BIOL 502 Population Genetics Spring 2017

Week 4 Mutation and Neutral Theory

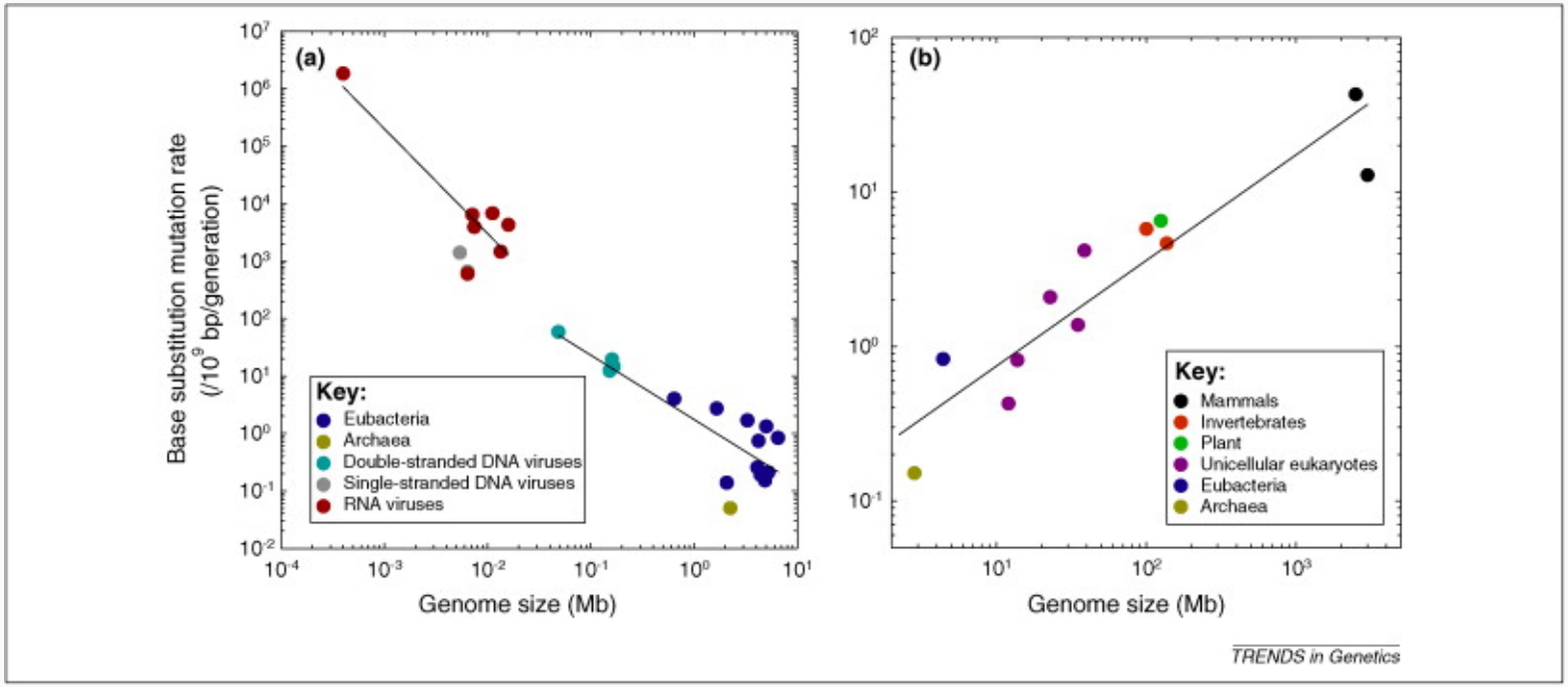
Arun Sethuraman

California State University San Marcos

Table of contents

Mutation Rates

Mutation Rate Variation across the tree of life - Lynch 2010 DOI: 10.1016/j.tig.2010.05.003



