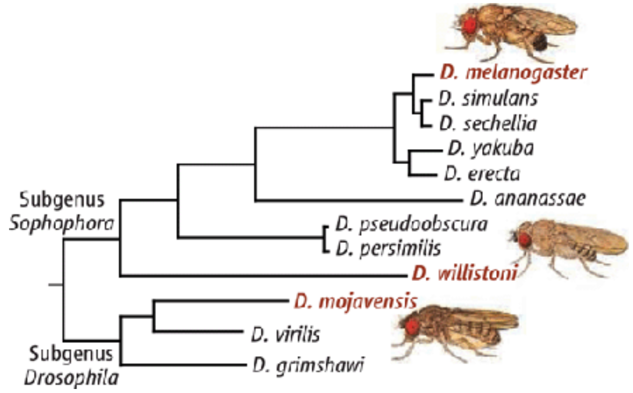


And now many others

Yeast (S. cerevisiae)



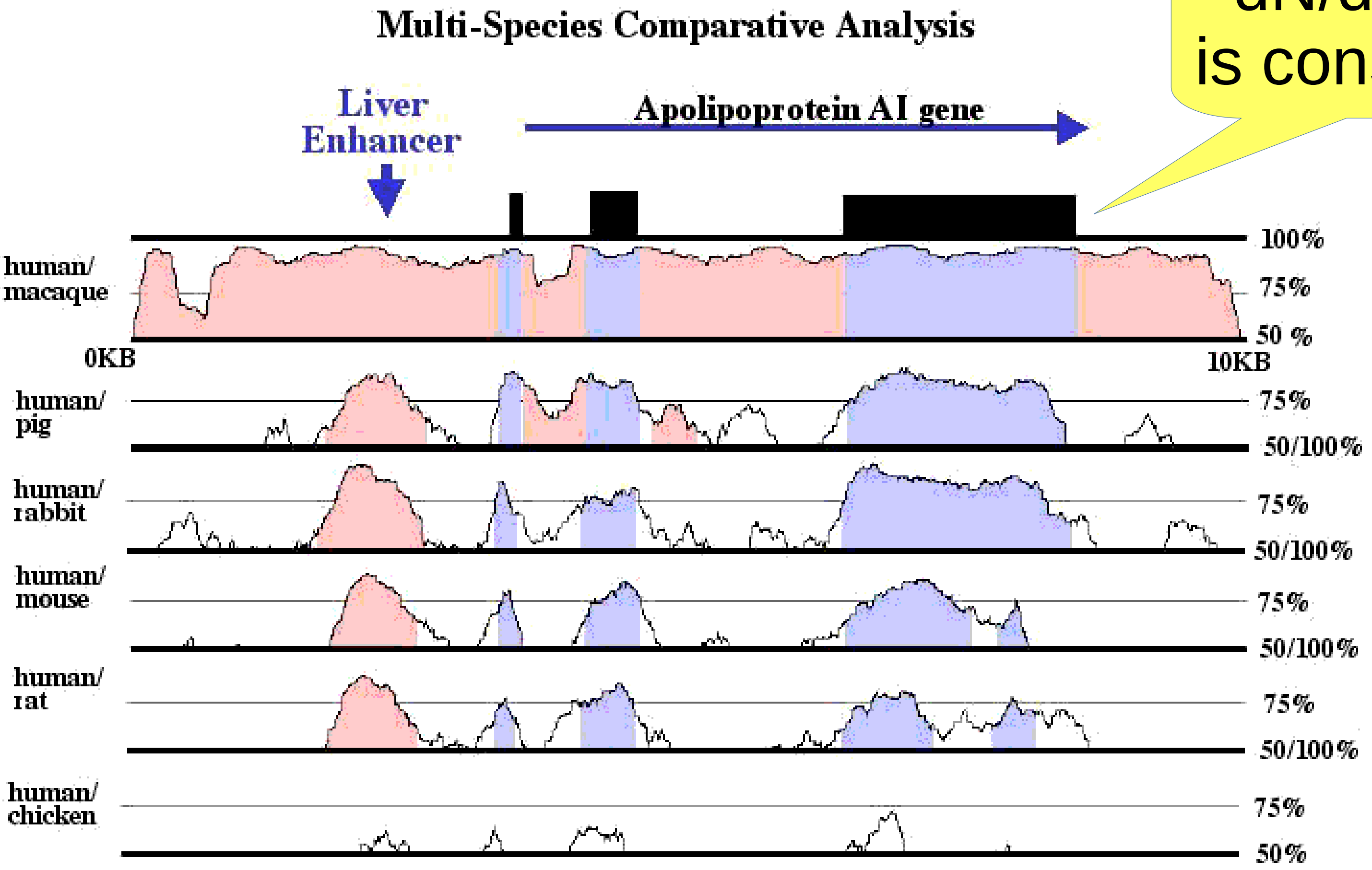
Fruit Flies (D. melanogaster)



Close vs. Distantly Related Species

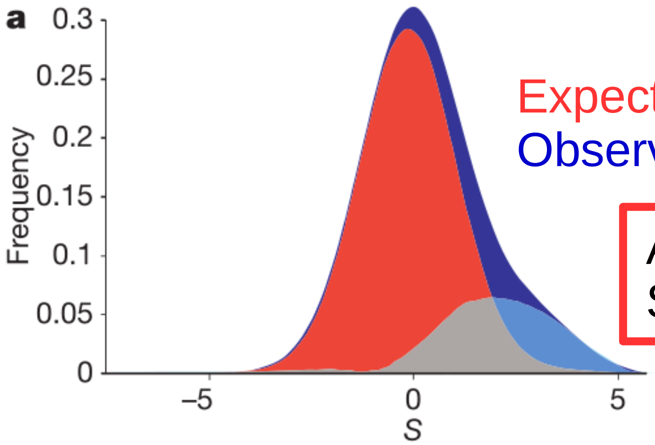
Close: easy to align, low signal to noise
Distant: misses some functional sequences

dN/dS < 1
is conserved



How do we define "conserved" noncoding?

Defining conserved sequences using Mouse and Human



Expected

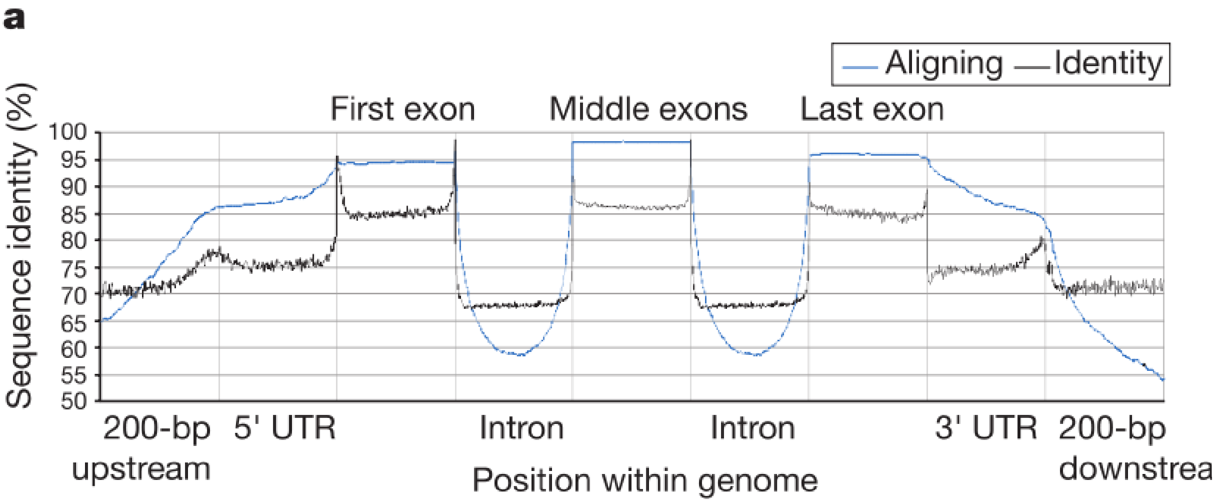
Observed

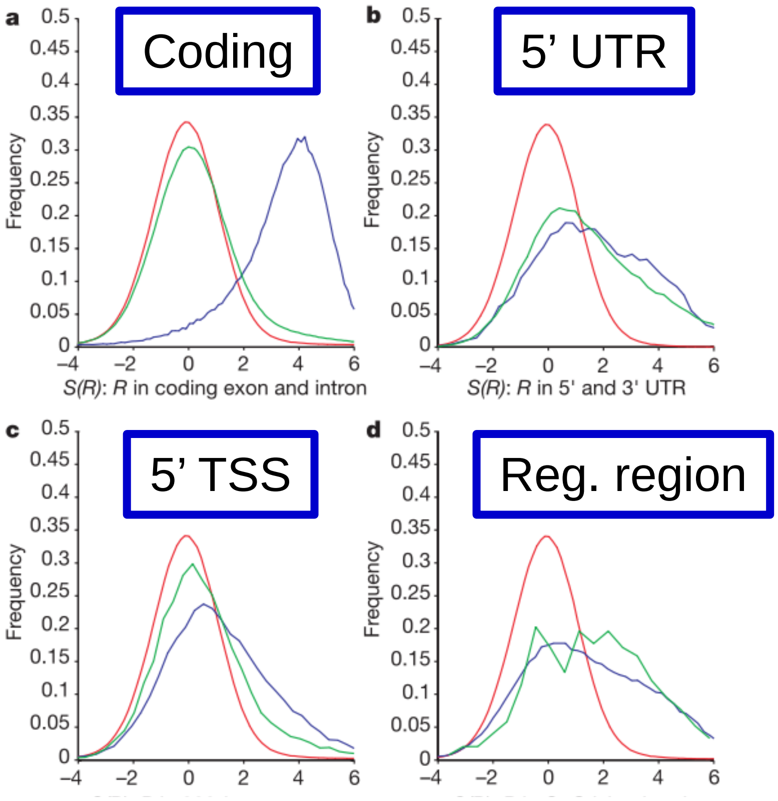
Ancestral repeats
Synonymous sites

Assumed neutral

S = conservation score (normalized substitution rate)
Observed – expected: 5% of the human genome is conserved, only 18% of conserved sequences are coding

Patterns of Substitution between Mouse and Human





Ancestral repeats
Synonymous sites

S(R) = conservation score

Problem: resolution and defining regions

Which region is conserved? (boundary)

```
ATGCATGCATGATATGCGCGTGCTTACCAGCTCGCGCGGTTATCGTCGCGC
...................T........T...T.G.T..A.T...A..G.T

ATGCATGCATGATATGCGCGTGCTTACCAGCTCGCGCGGTTATCGTCGCGC
...................T........T...T.G.T..A.T...A..G.T
```

How small of a region can be found? (resolution)

Excercise

P_{ii} = 0.45 when substitution rate = 1 subs/site, what's the probability of

three identical sites: $0.45^3 = 3.4 \times 10^{-4}$

25 identical sites: $0.45^{25} = 2.1 \times 10^{-9}$

How many 3 bp sites would you expect in the human genome $3\text{x}10^9$? $3\text{x}10^9 * 3.4 \times 10^{-4} = 10^6$

Problem: resolution and defining regions

Hidden Markov Models (HMMs)

Which region is conserved?

```
ATGCATGCATGATATGCGCGTGCTTACCAGCTCGCGCGGTTATCGTCGCGC
...................T........T...T.G.T..A.T...A..G.T

ATGCATGCATGATATGCGCGTGCTTACCAGCTCGCGCGGTTATCGTCGCGC
...................T........T...T.G.T..A.T...A..G.T
```

How small of a region can be found?

