BIOL 502 Population Genetics Spring 2017

Week 6 Selection

Arun Sethuraman

California State University San Marcos

Table of contents

Instances of natural selection in humans

Allison, A.C. Br Med J. 1954 Feb 6; 1(4857): 290294.

PROTECTION AFFORDED BY SICKLE-CELL TRAIT AGAINST SUBTERTIAN MALARIAL INFECTION

BY

A. C. ALLISON, D.Phil., B.M.*

(From the Clinical Pathology Laboratory, the Radcliffe Infirmary, Oxford)

The main problem can be stated briefly: how can the sickle-cell gene be maintained at such a high frequency among so many peoples in spite of the constant elimination of these genes through deaths from the anaemia? Since most sickle-cell anaemia subjects are homozygotes, the failure of each one to reproduce usually means the loss of two sickle-cell genes in every generation. It can be estimated that for the lost genes to be replaced by recurrent mutation so as to leave a balanced state, assuming that the sickle-cell trait—that is, the heterozygous condition—is neutral from the point of view of natural selection, it would be necessary to have a mutation rate of the order of 10^{-1}. This is about 3,000 times greater than naturally occurring mutation rates calculated for man and, with rare exceptions, in many other animals—3.2×10^{-5} in the case of haemophilia (Haldane, 1947). A mutation rate of this order of magnitude can reasonably be excluded as an explanation of the remarkably high frequencies of the sickle-cell trait observed in Africa and elsewhere.

Bersaglieri et al. 2004, AJHG DOI: 10.1086/421051

AJHG

All Content

The American Journal of Huma

Explore | Online Now | Current Issue | Archive | Journal Information | For Authors

< Previous Article

Volume 74, Issue 6, p1111–1120, June 2004

Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene

Todd Bersaglieri, Pardis C. Sabeti, Nick Patterson, Trisha Vanderploeg, Steve F. Schaffner, Jared A. Drake, Matthew Rhodes*, David E. Reich, Joel N. Hirschhorn

* Current affiliation: Cornell University, Ithaca, NY.

Open Archive DOI: http://dx.doi.org/10.1086/421051

Article Info

Altmetric 86

Summary | Full Text | Images | References

Bersaglieri et al. 2004, AJHG DOI: 10.1086/421051

Table 1

Frequencies in Different Populations of Two Alleles Associated with Lactase Persistence

		Frequency (%) for	
Population Group (Region and/or Country)	No. of Chromosomes	**−13910T**	**−22018A**
European American	48	77.2	77.1
African American	100	14.0	13.3
East Asian	35	0	0
Yoruba (Nigeria)	50	0	0
Bantu Northeast (Kenya)	24	0	0
San (Namibia)	14	0	0
Bantu (South Africa)	16	0	0
Mozabite (Mzab, Algeria)	60	21.7	21.7
Bedouin (Negev, Israel)	98	3.1	4.1
Druze (Carmel, Israel)	96	2.1	2.1
Palestinian (Central Israel)	102	3.9	3.9

Bustamante et al. 2005 Nature DOI: 10.1038/nature04240

Vol 437|20 October 2005|doi:10.1038/nature04240 nature

LETTERS

Natural selection on protein-coding genes in the human genome

Carlos D. Bustamante[1], Adi Fledel-Alon[1], Scott Williamson[1], Rasmus Nielsen[1,2], Melissa Todd Hubisz[1], Stephen Glanowski[3], David M. Tanenbaum[3], Thomas J. White[4], John J. Sninsky[4], Ryan D. Hernandez[1], Daniel Civello[4], Mark D. Adams[5], Michele Cargill[4]* & Andrew G. Clark[6]*

Bustamante et al. 2005 Nature DOI: 10.1038/nature04240

Table 1 | Molecular functions and biological processes showing as an excess of positively or negatively selected genes

Category	*P*-value	Number where CI > 0	Number where CI < 0	*N*
Biological process				
Apoptosis	0.00336	12	10 (5)	99 (53)
Cell structure and motility	**0.00008**	**4**	**27 (8)**	**176 (101)**
Ectoderm development	**0.02805**	**1**	**12 (7)**	**98 (61)**
Gametogenesis	0.03411	5	4 (1)	41 (23)
General vesicle transport	**0.00016**	**0**	**14 (4)**	**40 (20)**
Intracellular protein traffic	**0.01151**	**8**	**32 (10)**	**159 (83)**
mRNA transcription	0.00002	29	34 (17)	333 (185)
Natural-killer-cell-mediated immunity	0.03299	1	3 (2)	19 (9)
Nucleoside, nucleotide and nucleic acid metabolism	0.00467	38	65 (28)	568 (311)
Sensory perception	0.04577	9	9 (4)	101 (56)
Molecular function				
Actin family cytoskeletal protein	**0.00008**	**4**	**23 (10)**	**104 (60)**
Cytoskeletal protein	**0.00000**	**7**	**36 (12)**	**205 (118)**
Defence/immunity protein	0.00965	10	12 (5)	89 (54)
Extracellular matrix	**0.01478**	**3**	**15 (11)**	**103 (78)**
Homeotic transcription factor	0.02586	2	2 (0)	28 (8)
Immunoglobulin receptor family member	0.04558	7	3 (1)	36 (27)
Kinase modulator	**0.00538**	**0**	**9 (3)**	**28 (14)**
KRAB box transcription factor	0.00004	17	10 (5)	168 (108)
Membrane traffic protein	**0.02701**	**2**	**17 (6)**	**70 (32)**
Microtubule-binding motor protein	**0.03109**	**1**	**7 (1)**	**31 (19)**
Microtubule family cytoskeletal protein	**0.01373**	**2**	**9 (1)**	**52 (34)**
Non-motor actin-binding protein	**0.01691**	**3**	**12 (3)**	**58 (27)**
Nuclear hormone receptor	0.00143	3	0 (0)	10 (7)
Protein kinase	0.04366	6	4 (1)	94 (36)
Receptor	0.00211	31	37 (18)	343 (207)
RNA helicase	**0.03948**	**0**	**8 (4)**	**33 (18)**
Transcription factor	0.00000	39	41 (19)	428 (240)
Voltage-gated potassium channel	**0.03943**	**0**	**5 (2)**	**23 (11)**
Zinc finger transcription factor	0.00074	20	19 (9)	229 (141)

Classification is based on Panther classification. The *P*-value is the uncorrected value from the Mann–Whitney *U*-test. A total of 139 different molecular functions and 133 biological processes were tested; non-significant categories are listed in Supplementary Table 1, as are significant categories with considerable overlap with those shown in the Table. Parentheses for all count data denote the IPS data set. Bold text indicates negatively selected genes; non-bold text indicates positively selected genes.

Selection

Charles Darwin

On Natural Selection

"One general law, leading to the advancement of all organic beings, namely, multiply, vary, let the strongest live and the weakest die."

"Owing to this struggle for life, variations, however slight and from whatever cause proceeding, if they be in any degree profitable to the individuals of a species, in their infinitely complex relations to other organic beings and to their physical conditions of life, will tend to the preservation of such individuals, and will generally be inherited by the offspring. The offspring, also will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive. I have called this principle, by which each slight variation, if useful is preserved, by the term Natural Selection"

Natural Selection - paraphrased

- More offspring are produced than can possibly survive and reproduce.
- Organisms differ in their ability to survive and reproduce, in part because of their genetic differences.
- In every generation, genotypes that promote survival in the current environment are present in excess at the reproductive age, and thus contribute disproportionately to the offspring of the next generation.
- The progressive genetic improvement in populations resulting from natural selection constitutes the process of `evolutionary adaptation`.

Definitions

Fitness

Some measure of difference in the ability of individuals in a population to survive and reproduce - can be fecundity, viability, survivorship (of gametes, offspring, adults), but ultimately manifests as an `absolute` change in genotypic composition in the next generation, and as a `relative` change in the genotype frequencies in the next generation.

Adaptation

Some trait that enhances the fitness of an individual or a population.

