

Computational Human Genetics

Itsik Pe'er

Department of Computer Science
Columbia University

Fall 2006

Reminder

- The structure and demography affect genetics of neutral populations



...but non-neutral sites are more interesting

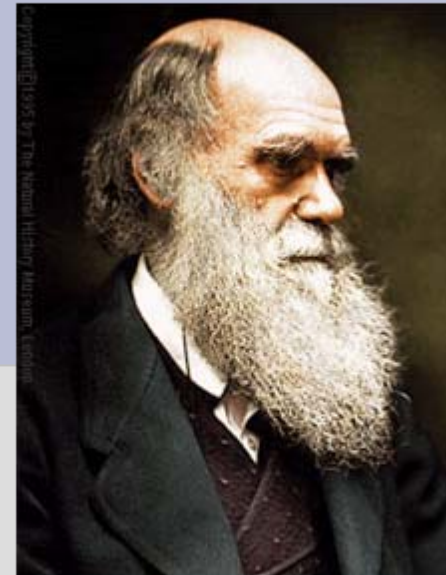
Meeting #8

Selection

Natural Selection

- What is selection:
 - General concepts
 - Types of selection
- Negative selection
- Positive selection:
 - Single site tests
 - Using linkage disequilibrium

Survival of the Luckiest



Charles Darwin (colourized B&W print)

- Darwin:
Variation in nature:
 - Created by mutation
 - Eliminated by selection: fixation
- Today:
Factors affecting fixation of a polymorphism:
 - Selection
 - Drift:
 - Chance
 - Frequency
 - Population size

Fixation under Neutrality

- Prob(allele a will become fixed) = frequency
- Prob(new mutation will become fixed) = $1/2N$
- Fixation rate = $2N\mu/2N = \mu$ = mutation rate

Fitness

- W : chance of reproducing

- s : selection coefficient:

$$W_{aa} = 1 + 2s \quad W_{Aa} = 1 + s \quad W_{AA} = 1$$

- In a large, constant, random-mating sample:

$$AA: p^2 \quad Aa: 2pq \quad aa: q^2$$

- Frequencies of reproduction:

$$AA: W_{AA}p^2 \quad Aa: W_{Aa}2pq \quad aa: W_{aa}q^2$$

- Average fitness:

$$\bar{W} = p^2W_{AA} + 2pqW_{Aa} + q^2W_{aa}$$

Fitness

- Expected frequencies at next generation:
AA: $W_{AA}p^2/W$ Aa: $W_{Aa}2pq/W$ aa: $W_{aa}q^2/W$

$$Exp(p') = p \frac{pW_{AA} + qW_{Aa}}{\bar{W}} = \frac{pW_A}{\bar{W}}$$

$$M(p) = Exp(\Delta p) = \frac{p(W_A - \bar{W})}{\bar{W}} = \frac{pq(W_A - W_a)}{\bar{W}} \approx pqs$$

$$V(p) = Var(\Delta p) = \frac{p'q'}{2N} \approx \frac{pq}{2N}$$

- Directional selection changes frequencies
- Selection is inefficient against rare alleles

Fixation under Selection: Formulae

- Probability of fixing an allele: $u(p,t)$

$$\frac{\partial u(p,t)}{\partial t} = \frac{1}{2}V(p)\frac{\partial^2 u(p,t)}{\partial p^2} + M(p)\frac{\partial u(p,t)}{\partial p}$$

- Define: $u(p) = \lim_{t \rightarrow \infty} u(p,t)$

$$0 = \frac{1}{2}V(p)\frac{d^2 u(p)}{dp^2} + M(p)\frac{du(p)}{dp} \quad \frac{du(p)}{dp} = e^{-\int \frac{2M(p)}{V(p)} dp} = e^{-4Nsp}$$

$$\text{Fix: } u(0)=0 ; u(1)=1$$

$$u(p) = \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}}$$

- Selection is ineffective if $s \ll 1/N$

New Allele under Selection

$$u\left(\frac{1}{2N}\right) = \frac{2s}{1 - e^{-4Ns}}$$

- Deleterious alleles will arise to fixation only in small populations

Time to Fixation of New Alleles

- Under neutrality:
=tMRCA= $4N_e$
- With positive selection:
 $(2/s)\ln(2N_e)$
- Positive selection is immediate!

