Development of the Z_α statistics for identifying regions of the genome under selection

UNIVERSITY OF Southampton

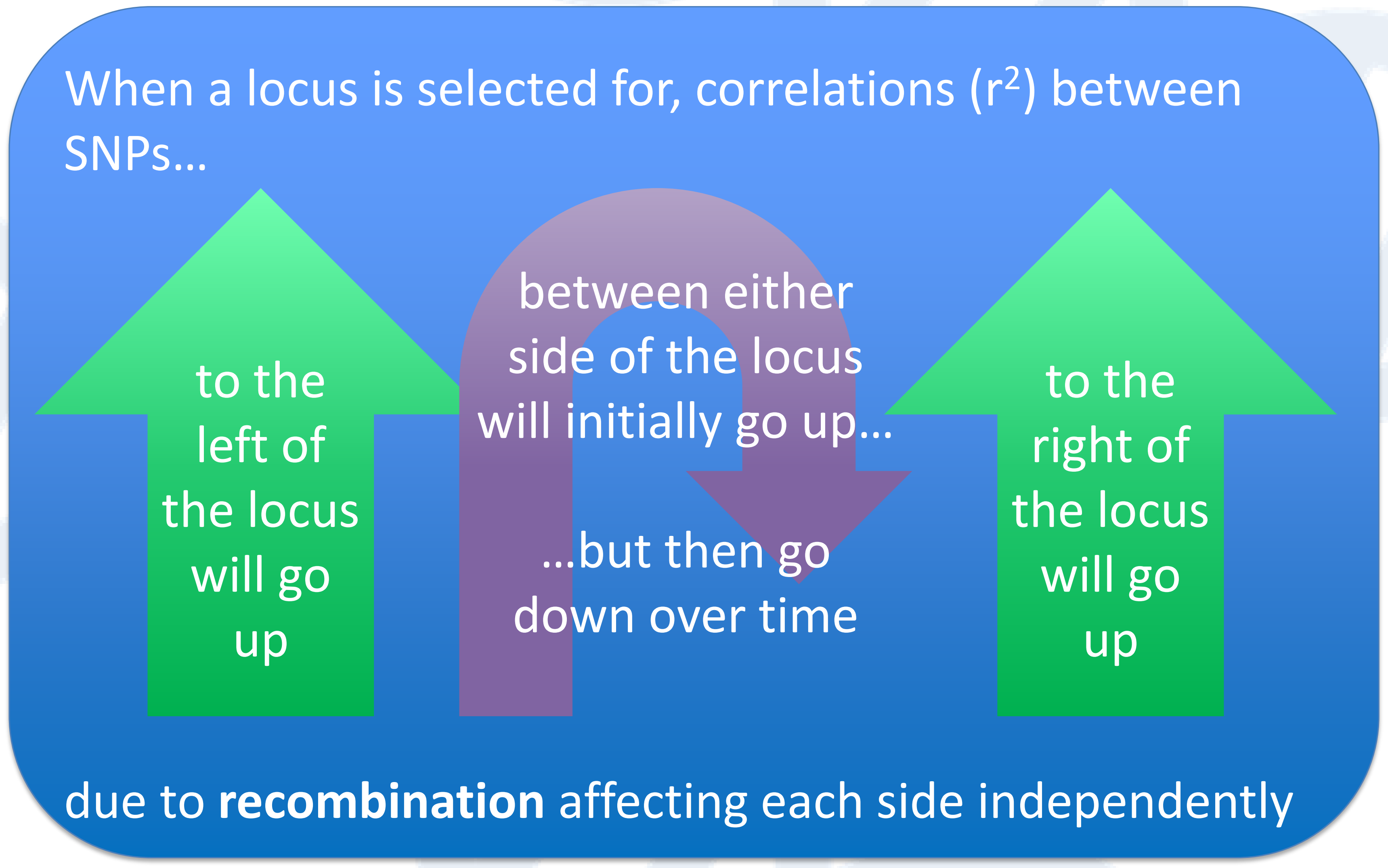
Clare Horscroft[1,3], Reuben Pengelly[1,3], Timothy J. Sluckin[2,3], Andrew Collins[1,3]