Quantifying Selection in Diploids

Selection in Diploids

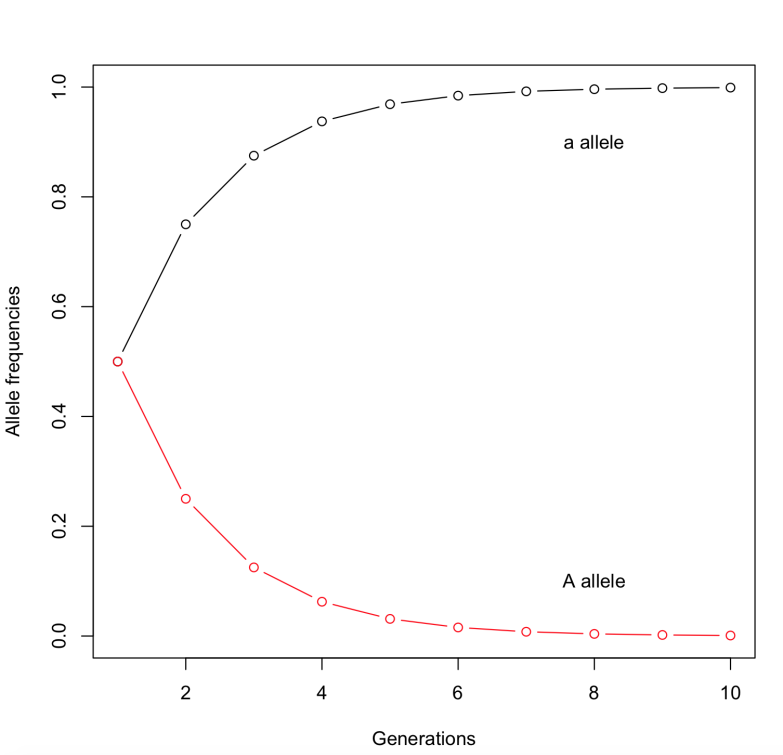
- Consider a randomly mating population at a single biallelic genetic locus, A, with two alleles A and a at frequencies p' and q' at generation t. Let p and q be the allele frequencies in generation $t-1$.
- Let's permit the fitness of each genotype to differ, such that w_{11}, w_{12} and w_{22} are the fitnesses of the AA, Aa and aa genotypes respectively.
- Fitness here can either be the absolute probability of survival of each genotype (absolute fitness), or the fitness of one genotype relative to another (relative fitness).

Selection in Diploids - contd.

- Expected genotype frequencies of AA, Aa and aa in generation $t-1$ are p^2, $2pq$ and q^2 respectively.
- After selection, in generation t, the expected genotype frequencies among surviving adults will be p^2w_{11}, $2pqw_{12}$ and q^2w_{22} respectively.
- If we define the mean fitness of this population to be $\bar{w} = p^2w_{11} + 2pqw_{12} + q^2w_{22}$, relative genotype frequencies are $\frac{p^2w_{11}}{\bar{w}}$, $\frac{2pqw_{12}}{\bar{w}}$ and $\frac{q^2w_{22}}{\bar{w}}$ respectively.
- Among surviving adults, AA genotypes produce all A gametes, and Aa produce half A and half a gametes, and the aa genotypes produce all a gametes, so allele frequencies can be written as: $p' = \frac{p^2w_{11}+pqw_{12}}{\bar{w}}$ and $q' = \frac{q^2w_{22}+pqw_{12}}{\bar{w}}$.
- In other words, $p' - p = \Delta p = \frac{pq[p(w_{11}-w_{12})+q(w_{12}-w_{22})]}{\bar{w}}$

Change in allele frequencies due to viability selection

```
w11=0 #fitness of AA
w12=0.5 #fitness of Aa
w22=1.0 #fitness of aa
freqa<-c(1:10) #define a allele freq vector
freqA<-c(1:10) #define A allele freq vector
freqA[1]=0.5 #freq of A in gen 1
freqa[1]=0.5 #freq of a in gen 1
#simulate 10 generations
for (i in 2:10){ #loop over generations
#compute relative fitness
wbar = freqA[i-1]^2*w11+
2*freqA[i-1]*freqa[i-1]*w12+freqa[i-1]^2*w22
#compute A allele frequency
freqA[i]=(freqA[i-1]^2*w11+freqA[i-1]
*freqa[i-1]*w12)/wbar
#compute a allele frequency
freqa[i]=1-freqA[i]
}
#Plot
plot(freqa, xlab="Generations",xlim=c(1,10),
ylim=c(0.0,1.0),ylab="Allele frequencies",type="b")
points(freqA,col="red",type="b")
text(8,0.1, "A allele")
text(8,0.9,"a allele")
```



Problem 5.5 - Insecticide Resistance

Resistance to organophosphate and carbamate insecticides in species of Culex and Anopheles mosquitoes has been shown to be mediated by 4 independent mutations in the gene $ace-1$. This is a dominant mutation, such that mosquitoes bearing resistance allele are 10 times as likely to survive (relative fitnesses 10:1). How long would it take the allele frequency to rise from 0.01 to 0.5?

