

Inferring the selective pressures acting on insertions and deletions in the great tit genome

Henry Barton

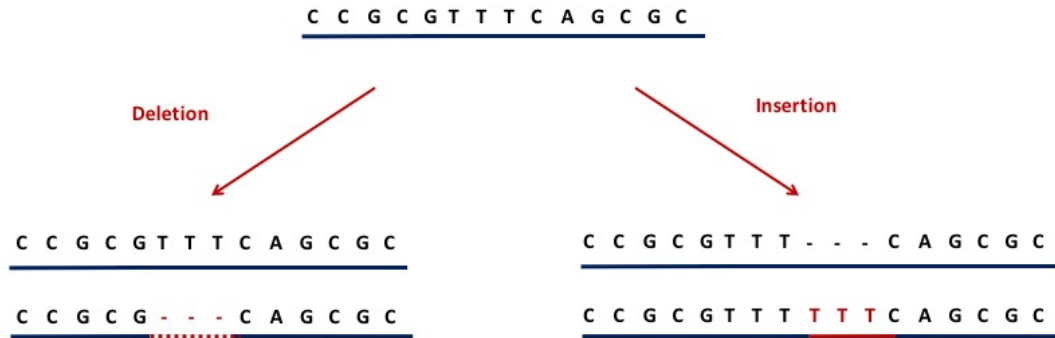
The University of Sheffield, UK
hbarton2@sheffield.ac.uk
https://henryjuho.github.io/hj_barton/

11/10/18

Introduction

Insertions and deletions

- ▶ short INDELs: sections of DNA $< 50\text{bp}$ that are deleted or inserted in a genome



INDELs often overlooked

- ▶ Disproportionately occur in repetitive sequence
- ▶ Hard to align
- ▶ Often occur in hotspots
- ▶ 1/8 as frequent as SNPs in humans

(Earl et al., 2014; Montgomery et al., 2013)

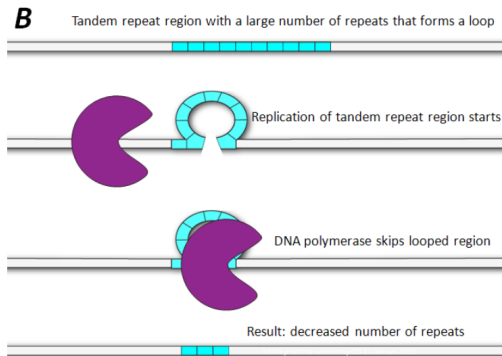
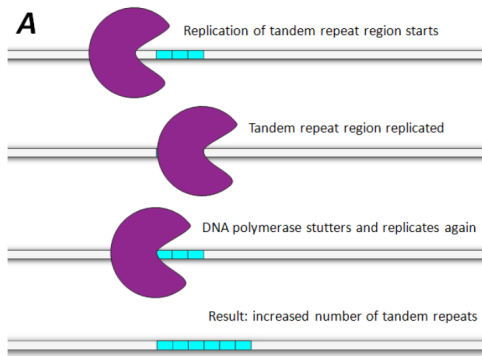
The importance of INDELs in genome evolution

- ▶ Contribute more to sequence divergence, in terms of the number of base differences, than SNPs
- ▶ Influence genome size:
 - ▶ low deletion rate → large genomes?
 - ▶ high deletion rate → compact genomes?

(Britten, 2002; Nam and Ellegren, 2012; Ometto et al., 2005; Sun et al., 2012)

INDEL mutation

- ▶ Deletion bias in most organisms
- ▶ Polymerase slippage can explain majority of short INDEL events

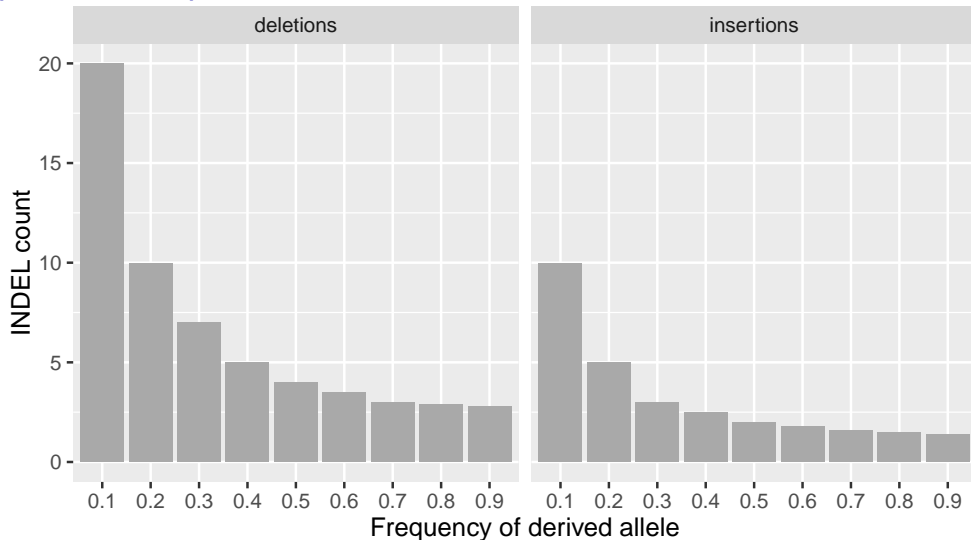


INDEL selection

- ▶ Deletions
 - ▶ lower mean allele frequency
 - ▶ more deleterious
 - ▶ two breakpoints
- ▶ Insertions may be favoured:
 - ▶ elevated fixation probability
 - ▶ biased gene conversion
 - ▶ minimum intron size
- ▶ Polarisation error

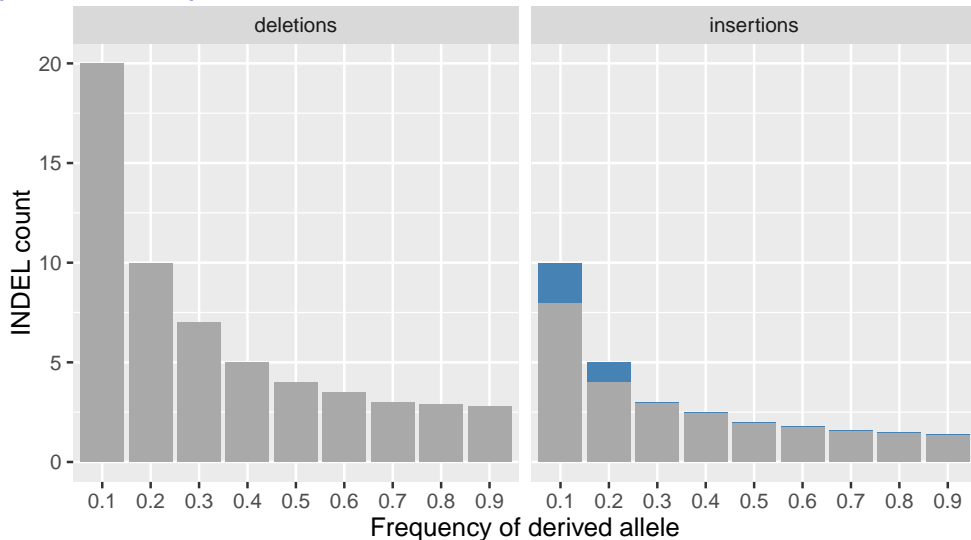
(Leushkin and Bazykin, 2013; Ometto et al., 2005; Petrov, 2002; Sjödin et al., 2010)

Importance of polarisation error



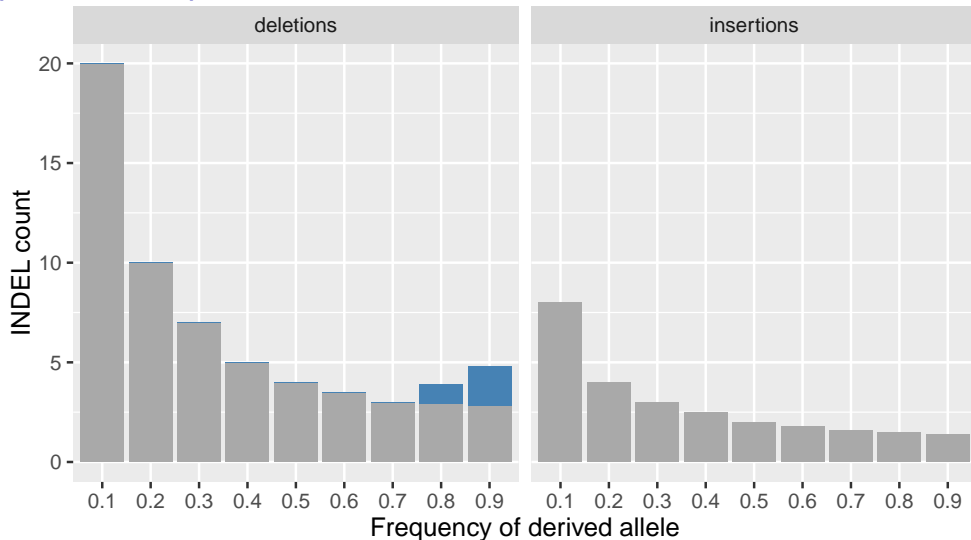
(see Hernandez et al., 2007)