Natural Selection

- What is selection:
 - General concepts
 - Types of selection
- Negative selection
- Positive selection:
 - Single site tests
 - Using linkage disequilibrium

Direction of Selection

- Neutral: No selection
- Negative: Against new variants
- Positive: In favor of new variants
- Balancing: Maintains both variants
- Against heterozygotes: Against rare variants

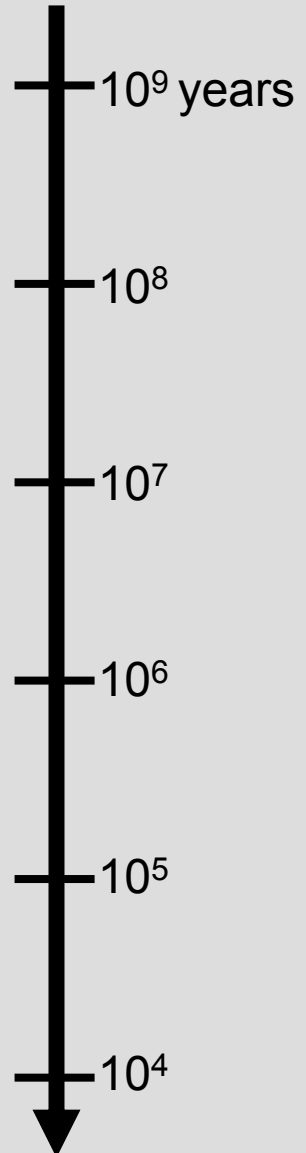
Timescales of Selection

Between species

Along the pan-human lineage

Between populations

Within populations



Natural Selection

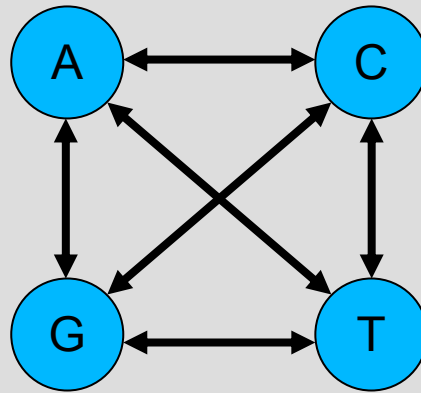
- What is selection:
 - General concepts
 - Types of selection
- Negative selection
- Positive selection:
 - Single site tests
 - Using linkage disequilibrium

The Neutral Theory

(Kimura 1968, Jukes & King 1969)

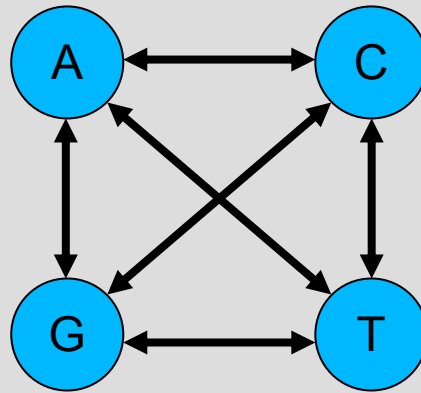
- Most of evolution is neutral
- Proofs:
 - Substitution rate = mutation rate
 - Many mutations have no observed effect
 - At the observed mutation rate, if variation was functional, we would all be dead

Neutral Substitution Matrix



$$\begin{bmatrix} 1-3p & p & p & p \\ p & 1-3p & p & p \\ p & p & 1-3p & p \\ p & p & p & 1-3p \end{bmatrix}^t$$

Neutral Substitution Matrix



$$\begin{bmatrix} 1 - \sum p & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & 1 - \sum p & p_{CT} & p_{CT} \\ p_{GA} & p_{GC} & 1 - \sum p & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & 1 - \sum p \end{bmatrix}^t$$