1. Genetic Epidemiology and Bioinformatics, Faculty of Medicine, University of Southampton 2. Mathematical Sciences, University of Southampton 3. Institute for Life Sciences, University of Southampton

Introduction

- Z_α is a Linkage Disequilibrium (LD) based statistic for finding evidence of selection in the genome
- Aim: Develop and apply the Z_α family of statistics first published by Jacobs *et al.* (2016) [1]

Methods – the Z_α Statistic



When a locus is selected for, correlations (r²) between SNPs…

to the left of the locus will go up

between either side of the locus will initially go up…

…but then go down over time

to the right of the locus will go up

due to **recombination** affecting each side independently

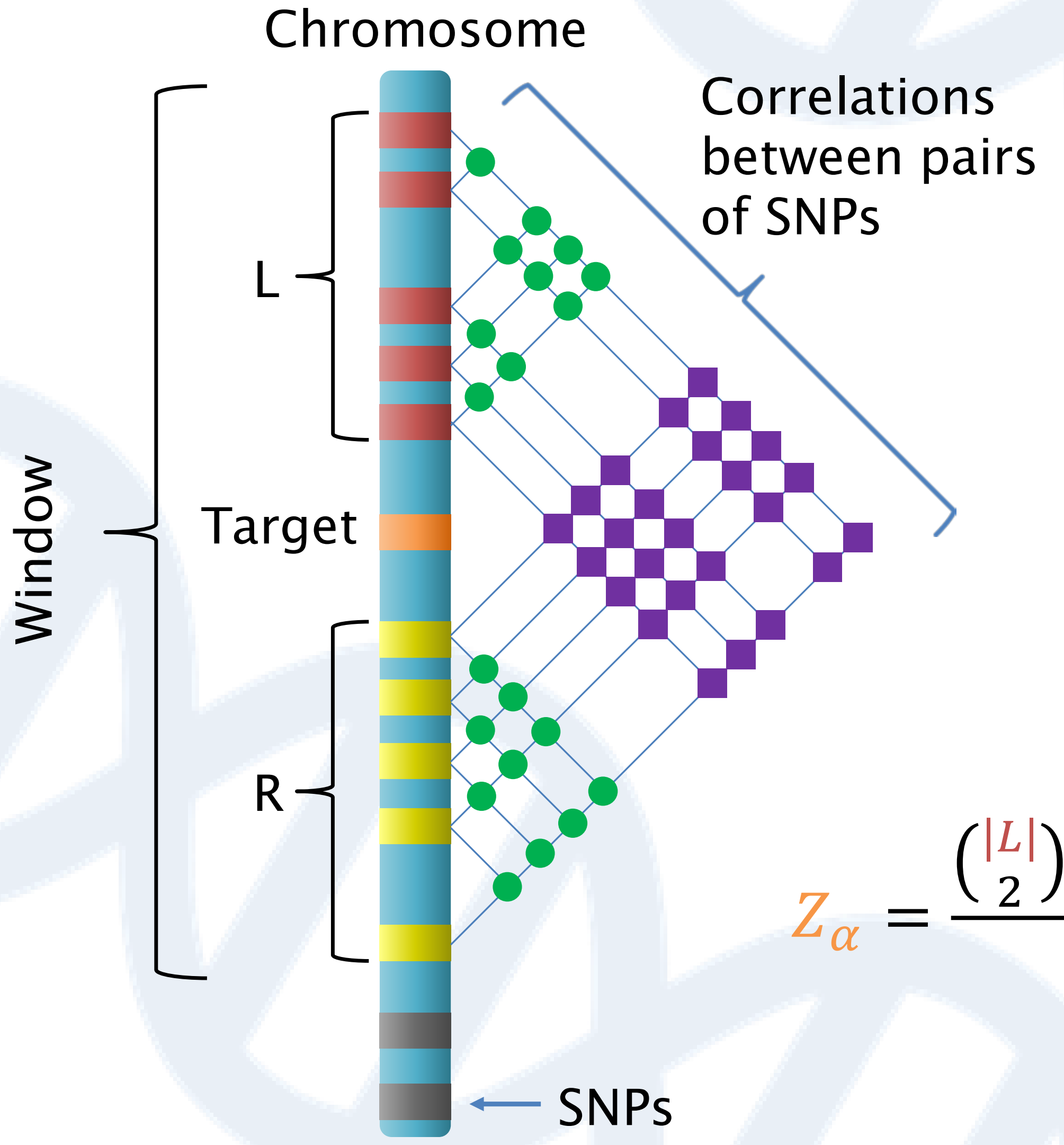


Figure 1:

- Statistics are calculated for a target SNP
- A window size is selected and SNPs within the window each side are counted, |L| to the left and |R| to the right
- Correlations (r²) are calculated between each pair of SNPs, depicted here as green circles and purple squares
- $\binom{|L|}{2}$ means the number of pairs on the left side

$$Z_\alpha = \frac{\binom{|L|}{2}^{-1}\sum_{i,j\in L} r_{i,j}^2 + \binom{|R|}{2}^{-1}\sum_{i,j\in R} r_{i,j}^2}{2}$$

$$Z_\beta = \frac{\sum_{i\in L, j\in R} r_{i,j}^2}{|L||R|}$$

- Z_α will be elevated around regions undergoing selection
- Z_β will be elevated around regions undergoing selection, but then decrease as the selected allele reaches fixation
- Thus Z_α/Z_β is a useful statistic to ascertain the stage of the selective process
- Adjusting for expected LD should enhance the ability of Z_α to distinguish between regions of the genome with and without selective events

Adjusting for expected LD

Generate an LD profile from independent data. Returns an expected r² value for given genetic distances between pairs of SNPs

Calculate expected Z_α ($E[Z_\alpha]$) for each SNP using r² values from the LD profile

Combine the statistics, e.g. by using $Z_\alpha/E[Z_\alpha]$

zalpha is now available as a free R package on CRAN!