Importance of polarisation error



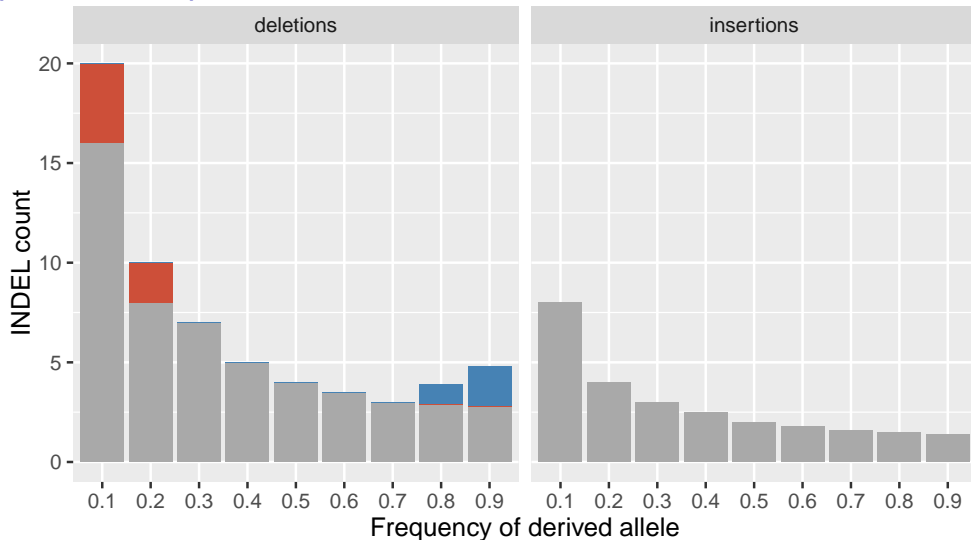
(see Hernandez et al., 2007)

Importance of polarisation error



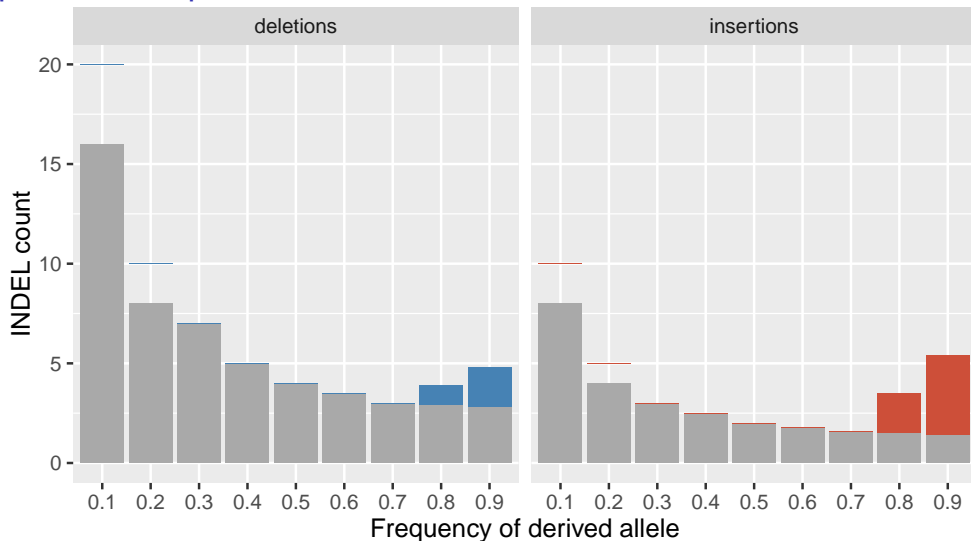
(see Hernandez et al., 2007)

Importance of polarisation error



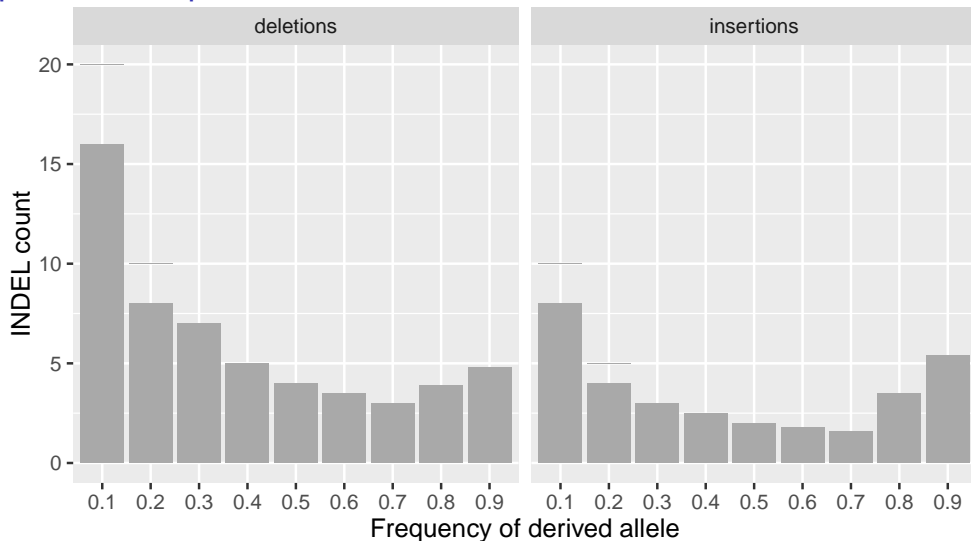
(see Hernandez et al., 2007)

Importance of polarisation error



(see Hernandez et al., 2007)

Importance of polarisation error



(see Hernandez et al., 2007)

Aims

Overcome confounding affect of polarisation error

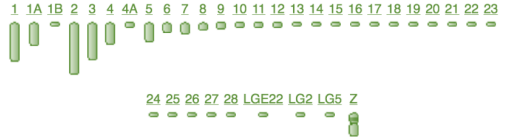
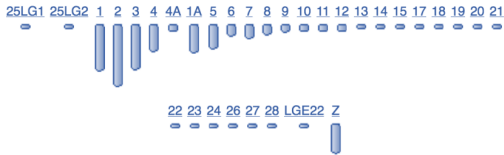
Quantify how natural selection shapes INDEL diversity in the great tit (*Parus major*)

1. within coding regions
2. in non-coding regions



Advantages of an avian system

- ▶ Conserved karyotype and synteny - good for alignments
- ▶ Genomes consist of few large macrochromosomes and many small microchromosomes
- ▶ Results in a highly dynamic recombination landscape - power to associations with recombination



(van Oers et al., 2014; Stapley et al., 2008)

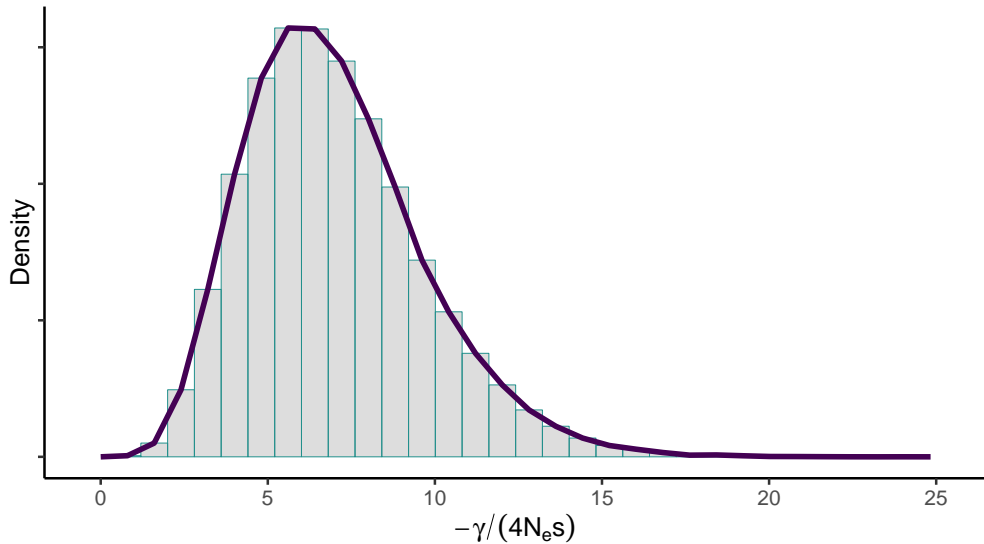
Tackling polarisation error

A novel maximum likelihood approach

- ▶ 'anavar'
- ▶ takes the unfolded site frequency spectrum
- ▶ estimates for both insertions and deletions:
 - ▶ mutation rate ($\theta = 4N_e\mu$)
 - ▶ **the distribution of fitness effects (DFE)**
 - ▶ polarisation error (ϵ)
- ▶ Controls for demography using neutral sites (Eyre-Walker et al., 2006)
- ▶ Applicable to both INDELs and SNPs or a combination