Types of Mutations

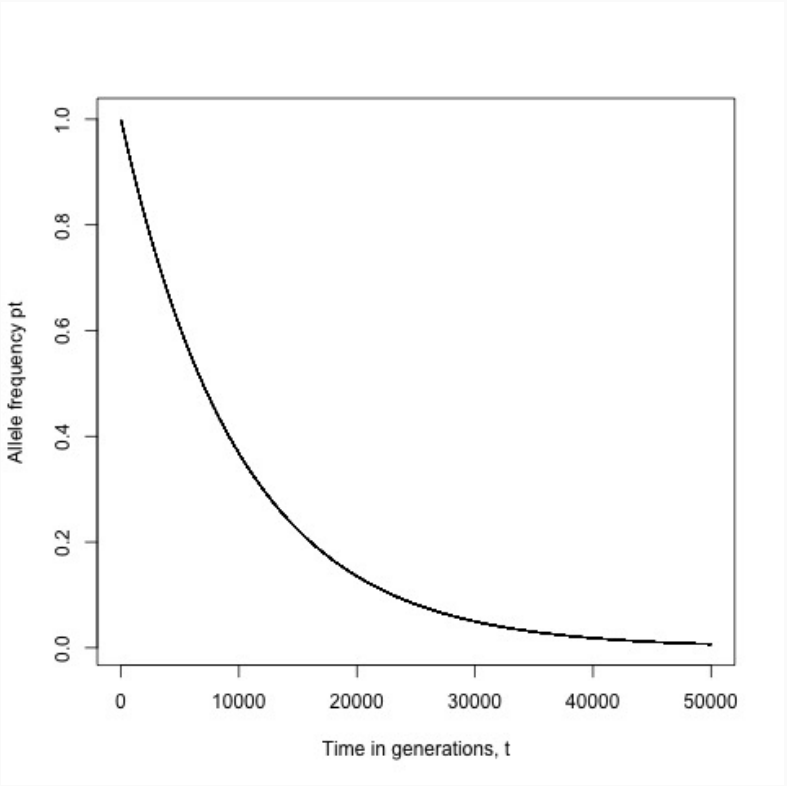
- Point Mutation - Substitutions
  - Transversions/Transitions
  - Silent/synonymous
  - Non-synonymous
- Frameshift Mutations
  - Insertions
  - Duplication
  - Deletions
  - Inversion
  - Missense
  - Nonsense
  - Translocations

Mutation Pressure - Irreversible

- Consider a gene with two alleles, A and a, such that A allele mutates into the a allele with the probability of μ. Let frequency of A and a in generation t be p_t and q_t respectively, such that $p_t + q_t = 1$.
- Assume NO DRIFT (infinitely large population), and NO BACK MUTATIONS!
- The frequency p_t thus depends on the frequency of the A allele in generation $t-1$, i.e. p_{t-1}, as well as the mutation pressure, the number of A alleles in generation $t-1$ that did not mutate.
- $p_t = p_{t-1}(1-\mu)$
- Similarly, $p_{t-1} = p_{t-2}(1-\mu)\ldots$.
- Hence $p_t = p_0(1-\mu)^t$

Mutation Pressure - Irreversible

```
p0<-1.0 #Starting allele frequency of A allele
g<-50000 #Number of generations
mut<-1e-4 #Mutation rate
pt<-c(1:g) #Define vector pt
#for allele frequencies
for(t in 1:g){
pt[t] = p0*(1-mut)^t;} #Simulate
plot(pt,type="l",xlab="Time in generations,
t", ylab="Allele frequency pt") #Plot
```

