Overview

- Tests of neutrality:
  - dN/dS example
  - HKY example
  - McDonald/Kreitman
  - Tajima's D
  - Branch-length comparison
  - Conservation
- How much of the genome is functional?

One-minute responses

- For effective population size, how do you know whether to use the whole population or just individuals that could realistically interbreed? – *see upcoming section on population subdivision and gene flow!*

- Reference for cichlids? *Several papers by Michio Hori, but all behind firewalls as far as I can tell, alas*

- Real examples of tests? *Coming up*

- Omitting derivations lets us move faster but sometimes formulae seem to come from nowhere. *It's true. I'll try to strike a balance.*

A live example: dN/dS

- Endo et al. 1996 analyzed 3595 “gene groups” (sets of alignable coding sequences across species) from 1990’s databases
  - They added anything to a gene group that confidently aligned with it
  - They computed pairwise dN/dS within each group
  - “Positive selection” detected when more than half the pairwise comparisons had $dN/dS > 1$
  - Only 17 gene groups showed positive selection (0.45%)
  - 9/17 were pathogen surface proteins exposed to immune system
- Issues with this approach?

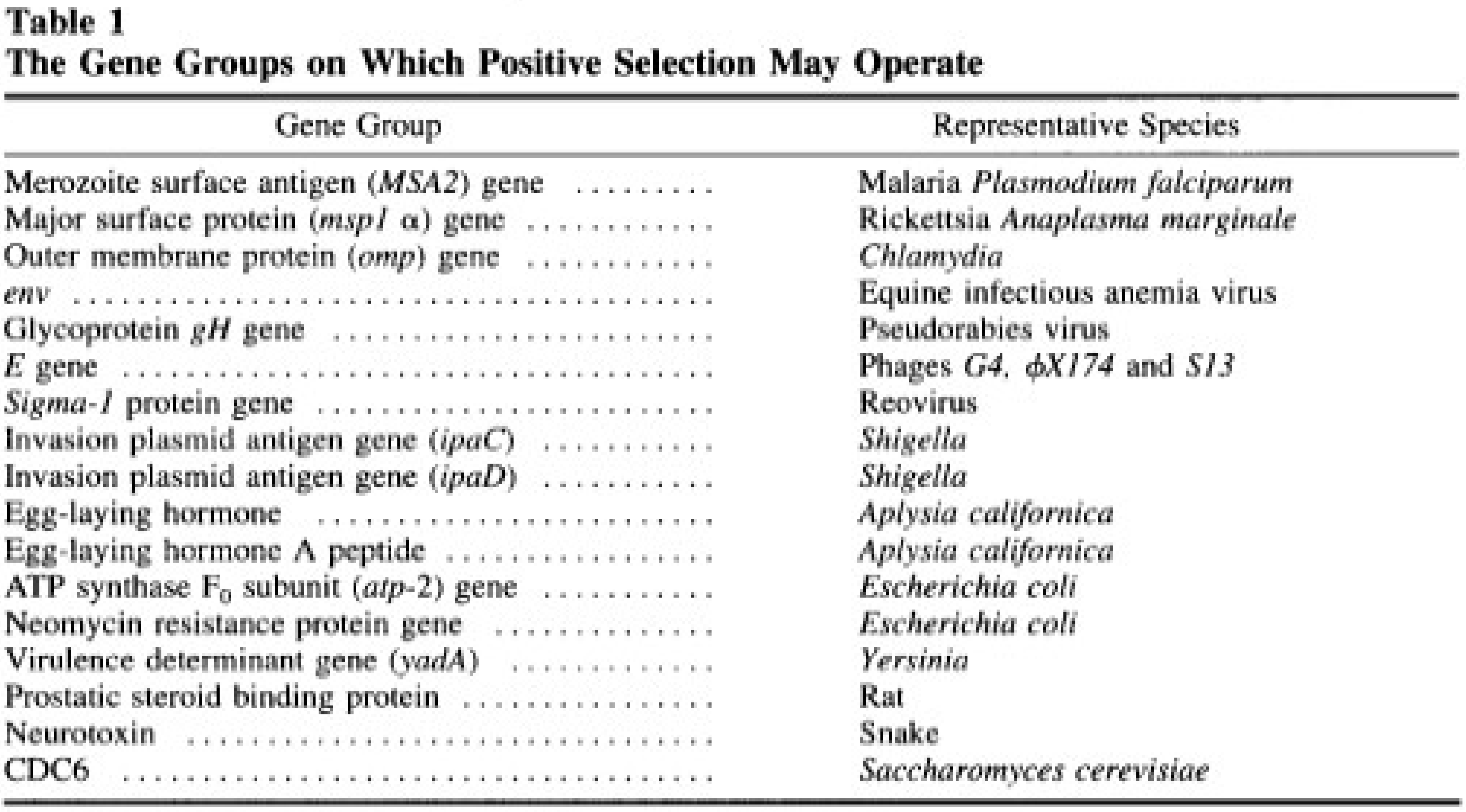


Table 1
The Gene Groups on Which Positive Selection May Operate

Gene Group	Representative Species
Merozoite surface antigen (*MSA2*) gene	Malaria *Plasmodium falciparum*
Major surface protein (*msp1* α) gene	Rickettsia *Anaplasma marginale*
Outer membrane protein (*omp*) gene	*Chlamydia*
env	Equine infectious anemia virus
Glycoprotein *gH* gene	Pseudorabies virus
E gene	Phages *G4*, *ϕX174* and *S13*
Sigma-1 protein gene	Reovirus
Invasion plasmid antigen gene (*ipaC*)	*Shigella*
Invasion plasmid antigen gene (*ipaD*)	*Shigella*
Egg-laying hormone	*Aplysia californica*
Egg-laying hormone A peptide	*Aplysia californica*
ATP synthase F_0 subunit (*atp*-2) gene	*Escherichia coli*
Neomycin resistance protein gene	*Escherichia coli*
Virulence determinant gene (*yadA*)	*Yersinia*
Prostatic steroid binding protein	Rat
Neurotoxin	Snake
CDC6	*Saccharomyces cerevisiae*

From Endo et al. (1996) Mol Biol Evol 13: 685-90.

A live example: HKA

- Hudson, Kreitman and Aguade 1987 (original paper for this test)

Locus	Adh 5' flanking region	Adh locus
Differences between species	210	18
Differences within species	9	8

- Within-species numbers come from 82 *D. melanogaster* samples
- Between-species come from one *D. melanogaster* and one *D. sechellia*
- Authors attributed this to balancing selection on the coding sequence

HKA assumptions

- HKA assumes:
  - The “neutral” comparison gene is really neutral
  - Mutation rate constant for each gene (doesn’t need to be equal between genes)
  - No large changes in population size
  - Divergence time of the two loci is the same (no “ancestral polymorphism”)
- Measure statistical significance with a χ^2 test

McDonald/Kreitman test

- Call within-species comparisons w and between-species b
- Under neutrality:
- $dS_b/dS_w = dN_b/dN_w$
- Deviation from this indicates some kind of selection
- Generally used as a test for adaptive evolution
- Criticized for being vulnerable to weakly deleterious mutations
  - Weakly deleterious mutations contribute to dN_w but not dN_b
  - Obscures presence of adaptive evolutuion

Tajima's D

- Two estimates of population diversity:
  - Based on number of variable sites
  - Based on mean pairwise differences

- Each yields an estimate of $\theta = 4N_e\mu$

- In a neutral situation these estimators should agree

Estimator based on variable sites

- Called π or Watterson's estimator
- Under a neutral infinite sites model:
  - For a number of sampled sequences k
  - And a given $\theta = 4N_e\mu$
  - Expected number of mutated sites is expected branch length of the coalescent
- Let's derive this

Estimator based on variable sites

- Length of a time interval is $2N_e/[k(k-1)/2]$
- Branch length in that interval is k times this
- Total branch length is sum over intervals
- Pull out k term: $a = \sum_{k=1}^{n-1} \frac{1}{k}$
- Expected mutations is total branch length times μ
- $S = 4N_e\mu \times a$
- $4N_e\mu = \frac{S}{a}$
- This estimator is often called θ_S