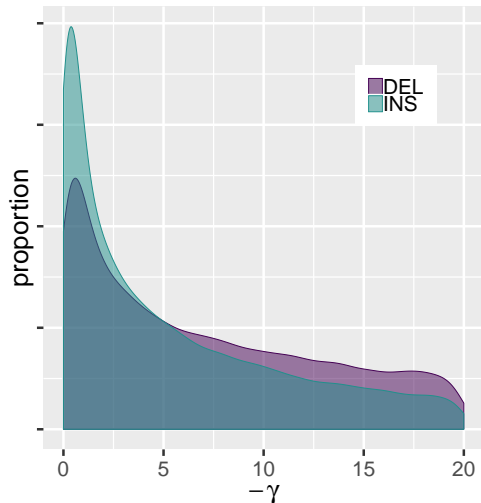
(Barton and Zeng, MBE, 2018)

What is the DFE?

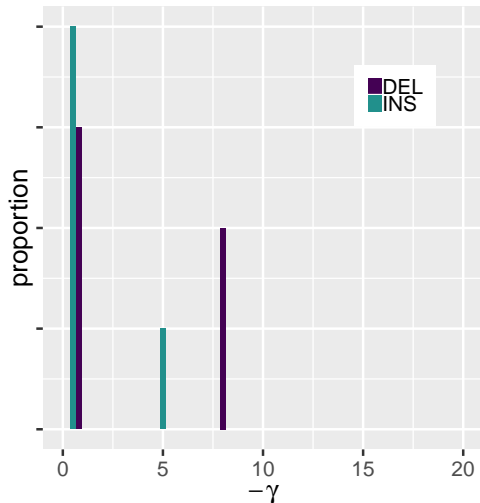


The model can describe the DFE in two ways

Continuous DFE

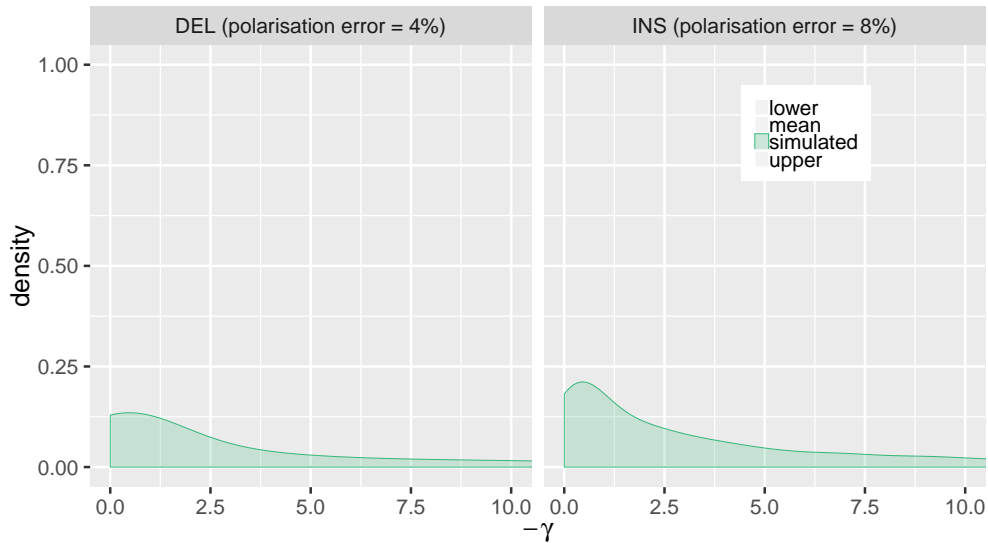


Discrete DFE

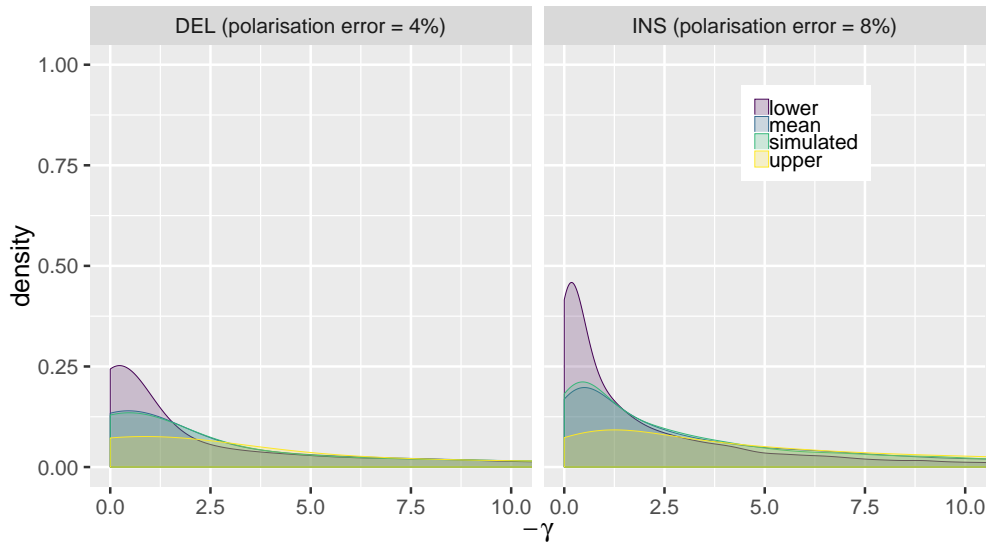


(Barton and Zeng, MBE, 2018)

Predicts the DFE well, with polarisation error



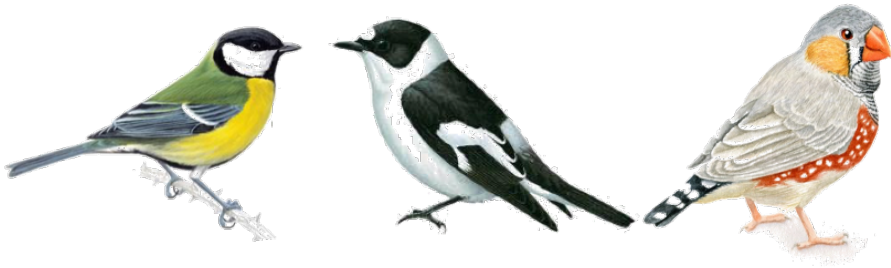
Predicts the DFE well, with polarisation error



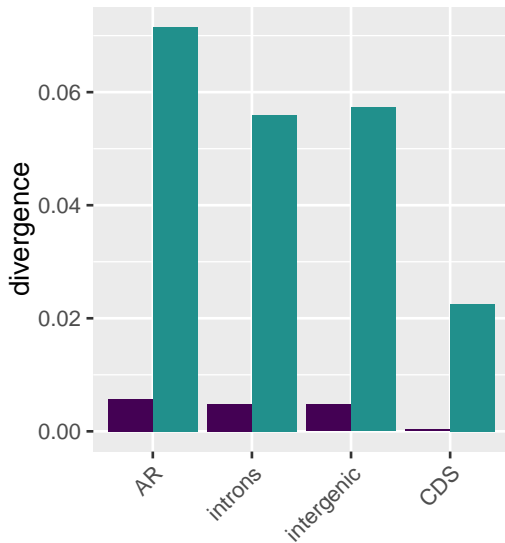
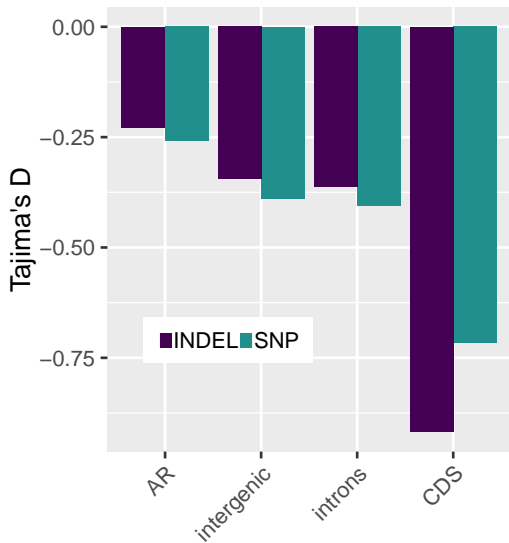
Applying the model to the great tit