Negative Selection vs. Function



Similarity: 70%

Coding: 85%

UTR: 75%

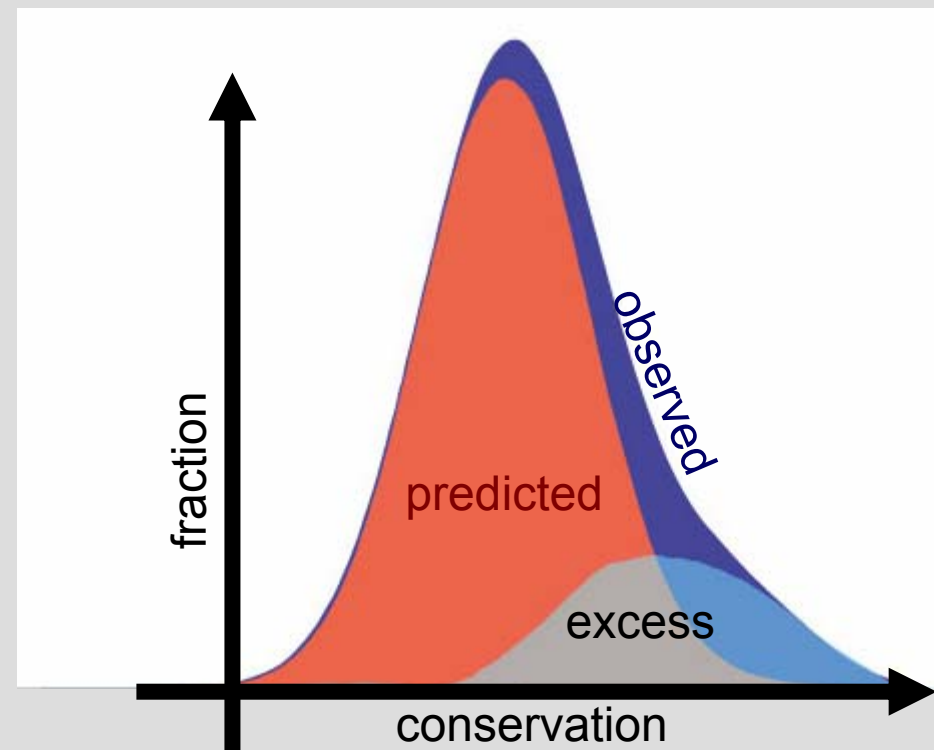
Regulatory: 75%

Introns: 70%



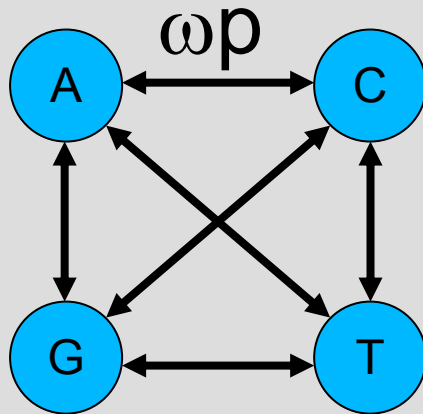
Selected Part of the Genome

- 5% of the genome under negative selection
- Includes essentially all known-function regions
- At the ultraconserved tail:
 - ~500 segments >200bp identical human-mouse +
 - Most of them noncoding



Selection in Genes

- $\omega = K_a / K_s$ – ratio of non-synonymous to synonymous substitution
- $\omega < \Rightarrow > 1$: negative/neutral/positives selection



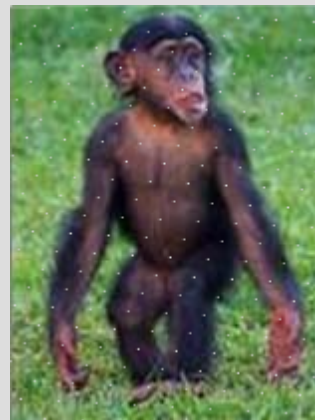
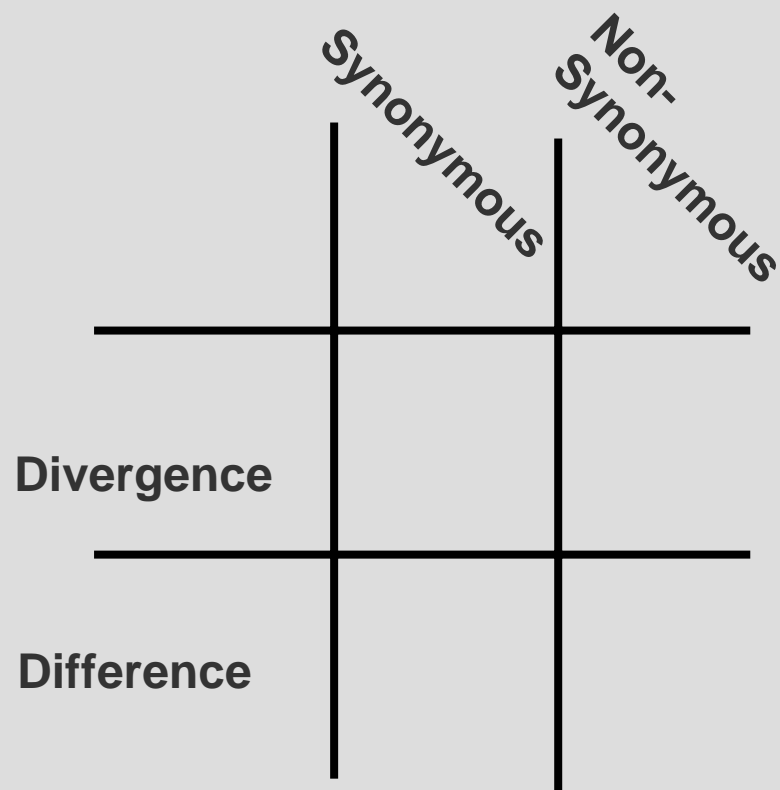
Typically: $\omega \approx 0.1-0.25$

