Incorporating linkage into genotype-environment association studies

Tom R. Booker*, **Samuel Yeaman**† **and Michael C. Whitlock***
*University of British Columbia, †University of Calgary

ABSTRACT Here is a really concise and nicely written summary of the paper, highlighting the main findings and take home messages.

I'm just using the GENETICS template because it looks nice!

KEYWORDS Local Adaptation, Population Genetics, Environmental Genomics

Introduction

With the advent of high-throughput sequencing technologies and methods to analyse the resulting data.
Understanding the genetic architecture of adaptation is
There are a number of factors that potentially contribute tot he architecture of local adaptation
A
N
Going forwards,

Correlated coalescent histories for linked sites under neutrality

Linked sites do not evolve independently. If the rate of recombination is low relative to the rate of migration, we postulate that within a certain genetic distance, c, there will be strong autocorrelation in the coalescent histories among sites. If that is the case, SNPs that are within c Morgans of each other may behave effectively as independent draws from the same distribution. Under that assumption, calculating the correlation of allele frequencies with at SNP i provides a test of the hypothesis, is this particular genomic region correlated with the.

Under that assumption, all of the neutral SNPs present within a distance c of each other provide an independent test of the hypothesis, "is the genetic variation in this particular region associated with the environment?".

There are two reasons why tightly linked markers may exhibit similarly strong associations with the environment. Firstly, alleles that contribute to local adaptation that establish may generate LD, so closely linked markers. Secondly, theory suggests that when local adaptation is facilitated by a slleles of small effect architecture, there is a selective advantage for those alleles to cluster in the genome. This effect may manifest itself as multiple alleles

Why do we expect linkage disequilibrium under local adaptation?

Theoretical studies of local adaptation suggest that we should expect regions of the genome subject to spatially varying selection pressure to exhibit elevated linkage disequilibrium relative to the genomic background. There are several possible reasons why this might be the case. Firstly, a locus subject to strong spatially varying selection can act as a barrier to gene flow in that paricular region of the genome and generate LD with surrounding sites (Barton and Bengston et al). Such theoretical studies have typically only examined the

Manuscript compiled: Thursday 27th August, 2020

role of alleles with spatially antagonistic effects on fitness. Conditionally neutral alleles, which conceivably contribute to local adaptation, have not received the same theoretical treatment. However, if selection acting on conditionally selected sites were strong it would presumably generate LD in the regions where they arose. Second, there is a selective advantage for alleles that are involved in local adaptation to aggregate in regions of low recombination so favourable combinations of alleles are inherited together (Yeaman papers). For that reason we might expect multiple causal alleles in particular genomic regions, for example, in the protein-coding and regulatory regions of genes. In sunflowers, for example, there is evidence that suppressed recombination among alleles involved in local adaptation has arisen via large chromosomal inversions (Owens paper). In either case, we expect that strongly selected loci to exhibit elevated LD relative to the genomic background.

In this study, we propose using combining information across tightly linked sites to identify regions of the genome under selection. Typically GEA studies examine patterns of genetic variation across a landscape at many polymorphic sites. Typically single nucleotide polymorphisms (SNPs) are used.

Materials and Methods

The Weighted-Z Analysis

The weighted-Z test combines *p*-values from multiple independent tests into a single score with each test given a weight that is proportional to the inverse of its error variance (Whitlock 2004).

At a given polymorphic site, we denote the average frequency of the minor allele across populations as $\overline{p}$ ($\overline{q}$ corresponds to the major allele). The product $\overline{pq}$ provides an estimate of the variance in allele frequencies among populations, so is appropriate as a weight.

Here we propose the weighted-Z analysis (hereafter the WZA) as a means to combine information across sites. Each polymorphic site (typically single nucleotide polymorphisms) is used to calculate a In the context of GEA we combine information in For genomic region, k, which contains n polymorphic sites we calculate the following,

$$Z_{w,k} = \frac{\sum_{i=1}^{n} \overline{p_i q_i} z_i}{\sqrt{\sum_{i=1}^{n} (\overline{p_i q_i})^2}} \tag{1}$$

where $\overline{p_i}$ and $\overline{q_i}$ are the average allele frequencies across demes for polymorphism i and z_i is the standard normal deviate.

The top-candidate test

Yeaman et al (2016) proposed a method for combining information across sites in genotype-environment association studies. The top-candidate test, as Yeaman et al (2016) called it, attempts to identify regions of the genome involved in local adaptation under the assumption that alleles in such regions will tend to generate LD with neighbouring sites so multiple linked markers may exhibit a significant correlation with important environmental variables. First, the genome-wide distribution of SNPs is examined to identify outliers. SNPs with *p*-values in the 99th percentile genome-wide are classified as outliers. Then, the frequency of outlier SNPs in analysis windows is calculated. In Yeaman et al (2016), sequence up and downstream of genes were used as analysis windows but since we make use of it in this paper, we provide

There are philosophical reasons as to why the WZA should be preferred. First, the Top-candidate test assumes that there is a fraction of the genetic markers analysed that are tagging causal variants (i.e. that there are true positives in the dataset). This is undesirable, because there may well be no detectable variation that contributes to local adaptation present, i.e. the study may simply be underpowered. Secondly, the test gives equal weight to all markers. However, alleles at different frequencies possess different levels of information about population history. A final related point is that all SNPs that have exceeded the significance threshold are treated identically. For example, with a significance threshold of 0.01, genomic regions with only a single outlier are treated in the same way whether that outlier has a *p*-value of 0.009 or 10^{-10}.