Sample and pipeline

- ▶ 10 european great tit males (Corcoran et al., 2017)
- ▶ high coverage (44x)
- ▶ variant calling with GATK
- ▶ multispecies alignment between zebra finch, flycatcher and great tit
- ▶ parsimony based polarisation

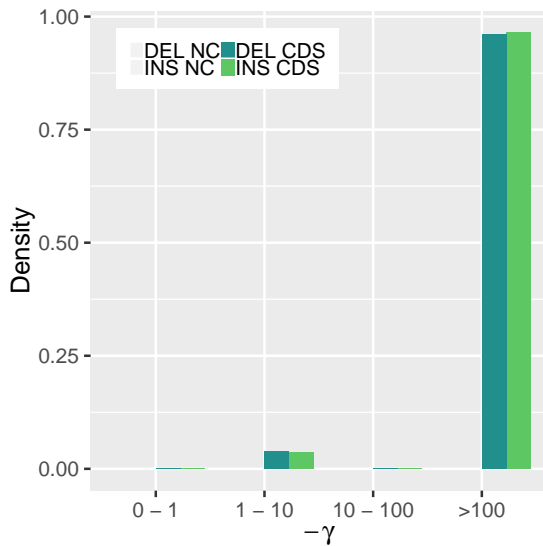


Summarising the data set

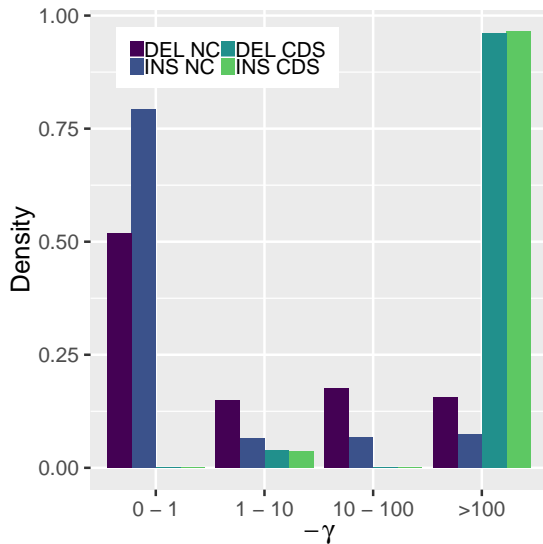


INDELs genome wide

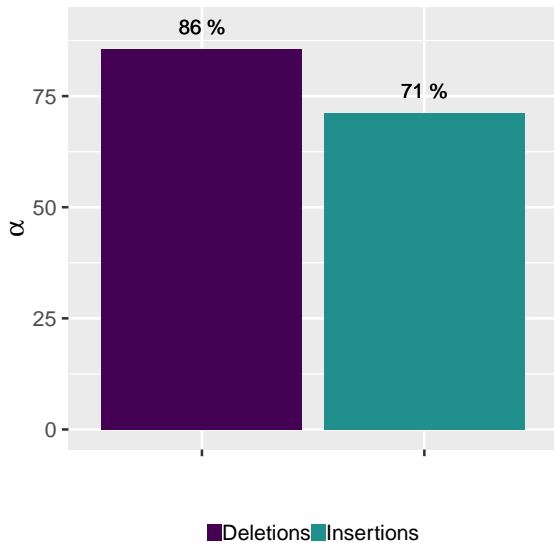
Polymorphic INDELs predominantly strongly deleterious