Positive selection:
Host-defense
Olfaction
Reproduction

- Also: 20% rejection of NS changes

Divergence - Difference

- Compare between-species differences (K_a/K_s)
- Use within-species divergence to control



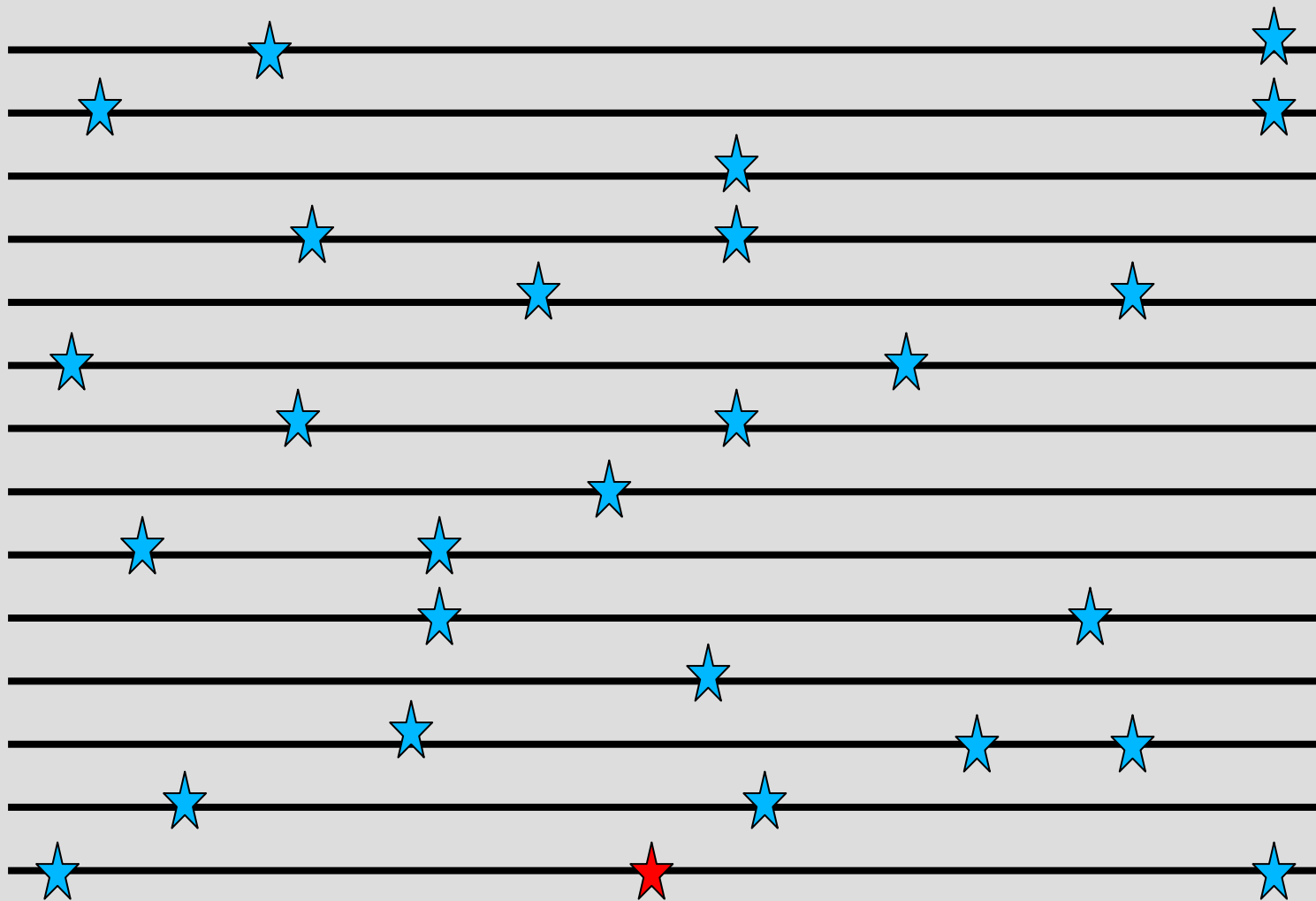
Natural Selection

- What is selection:
 - General concepts
 - Types of selection
- Negative selection
- Positive selection:
 - Single site tests
 - Using linkage disequilibrium