More species = higher substitution rate

Excercise

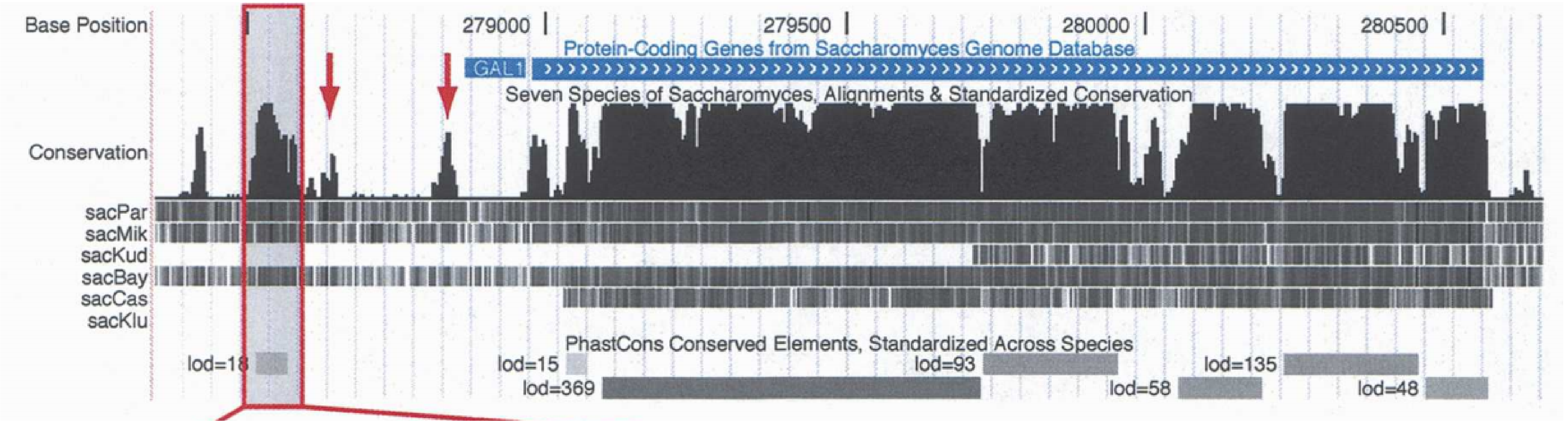
P_{ii} = 0.45 when substitution rate = 1 subs/site, whats the probability of

three identical sites: $0.45^3 = 3.4 \times 10^{-4}$

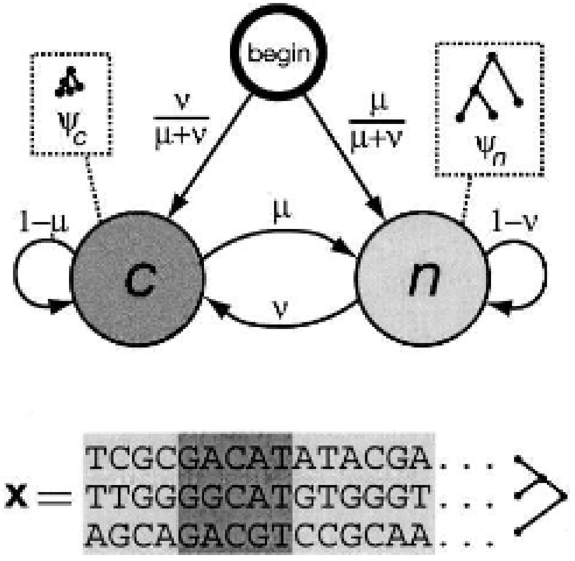
25 identical sites: $0.45^{25} = 2.1 \times 10^{-9}$

How many 3 bp sites would you expect in the human genome $3\text{x}10^9$? $3\text{x}10^9 * 3.4 \times 10^{-4} = 10^6$

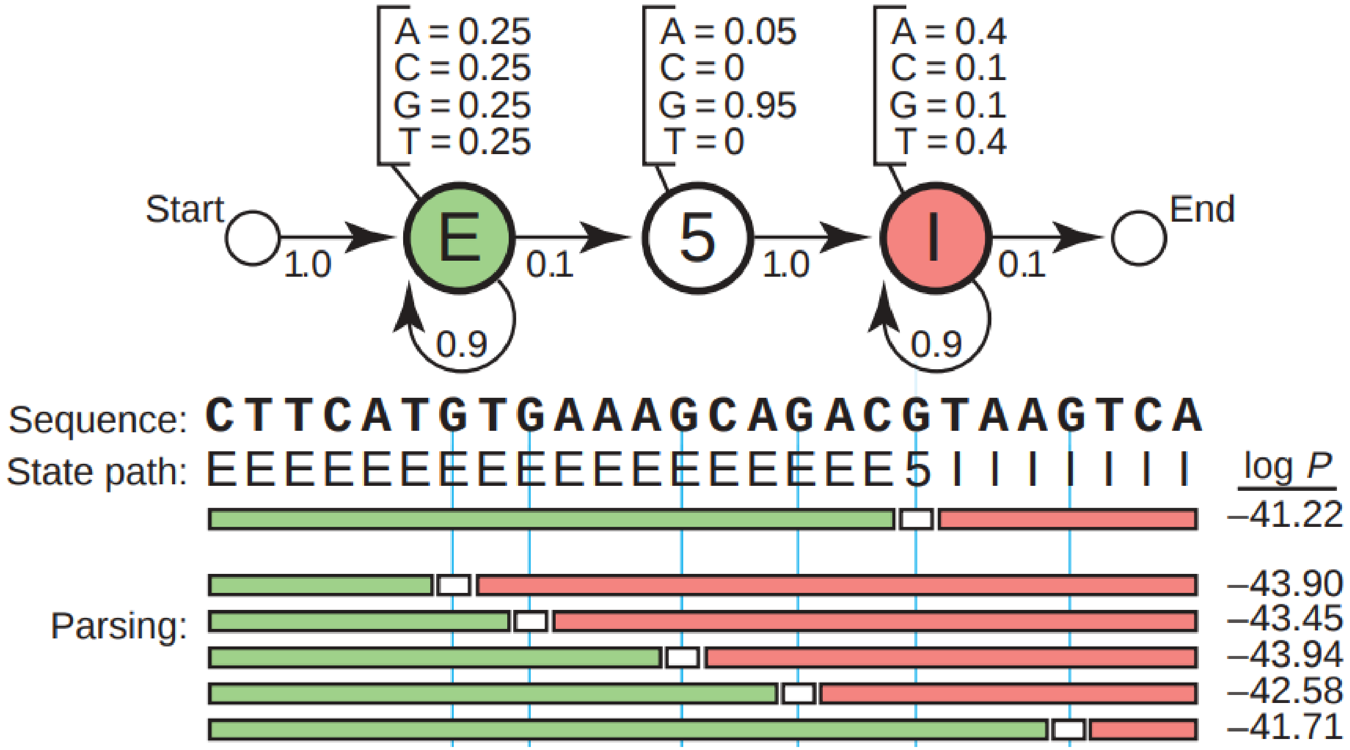
From conservation to annotation: Hidden Markov Models


Base Position
279000
279500
280000
280500
Protein-Coding Genes from Saccharomyces Genome Database
GAL1
Seven Species of Saccharomyces, Alignments & Standardized Conservation
Conservation
sacPar
sacMik
sacKud
sacBay
sacCas
sacKlu
PhastCons Conserved Elements, Standardized Across Species
lod=18
lod=15
lod=93
lod=135
lod=369
lod=58
lod=48

- Not enough data to provide meaningful scores at single nucleotide resolution
- Functional elements are block-like (but what about window size)
- Hidden Markov Models are Markov processes with a hidden (unobserved) state (in this case conserved/ unconserved)

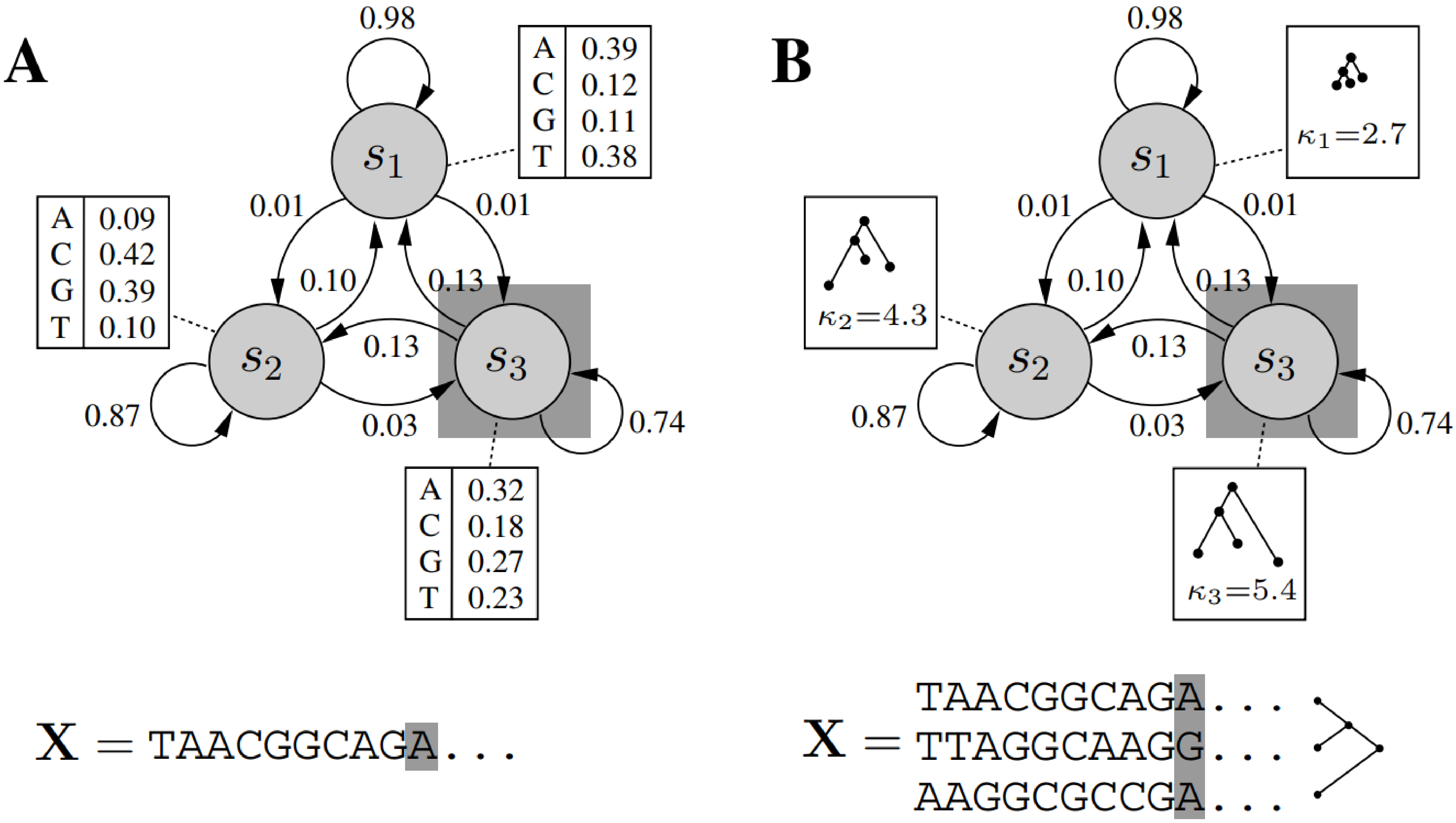


Brief intro to HMMs



- Identify exons (E) and intron (I) and 5' splice sites (5)
- Markov chain model in space (genome sequence)
- States are unobserved (hidden), emissions are observed
- Boundaries are inferred by the most likely path