Polymorphic INDELs predominantly strongly deleterious

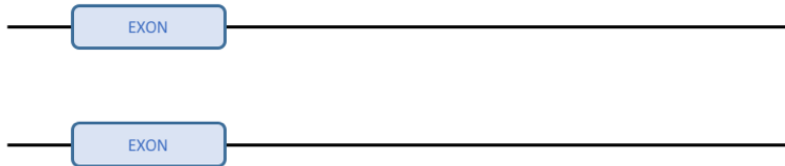


Coding INDEL fixations largely beneficial

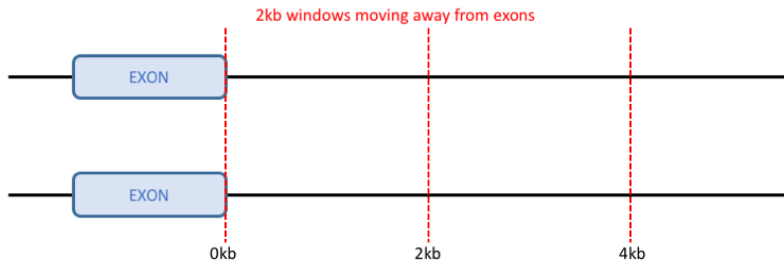


Moving away from coding regions

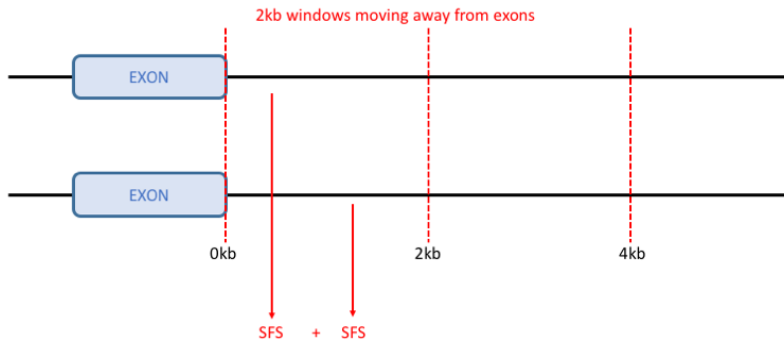
Approach



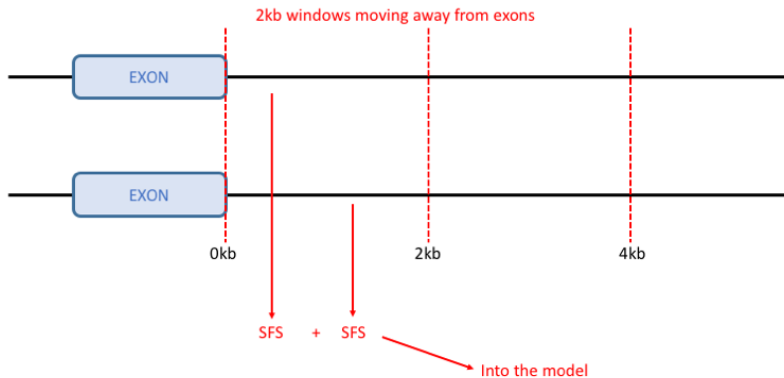
Approach



Approach



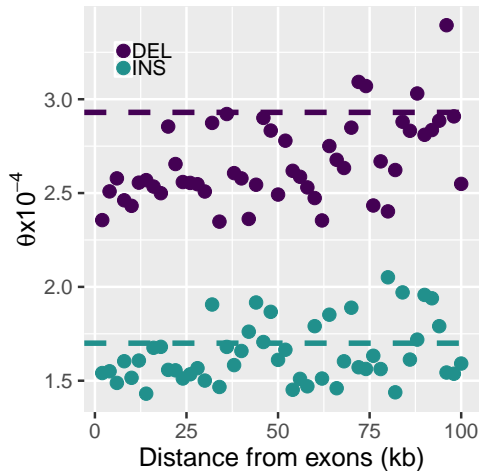
Approach



Diversity increases with distance

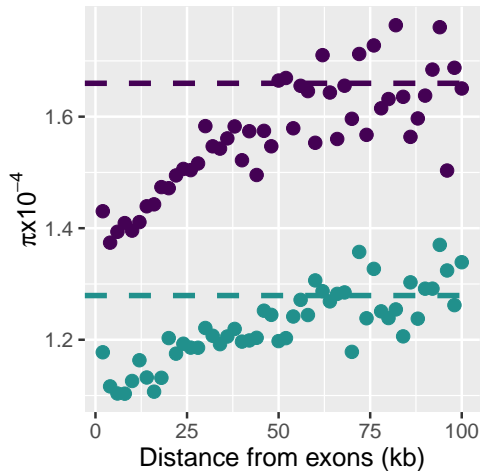
Ins: $\rho = 0.28$ $p < 0.05$

Del: $\rho = 0.47$ $p < 0.01$



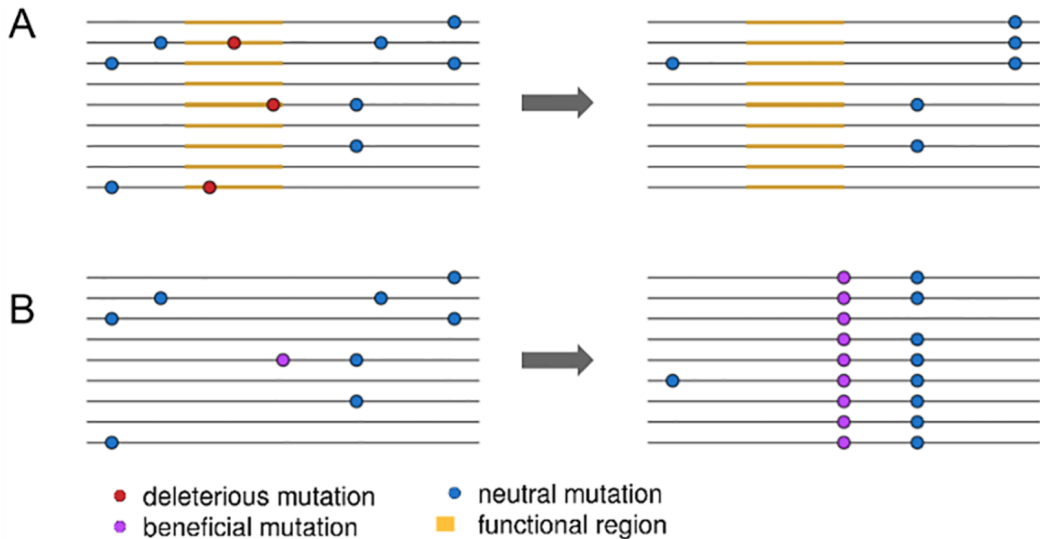
Ins: $\rho = 0.84$ $p < 0.01$

Del: $\rho = 0.79$ $p < 0.01$



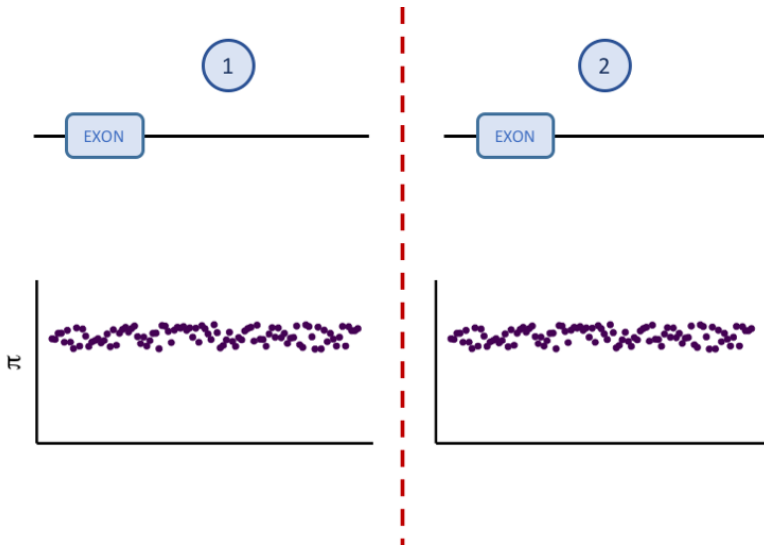
Linked selection?

Linked selection

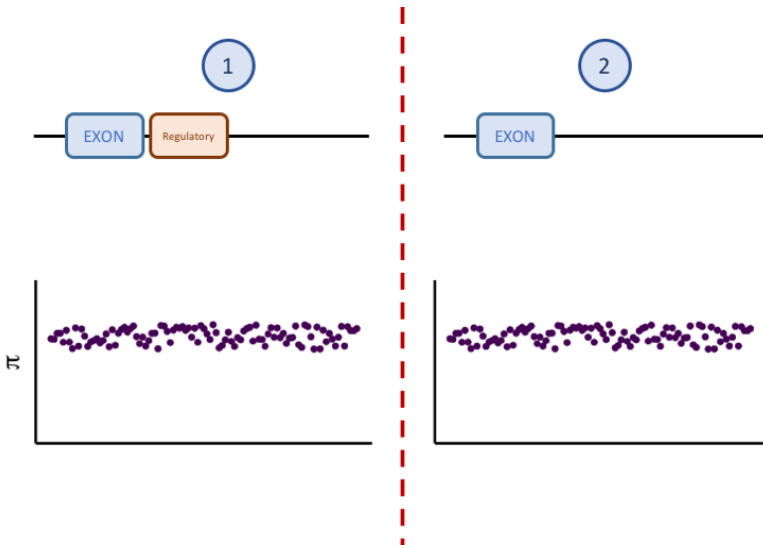


(Cropped from fig 1. Josephs and Wright, 2016)