zalpha on CRAN: https://cran.r-project.org/package=zalpha

zalpha on GitHub: https://github.com/chorscroft/zalpha

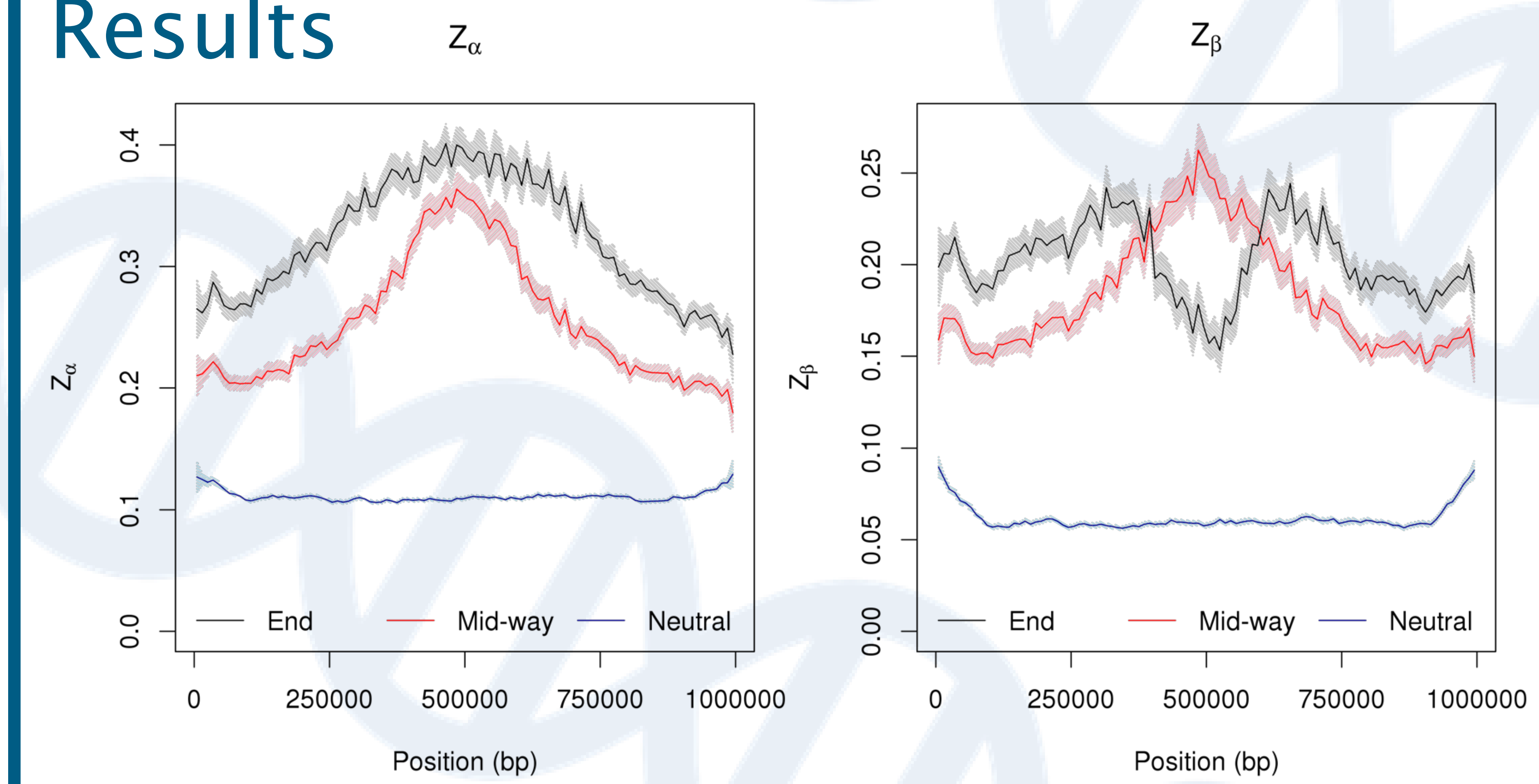
Methods – Simulations

Simulation models using SLiM v3.3

1) Only **neutral** variation, mutation rate 1e^{-7}, constant recombination rate 1e^{-8}, 10,000 individuals for 1,500 generations, 1 Mb chromosomes

2) Selected variant in the centre of the region at generation 1,000 with fitness effect 0.05; statistics calculated **mid-way** through the sweep and at the **end** (50% and 90% frequency in the population)