Phylogenetic Hidden Markov Model: Markov chains in time and space

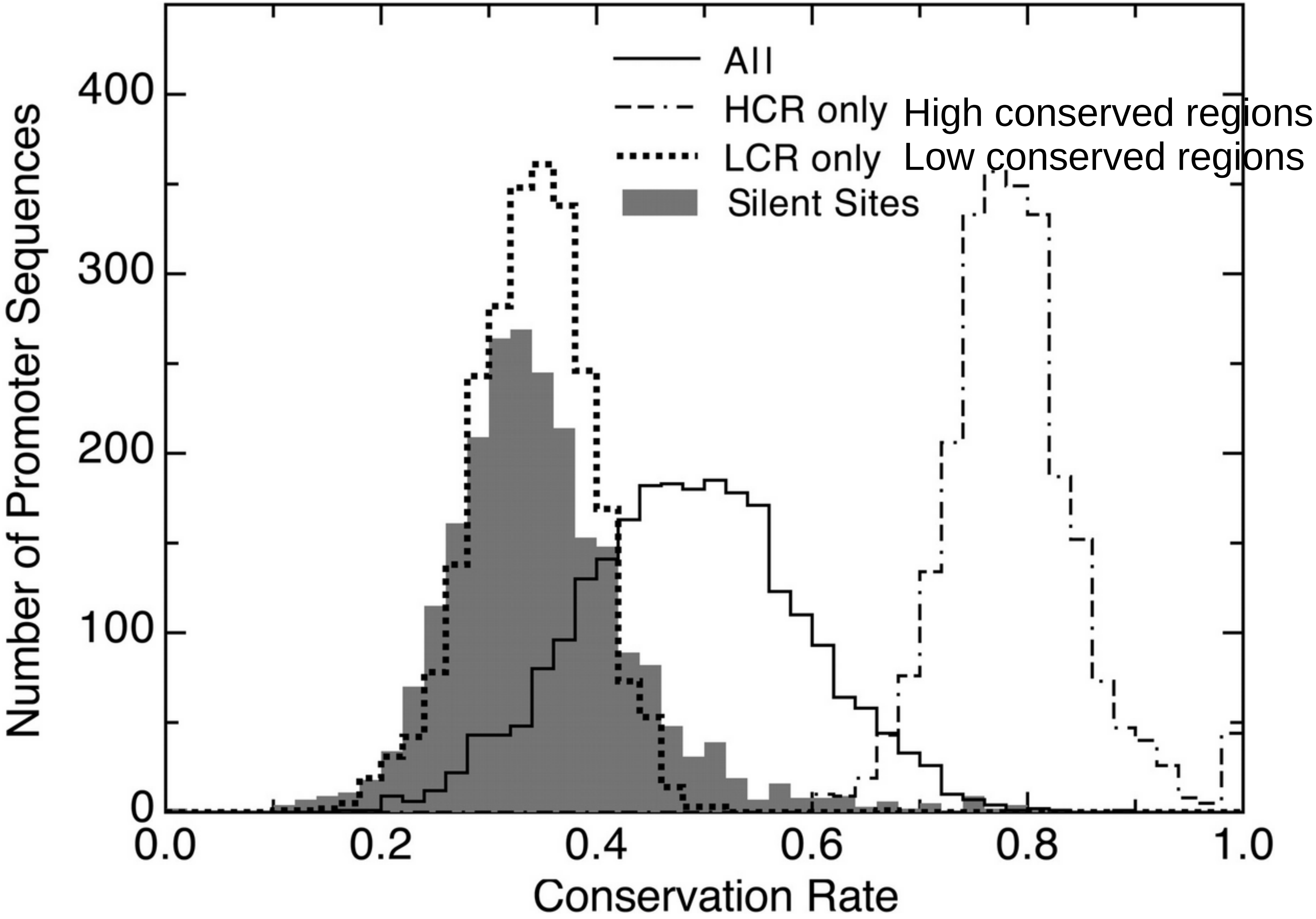

A
B
0.98
s_1
s_2
s_3
0.01
0.01
0.10
0.13
0.13
0.03
0.87
0.74
$\kappa_1 = 2.7$
$\kappa_2 = 4.3$
$\kappa_3 = 5.4$
$\mathbf{X} = \mathtt{TAACGGCAGA}\ldots$
$\mathbf{X} = \mathtt{TAACGGCAGA}\ldots$
$\mathtt{TTAGGCAAGG}\ldots$
$\mathtt{AAGGCGCCGA}\ldots$

A	0.39
C	0.12
G	0.11
T	0.38

A	0.09
C	0.42
G	0.39
T	0.10

A	0.32
C	0.18
G	0.27
T	0.23

Silent Sites Compared to 2 state HMM

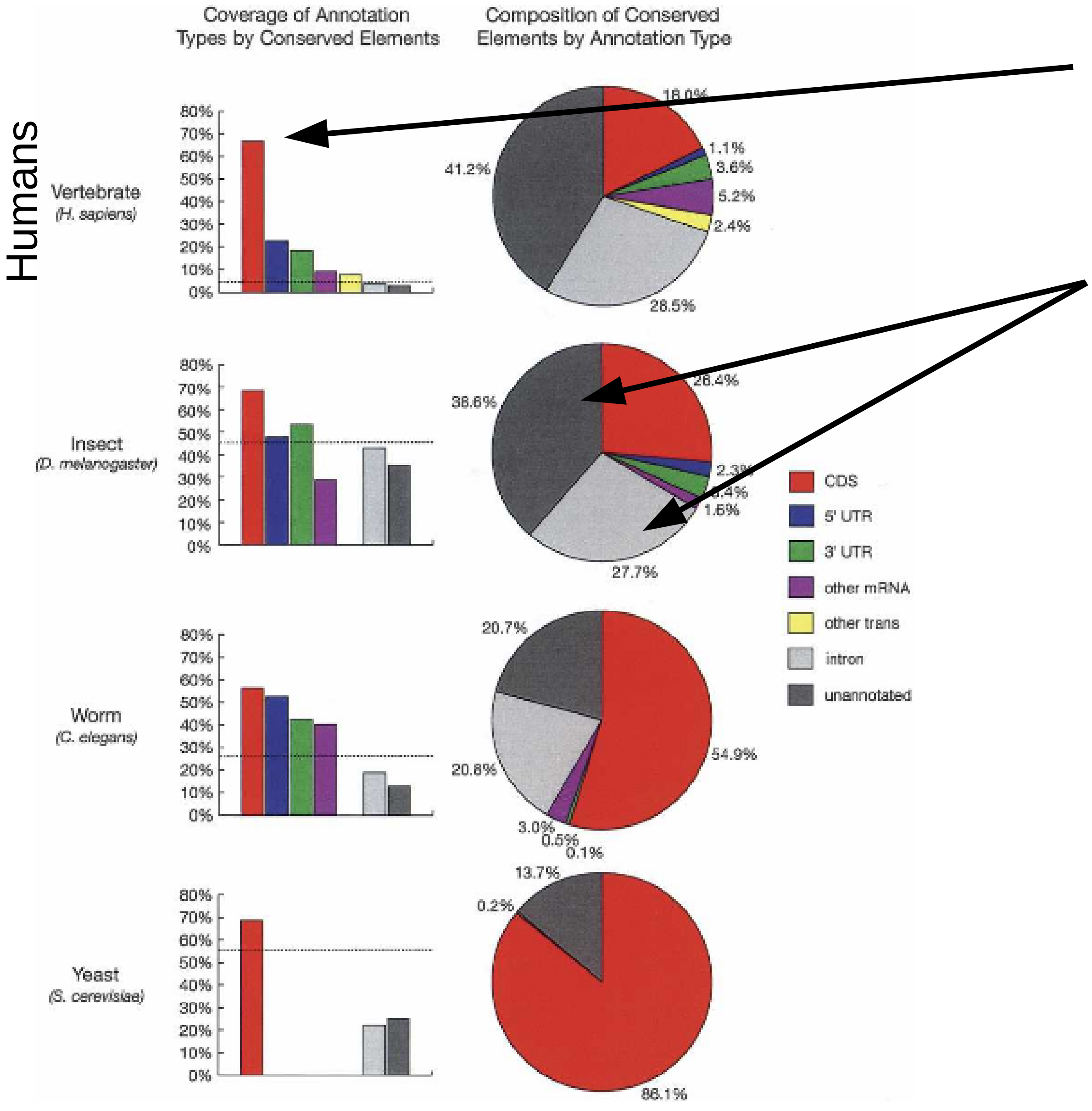


High conserved regions

Low conserved regions

Chin et al. 2005

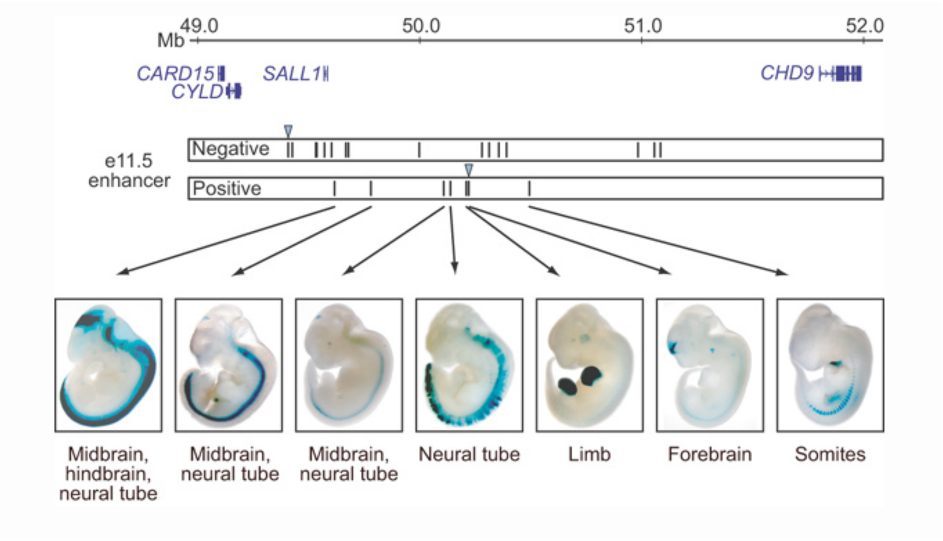
Annotation of Conserved Sequences using a PhyloHMM



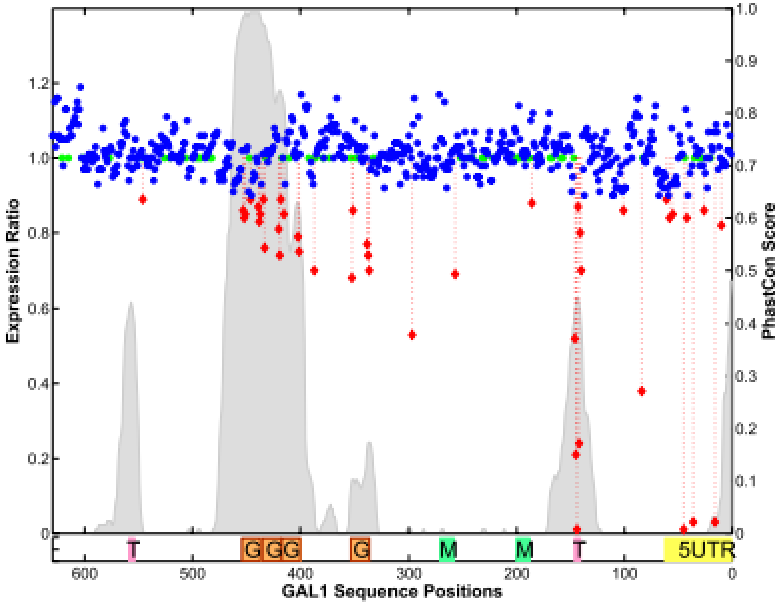
- Most coding sequences are conserved
- Most conserved sequences are noncoding in animals
- Current evidence shows most conserved noncoding sequences are regulatory elements