Estimator based on mean pairwise differences

- Define mean number of differences between pairs of sequences as π
- This is an estimate of θ (per locus!) because the expected differences between a pair are $2N \times 2\mu$
- Usually called θ_π

Tajima's insight

- We have two different estimators of θ
- In a pure Wright-Fisher situation they should be approximately equal
- They are differently sensitive to deviations:
  - θ_S is much more impressed by rare alleles than θ_π
- $d = \theta_\pi - \theta_S$
- Test statistic "Tajima's D" $= \frac{d}{\sigma(d)}$
- $\sigma(d)$ is standard deviation of D

Behavior of Tajima's D reflects the coalescent

- Remember $d = \theta_\pi - \theta_S$
- $D = 0$ interpretation?
- $D < 0$ interpretation?
- $D > 0$ interpretation?

Behavior of Tajima's D reflects the coalescent

- $D = 0$ neutrality
- $D < 0$ population growth, directional selection
- $D > 0$ population shrinkage, balancing selection
- Significance value usually obtained by simulation
- A rough rule of thumb: significant if more than +2 or less than -2
- Concern: population subdivision?

Conservation as a measure of (purifying) selection

- Regions that are very similar among species might be:
  - Functional and under purifying selection
  - Recent copies of something functional (but might not be any longer)

- Regions that are not similar might be:
  - Not functional
  - Functional in only some species, or different functions in different species
  - Functional, but only a few sites are conserved
  - Functional, but rapidly shifting between species (reproductive proteins)
  - Functional, but undergoing concerted evolution