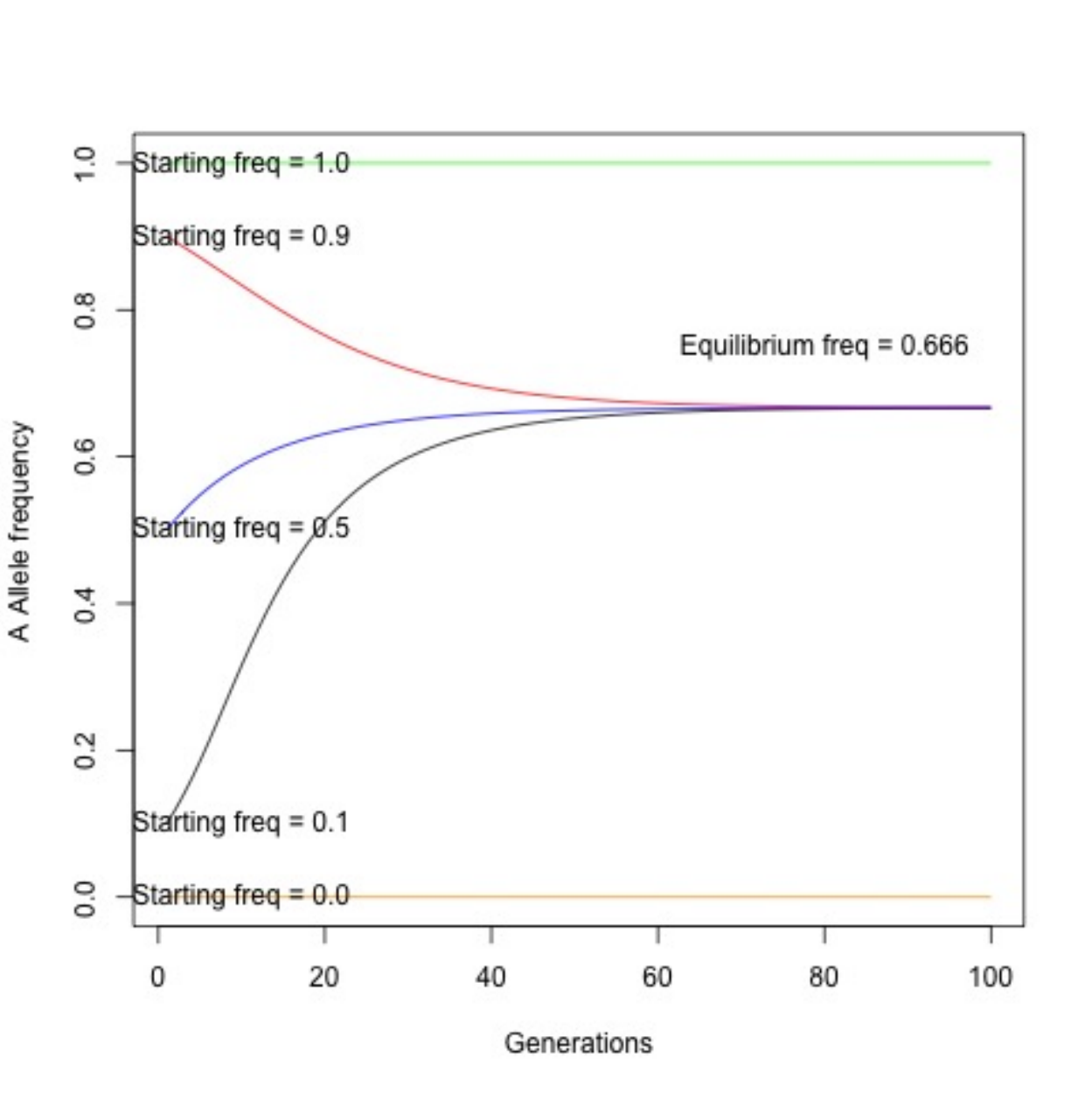
Heterozygote Superiority

Overdominance

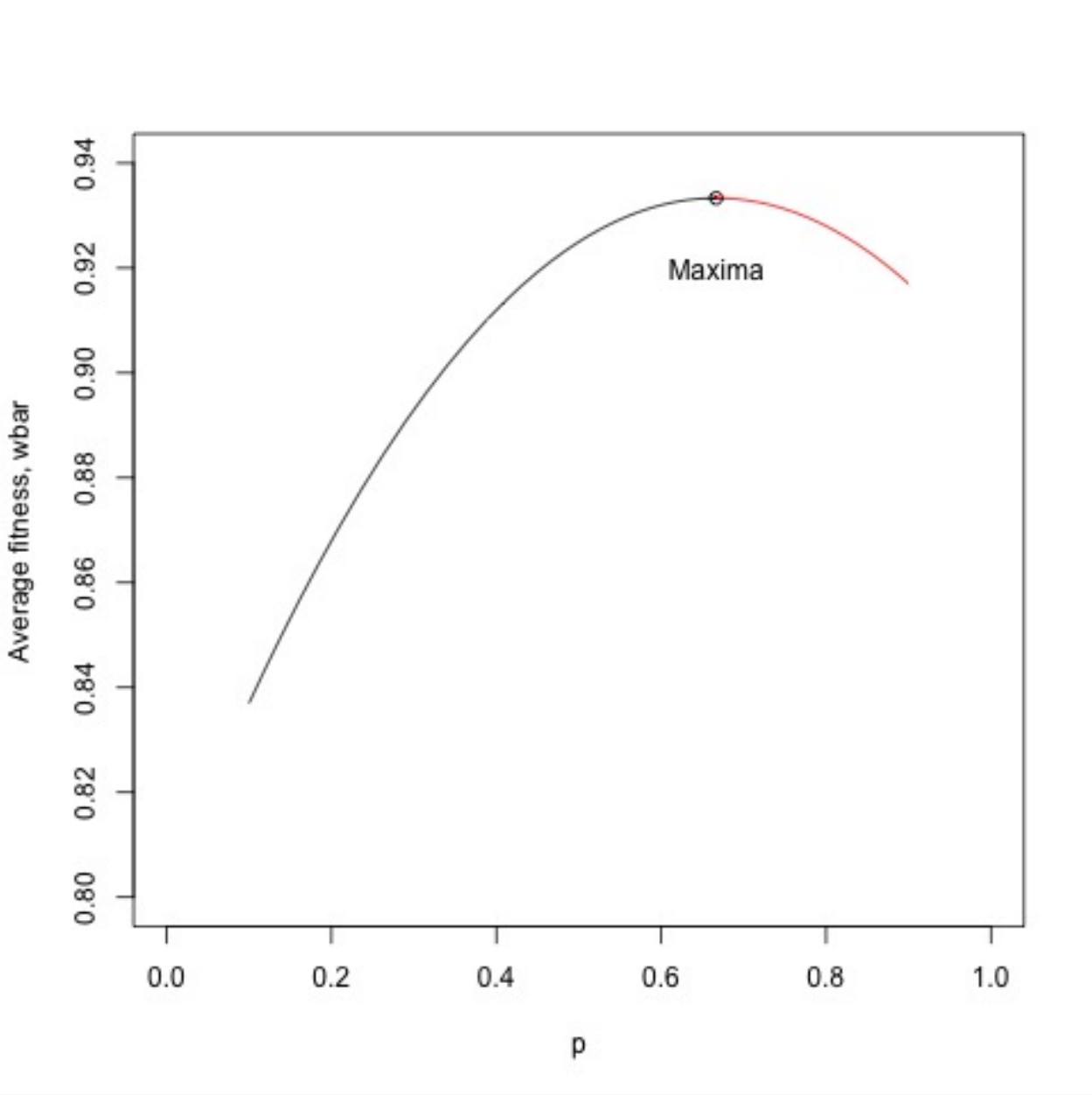
If heterozygotes have higher fitness than either homozygote, the situation is called *overdominance* or *heterozygote superiority*, i.e. $w_{12} > w_{11}$ and $w_{12} > w_{22}$.

- Setting $\Delta p = 0$, we can compute the equilibrium allele frequencies, giving $\hat{p} = \frac{w_{12} - w_{22}}{2w_{12} - w_{11} - w_{22}}$.
- If we set $w_{11} = 1 - s, w_{12} = 1$ and $w_{22} = 1 - t$, we get $\hat{p} = \frac{t}{s+t}$.

Equilibrium Frequencies

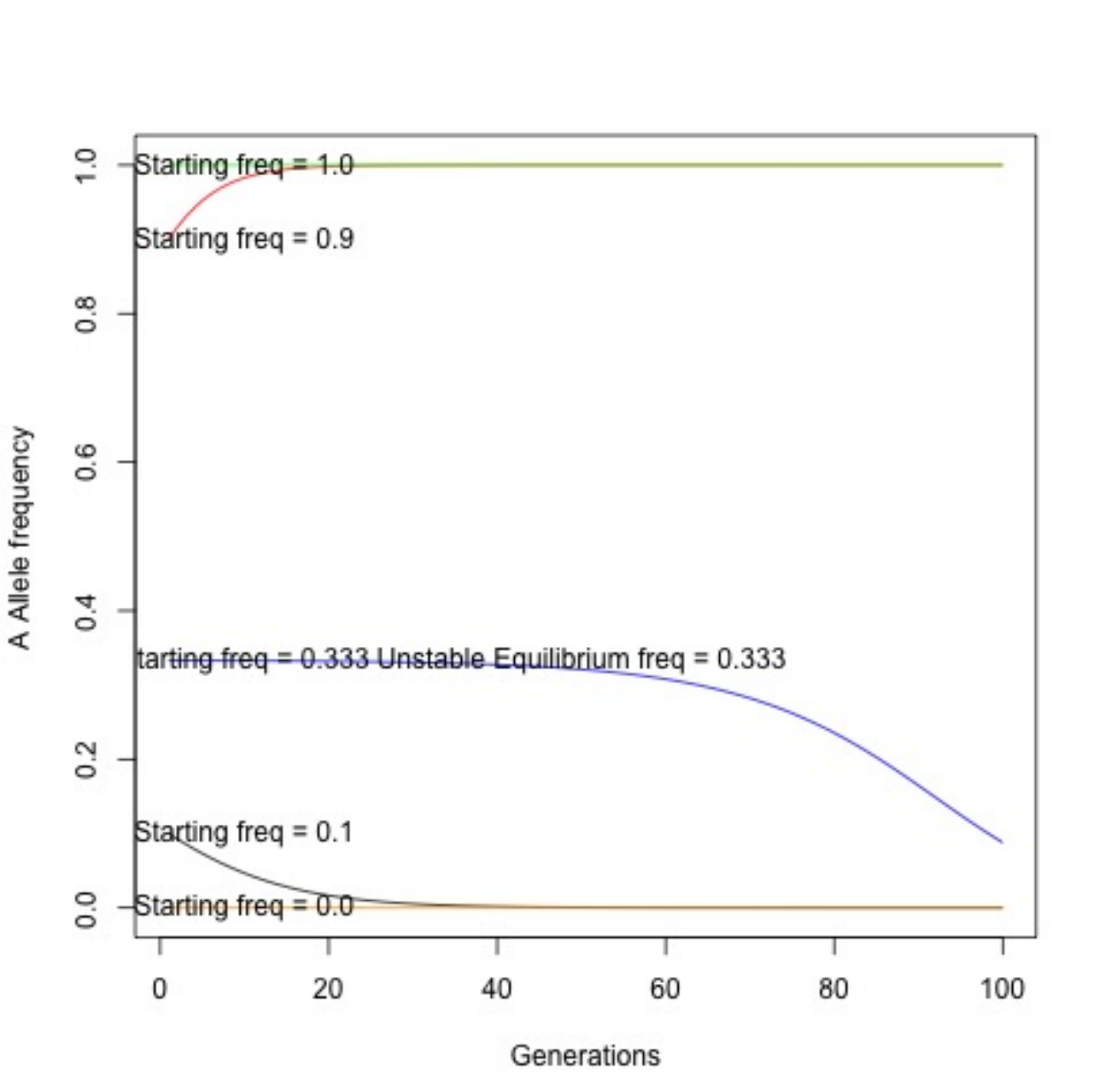


Equilibrium Fitness - Adaptive Landscape



Other Situations - Homework

Heterozygote inferiority ($w_{12} < w_{11}$ and $w_{12} < w_{22}$)



Mutation-Selection Balance

Mutation-Selection Balance

- Recall - most mutations are deleterious, but some persist (probability of loss is ≈ 0.366, but probability of maintaining a low density variant is still $1 - 0.366 = 0.67$, which is fairly high.
- Especially if mutation acts to "load" new mutations continuously, selection can act to remove some of this, but not all.
- So now if μ is the mutation rate per generation per base, and w_{AA}, w_{Aa}, w_{aa} are fitnesses of different genotypes, then we can still write $p' = \frac{p^2 w_{AA} + pq w_{Aa}}{\bar{w}}(1 - \mu)$
- If we write the relative fitnesses as $w_{AA} = 1, w_{Aa} = 1 - hs, w_{aa} = 1 - s$, then we define h as the `degree of dominance` of the a allele, and s as the `selection coefficient`.
- If $h = 0$, then a is completely recessive.
- If $h = 1$, then a is a dominant allele, since AA and Aa have identical fitness.
- If $h = \frac{1}{2}$, we call this semi-dominance.