Siepel et al. 2005

Expression assays of conserved noncoding sequences



Enhancer reporter assays



Saturation mutagenesis promoter assays

Pennacchio et al. 2006

Yun et al. 2012

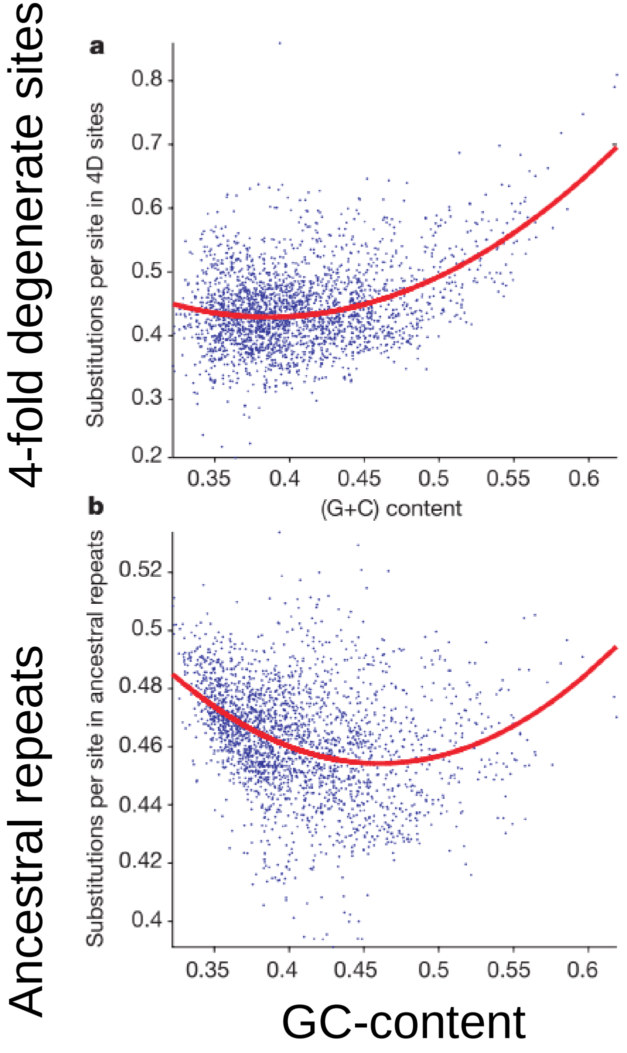
Problem:
GC content ~ substitution rate

GC-content problem:

- which regions have substitution rate less than 'neutral' rate
- the neutral rate depends on GC
- we must account for this

GC-content explanations:

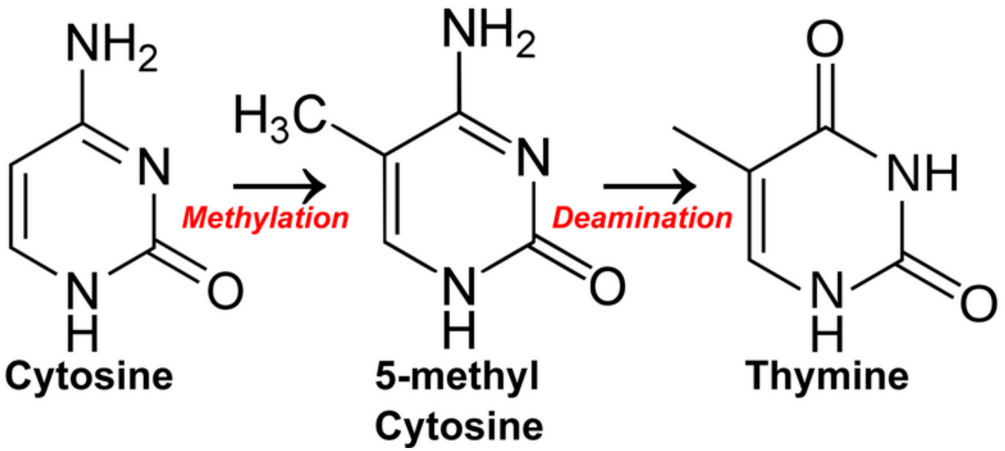
- model incorrect
- biology (mutation vs selection)
- Substitution rate = mutation rate * fixation (selection) * time

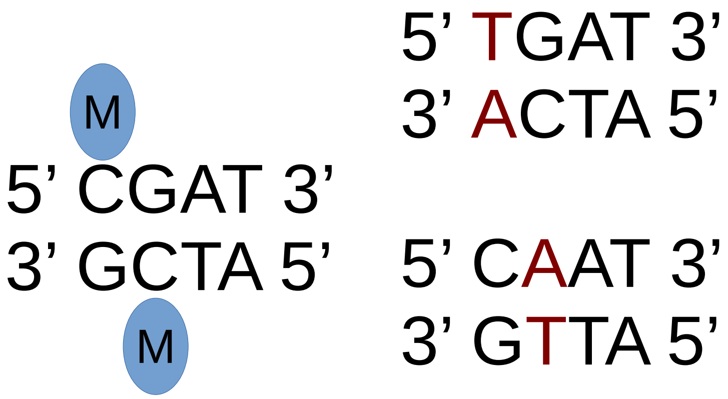


Mutation rate variation

- Transitions vs. Transversions – transitions occur twice as often as transversions
- CpG - Spontaneous deamination of 5-methylcytosine results in thymine and ammonia, 20x higher rate of transition: G/C to A/T
  - 28% of mutations are transitions at CpG sites but only 3.5% of sites are CpG
- Genomic position (5-10%)
- Age, sex (2 – 10 fold)
- Repeats (polynucleotides, microsatellites)

CpG = 5'—C—phosphate—G—3', rather than CG base pairing





5’ TGAT 3’
3’ ACTA 5’

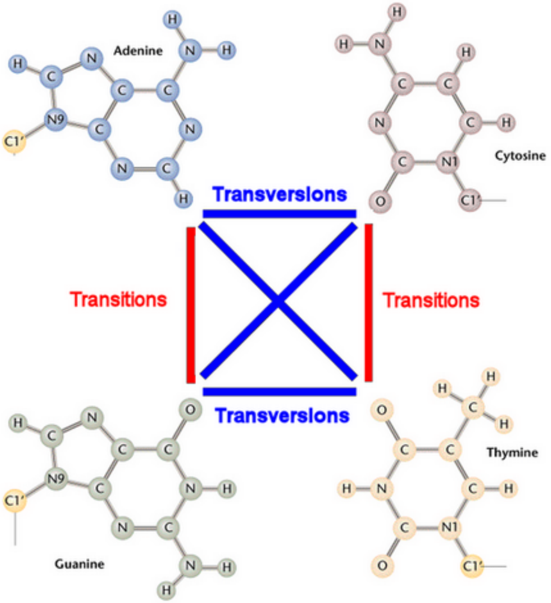
5’ CGAT 3’
3’ GCTA 5’

5’ CAAT 3’
3’ GTTA 5’

Mutation rate variation

- Transitions vs. Transversions – transitions occur twice as often as transversions
- CpG - Spontaneous deamination of 5-methylcytosine results in thymine and ammonia, 20x higher rate of transition: G/C to A/T
  - 28% of mutations are transitions at CpG sites but only 3.5% of sites are CpG
- Genomic position (5-10%)
- Age, sex (2 – 10 fold)
- Repeats (polynucleotides, microsatellites)

CpG = 5'—C—phosphate—G—3', rather than CG base pairing



Age: disease mutations increase with age

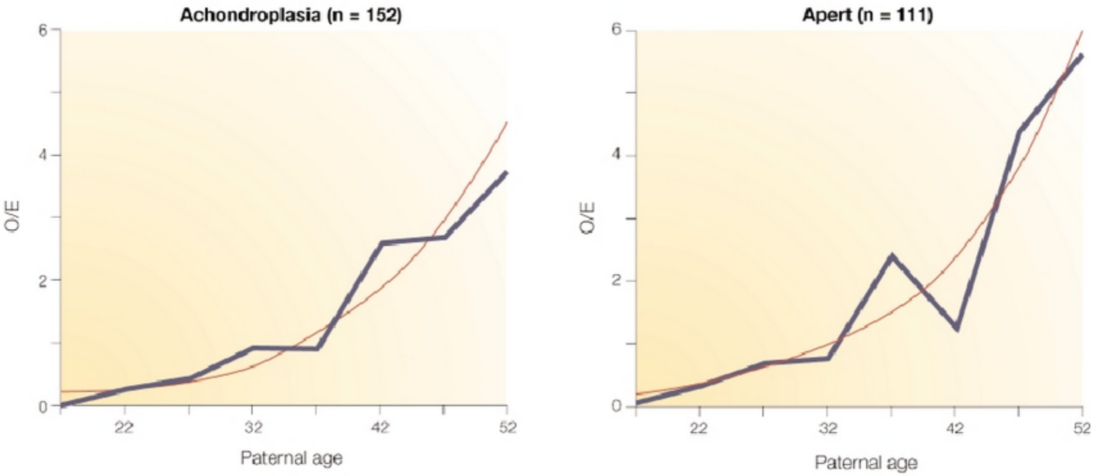
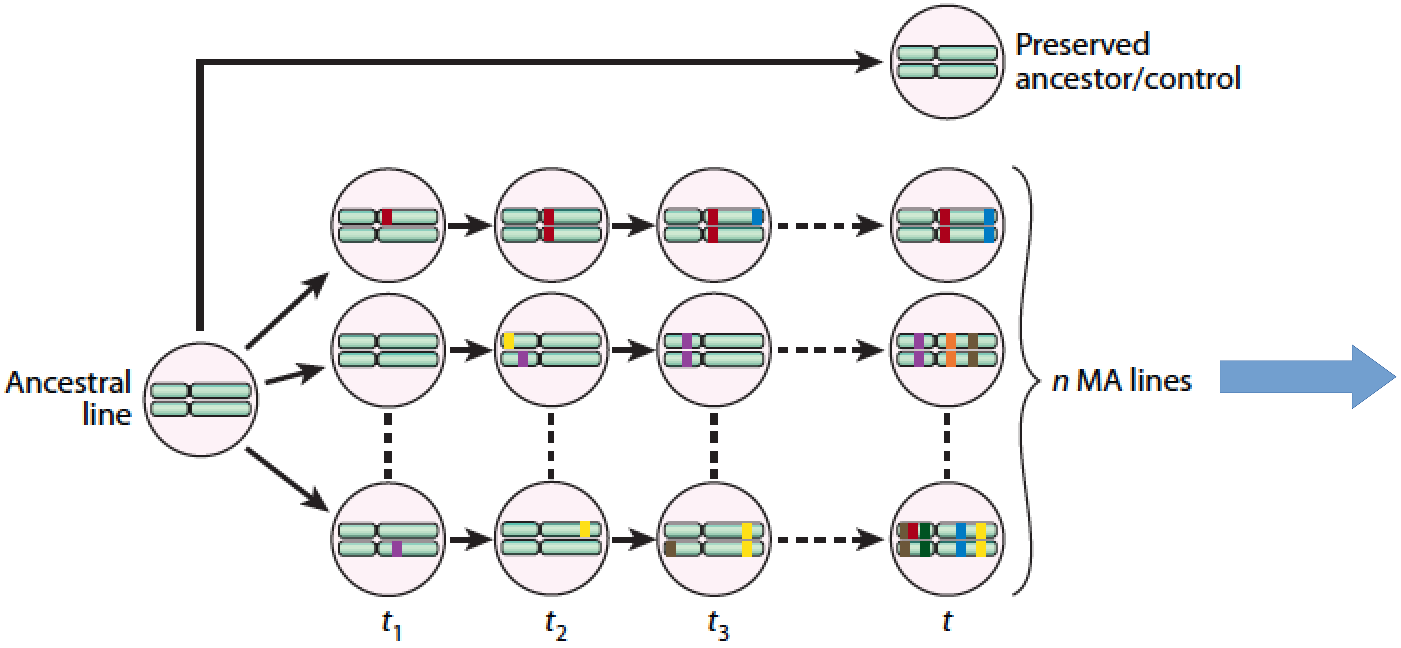


Figure 3 | **Relative frequency of de novo achondroplasia and Apert syndrome for different paternal ages.** The ordinate is the ratio of the observed number of mutations (O) to the number expected (E), if all paternal ages are associated with the same frequency of mutation. The blue line gives the actual data; the red line is the best-fitting exponential curve. (Figure adapted from REF. 4.)