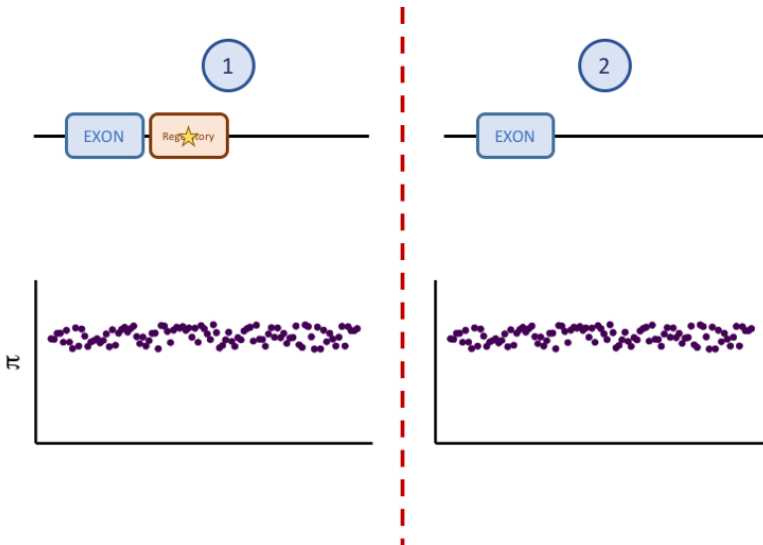
Two possible explanations



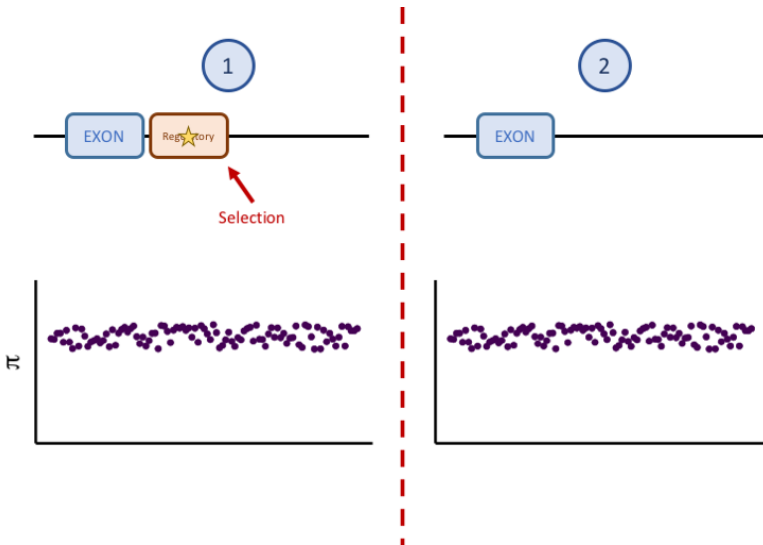
Two possible explanations



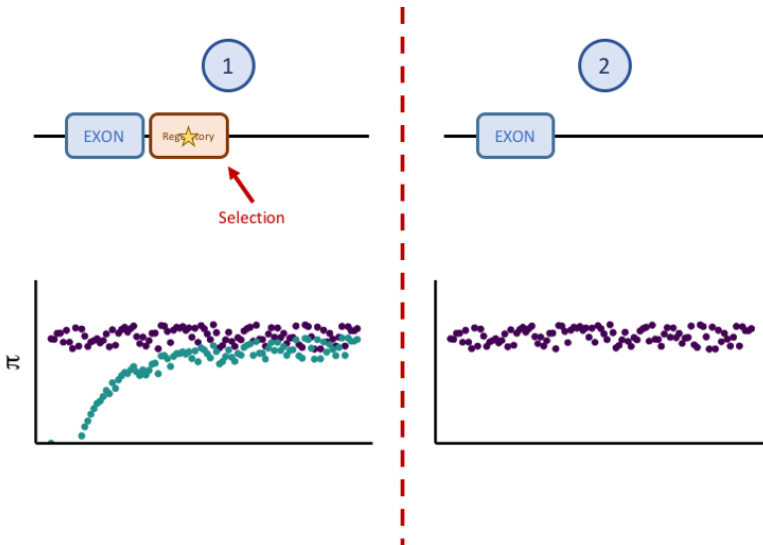
Two possible explanations



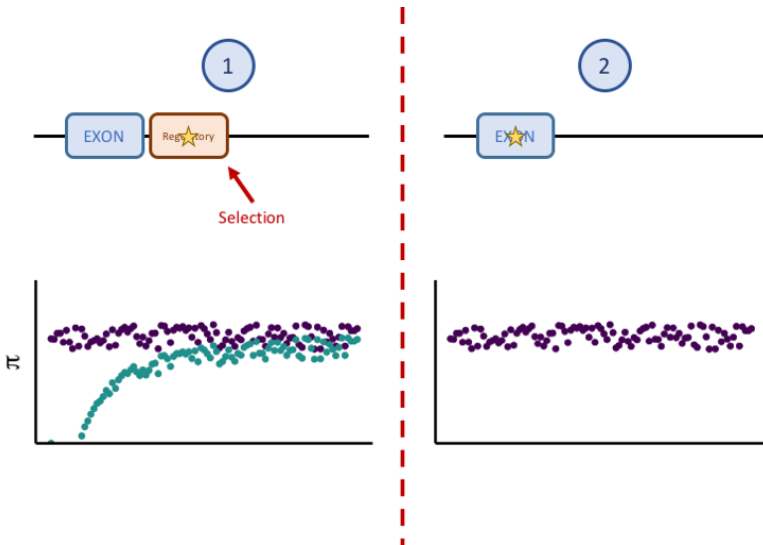
Two possible explanations



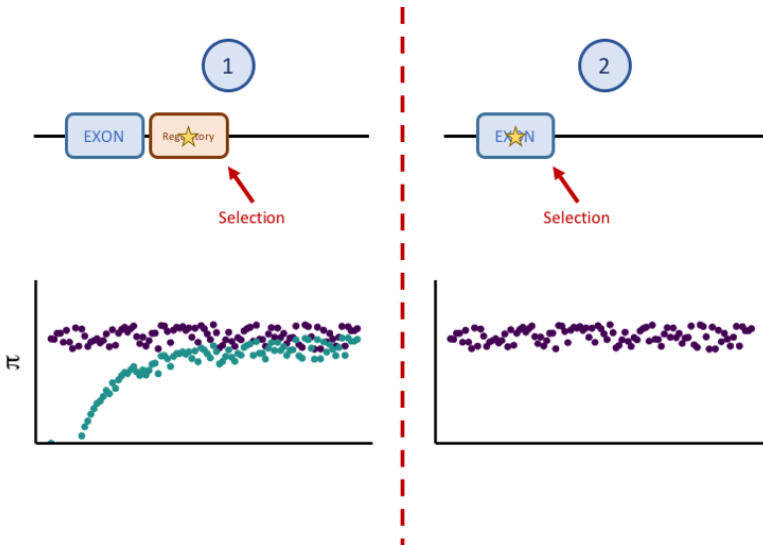
Two possible explanations



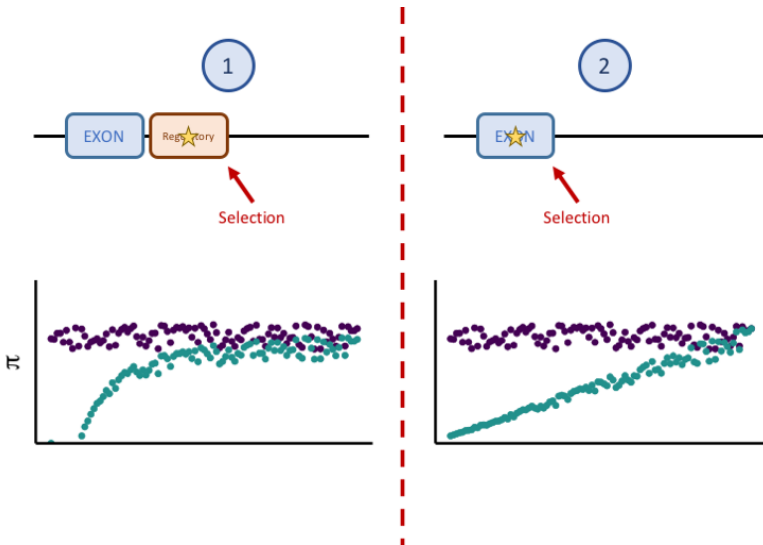
Two possible explanations



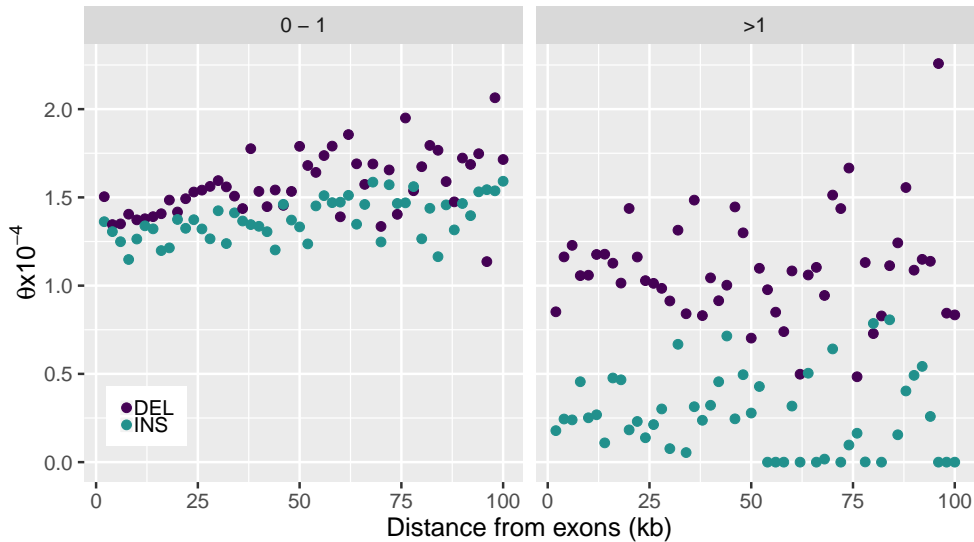
Two possible explanations



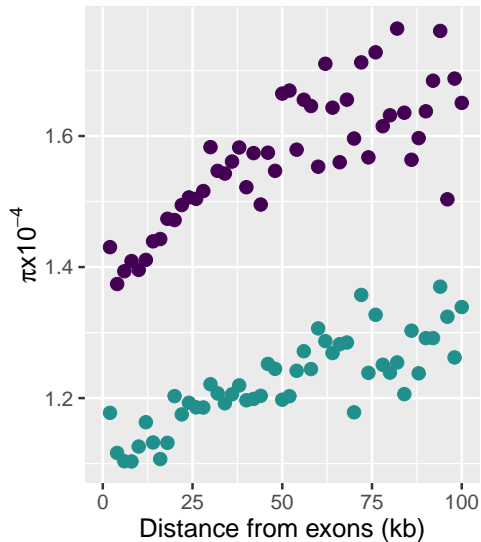
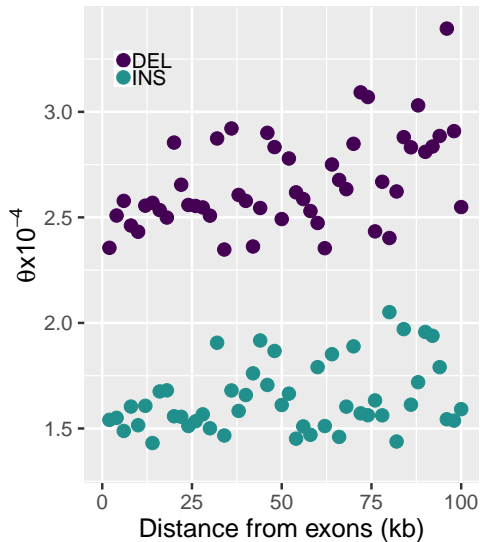
Two possible explanations



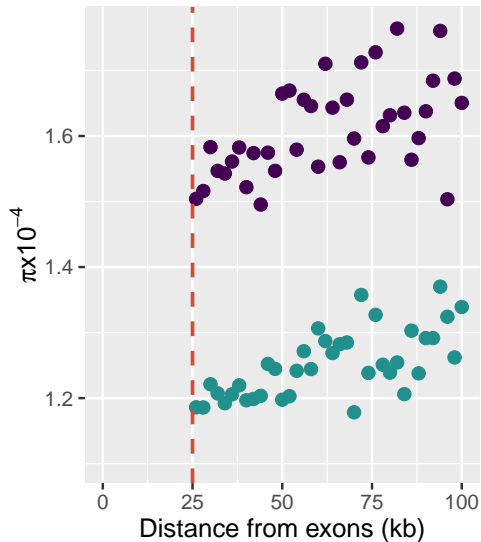
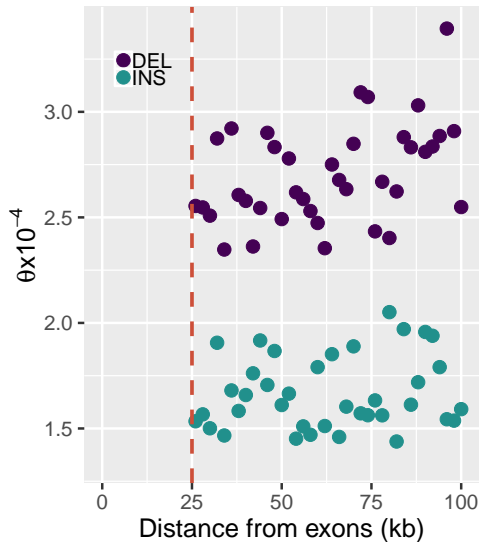
Driven by neutral variation