Mutation-Selection Equilibrium

- When selection is balanced by recurring mutations, we obtain a globally stable equilibrium, such that $p' = p$.
- Solving, we obtain, if $h = 0$, $\bar{q} = \sqrt{\frac{\mu}{s}}$.
- If the harmful allele shows partial dominance ($h > 0$), an excellent approximation is obtained at $\bar{q} = \frac{\mu}{hs}$.

Haldane-Muller Principle

Mutational Load

The reduction in fitness of a population due to recurrent mutations only depends on the mutation rate, and not on the relative fitness of mutant genotypes.

- If a new "harmful" mutation arises in the population (i.e. the allele reduces fitness), then selection can act to purge this easily (and completely), and so $\hat{q} = 0$, and $\bar{w} = 1$.
- If there are recurrent mutations, however, equilibrium frequency of the mutation is greater than 0.
- When $h = 0$ (completely recessive), average fitness of the population at equilibrium $= 1 - \hat{q}^2 s = 1 - \frac{\mu}{s} s = 1 - \mu$
- The reduction in average fitness due to mutation $= 1 - (1 - \mu) = \mu$, also called *mutation load*.
- What *a* is partially dominant ($h > 0$), then mutation load is $\approx 2\mu$, since average fitness at equilibrium is $1 - 2\hat{p}\hat{q}hs - \hat{q}^2 s \approx 1 - 2\mu$.

Other types of Selection

Types of Selection

Fecundity Selection

Differences in fitness between genotypes result from differential ability of mating pairs to produce offspring.

Diversifying Selection

Selection that favors extreme phenotypes, i.e. type of selection in which genotypes are favored merely because they are different. Correspondingly, diversifying selection maintains large allelic diversity. E.x. MHC genes maintained by the resistance to parasitic microbes.

Gametic Selection

Selection at the haploid phase of the life cycle (e.g. sperm/egg/gametophyte).

Types of selection

Epistasis

When a phenotype is controlled by multiple loci, and gene-gene interactions, and thus overall fitness of a phenotype is a function of allele frequencies of both loci, and their relative fitness.

Sexual Selection

Competition for mates as a source of selection, thus selecting for certain allelic combinations (even if these combinations may be detrimental to population fitness) - e.x. plumage displays in birds, horn size in sheep.

Kin Selection

Positive selection for certain alleles takes place indirectly through enhanced reproduction of the genetic relatives of carriers of the alleles, rather than directly through an increased fitness of the carriers - e.x. altruistic behavior in eusocial insects.

Selection and Drift

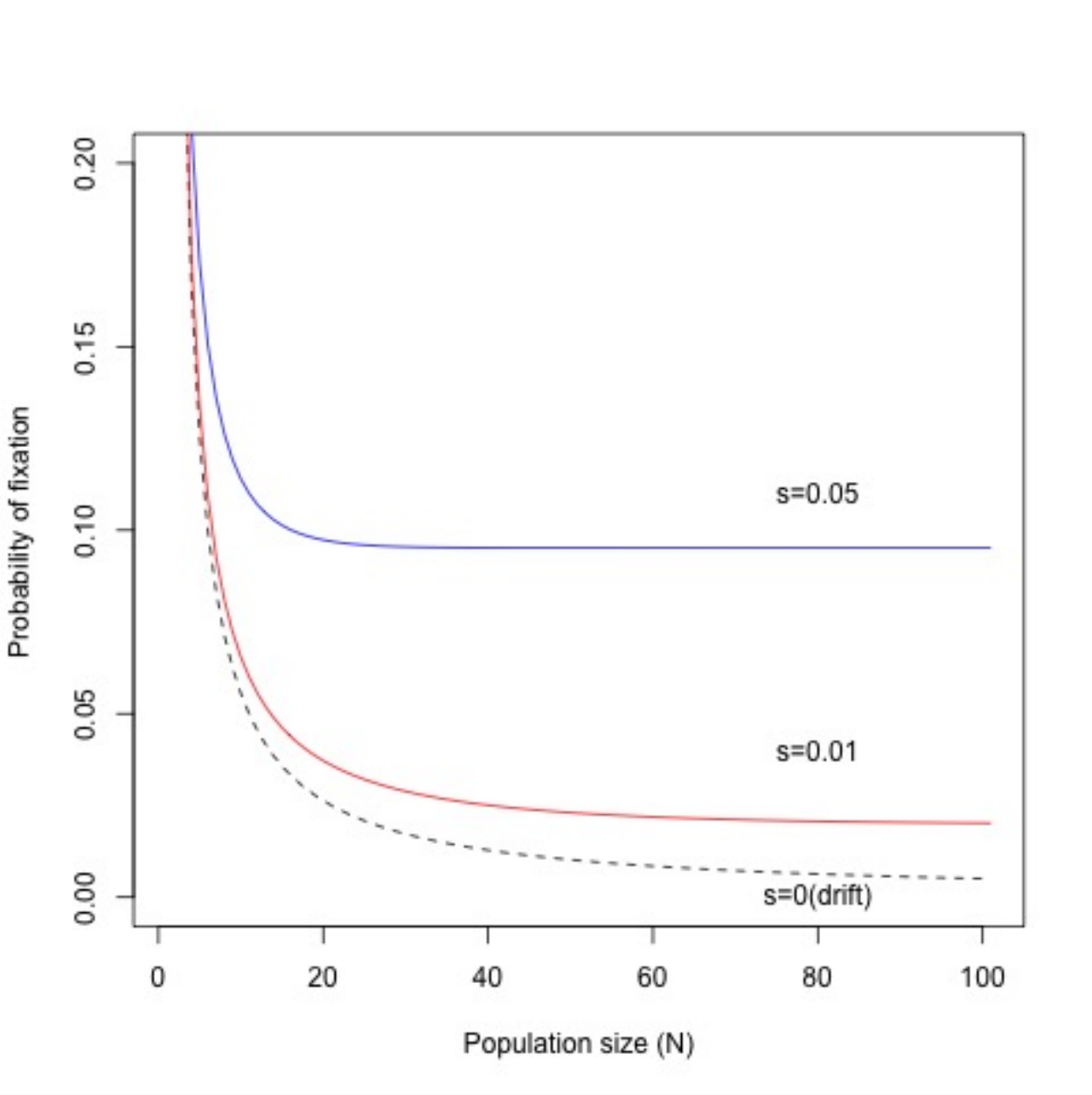
Selection in a finite population

- Selection produces predictable changes in allele frequencies, whereas drift produces random changes from one generation to the next - how do these opposing processes interact?
- In an infinite population, recall, that selection always results in increase/decrease of allele frequencies, depending on its fitness, whereas the effect of drift would be minimal.
- Consider a finite population, of size $2N$ at a diploid, bi-allelic locus, such that there are two alleles, A and a at frequencies p and q respectively.
- Assume that fitness is "additive", such that the selection coefficient of the new mutant allele (a) is s, then the relative fitnesses of AA, Aa, aa are 1, $1+s$, and $1+2s$ respectively.
- Under this assumption, Kimura (1957, 1962) derived the fixation probability of a new mutant allele (with initial frequency p) to be $\approx \frac{1-e^{-4N_e sp}}{1-e^{-4N_e s}}$.