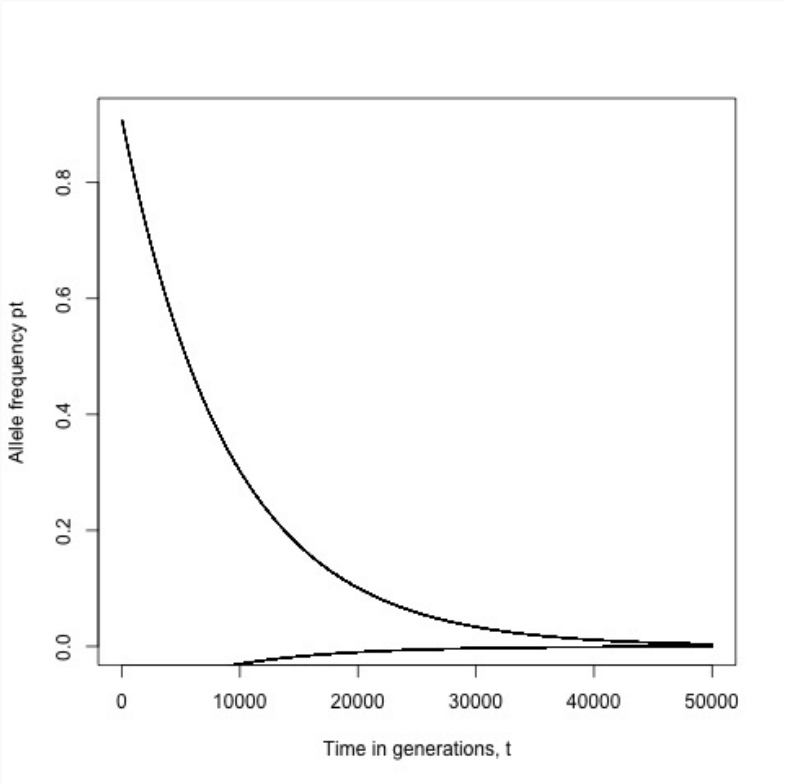


Mutation pressure - Reversible

- Now consider reversible mutations, such that the probability of a back mutation, i.e. a to an A allele at this same locus is ν.
- Now the frequency p_t of the A allele in generation t depends on the frequency of A alleles in the $t-1$ generation which do not mutate, and the frequency of a alleles that do mutate.
- Hence $p_t = p_{t-1}(1-\mu) + q_{t-1}\nu = p_{t-1}(1-\mu) + (1-p_{t-1})\nu$
- Using the same logic as before, we get the recursion: $p_t - \frac{\nu}{\mu+\nu} = (p_0 - \frac{\nu}{\mu+\nu})(1-\mu-\nu)^t$
- If we consider a very long period of time, such that t is large, such that $(1-\mu-\nu)^t$ is 0.
- The allele frequency reaches an equilibrium, $\hat{p} = \frac{\nu}{\mu+\nu}$

Mutation pressure - Reversible

```
p0<-1 #frequency of A allele
for(t in 1:g){
pt[t]=(p0-1e-5/(1e-4+1e-5))*(1-1e-4-1e-5)^t;}
#Simulate
plot(pt,type="l",xlab="Time in generations,
t", ylab="Allele frequency pt") #plot
p0<-0 #frequency of A allele
for(t in 1:g){
pt[t]=(p0-1e-5/(1e-4+1e-5))*(1-1e-4-1e-5)^t;}
#Simulate
points(pt,type="l") #plot
```



Mutation-Drift Equilibrium

Probability of Fixation

- Now let's relax the assumption of no drift (i.e. the population has a finite size, $2N$ comprised of A and a alleles).
- So if μ is the mutation rate on average of any allele mutating into another, if a population is of size $2N$, then probability of picking two A alleles (IBS) to form the next generation is $\frac{1}{2N}$.
- In other words, probability of "losing" the A allele in one generation $= (1 - \frac{1}{2N})^{2N} \approx e^{-1} = 0.368$.
- This number is fairly high - showing that the risk of any allele being lost by chance is substantial.

Neutral Theory

Neutral Theory

Neutral Theory

Mutation introduces new alleles into a population, and drift determines if the neutral alleles are ultimately fixed or lost, though most new mutations are lost. At equilibrium, there is a balance between mutation and drift.

Infinite Alleles Model

Every new mutation creates a new allele that does not already exist in the population.

Infinite Sites Model

A mutation occur at unique sites - i.e. there are no recurrent mutations.