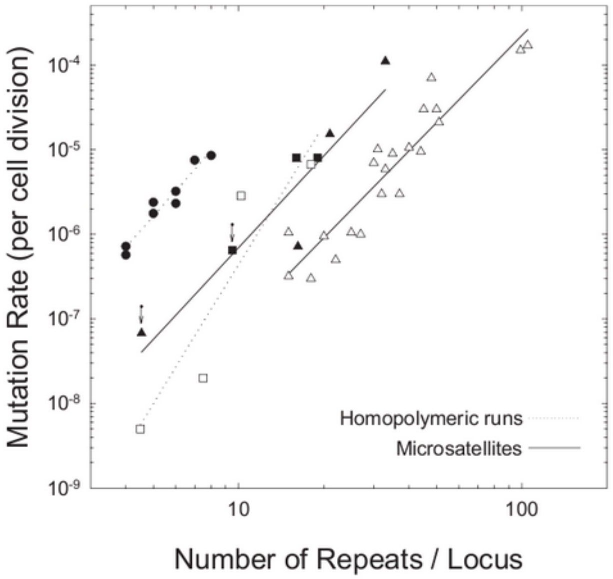
Mutation accumulation (MA): measuring mutational biases/rates



Sequencing

Microsatellite = Short repeats, e.g. ACACAC = AC_3 (2-6 bp motif)
Homopolymeric run = AAAAA = A_5

Mutation accumulation



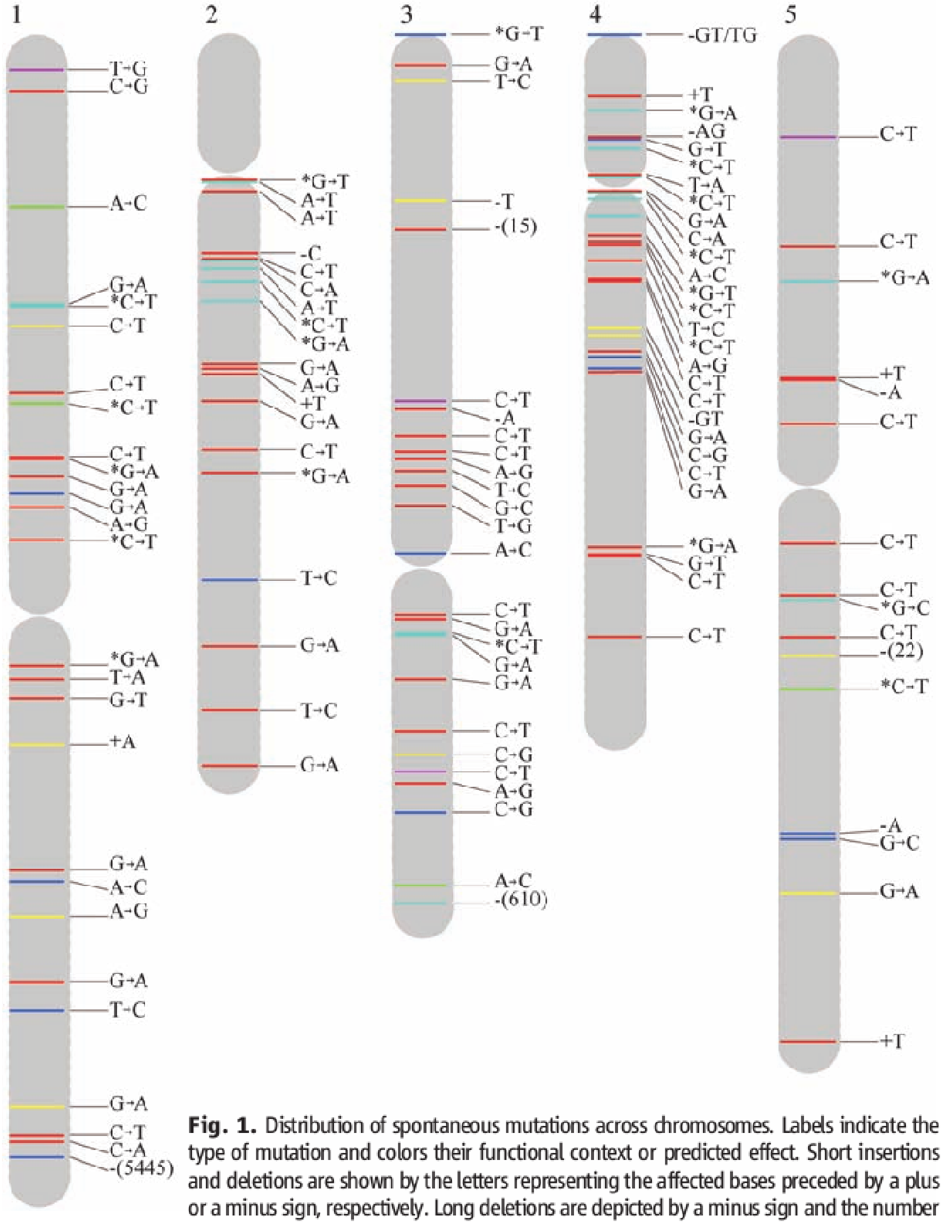


Fig. 1. Distribution of spontaneous mutations across chromosomes. Labels indicate the type of mutation and colors their functional context or predicted effect. Short insertions and deletions are shown by the letters representing the affected bases preceded by a plus or a minus sign, respectively. Long deletions are depicted by a minus sign and the number

- •G/C to A/T 2.9-fold higher than reverse!
- •Predicts 74% AT content, more than obs
- •Does selection act on GC content? Maybe, but there are other explanations

What is the cause of variation in CpG across the genome

- CpG dinucleotides occur with a much lower frequency in vertebrate genomes than would be expected due to random chance.

- Human genome: 42% GC content, CpG expected to occur 0.21 * 0.21 = 4.41% of the time. Observed frequency of CpG dinucleotides is 1%.

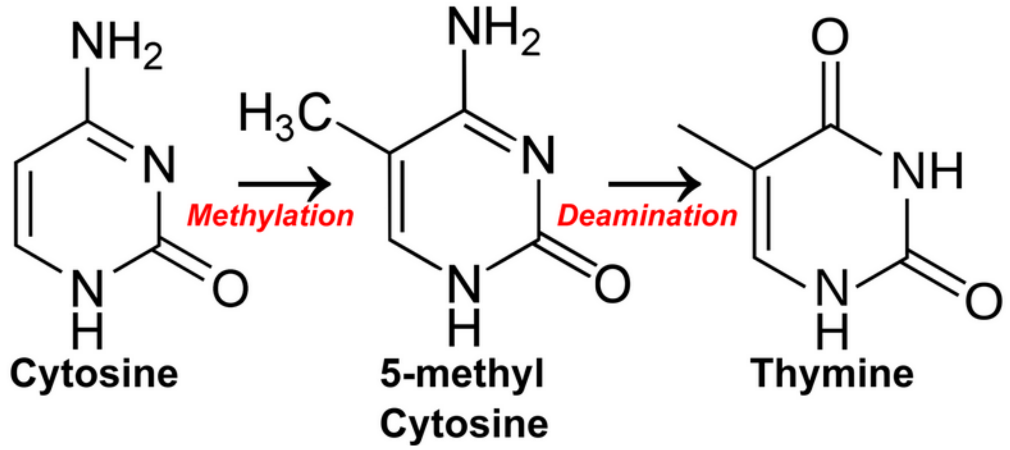
- CpG islands (300-3000 bp) are regions with a high frequency of CpG sites

- found in or near approximately 40% of promoters of mammalian genes

- methylated CpG sites in CpG islands of promoters causes stable silencing of genes

Methylation and the epigenome

CpG = 5'—C—phosphate—G—3',
rather than CG base pairing



M
5’ CGAT 3’
3’ GCTA 5’
M

5’ TGAT 3’
3’ ACTA 5’

5’ CAAT 3’
3’ GTTA 5’

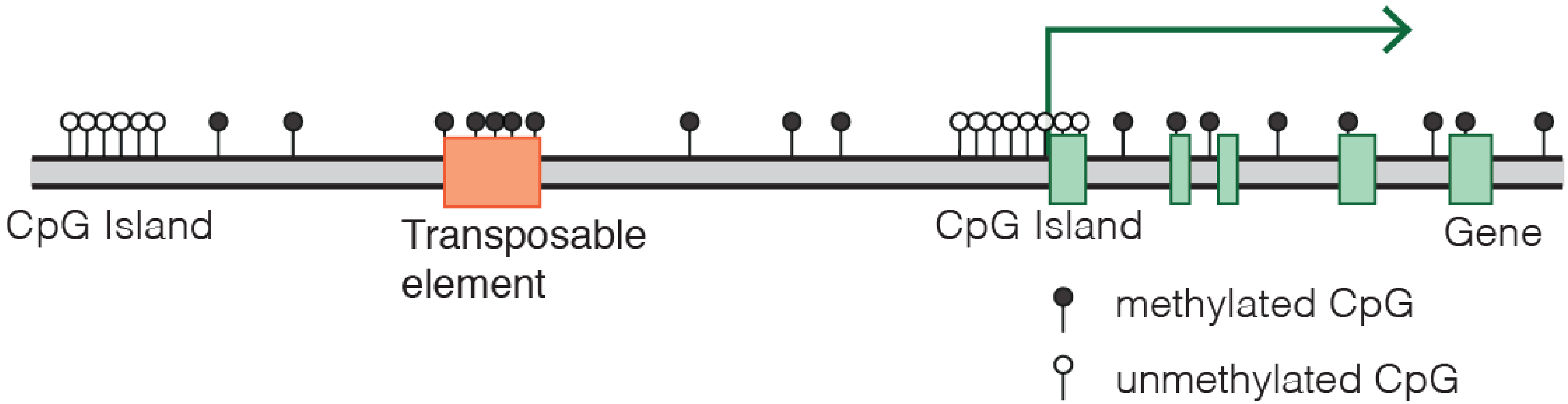
G:C to A:T mutations
CpG sites are destroyed!