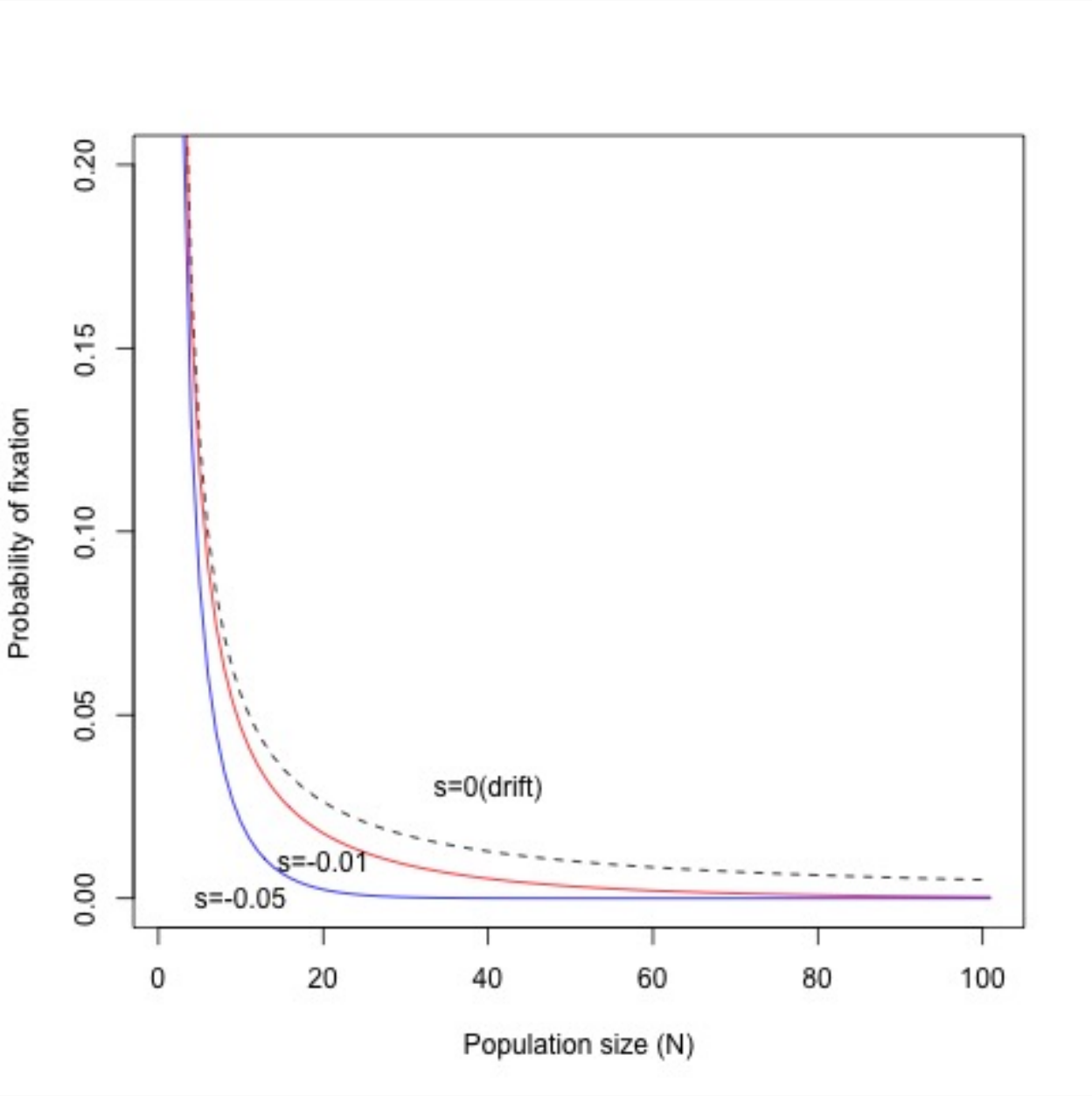
Selection-Drift Equilibrium

- Initial frequency, $p = \frac{1}{2N}$.
- So $Pr(fix) \approx \frac{1-e^{-2\frac{N_e}{N}s}}{1-e^{-4N_e s}}$
- At equilibrium, if we set $N_e = N$, we get $Pr(fix) = \frac{1-e^{-2s}}{1-e^{-4Ns}}$
- If the mutation is advantageous ($s > 0$), if s is small, then if N is large, $Pr(fix) \approx 2s$.
- If N is small, and $4Ns$ is small (i.e. if s is also small), then $Pr(fix) \approx \frac{1}{2N}$.
- How about if the mutation is deleterious? ($s < 0$)

Probability of Fixation - Advantageous allele



Probability of Fixation - Deleterious allele



Methods to detect selection

Kinds of selection

Purifying selection

Deleterious mutations are eliminated to preserve the function of the protein or DNA sequence.

Positive selection

Also called Darwinian selection, where rare favorable mutations are selected for, resulting in the substitution of the new mutation for the previous best allele.

Balancing selection

Selection that acts to maintain two or more variants at a locus - e.x. overdominance.

Synonymous versus non-synonymous substitutions

- Assume that the number of synonymous substitutions is dS, and number of non-synonymous substitutions is dN.
- If on an average, a population has a greater number of non-synonymous substitutions than synonymous substitutions (i.e. $dN > dS$), we would hypothesize that the population has an apparent excess of replacement polymorphisms, and hence has experienced positive Darwinian selection for these new mutations, so $\frac{dN}{dS} > 1$.
- If on the other hand, $dN < dS$, then the population on an average has experienced negative purifying selection against replacement polymorphisms, so $\frac{dN}{dS} < 1$.
- If $dN = dS$, then the population is said to be evolving neutrally.

Formal interpretation of dN/dS ratios

Neutrality of non-synonymous mutations

At a single synonymous site (already neutral), number of mutations fixed per generation is: dS = mutations fixed/generation = (mutations arising per generation)x(probability of fixation) $= 2N\mu \times \frac{1}{2N} = \mu$.

At a non-synonymous site (assumed neutral here), similarly, $dN = \mu$.

So $\frac{dN}{dS} = \frac{\mu}{\mu} = 1$.

A fraction f of non-synonymous mutations are deleterious

Here $dS = \mu$, and if a fraction f of dN are deleterious, then $dN = f\mu + (1 - f)0 = f\mu$.

So $\frac{dN}{dS} = \frac{f\mu}{\mu} = f$. Since $f < 1.0$, $\frac{dN}{dS} < 1$.

Interpretation - contd.

Fraction f of all non-synonymous mutations are non-deleterious, fraction θ are advantageous

Again, $dS = \mu$.

Of the non-synonymous mutations, $1 - f$ are deleterious, and do not fix.

A fraction $f(1-\theta)$ are neutral and fix at the rate of μ per generation.

A fraction $f\theta$ are advantageous, that arise at the rate of $2N\mu$ per generation and fix with a probability s, where s is the selection coefficient.

$$dN = (1-f)0 + f(1-\theta)\mu + f\theta 2N\mu s$$

$$\frac{dN}{dS} = \frac{f(1-\theta)\mu + f\theta 2N\mu s}{\mu} = f(1-\theta) + f\theta 2N\mu s$$

