Exploring selective scanning with Dz statistic: simulation and empirical studies

Alouette Zhang

Department of Human Genetics, McGill University, Canada

zhiwei.zhang@mail.mcgill.ca

January 28, 2025

Abstract

Selective sweeps are significant evolutionary events where beneficial variants increase rapidly in frequency and reach fixation in the population. This affects surrounding variants via hitchhiking that leads to reduced diversity and excess linkage disequilibrium (LD). Various methods exploit an ensemble of these features to detect sweeps. Here, we consider the effect of hard sweeps on Dz, a measure of LD between rare variants which was recently shown to be informative about human demography. We show the theoretical expectation of Dz can leverage both the reduced diversity and the excess LD around the sweep site for sweep detection and is more informative about recent and ancient sweeps compared to other commonly used LD statistics. Dz still retains adequate power for up to xxx thousand years ago in simulations with human-like parameters. Additionally, we recapture well-studied selective sweeps in different populations in the 1000 Genomes Project data using a window scan incorporating the expectation of Dz. Our results demonstrate Dz as a potential statistic for sweep detection that takes advantage of genetic patterns not considered by most current methods.

Keywords: selective sweep, sweep detection, two-locus statistics, linkage disequilibrium

Contents

1 Introduction

A selective sweep refers to the rapid spread of beneficial alleles in a population by directional selection. When the selective sweep is driven by one de novo beneficial mutation, it is known as a "classical" or "hard" selective sweep. Around the sweep, genetic diversity is reduced, and more variants segregate at high or low frequency, leading to a shift in the site frequency spectrum (SFS).
Recent interest raised revolving two-locus statistics, due to its sensitivity to deep coalescent.
Linkage disequilibrium (LD), the non-random association between two or more loci, also considerably increases between the hitch-hiked neutral variants on the same side of the sweep site (McVean, 2006). The patterns and span of the hitch-hiked regions are used to infer the location, onset, and strength of the selective sweep.
Several studies have utilized LD in detecting sweep sweeps, for its advantage of being less affected by background selection.
Now studies have been incorporated an ensemble of summary statistics into Machine Learning methods. This leads to the importance of identifying statistics that measure difference aspects of the genotypes to aid in the discovery of selective sweeps. Excess of rare variants are signature of the selective sweep, but computing LD among rare variants subject to high variance since the allele frequency is used as the denominator. Studies often exclude rare variant (Minor Allele Frequency ¡ 0.05) during the calculation, so the LD between rare variants is less considered during sweep detection. A recent study proposes a stable LD statistics for rare variants by calculating the composite LD and MAF among all rare variants for the same gene region.
McVean examined the LD patterns between a pair of neutral loci near the selective sweep.
Kim and Nielsen proposed the ω statistics to calculate pairs of neutral alleles on the same side of the selective sweep. OmegaPlus is a software that utilized the ω statistics to efficiently calculate ().
iLDS that examined the difference of LD between non-synonymous and synonymous variants under selective sweeps. iLDS discovered the difference is reversed between the common and rare variants.
Each summary statistic is subject different limitations. Machine learning method combine an ensemble of summary statistics in "logistic regression".
Change in population size can elevates LD that resembles sweeps.

2 Methods

This section describes the experimental design, materials, and procedures used in the study. It should be detailed enough to allow replication of the research.

2.1 Hard selective sweep simulations

2.1.1 Selective sweep at fixation

We simulated a 10 Mb region for a population of 50 randomly mating diploid individuals. Simulations adopted a scaled recombination rate $4N_eL\rho = 500$ and a scaled mutation rate $4N_e\mu = 0.002$ [1], where N_e is the effective population size, L is the length of the simulated region, ρ is the recombination rate per nucleotide per generation, and μ is the mutation

rate per nucleotide per generation. Prior to simulating the selective sweep, the population followed a burn-in period of xx generations using msprime [2] version 1.2.0 to establish coalescence. Then the region was simulated using fwdpy11 [**<empty citation>**] version 0.22.0 with a single beneficial mutation in the middle with a scaled selection rate $2N_e s = 500$ [1], where s is the selective coefficient. The simulations were saved as tree sequence data and added neutral mutations using tskit [**<empty citation>**] version 0.5.6. Then, we saved the genotypes of 50 individuals exported utilizing as the Variant Call Format (VCF) for two-locus statistics inference.

2.1.2 Selective sweep post fixation

3 Results

Present the main findings of the study. Use tables and figures to support your results.

Condition	Measurement 1	Measurement 2
A	10	20
B	15	25
C	20	30

Table 1: A sample table showing data.

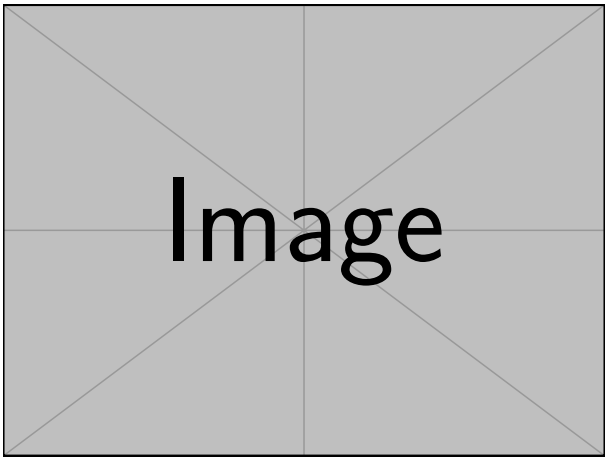

Figure 1: A sample figure showing an experimental result.

4 Discussion

Something something [3] Interpret the results, discussing their implications and how they relate to previous research. Mention any limitations of the study and suggest future research directions.

5 Conclusion

Summarize the main findings and their significance. This section should be concise and highlight the key contributions of the research.

6 References

Add your references here, formatted according to the citation style you're using. In LaTeX, references are often managed with BibTeX. Here's an example using the bibliography file 'references.bib'.

References

[1] Y. Kim and R. Nielsen. "Linkage Disequilibrium as a Signature of Selective Sweeps". In: *Genetics* 167 (2004), pp. 1513–1524.

[2] F. Baumdicker et al. "Efficient ancestry and mutation simulation with msprime 1.0". In: *Genetics* 220 (2022).

[3] A. E. Gill. "Some Simple Solutions for Heat-Induced Tropical Circulation". In: *Quart. J. R. Met. Soc.* 106 (1980), pp. 447–462.