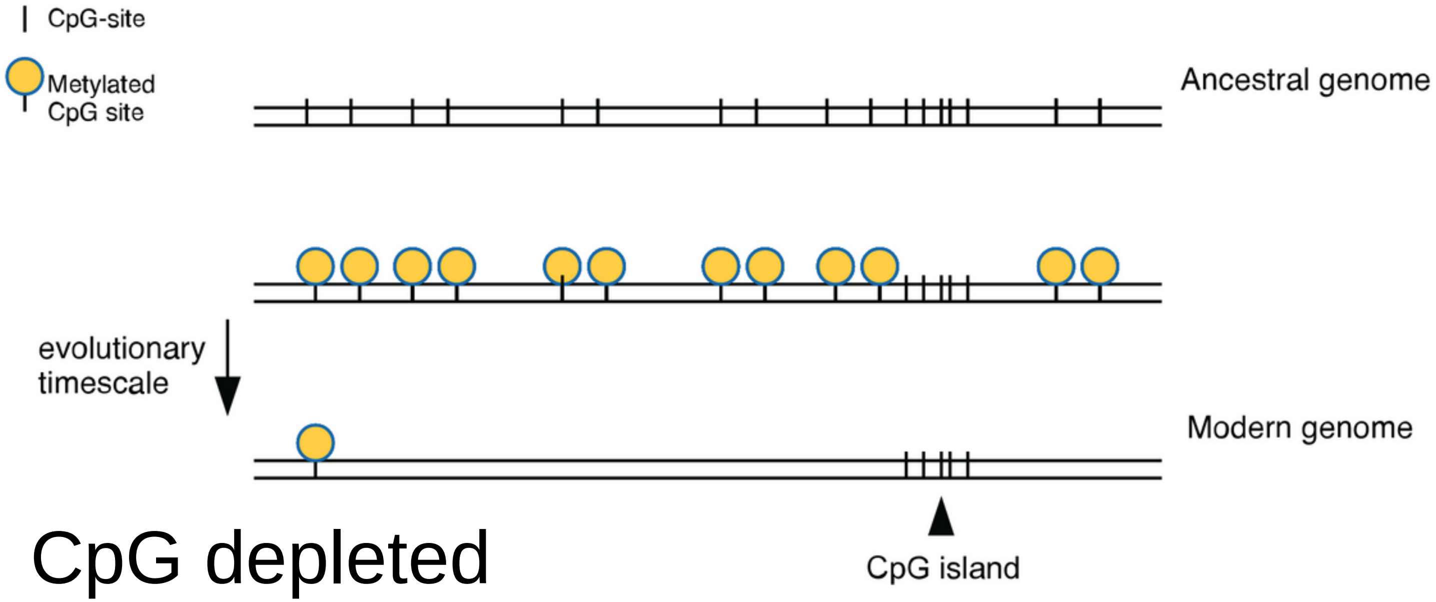
Why might selection be involved in maintenance of CpG sites

- Methylation causes gene silencing, in cancer, imprinting, X-chromosome inactivation, repression of transposable elements
- DNA methylation patterns are largely erased and then re-established between generations in mammals. Methylation is essential for differentiation.

Methylation and CpG sites

- Most CpG sites are methylated, but most CpG islands are un-methylated. In vertebrates, around 60–80% of CpG are methylated in somatic cells.

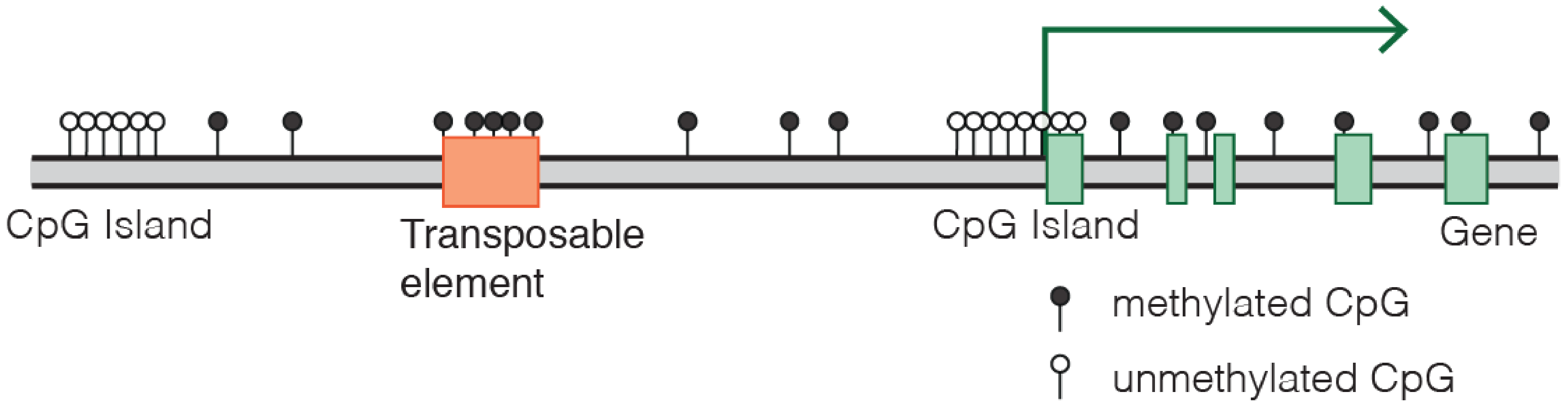


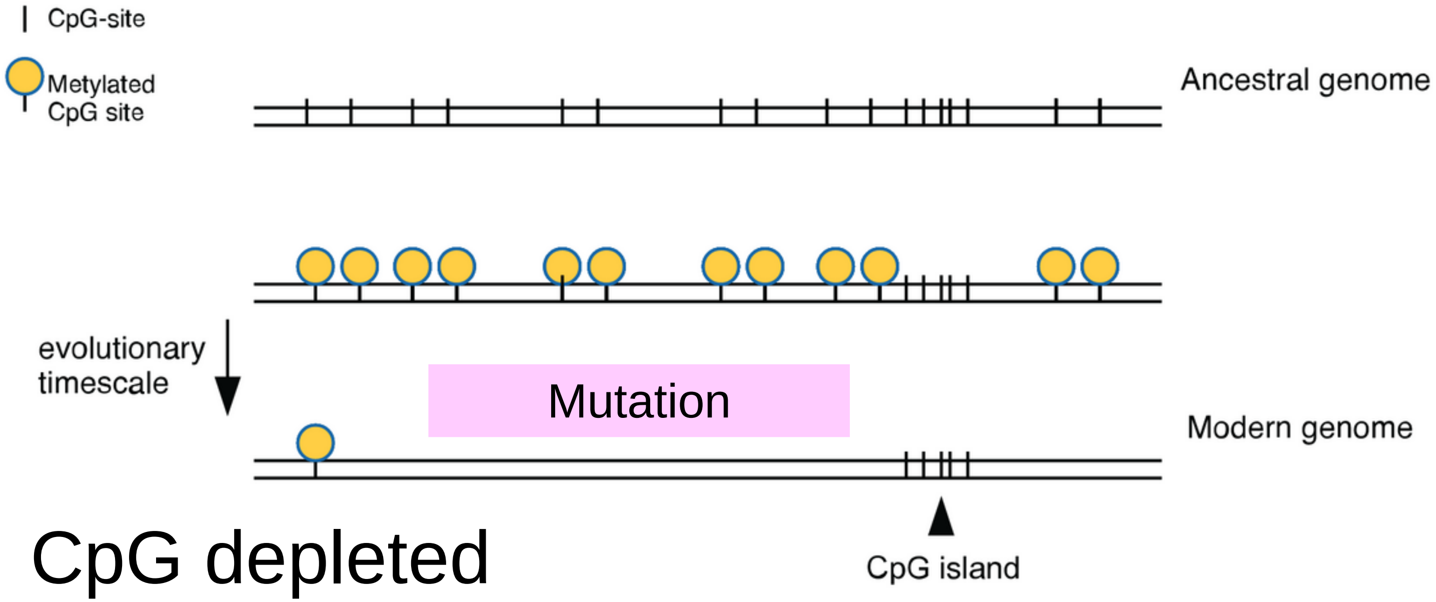


CpG depleted

Methylation and CpG sites

- Most CpG sites are methylated, but most CpG islands are un-methylated. In vertebrates, around 60–80% of CpG are methylated in somatic cells.



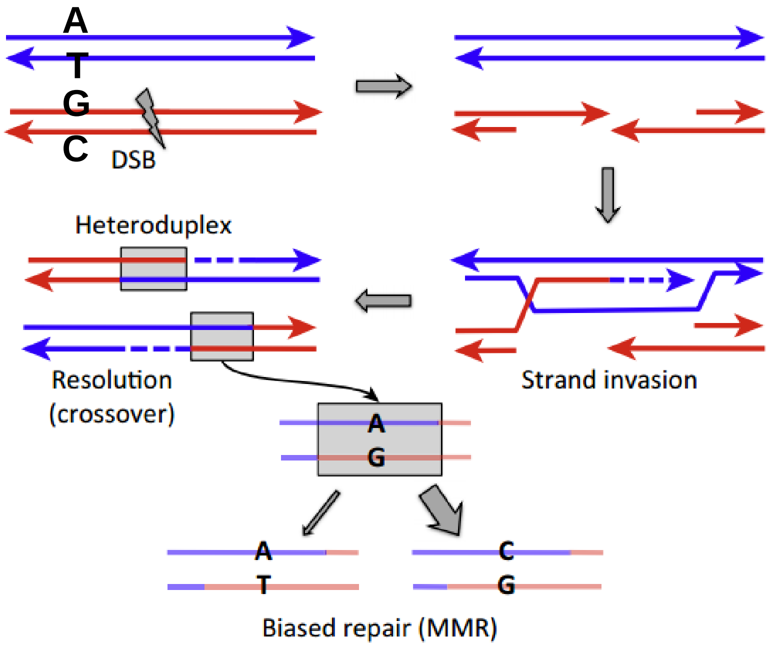


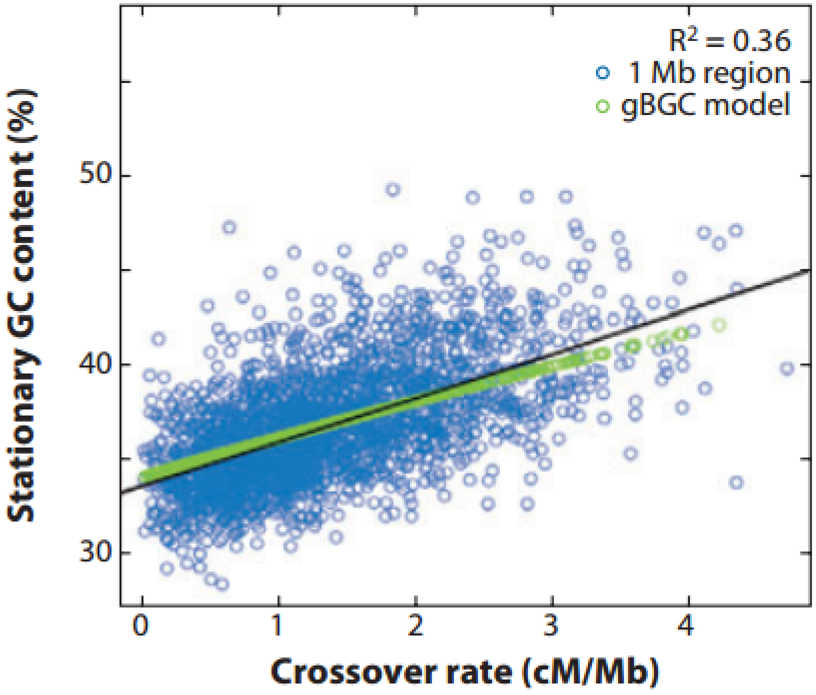
Mutation

CpG depleted

Selection

Problem: Biased Gene Conversion causes GC bias (AT to GC)





- Mismatches are more often repaired to C:G rather than A:T
- Recombination occurs in hotspots
- Recombination hotspots evolve rapidly
- Biased gene conversion occurs in bursts (non-equilibrium)

Recombination and predicted equilibrium GC frequency

Predicted

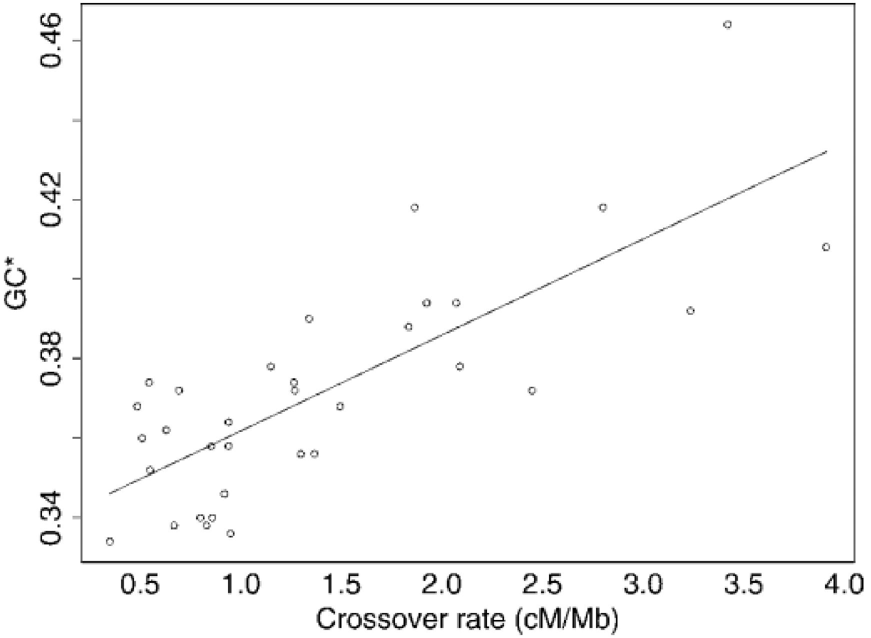


FIG. 1.—Relationship between the base composition toward which a locus is evolving (*GC**) and its crossover rate (cM/Mb; sex-averaged). N = 33 loci from 12 human autosomes. $r^2 = 0.61$, $P < 2.10^{-16}$ (Student *t*-test).