Positive Selection in Humans

- Hallmark: rapid increase in frequency
- Looking for sites/regions/families
- Tests:
 - Reduction in diversity
 - Common derived alleles
 - Population differences
 - Extended linkage disequilibrium

Selective Sweep



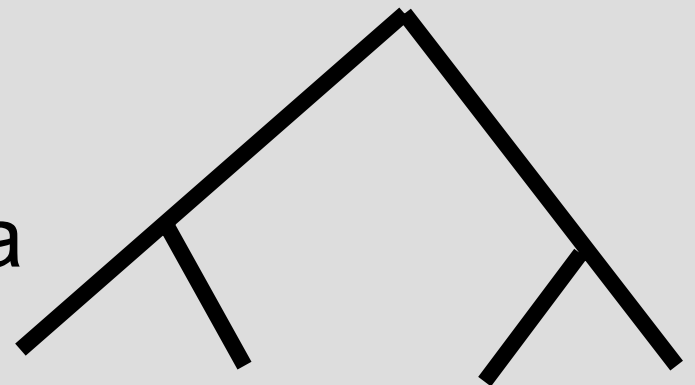
Post-Sweep Diversity

- Expectation: reduction in diversity
- Compare to bottleneck:
 - Only local effect, taper off at flanks
- Compare to low mutation rate:
 - Variants exist, but rare (careful of errors...)