Results



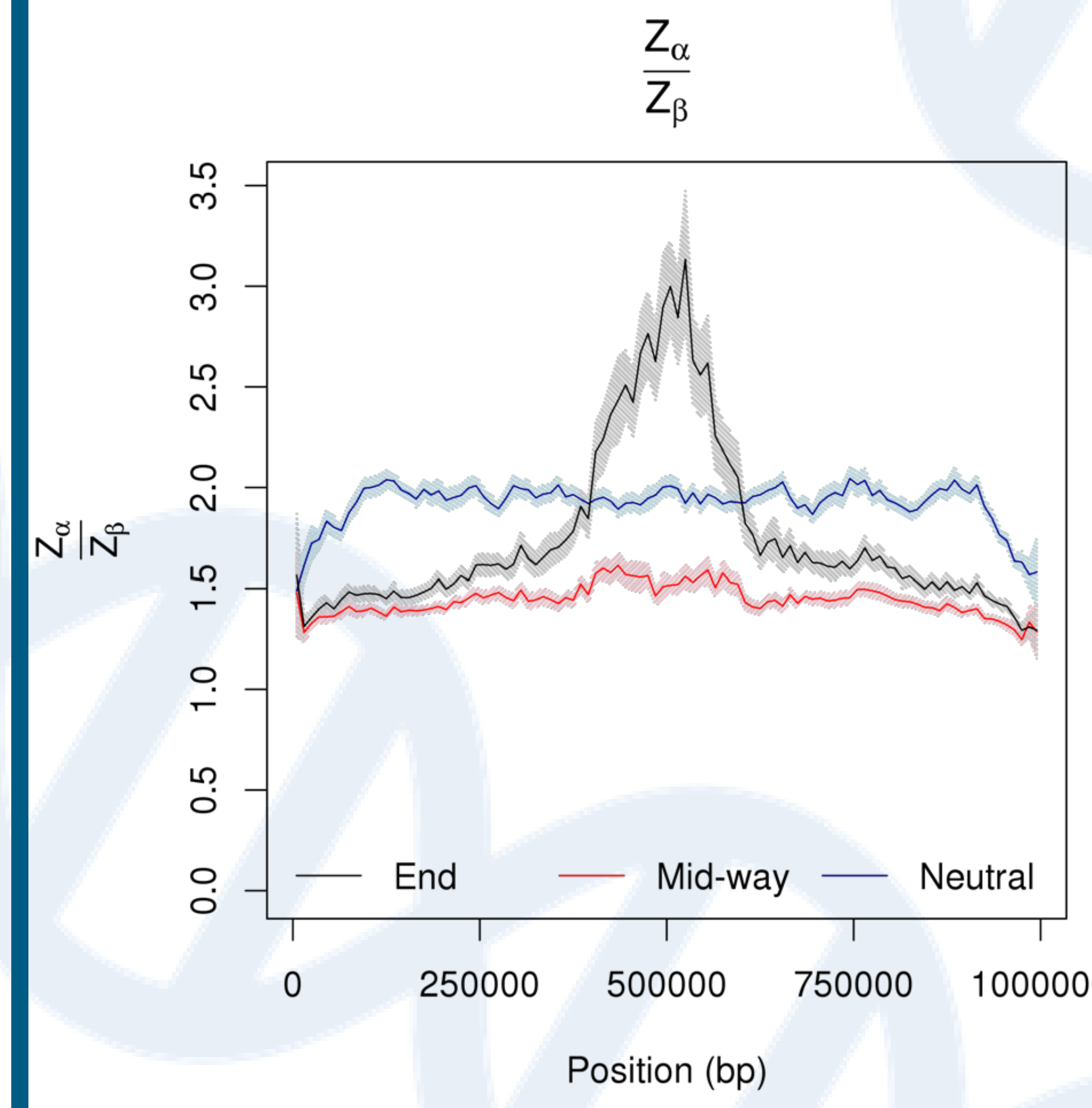


Figure 2:

- These graphs show the results of running the statistics Z_α, Z_β and Z_α/Z_β over a sample of 1,000 chromosomes from 100 simulations of the **neutral** model, and samples from **mid-way** through and near the **end** of 100 simulations of the selected model.
- The graphs show averages of the statistics in 10 Kb bins along the chromosome, including 95% confidence intervals as shaded areas.
- The three figures show different behaviours for the **neutral** sample, the sample taken **mid-way** through a selective sweep, and the sample near the **end**.
- (Not shown) adjusting for expected LD given genetic distances between SNPs improves the performance of the statistics i.e. they are able to differentiate between neutral selected sites more often than without the adjustment

Conclusions

Z_α and related statistics can distinguish regions of the genome with a selective event from those without

Combining statistics can lead to further insights (e.g. the stage of the sweep)

Adjusting for expected LD further improves performance

Work to confirm the statistic performs well in other simulated scenarios and in real-life datasets with known selected regions is underway

www.southampton.ac.uk/medicine

[1] Jacobs, G.S., T.J. Sluckin, and T. Kivisild, *Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps.* Genetics, 2016. **203**(4): p. 1807.