Observed

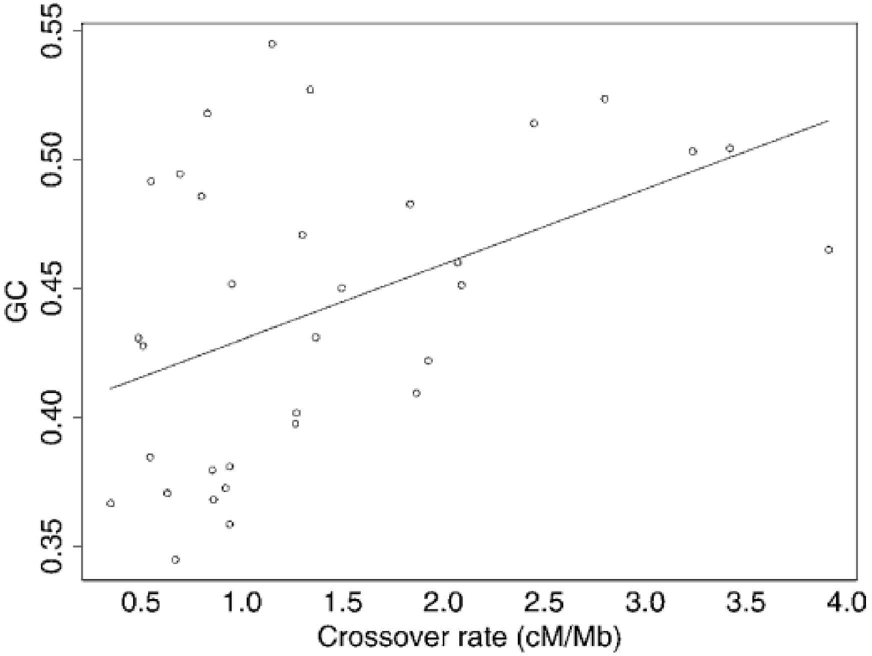


FIG. 2.—Relationship between the present base composition of a loci (*GC*) and its crossover rate (cM/Mb; sex-averaged). N = 33 loci from 12 human autosomes. $r^2 = 0.21$, $P < 2.10^{-16}$ (Student *t*-test). We

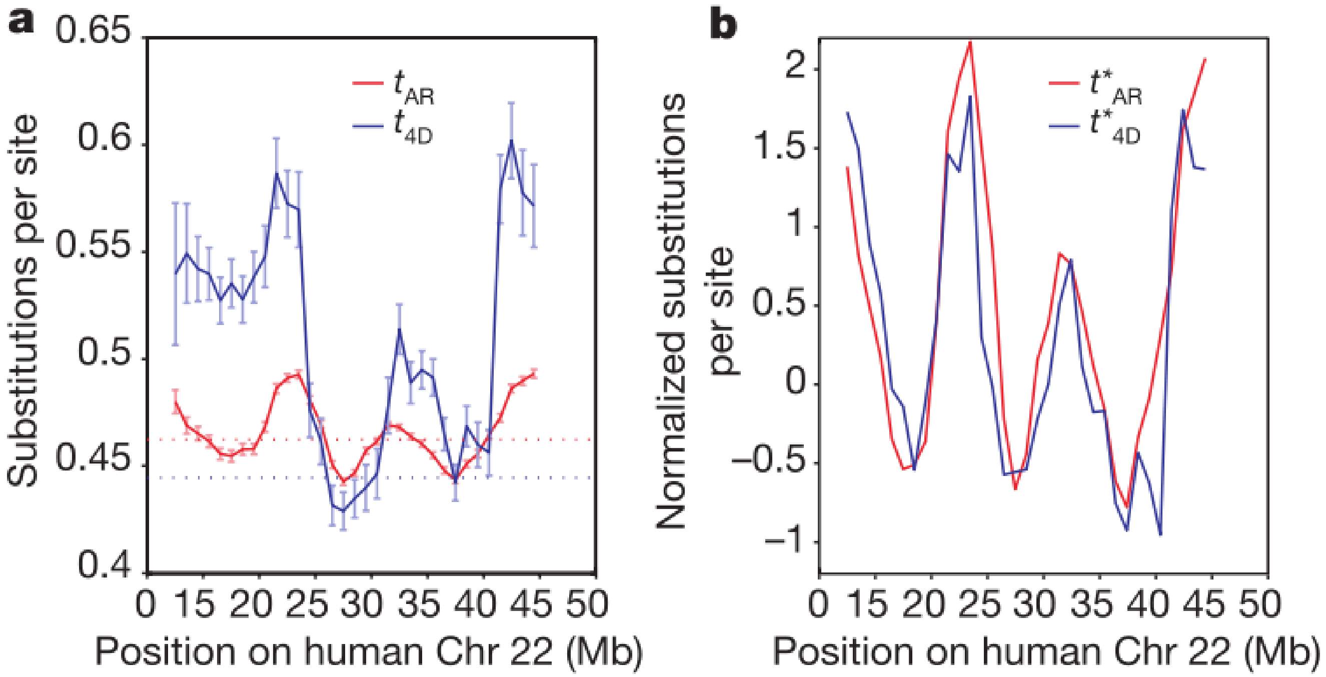
GC bias can be modeled as mutational (ok)
GC content is not at equilibrium! (not ok)

Accounting for GC variation across the genome

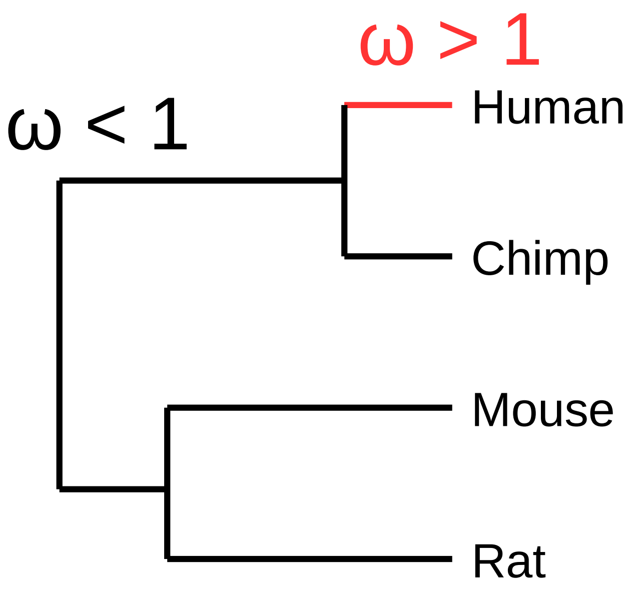
- Substitution models assume independence of sites (model is wrong for CpG)
- Substitution models can handle mutational biases, and work reasonable well

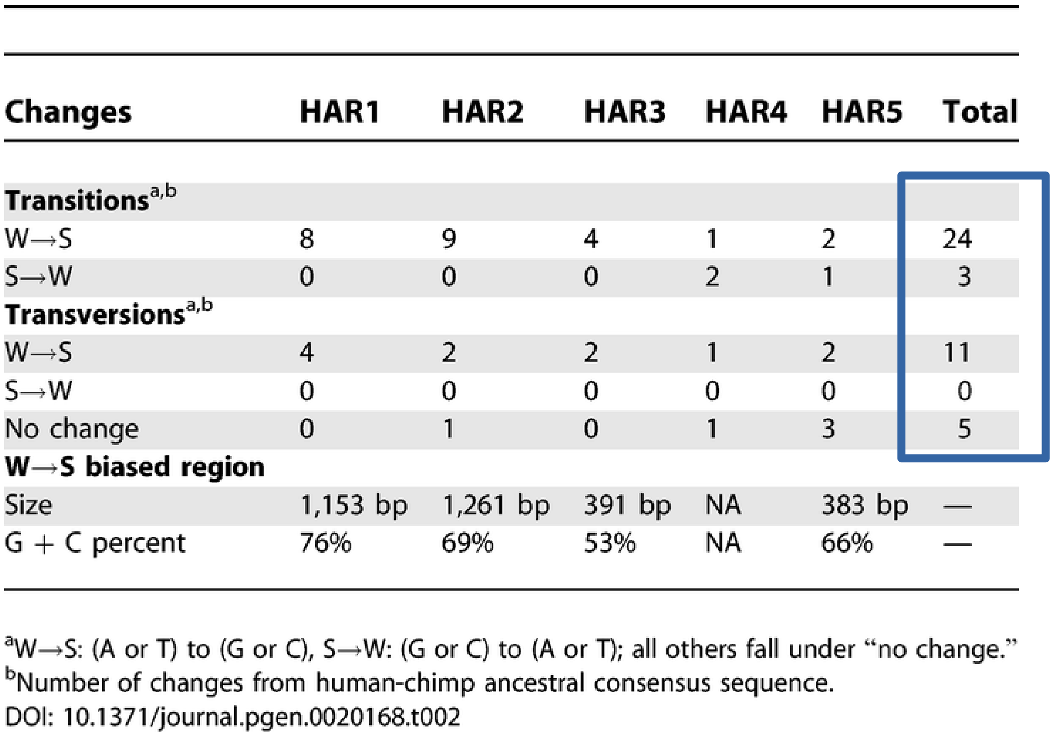
Uncorrected for GC

Corrected for GC



Human accelerated regions (HARs) recombination or selection?





Changes	HAR1	HAR2	HAR3	HAR4	HAR5	Total
Transitions[a,b]						
W→S	8	9	4	1	2	24
S→W	0	0	0	2	1	3
Transversions[a,b]						
W→S	4	2	2	1	2	11
S→W	0	0	0	0	0	0
No change	0	1	0	1	3	5
W→S biased region						
Size	1,153 bp	1,261 bp	391 bp	NA	383 bp	—
G + C percent	76%	69%	53%	NA	66%	—

[a] W→S: (A or T) to (G or C), S→W: (G or C) to (A or T); all others fall under "no change."