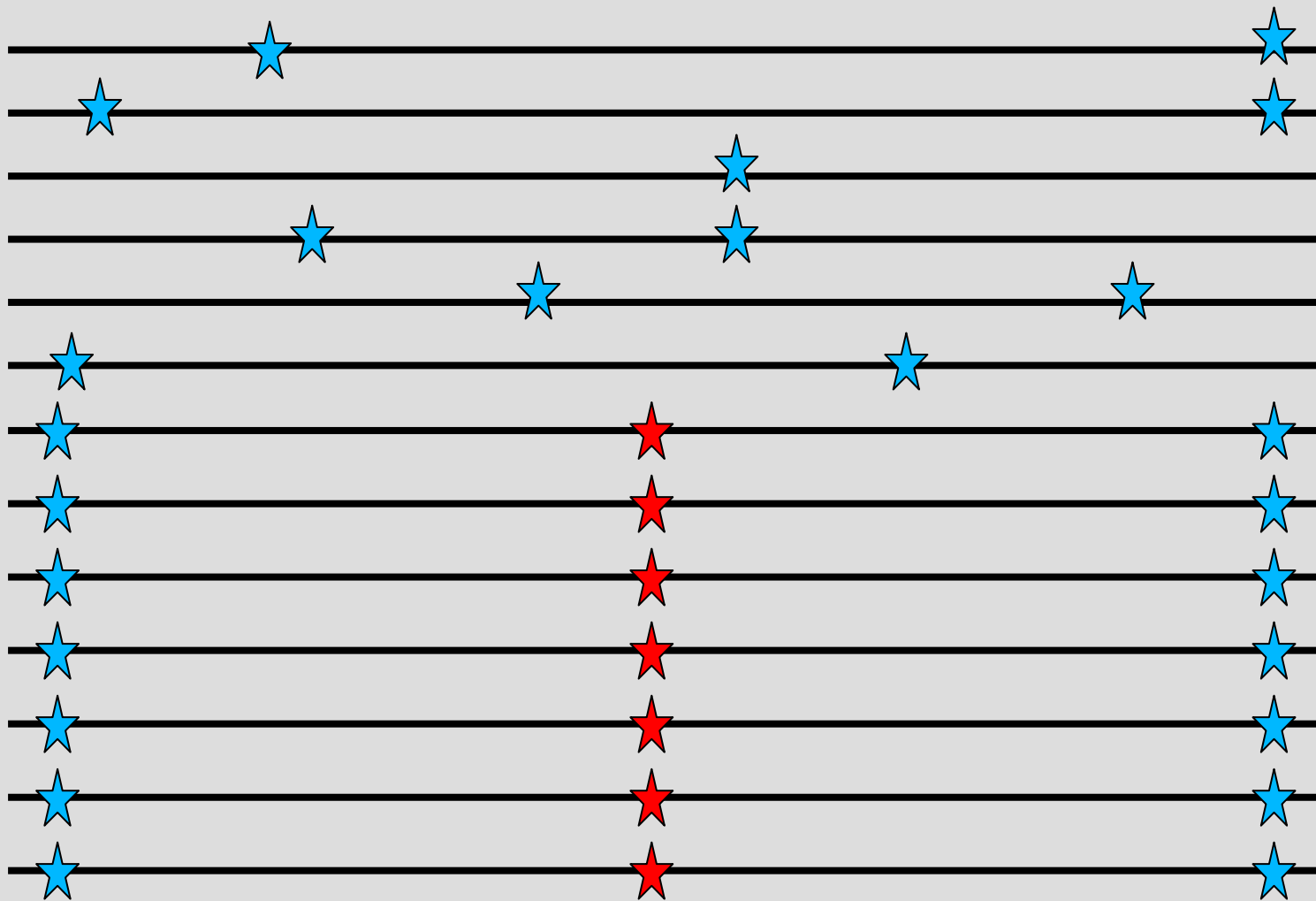
Testing Allele Frequencies

- Under the null:
$$\text{Exp}(\#\text{SNPs}/\sum 1/i) = 4N\mu$$
- Estimate $\theta = 4N\mu$ as avg. #pairwise differences
- Tajima's D statistic: $(\hat{\theta} - \text{Exp}(\theta)) / \text{sqrt}(\text{var}(\theta))$

- Effective for sweeps $< 250\text{kya}$



Partial Selective Sweep



High-Frequency Derived Alleles

- Erases frequent-ancestral correlation
- H-statistic:
Same as D, but estimates θ by up-weighting high-frequency derived alleles
- Effective for sweeps $< 80\text{kya}$