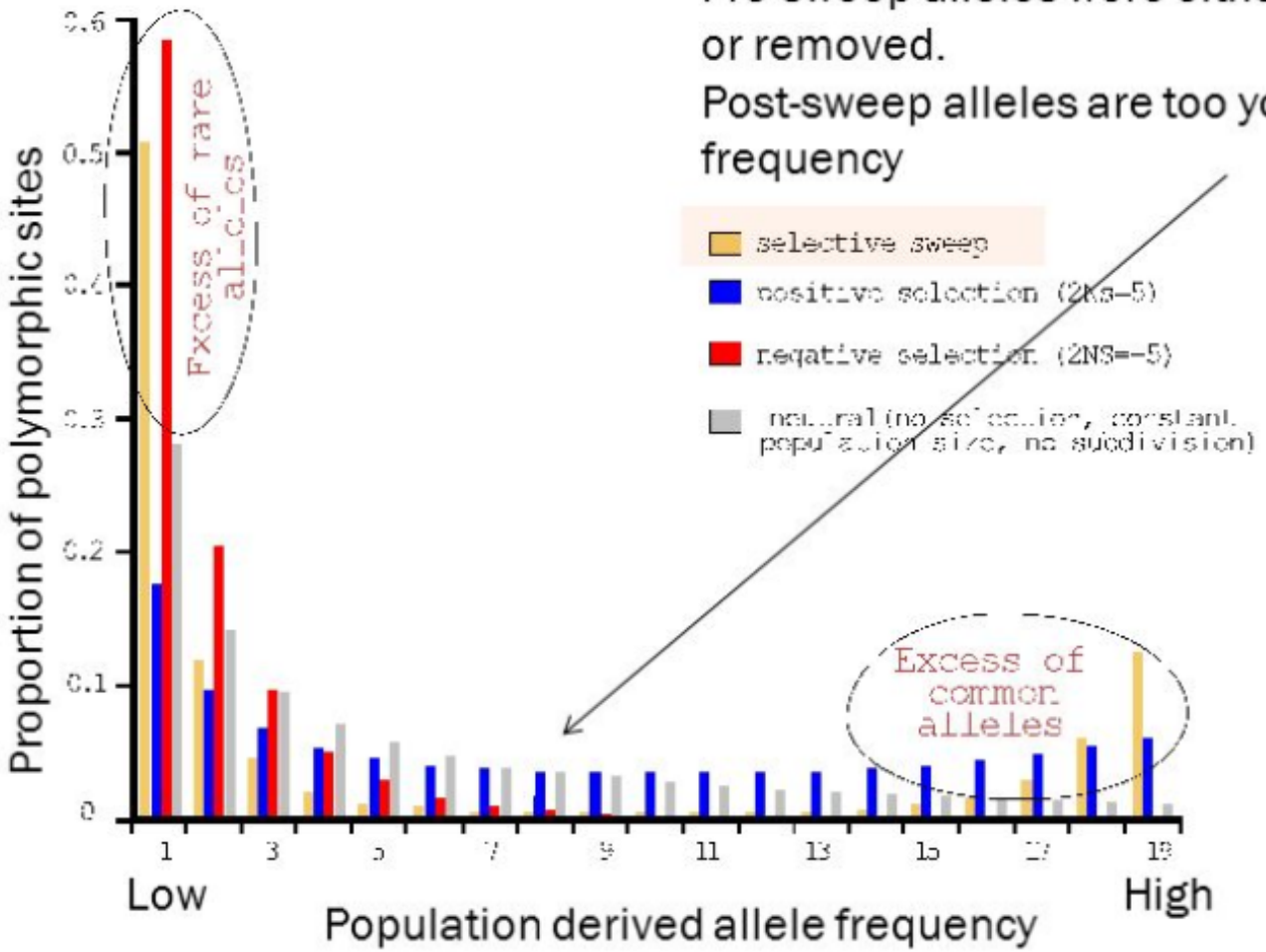
This can be > 1, especially if θ is large.

So $\frac{dN}{dS} > 1$.

Expected Allele Frequency Distributions

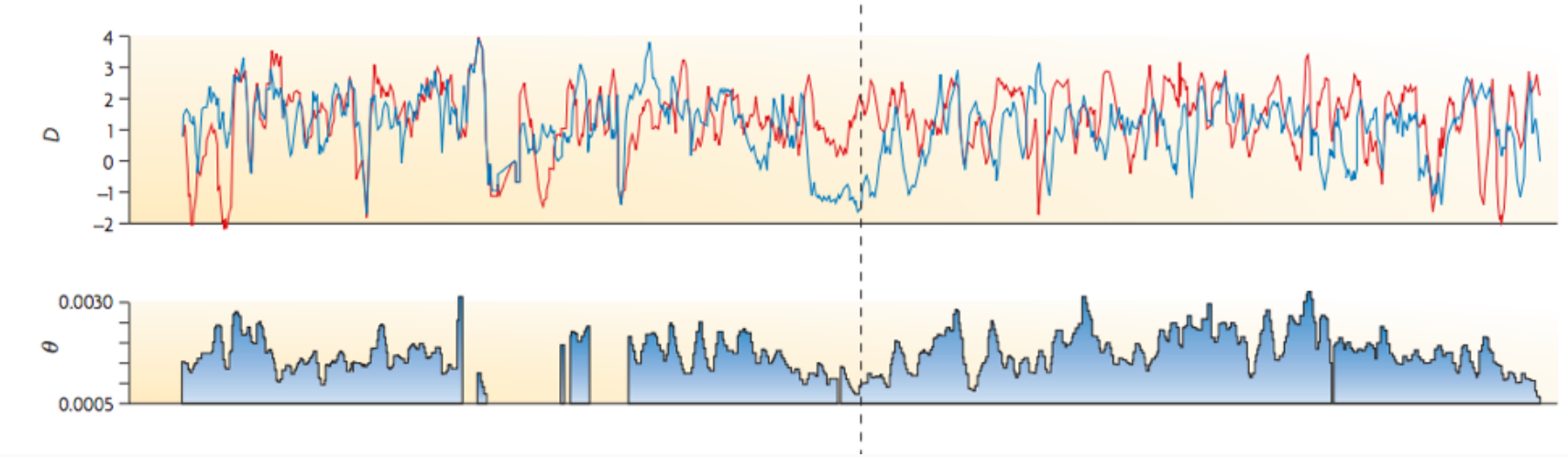
- Ewens (1972) worked out the equilibrium distribution of allele frequencies and showed that the expected number of different alleles, k in a sample of size n is: $E(k) = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \ldots + \frac{\theta}{\theta+(n-1)}$
- This is also called the site frequency spectrum, or allele frequency spectrum.
- At a neutral locus (drifting), we would anticipate that the SFS has a unimodal distribution, with the mode occurring at the most commonly observed allele, and a long tail.
- If selection acts to "skew" this distribution (away from neutral expectations), we can use those skews as a signature of selection.
- For e.x. if the most common allele is observed at a greater frequency than expected under neutrality, and correspondingly rare alleles are rarer, showing an "excess of rare alleles", i.e. many alleles whose frequencies are lower.
- This could be an indicator of purifying or negative selection, wherein some alleles are kept at lower frequencies than expected under neutrality.

Example


Few medium frequency derived alleles.
Pre-sweep alleles were either swept to high frequency or removed.
Post-sweep alleles are too young to reach medium frequency
selective sweep
positive selection (2Ns=5)
negative selection (2Ns=-5)
Excess of rare alleles
Excess of common alleles
Proportion of polymorphic sites
Low
High
Population derived allele frequency

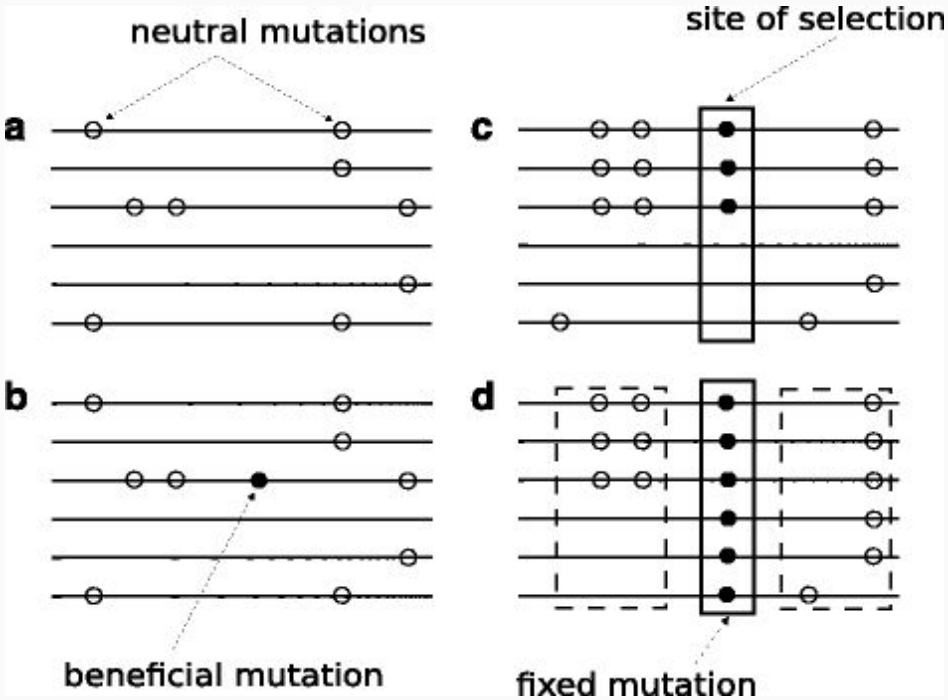
SFS Based Test of Tajima's D

- Recall, Tajima's D is a measure of relative diversity in rare alleles.
- If there is an excess of rare alleles (as previously shown to occur immediately after a bottleneck), due to a selective sweep (positive Darwinian selection), then Tajima's D would be expected to be negative.
- Figure from Nielsen et al. 2007, Nature Reviews Genetics, DOI: 10.1038/nrg2187, showing the values of Tajimas D, and $\theta = 4N_e$ estimated across the LCT (lactase) gene in humans (Asian, and European populations).