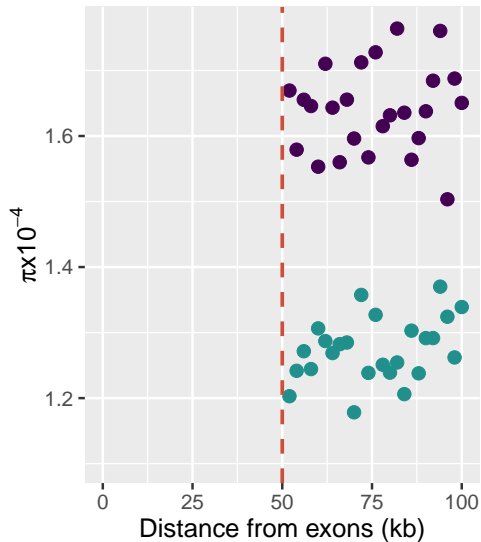
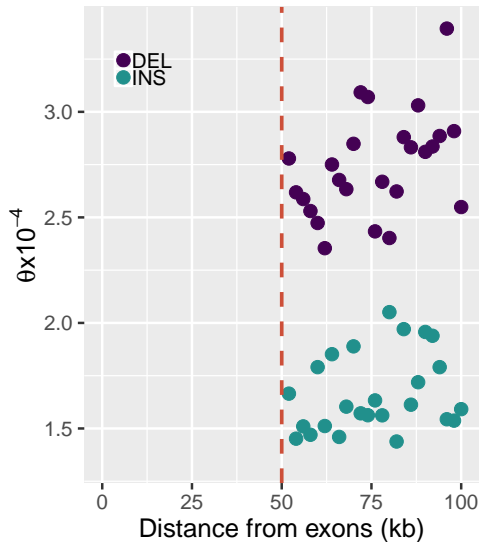
How far does the correlation persist?



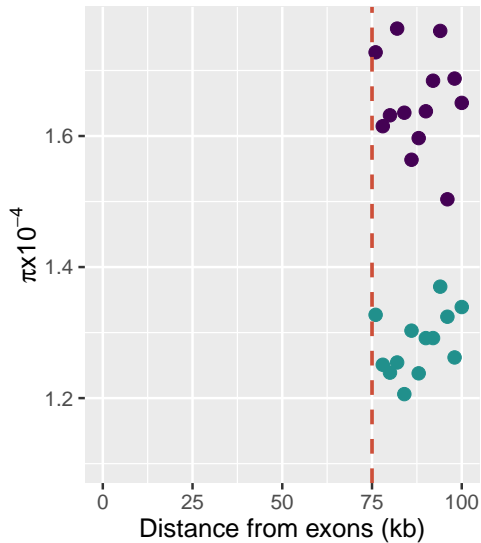
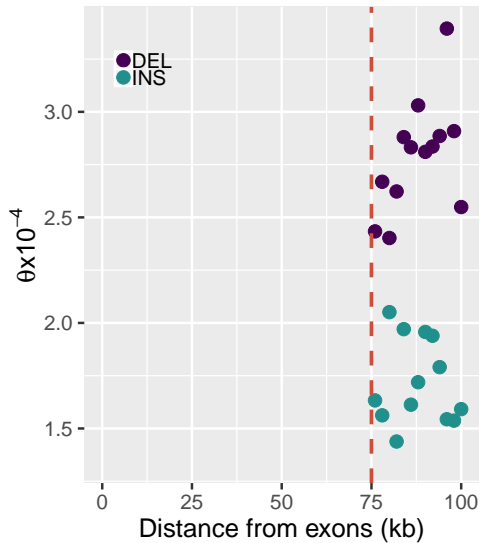
How far does the correlation persist?



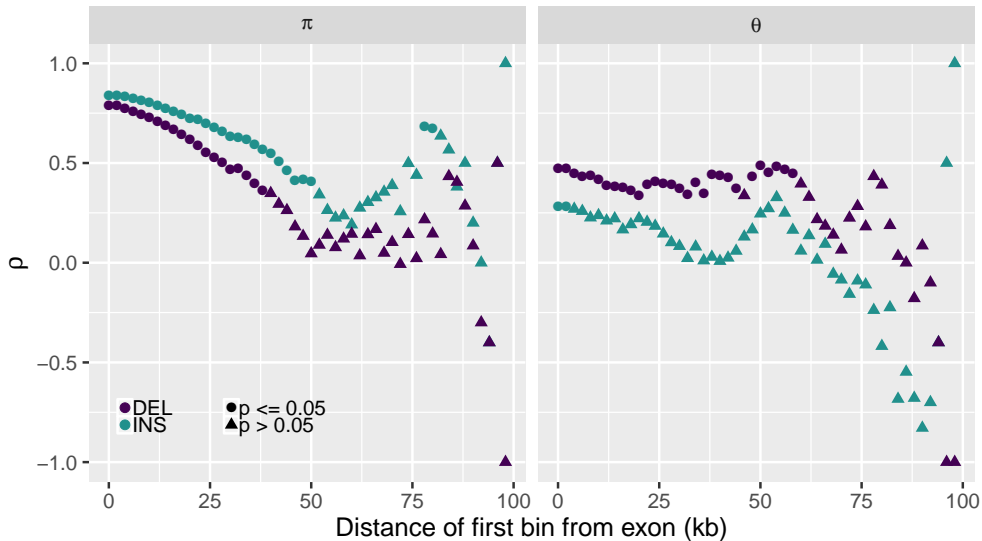
How far does the correlation persist?



How far does the correlation persist?



How far does the correlation persist?



Recombination should modulate the impact of linked selection

Getting the data

```
ATCGGGTCGATTTCGATTGTACCGTAACTCTCTCGCGCGCGCGCGCGCATATA  
ATCGGGTCGA- - -CGATTGTACCGTAACTCTCTCGCGCGCGCGCGCGCA- - -
```

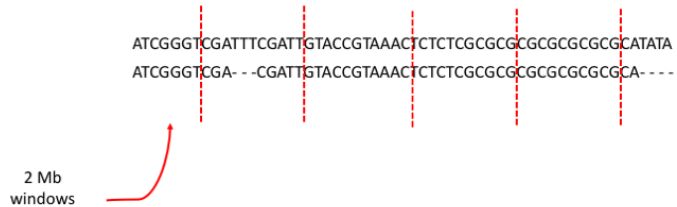
Getting the data



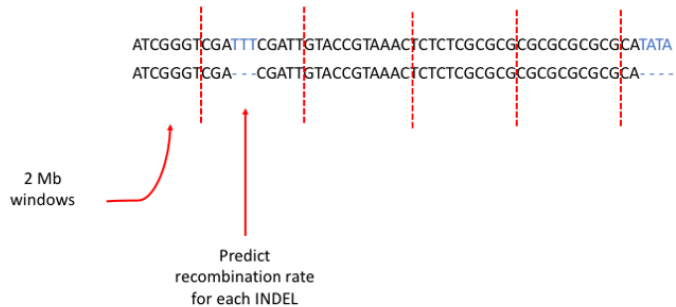
Diagram illustrating sequence alignment. Two DNA sequences are shown, with vertical red dashed lines indicating alignment positions. The top sequence is ATCGGGTCGATTTCGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCATATA. The bottom sequence is ATCGGGTCGA- -CGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCA- - -. The alignment shows that the sequences are identical up to the 18th position, after which they diverge.

```
ATCGGGTCGATTTCGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCATATA
ATCGGGTCGA- -CGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCA- - -
```