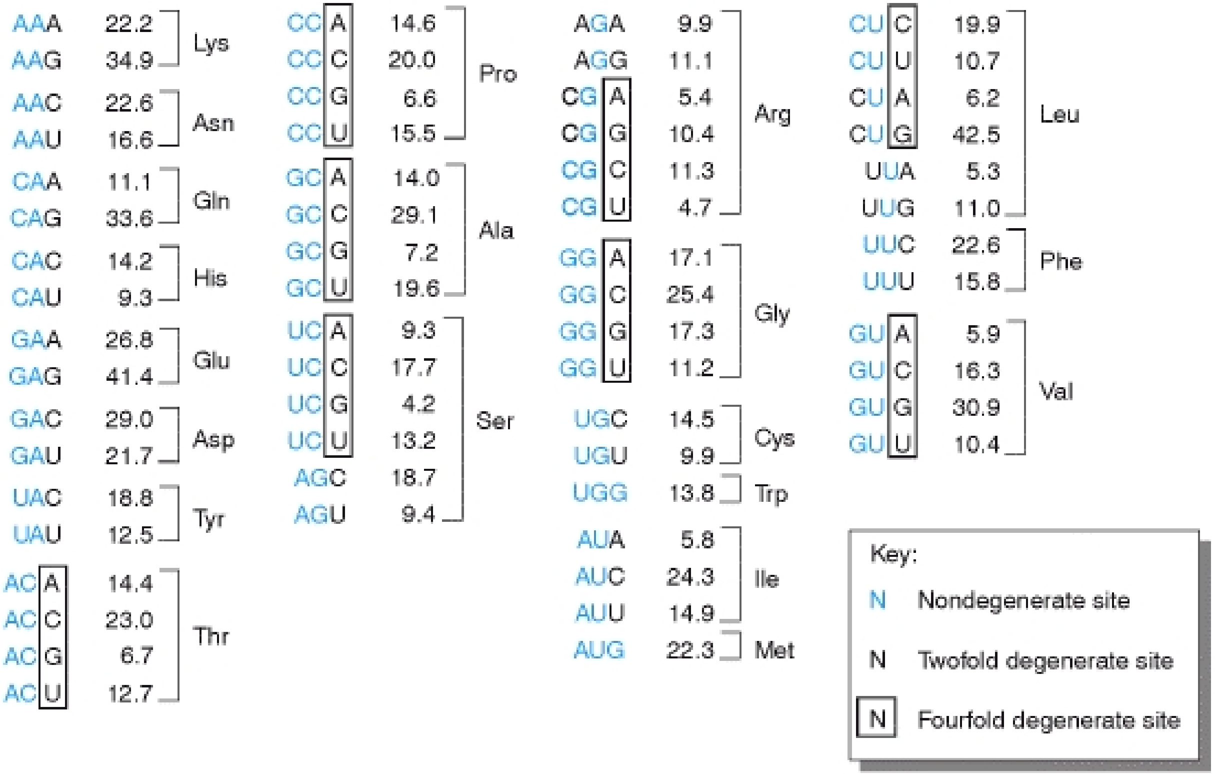
[b] Number of changes from human-chimp ancestral consensus sequence.

DOI: 10.1371/journal.pgen.0020168.t002

- 202 HARs identified by likelihood ratio test
- 1.5% overlap coding regions

- Biased gene conversion (BGC) promotes AT to GC changes (W)
- HARs may evolve fast due to selection OR recombination hotspots

Problem: Codon Bias



AAA	22.2	Lys
AAG	34.9	
AAC	22.6	Asn
AAU	16.6	
CAA	11.1	Gln
CAG	33.6	
CAC	14.2	His
CAU	9.3	
GAA	26.8	Glu
GAG	41.4	
GAC	29.0	Asp
GAU	21.7	
UAC	18.8	Tyr
UAU	12.5	
ACA	14.4	Thr
ACC	23.0	
ACG	6.7	
ACU	12.7	
CCA	14.6	Pro
CCC	20.0	
CCG	6.6	
CCU	15.5	
GCA	14.0	Ala
GCC	29.1	
GCG	7.2	
GCU	19.6	
UCA	9.3	Ser
UCC	17.7	
UCG	4.2	
UCU	13.2	
AGC	18.7	
AGU	9.4	
AGA	9.9	Arg
AGG	11.1	
CGA	5.4	
CGG	10.4	
CGC	11.3	
CGU	4.7	
GGA	17.1	Gly
GGC	25.4	
GGG	17.3	
GGU	11.2	
UGC	14.5	Cys
UGU	9.9	
UGG	13.8	Trp
AUA	5.8	Ile
AUC	24.3	
AUU	14.9	
AUG	22.3	Met
CUC	19.9	Leu
CUU	10.7	
CUA	6.2	
CUG	42.5	
UUA	5.3	
UUG	11.0	
UUC	22.6	Phe
UUU	15.8	
GUA	5.9	Val
GUC	16.3	
GUG	30.9	
GUU	10.4	

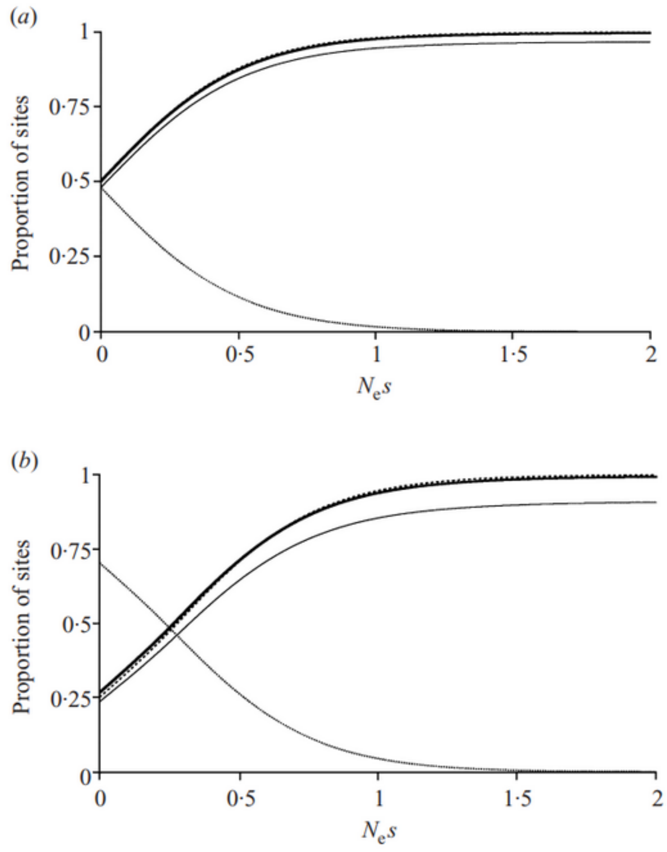
Key:

N	Nondegenerate site
N	Twofold degenerate site
N	Fourfold degenerate site

- GC ending codons are generally more common (preferred)
- Codon bias is higher in highly expressed genes (selection)
- GC bias is not present in noncoding regions (not just mutation)
- Preferred codons have more abundant tRNAs
- Explanation: selection on translation accuracy, speed

Models of codon usage bias

Weak selection Ns ~ 1



$$\pi_{x_i}\mu_{xy_i}f_{xy_i} = \pi_{y_i}\mu_{yx_i}f_{yx_i}$$

At stationarity, flux from X to Y = flux Y to X
pi = equilibrium frequency
u = mutation rate
f = fixation probability

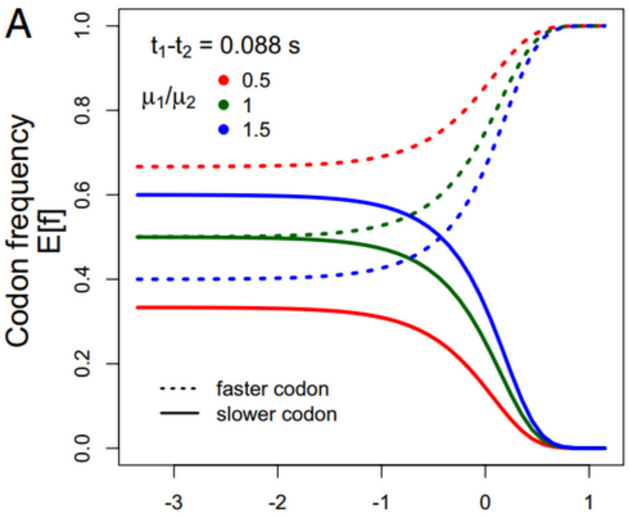
Exercise:
If there are two types of codons, x and y, and if $u_{yx} = 2*u_{xy}$ and no selection, what is π_x?

$\pi_x\, u_{xy} = \pi_y\, u_{yx}$
$\pi_x = \pi_y\, 2u_{xy} / u_{xy}$
$\pi_x + \pi_y = 1$
$\pi_x = (1-\pi_x)\, 2u_{xy} / u_{xy}$
$\pi_x = (1-\pi_x)\ 2$

$\pi_x = 2 – 2\pi_x$
$3\pi_x = 2$

$\pi_x = 2/3$

Gene expression and explanations

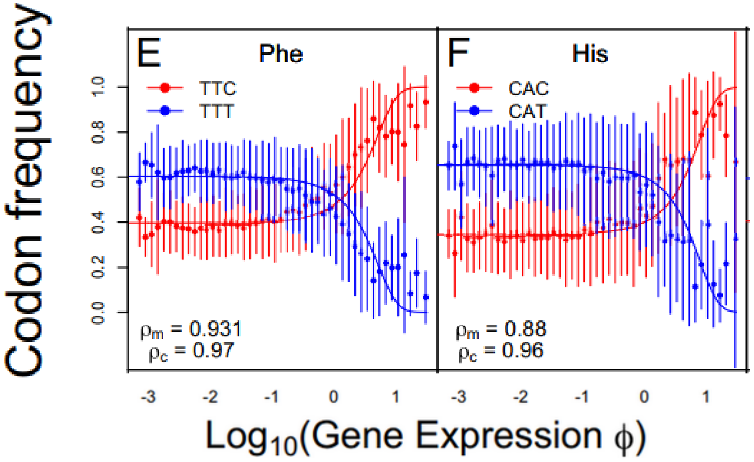


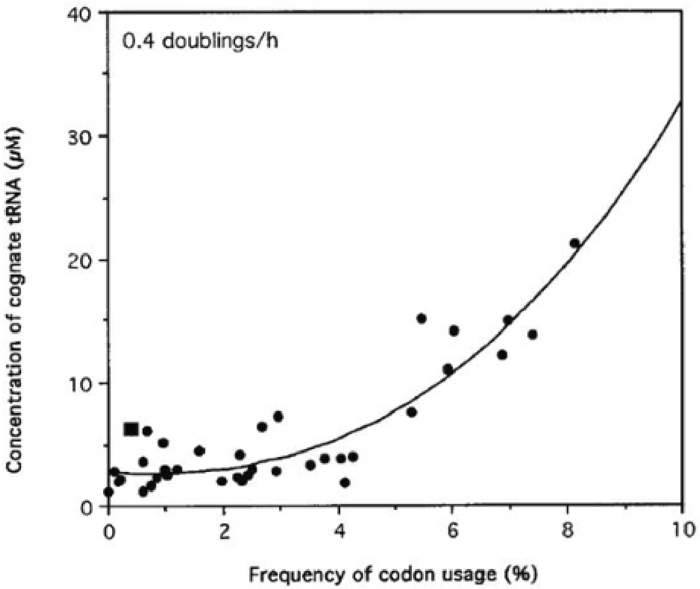
Mutation

- GC content
- nucleotide neighborhood

Selection

- efficiency
- accuracy





Shah and Gilchrist 2011

Exercises

1) Why are GC ending codons at high frequency in highly expressed genes, but low frequency in low expressed genes?

2) Are close or distantly related species better for identifying short functional sequences?

3) What is the most likely function of conserved noncoding sequences?

4) What causes variation in the mutation rate across the genome?

5) Why is the frequency of CpG sites lower than expected?

6) Why are CpG islands unmethylated?

7) Why do CpG sites violate assumptions of nucleotide substitution models?

8) What is the expected dN/dS ratio for a psuedogene – a duplicate gene that has become non-functional?

9) How does recombination influence the substitution rate?

10) In humans most coding sequences are (conserved/unconserved) and most conserved sequences are (Coding/Noncoding)