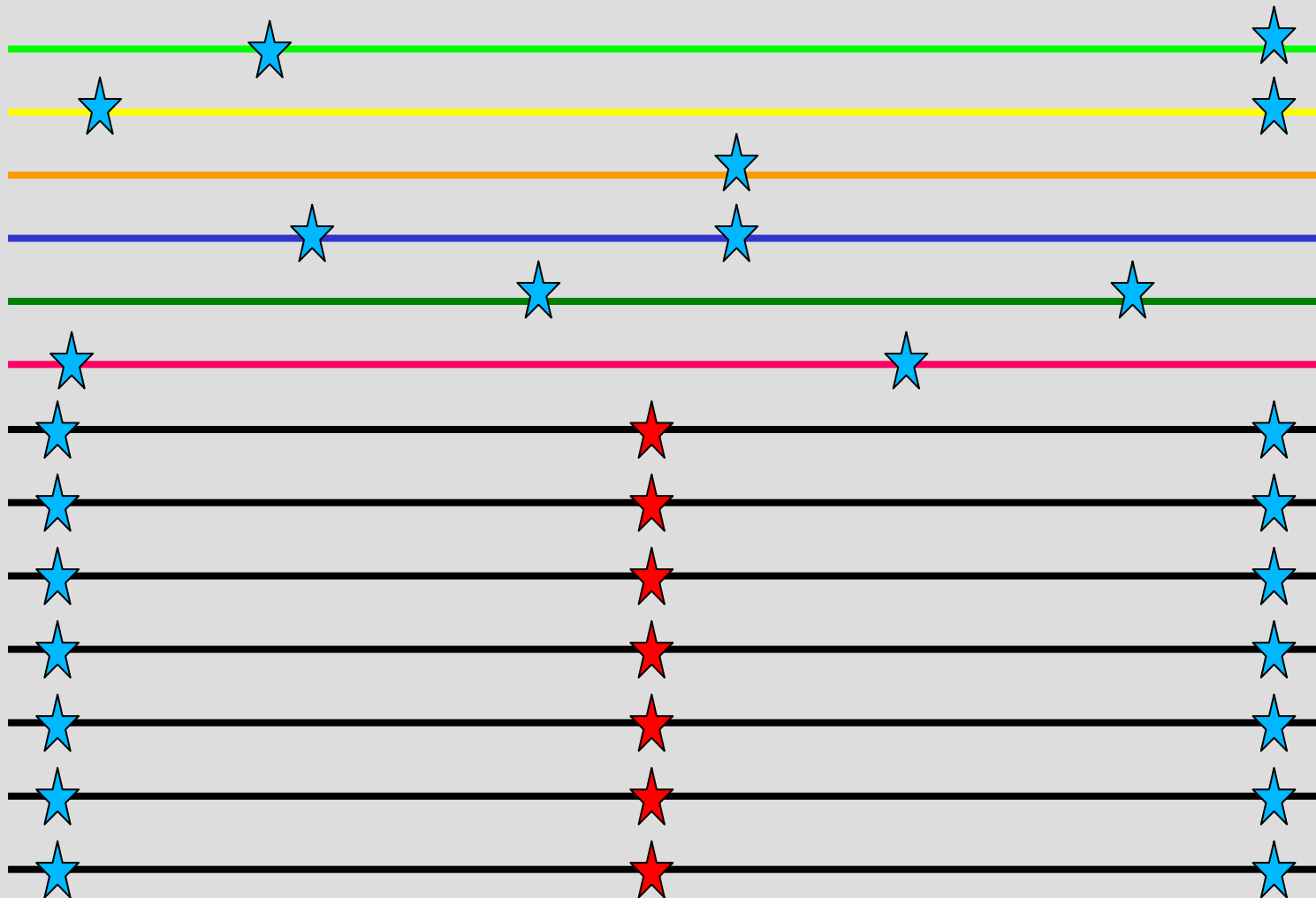
Population Differences

- Assumption:
 - Selective constraints differ by region
- Examples:
 - Lactase
 - Sickle-cell anemia
- Method:
 - F_{st} = % variance within sub-populations