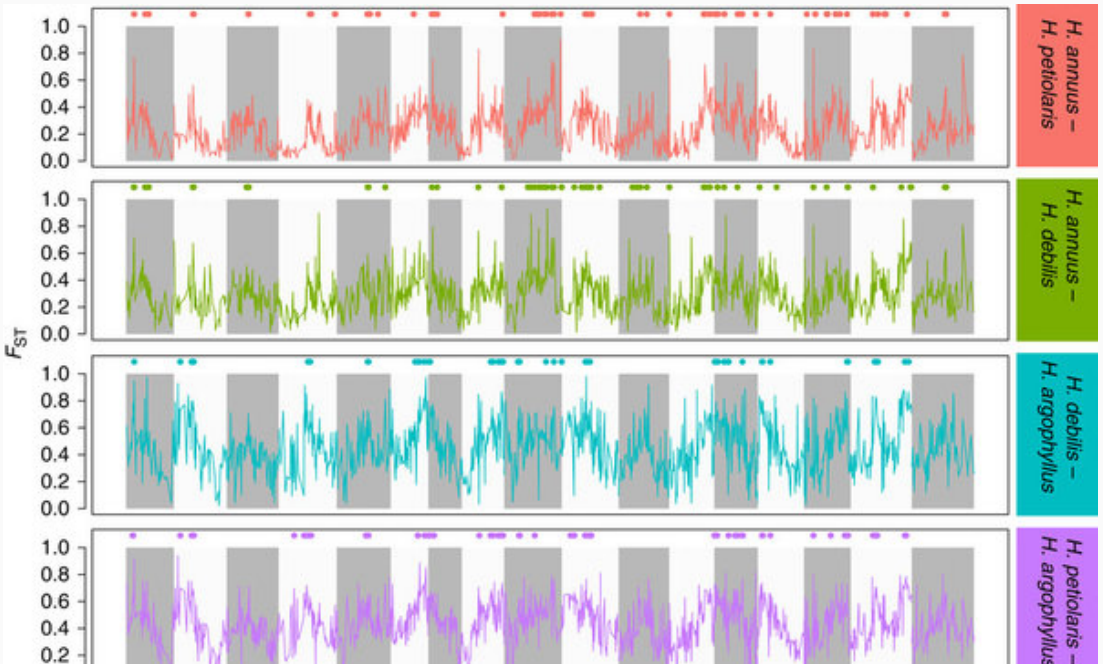
Linkage Disequilibrium

- If there's a selective sweep for a beneficial allele, all linked variants are also 'swept' with it, in a process called *genetic hitchhiking*.
- So LD would increase at a site that has undergone a recent selective sweep.
- This LD will then be broken up over evolutionary time by recombination.



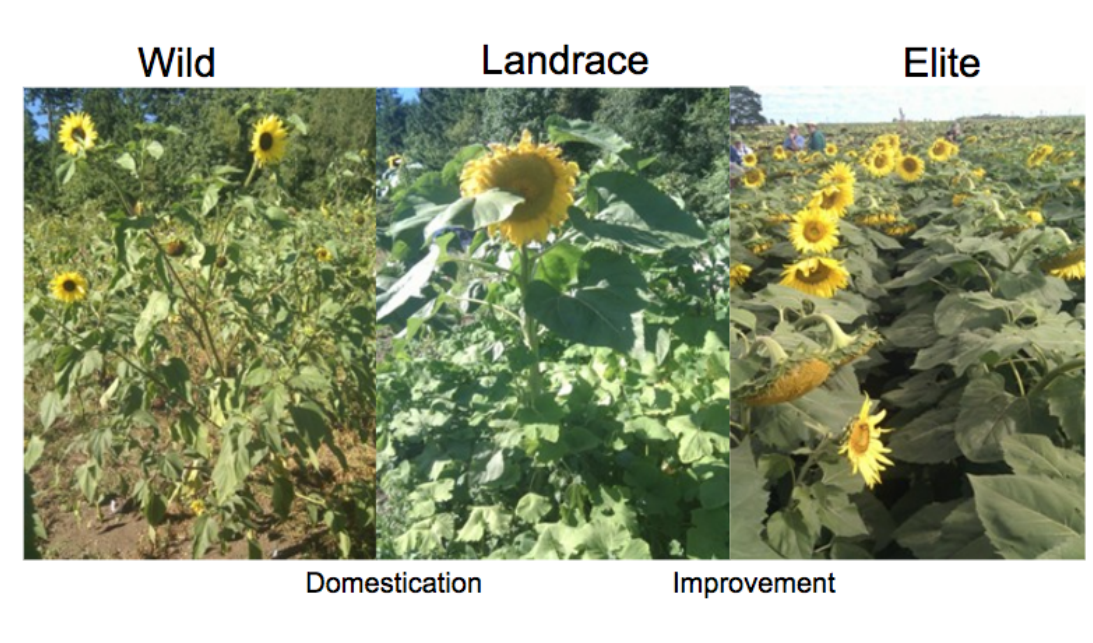
Population differentiation

- Selection increases the degree of differentiation when comparing two populations.
- For example, a gene that has undergone a selective sweep or purifying selection in one population, and not in the other, will appear to be highly differentiated among the two populations.
- Example - selection scan using F_{ST} in different populations of sunflowers to identify selection, Renaut et al. 2013, Nature Communications DOI:10.1038/ncomms2833



Identifying domestication genes in sunflowers

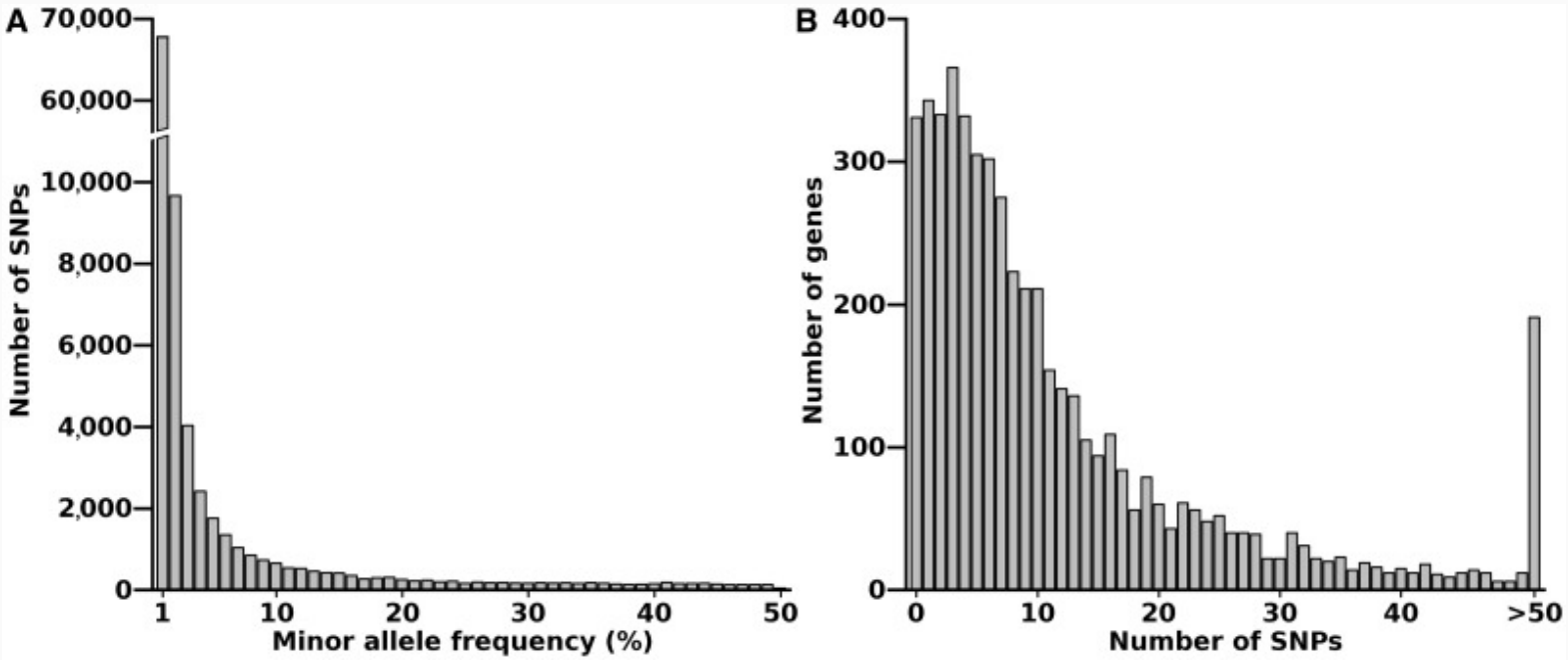
- Goal: To identify candidate genes for improvements in domesticated strains, specifically selecting for fatty acid biosynthesis.
- Approach: Sequence three populations of sunflowers - "wild", "domestication", and the "improvement" strains.
- Genome-wide scan for selection across these strains to identify domestication candidate genes, then use those to inform which genes to "modify" for improvement.