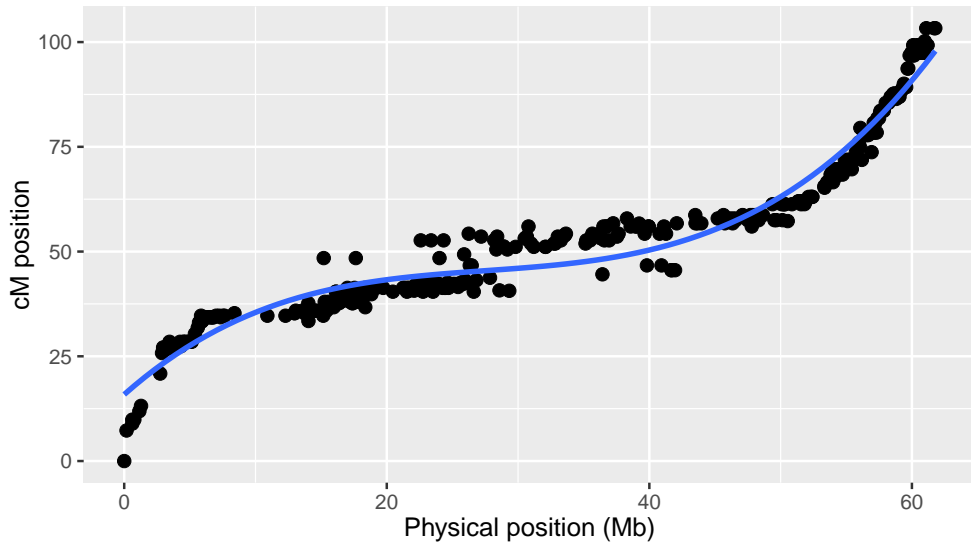
Getting the data



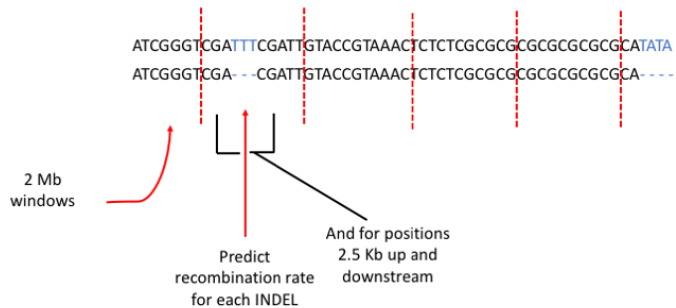
Getting the data



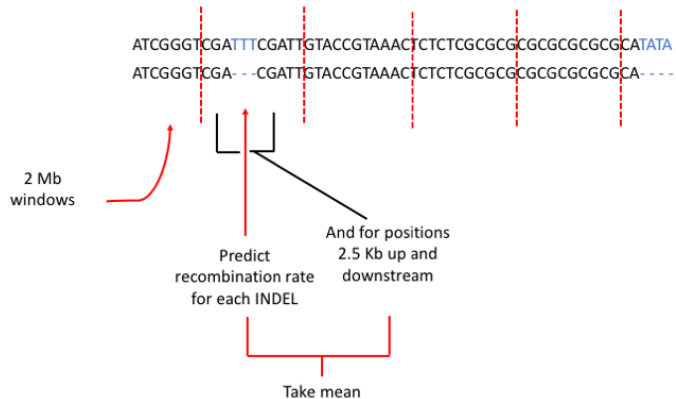
Getting the data - recombination rate



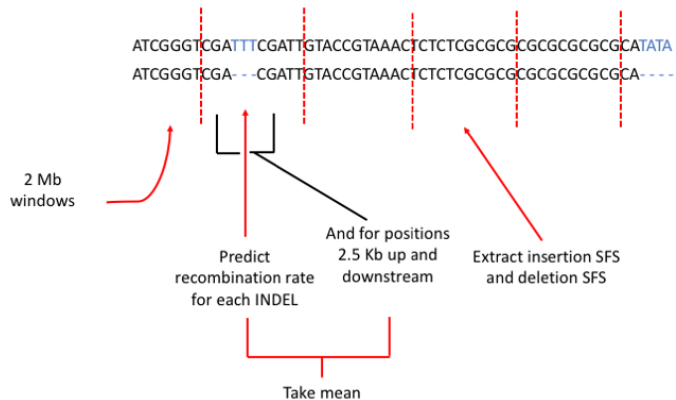
Getting the data



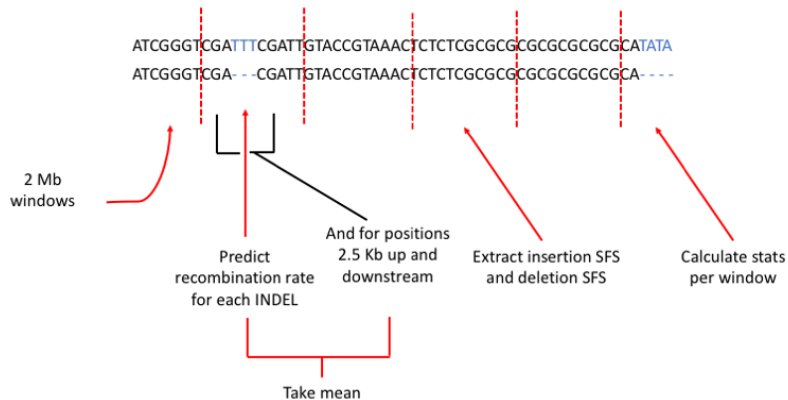
Getting the data



Getting the data



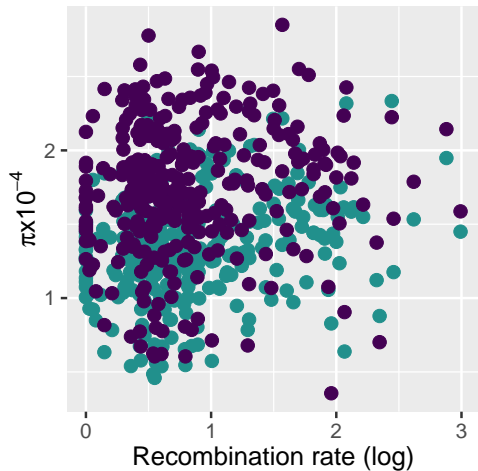
Getting the data



Association between diversity and recombination

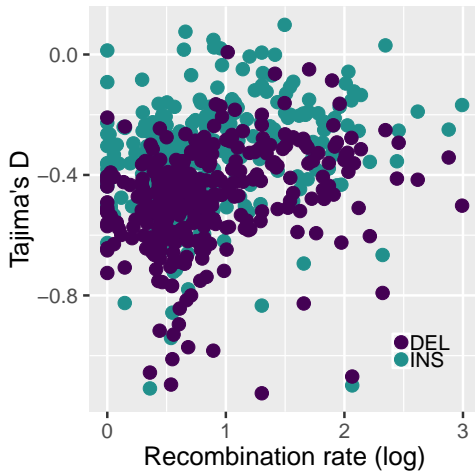
Ins: $\rho = 0.18$ $p < 0.05$

Del: $\rho = 0.12$ $p < 0.01$



Ins: $\rho = 0.3$ $p < 0.01$

Del: $\rho = 0.33$ $p < 0.01$



Round up

Conclusion

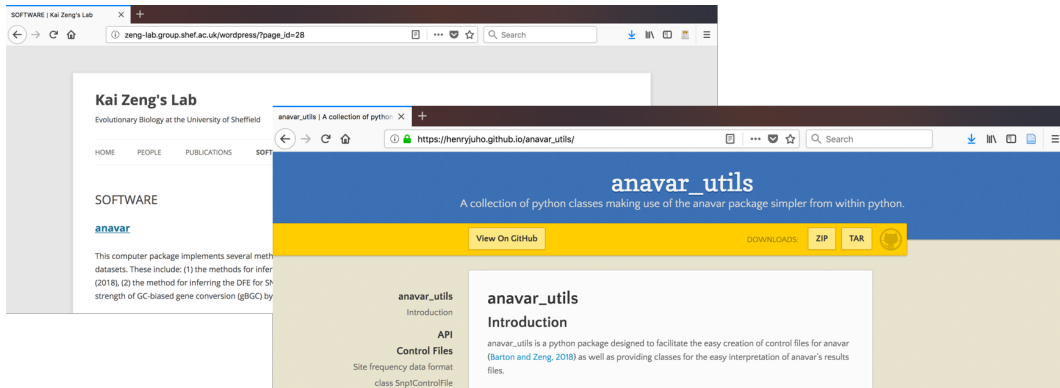
- ▶ INDELs in genes mostly extremely deleterious - 96%
- ▶ Remainder are weakly deleterious
- ▶ α estimate at 71% and 86% for insertions and deletions
- ▶ Regions adjacent to exons, and areas of low recombination, have reduced INDEL diversity - linked selection
- ▶ Extends over relatively large distance up to ~50kb

Next steps

- ▶ Interesting to investigate if reduced diversity is due to positive selection or purifying selection
- ▶ Look at whether efficacy of selection on INDELs is higher in regions with higher N_e

Plug for the model

- ▶ User friendly computer package - anavar - <http://zeng-lab.group.shef.ac.uk>
- ▶ Methods are applicable to both INDELs and SNPs or a combination
- ▶ Code for integration with python - https://henryjuho.github.io/anavar_utils/



Questions?