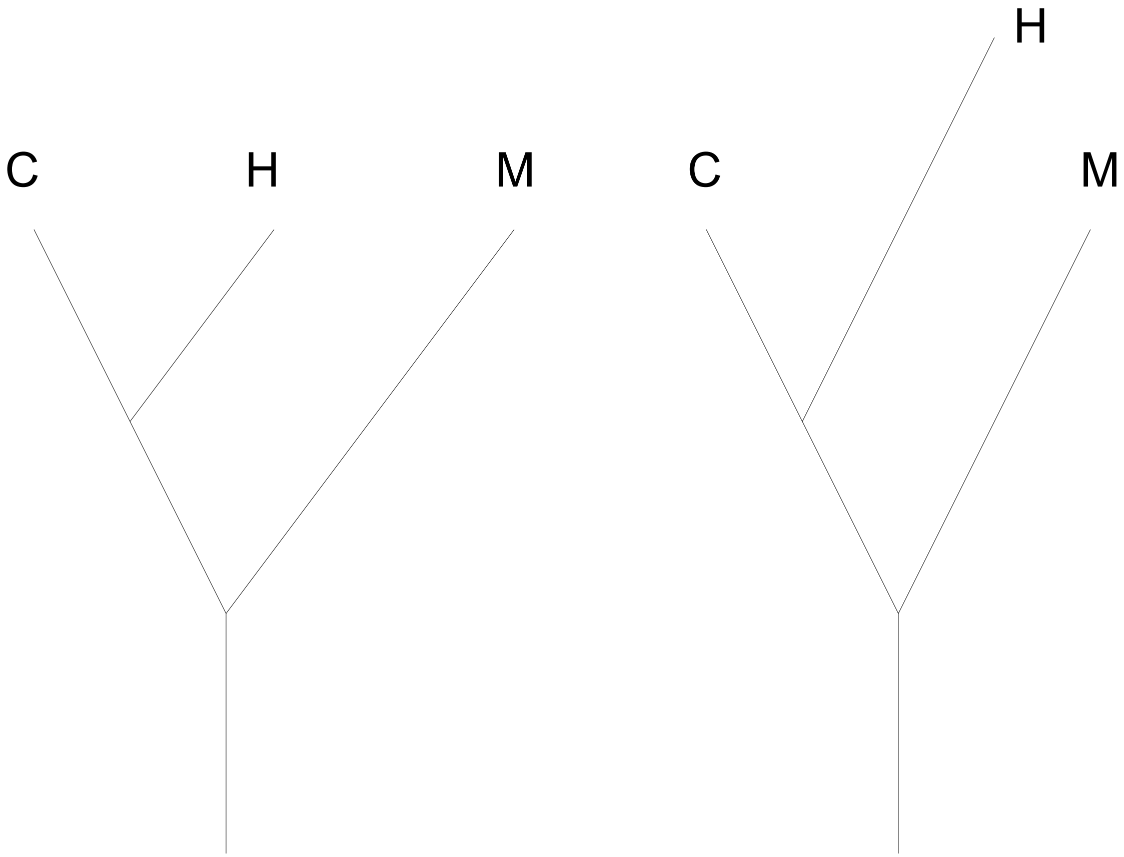
Abalone VERL protein

- Major component of egg vitelline envelope
- Must handshake with sperm lysin for fertilization
- Swanson et al. (2001) Mol Biol Evol:
  - dN/dS consistent with neutrality
  - Tajima's D not significantly different from 0 (and varied in different directions in the two species)
  - HKA not significantly different from neutrality
- Very odd for an utterly essential function!
- VERL may drift (with convergent evolution) while lysin chases it

Different branch lengths as measure of differing selection

Clark et al. (2004) Science 302: 1960-1963.

- Compared human and chimp with mouse as an outgroup
- Estimated branch lengths for many genes
- Looking for genes with longer branches in human than in chimp


C
H
M
H
C
M