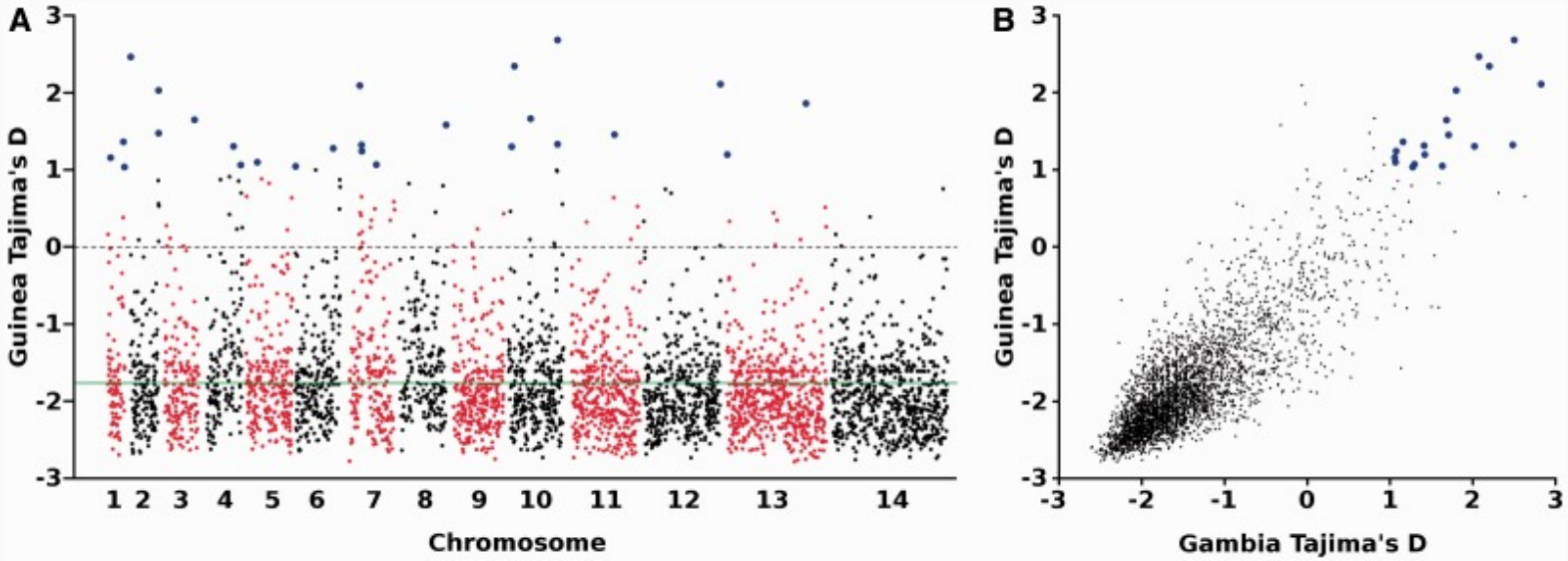
Example - Mobegi et al. 2014 DOI: 10.1093/molbev/msu106

- 100 *P. falciparum* isolates from Guinea were sequenced, compared to 52 isolates from The Gambia, to identify locally varying selective pressures, and their effects on drug pressure, host immunity, transmission opportunities between hosts.

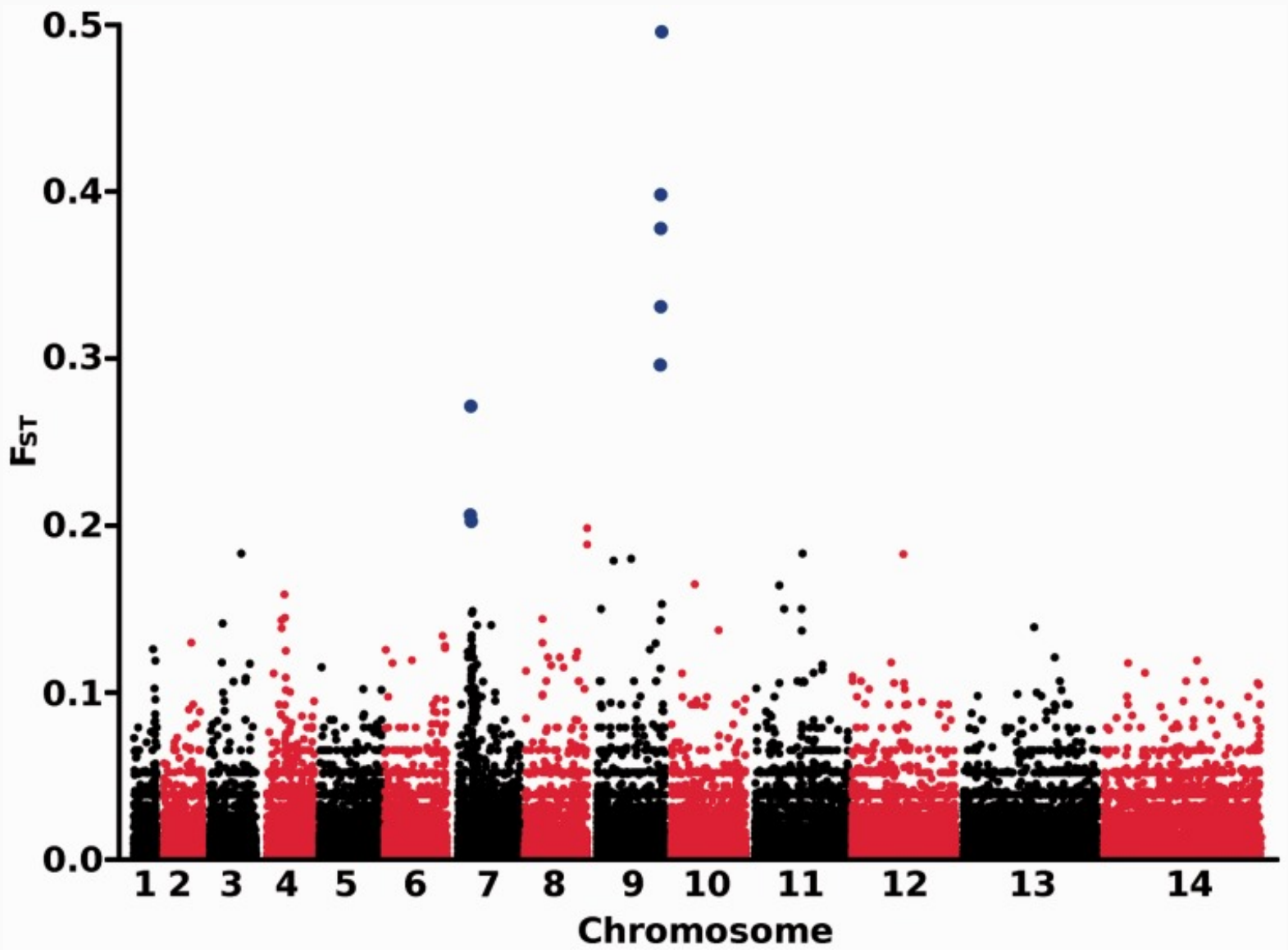
SFS


A
70,000
60,000
10,000
8,000
6,000
4,000
2,000
0
Number of SNPs
1
10
20
30
40
50
Minor allele frequency (%)
B
400
300
200
100
0
Number of genes
0
10
20
30
40
>50
Number of SNPs

Tajima's D


A
Guinea Tajima's D
3
2
1
0
-1
-2
-3
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Chromosome
B
Guinea Tajima's D
Gambia Tajima's D
-3 -2 -1 0 1 2 3

Fst



Questions?