



UMEÅ UNIVERSITET

FINDING FITNESS

Empirical and theoretical explorations
of inferring fitness effects from
population level SNP data

Bea Angelica Andersson

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorexamen framläggs till offentligt försvar i MA121, MIT, fredagen den 2 februari, kl. 09:00.

Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor, Tanja Slotte,

Stockholms Universitet, Stockholm, Sverige.

Organization

Umeå University
Department name

Document type

Doctoral thesis

Date of publication

12 January 2024

Author

Bea Angelica Andersson

Title

Finding fitness – Empirical and theoretical explorations of inferring fitness effects from population level SNP data

Abstract

The distribution of fitness effects (DFE) describes the likelihood that a new mutation has a specific effect on the fitness of an individual in a given population. The shape of the DFE is a result of several factors such as population size, mating system and selective environment, and can in turn influence the evolutionary potential of a species. The DFE has long been a field of intense research, but particularly since molecular methods enabled us to study of genetic variation in organisms empirically. This research has led to the development of several statistical methods that use population-level frequencies of single nucleotide polymorphisms (SNPs) to infer the DFE. However, these methods rely on assumptions about the data and the organism itself, which could potentially affect the accuracy of the inferences. In this thesis, I describe how two major factors – data quality and inbreeding – can affect the accuracy of DFE inferences. I also show how and when to (and when not to) use DFE inference methods based on SNP frequencies.

All genomic datasets contain inaccuracies and some level of uncertainty. The data sets are therefore often treated to remove the gaps or less reliable information, such as genotypes with low coverage. Some data sets need heavy filtering, which could reduce the amount of data available for analysis. We show that the choice of filter method affects the size of the final data set and the accuracy of the estimated DFE.

Many DFE estimation software assumes random mating within the study population. Unfortunately, this assumption induces some error when trying to estimate the DFE in inbred or selfing species. Some have assumed that this is a result of high rates of homozygosity in the data, and should only be a problem in populations with very high rates of selfing (>99%). We show that accuracy of the estimated DFE decreases already at relatively low rates of selfing (70%) and that removing homozygosity does not improve the accuracy, implying that another mechanism could be causing the error.

Keywords

Distribution of fitness effects, site frequency spectra, missing data filtering, downsampling, imputation, sample size, population structure, inbreeding, selfing, homozygosity, Arabidopsis, Pinus, DFE-alpha, SLiM, population genomics, adaptation.

Language

English

ISBN

print: 978-91-8070-268-3
PDF: 978-91-8070-269-0

ISSN**Number of pages**

45 + 4 papers