Brainstorm

- What could cause a long branch?
- If all human genes showed long branches, what could that mean?
- If only certain human genes showed long branches, what could that mean?

Accelerated evolution in the human lineage

Some ideas:

- Adaptive evolution in humans
- Deterioration in humans due to fixing bad mutations (bottlenecks?)
- Weaker selection on humans (technology?)
- Increased mutation rate in humans
- Decreased mutation rate in chimpanzees
- Shorter generation time in humans than chimpanzees

Humans and chimpanzees

Gene categories whose evolution has accelerated in human evolution:

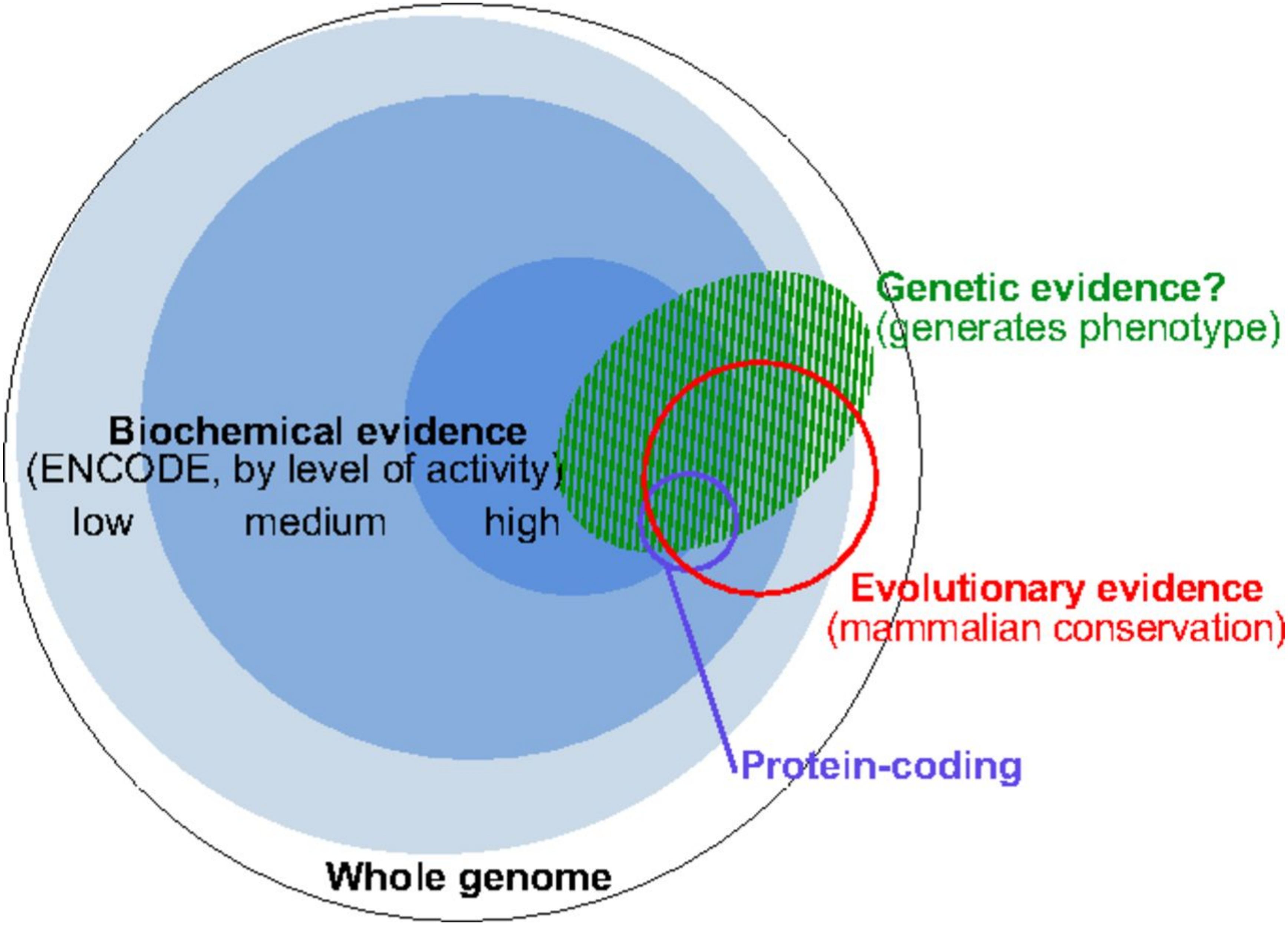
- Senses
- Digestion and food metabolism
- Reproduction, especially spermatogenesis
- Immune system and tumor suppression
- NOT brain function