Infinite Alleles Model

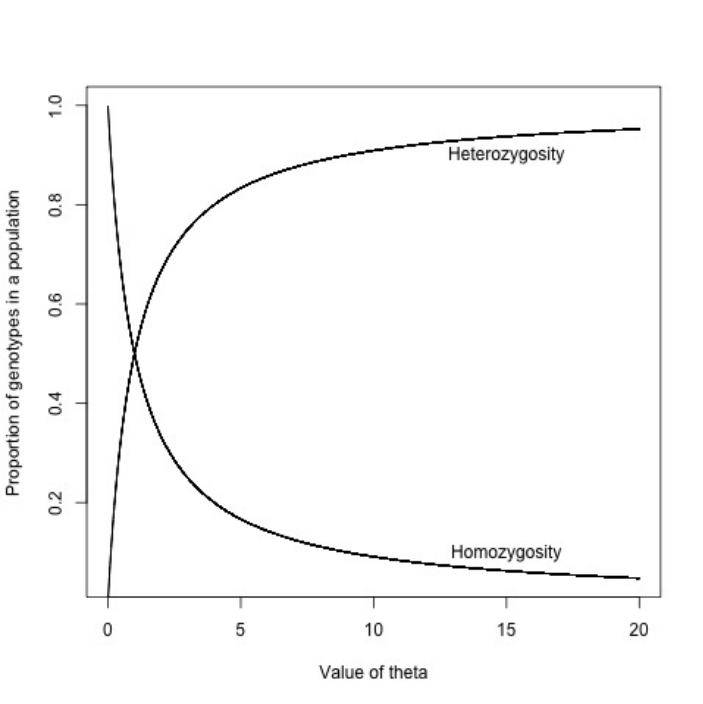
- Each mutation produces a new allele that is not already present in the population - so all homozygotes must be autozygotes (i.e. IBD).
- Homozygosity $=$ autozygosity.
- Recall, F_t is the probability that in generation t, two randomly chosen alleles from a population are IBD.
- Let μ be the mutation rate per generation.
- A homozygote $\alpha_i\alpha_i$ would then have a probability of $\frac{1}{2N}(1-\mu)^2$.
- A heterozygote $\alpha_i\alpha_j$ would have a probability of $(1-\frac{1}{2N})(1-\mu)^2 F_{t-1}$.
- So $F_t = (\frac{1}{2N})(1-\mu)^2 + (1-\frac{1}{2N})(1-\mu)^2 F_{t-1}$.

Redefining F

- Eventually, F_t will reach an equilibrium, when $F_{t-1} = F_t$.
- Solving, we get $\hat{F} = \frac{1}{1+4N\mu}$
- Here N can be interpreted as the effective population size, N_e, and if we write this quantity, $4N_e\mu = \theta$, then $\hat{F} = \frac{1}{1+\theta}$.
- Alternately, if the population is not homozygous, it has to be heterozygous, and so the heterozygosity can be written as $1 - \hat{F} = \frac{\theta}{1+\theta} = \frac{4N_e\mu}{1+4N_e\mu}$.

Heterozygosity vs Homozygosity

```
theta<-c(1:50000)
theta<-theta*4*1e-4
het<-c(1:50000)
hom<-c(1:50000)
hom<-1/(1+theta)
het<-1-hom
plot(theta,hom,xlab="Value of theta",
ylab="Proportion of genotypes in a population"
points(theta,het,type="l")
text(15,0.9,"Heterozygosity")
text(15,0.1,"Homozygosity")
```



Infinite Sites Model

Segregating Sites S

The number of nucleotide sites in the sample that are occupied by two or more nucleotides. $E(S) = \theta \sum_{i}^{n-1} \frac{1}{i}$

Nucleotide Mismatches Π

The number of nucleotide sites in the sample that differ between all individual pairs of aligned sequences. $E(\Pi) = \theta$

Tajima's D

If we define $a = \sum_{i=1}^{n-1} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1}$, $\hat{\theta}_W = \frac{S}{a}$, and $\hat{\theta}_T = \Pi$, then the difference between these two estimates can be defined as Tajima's D $= \frac{\Pi - \frac{S}{a}}{\sqrt{V(\Pi - \frac{S}{a})}} = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{V(\hat{\theta}_T - \hat{\theta}_W)}}$

Interpreting Tajima's D - Courtesy Arun Durvasala

- Consider a population that has undergone a recent bottleneck. We would expect that if we sample sequences after the bottleneck, most mutations observed are rare, i.e. appear in the population after the bottleneck.
- Remember these rare mutations can very well have occurred prior to the bottleneck as well, and could have drifted by chance through the bottleneck.
- Now let's consider some example alignments:
  __*__*_____
  _____-*__*-
  _____-*_____
  ________*-
- Here - indicates nucleotides that are identical, and * indicates mutations/differences.