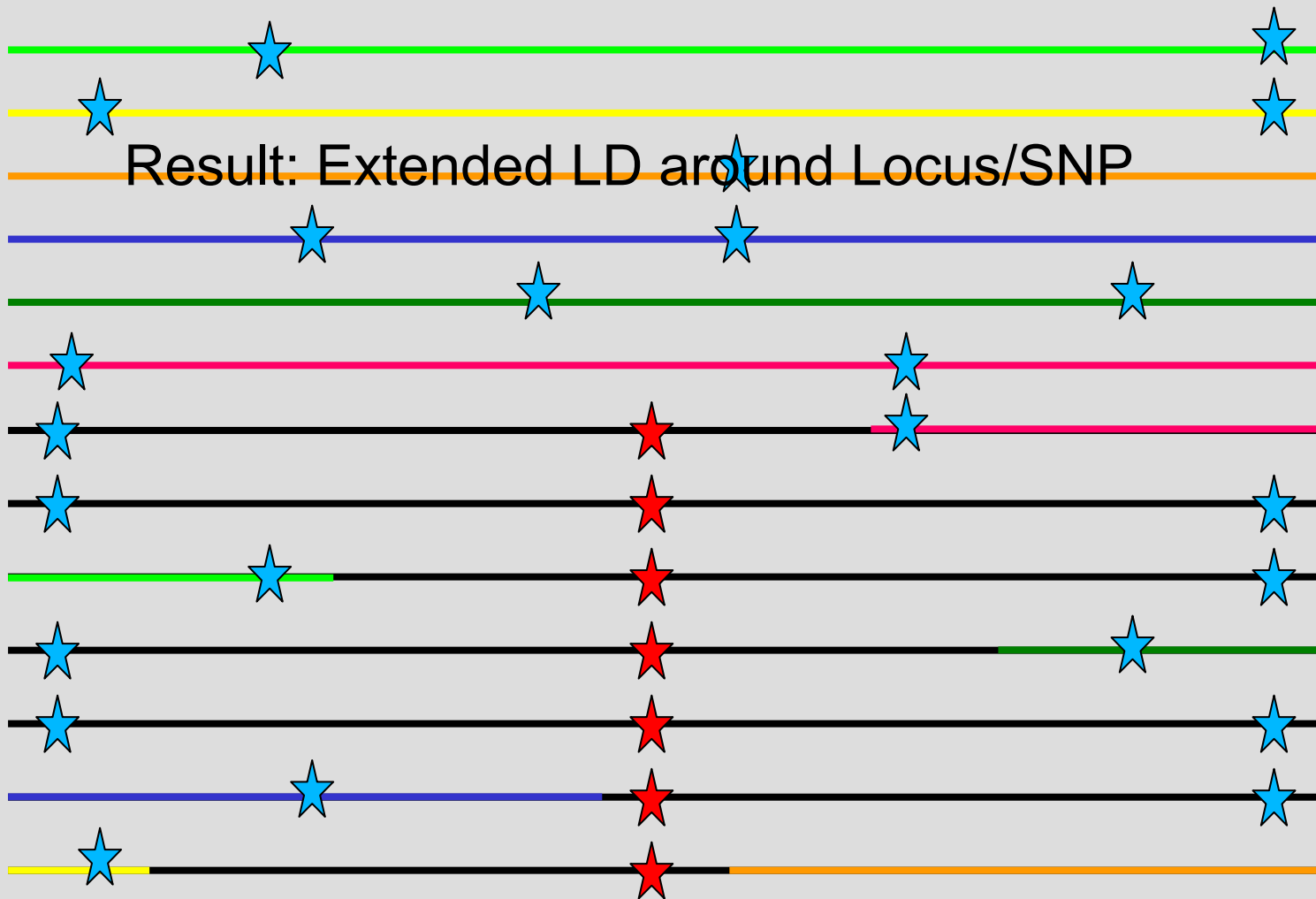
Natural Selection

- What is selection:
 - General concepts
 - Types of selection
- Negative selection
- Positive selection:
 - Single site tests
 - Using linkage disequilibrium

Selective Sweep in Progress



Selective Sweep in Progress



Do we see Positive Selection?

- ~500 100kb regions
- Majority:
Population specific
(Lack of power? Environments?)
- The usual suspects:
 - Immunity
 - Reproduction
 - Metabolism

Summary

- Variation is mostly random
- Functional variation is mostly deleterious
- Positive selection is a mechanism for rapid changes

Further Reading

- Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. Positive natural selection in the human lineage. *Science*. 2006 Jun 16;312(5780):1614-20.
- Modern Genetic Analysis, Griffiths Gelbart, Lewontin, Miller, online book, chapter 17, <http://bcs.whfreeman.com/mga2e/default.asp?s=&n=&i=&v=&o=&ns=0&uid=0&rau=0>
- Kimura M Evolutionary rate at the molecular level. *Nature*. 1968 Feb 17;217(5129):
- Kimura M On the probability of fixation of mutant genes in a population. *Genetics*. 1962 Jun;47:713-9.
- King JL, Jukes TH, Evolutionary rate at the molecular level *Science*. 1969 May 16;164(881):788-98.
- Bielawski JP, Yang, . Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J Struct Funct Genomics*. 2003;3(1-4):201-12
- McDonald JH, Kreitman M Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*. 1991 Jun 20;351(6328):652-4.
- Voight BF, Kudaravalli S, Wen X, Pritchard, A map of recent positive selection in the human genome. *PLoS Biol*. 2006 Mar;4(3):e72.
- Fay JC, Wu CI., Hitchhiking under positive Darwinian selection. *Genetics*. 2000 Jul;155(3):1405-13. Fay Wu
- Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989 Nov;123(3):585-95.
- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. Ultraconserved elements in the human genome. *Science* 2004 May 28;304(5675):1321-5

Extra Credit

1. Tay-sachs is lethal childhood disease. It is autosomal recessive deterministic. If the deleterious allele is currently 0.1, for a Hardy-Weinberg constant population of 1000, what is the expected probability 100 generations from now?

2. Critically read

Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, Hudson RR, Tishkoff SA & Lahn BT. Ongoing adaptive evolution of *ASPM*, a brain size determinant in *Homo sapiens*. *Science*, 309:1720 (2005)

Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, Tishkoff SA, Hudson RR & Lahn BT. *Microcephalin*, a gene regulating brain size, continues to evolve adaptively in humans. *Science*, 309:1717 (2005).

Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Cullen M, Mikkelsen TS, Roy J, Patterson N, Cooper R, Reich D, Altshuler D, O'Brien S, Lander ES. The case for selection at CCR5-Delta32. *PLoS Biol*. 2005 Nov;3(11):e378.

Summarize and evaluate evidence for recent selection in the genes discussed.

Project Suggestion I

- Splice sites are, in theory, more constrained than amino-acid coding sequences.
 - Evaluate negative selection on human-lineage splice sites by comparison to chimp and other mammals.
 - Correlate inferred selection to alternative splicing

Project Suggestion II

- Existing tests for positive selection typically study a single statistic to screen the genome.
 - Combine a haplotype-test (Voight et al, or Sabeti et al), an allele-frequency test (Tajima's D, or Fay & Wu's H) and a population-difference test (F_{ST}) to a unified score.
 - Simulate neutral data to get joint distributions of the statistics
 - Scan the genome for positive selection

Project Suggestion III

- Survey selection in special regions:
 - around ultraconserved segments
 - Around regions that are absent from the chimp genome

Report:

- Is there recent +/- selection in nearby areas?
- Are these regions different than the rest of the genome?