Flaws in this comparison?

- A single mutation could have a huge effect not seen in this test
- Coding regions only
- Some “mutations” are really polymorphisms, and their frequency depends on population size
  - Chimp long-term population size is larger than human, so this does not explain away human-specific increases
- Some false positives likely due to large number of comparisons

ENCODE controversy

- ENCODE study mapped:
  - transcription
  - transcription factor binding
  - chromatin structure
  - histone modification
- "These data enabled us to assign biochemical functions for 80% of the genome"
- (1.5% of the genome is coding sequence)
- ENCODE Project Consortium (2012) Nature 489: 57-74.



From Kellis et al. (2014) PNAS 111: 6131-6138

Could 80% of the genome be under selection?

Based on Kellis et al. (2014)

- Arguments for:
  - Pervasive evidence of biochemical activity
  - GWAS for phenotypes often lands in areas lacking known functional elements

- Arguments against:
  - Much of the genome is repeats: they may be “active” but are they meaningful?
  - Haldane argument: can a population afford selection on very many loci?
  - Lack of conservation–only 5% of genome strongly conserved in mammals
  - Low N_e of large mammals makes very weak selection ineffective

Haldane’s argument: “Genetic Load”

- Haldane argued that the cost of a harmful allele to a population is nearly independent of s:
  - Every copy added by mutation must eventually be removed by selection (a “selective death”)
  - Strongly harmful alleles hurt a few individuals a lot, then are gone
  - Weakly harmful alleles hurt each individual less, but hang around longer
- How many “selective deaths” can a population handle?
- Depends on reproductive excess

Weaknesses in this argument

- Hard selection:
  - Regardless of competition, unfit genotype tends to die (or fail to reproduce)
  - Too much of this threatens the population's survival
- Soft selection:
  - In the absence of competition, all genotypes are viable
  - "Unfit" genotypes have a competitive disadvantage in the presence of fitter ones
  - Does not reduce population viability
- Another issue: how do fitnesses interact at multiple loci? Can one "selective death" eliminate many harmful mutations at one swoop?

Small Neanderthal N_e

- Large “deserts” in European genome where no Neanderthal alleles found
- Two hypotheses:
  - Neanderthal alleles in these areas don’t work well in a modern human context
  - Small Neanderthal populations led to bad Neanderthal alleles which were weeded out

Monday

- Selection at multiple unlinked loci
- Interactions among loci
- A first look at linkage

One-minute responses

- Please:
  - Tear off a slip of paper
  - Give me one comment or question on something that worked, didn't work, needs elaboration, etc.