Interpreting Tajima's D

Example 1:

__*__*_____

_____*__*_

_____*_____

_____*_

Example 2:

_*_____

__*_____

_____*_____

_____*_

Example 1:

$\hat{\theta}_T = \frac{2+1+3+1+1+2}{6} = 1.67$

$\hat{\theta}_W = \frac{3}{1+\frac{1}{2}+\frac{1}{3}} = 1.63$

So in this case, the $\hat{\theta}$ estimates are similar

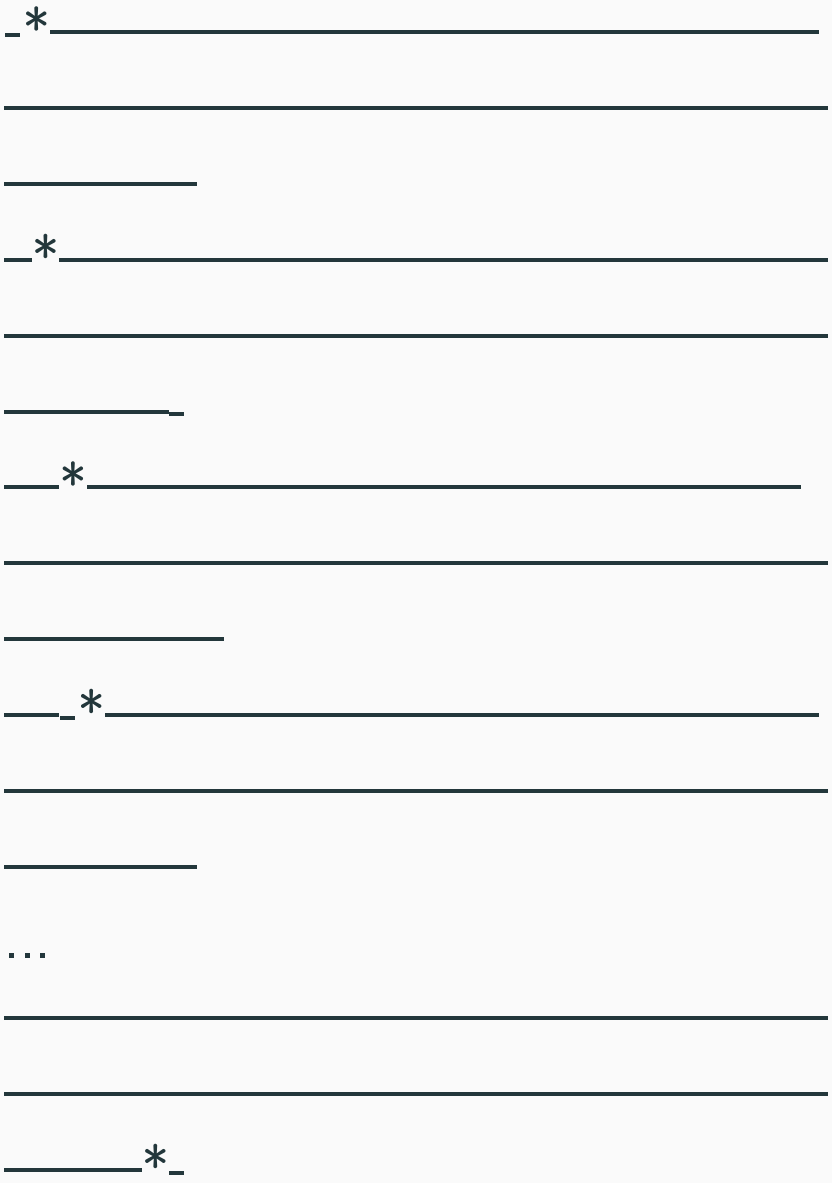
Example 2:

$\hat{\theta}_T = \frac{2+2+2+2+2+2}{6} = 2$

$\hat{\theta}_W = \frac{4}{1+\frac{1}{2}+\frac{1}{3}} = 2.2$ Here, because each mutation is a rare variant, $\hat{\theta}_T$ is lower than $\hat{\theta}_S$, although not pronounced because of small number of samples.

Effect of rare variants

Example 3:



Now $\hat{\theta}_T = \frac{2(99.5)}{99.5} = 2$

$\hat{\theta}_W = \frac{100}{1+\frac{1}{2}+\frac{1}{3}\cdots\frac{1}{99}} = 19.31$

- So as a result, if there are lots of rare variants across the sample, there is a huge difference between estimates of $\hat{\theta}_T$ and $\hat{\theta}_S$.
- Now think - the effects of a bottleneck on a population are that at the end of the bottleneck, most haplotypes will be the same.
- So any mutations that occur will be rare.
- Hence - Tajima's D will be negative.

Examples

Laurent et al. 2016 DOI:10.1111/mec.13385

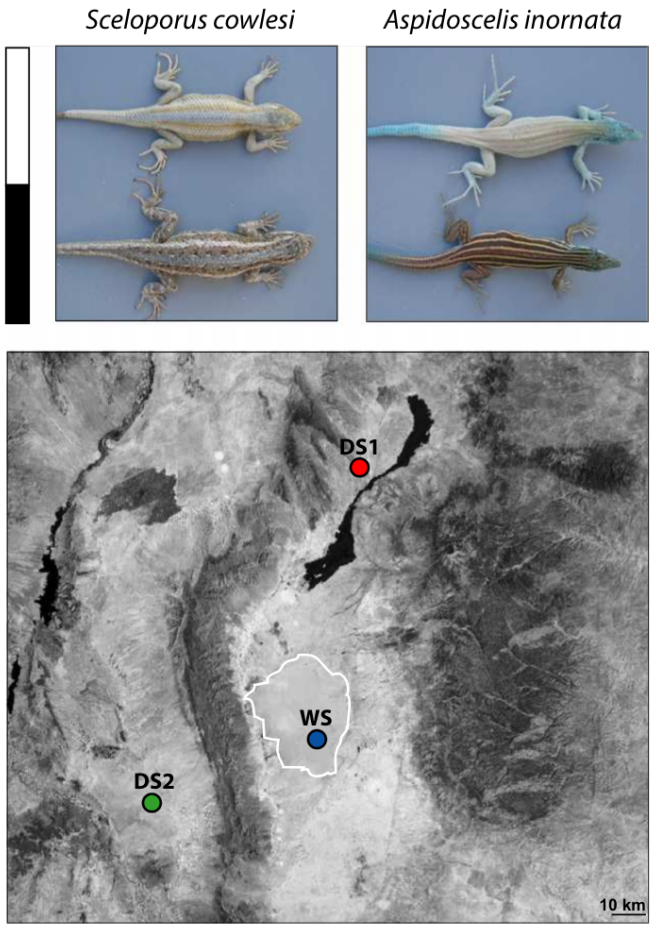


Fig. 1 Photographs and sampling localities for *Aspidoscelis inornata* and *Sceloporus cowlesi* from contrasting habitats. Blanched colour morphs are found at White Sands (indicated by the white bar) and dark colour morphs are found in the rest of the species' ranges (indicated by the black bar). Both species were sampled from the same three localities in New Mexico: White Sands National Monument (WS) in Otero County (blue), a dark soil Bureau of Land Management site (DS1) in Lincoln County (red) and a dark soil Jornada Long-term Ecological Research site (DS2) in Doña Ana County (green).

- Two species of white sand lizards
- *Mc1r* gene has been previously found to be associated with adaptation to white sands, i.e. blanched coloration.
- Quantified sequence variation across 50kb of the *Mc1r* gene among dark sand (DS) and white sand (WS) populations of both species.

Laurent et al. 2016

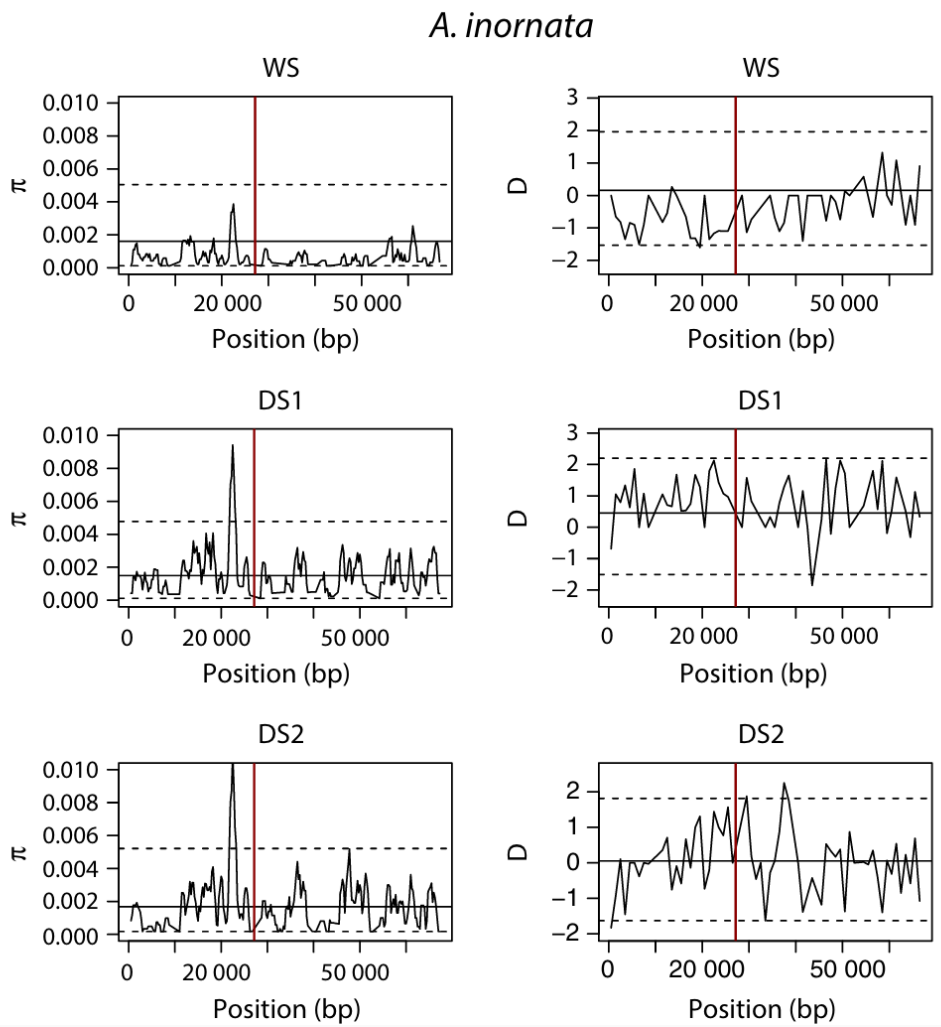


Fig. 6 Nucleotide diversity (π) and Tajima's *D* at the *Mc1r* locus in *Aspidoscelis inornata*. Sliding window profile of nucleotide diversity (π) and Tajima's *D* in the *Mc1r* region as calculated by VCFtools (window size 1000 bp and step size 250 bp). Sites that did not pass quality control were masked (using the '–mask' option). The solid horizontal lines represent the average values of these statistic calculated across all windows in the genomic background. The dashed horizontal lines represent the 0.025th and 0.975th quantiles of the same distribution. The red vertical solid line indicates the position of the nonsynonymous *Mc1r* mutation reported by Rosenblum *et al.* (2010).

Questions?