Simulating local adaptation

Simulations were performed in SLiM v3.4 (Messer and Haller). We simulated genomes with four chromosomes, modelled using $c = 0.5$ breakpoints. Local adaptation can act as a barrier to gene-flow, individuals that migrate from locations where they are well adapted into locations where they are disfavoured may not survive to propagate, so even freely recombining regions of the genome are linked to a degree and will influence evolution in regions of the genome that are freely recombining. To model local adaptation in unlinked regions of the genome, we modelled three cartoon chromosomes. The cartoon chromosomes were short sequences, 1,000bp long, that had a genetic map and net mutation rate (U_a) that was the as the focal chromosome.

We simulated three kinds of environment. The first was a reduced representation of climatic variation across British Columbia, Canada. We downloaded the map of degree days greater than 0 (DD0) for British Columbia from ClimateBC (website; REF). From the DD0 map, we extracted the data for a 99x99 grid using Dog Mountain, BC as the reference point in the South-West corner. We divided this map into a 14x14 grid. Each cell corresponded to an area of XXkm^2. We calculated the mean DD0 for each cell in the grid. We then converted the mean DD0 scores into Z-scores and rounded values up to the nearest third. These data were then used as the phenotypic optima for population models in SLiM.

We simulated local adaptation in four metapopulation models. The first is the island model, which represents an unstructured metapopulation.

Environments exhibit spatial autocorrelation. To model spatial autocorrelation in environment, we simulated a 2-dimensional stepping-stone model.

We simulated local adaptation across an We constructed a map using data for real climate variation from British Columbia, Canada. We downloaded the map of degree days greater than 0 for British Columbia from ClimateBC (website; REF). From the DD0 map, we extracted the data for a 99x99 grid using Dog Mountain, BC as the reference point in the South-West corner. We divided this map into a 16x16 grid and calculated the mean DD0 for each cell. We converted the means into Z-scores and rounded values up to the nearest third of a Z-score. These data were then used as the phenotypic optima for a structured population models in SLiM.

We simulated local adaptation using a model of stabilising selection. We modelled a genome of composed of four autosomes.

We used the standard expression for Gaussian stabilising selection,

$$W(z_{i,j}) = exp\Big[\frac{-(z_{i,j} - \theta_j)^2}{2V_s}\Big],$$

where z_i is the phenotype of the i^{th} individual in environment j, θ_j is the phenotypic optimum of environment j, and V_s is the variance of the Gaussian fitness function.

To test the performance of the weighted-Z analysis (WZA), we modelled populations adapting to various environments. Figure **??** shows a diagram of each of the populations simulated.

Covariance of phenotype and environment

In our simulations modelling stabilising selection, we used the covariance between phenotypes and environment as a measure of a gene's relevance for local adaptation. We calculated the average phenotypic effect of each gene as follows. We then used $\mathrm{Cov}(PB_g, env)/\mathrm{Cov}(PB, env)$ as a measure of a gene's contribution to local adaptation.

In our simulations, phenotypic variance (σ^2_P) was generated solely by genotypes, i.e. there were no environmental effects. Local adaptation generates variance in phenotypes between populations (σ^2_{PB}). As described above, the simulations incorporated a stochastic mutation model, so from replicate to replicate the effect size of alleles and their locations in the genome varied. As a result, the genes that contributed to local adaptation varied across simulation replicates. We therefore determined the contribution each gene made to local adaptation by calculating the proportion of phenotypic variance among populations explained by the SNPs in each gene. For each gene that contributes to phenotypic variation there are k causal SNPs each with a phenotypic effect of α_k. We use $\boldsymbol{\nu_g}$ to refer to the column vector of phenotypic effects for each of the k causal SNPs in gene g. In each population there are n diploid individuals and we have $\boldsymbol{M_d}$, an $n \times k$ matrix in which the

genotype of each individual at each causal SNP is coded as 0, 1 or 2 corresponding to aa, aA and AA genotypes, respectively. The contribution that each gene makes to the overall phenotype in each population is calculated as $C_{g,d} = \sum \boldsymbol{M_{g,d}} \boldsymbol{\nu_{g,d}}$. The variance in $C_{g,d}$ gives us a measure of the phenotypic variance between popuatlions generated by each gene ($\sigma^2_{PB,g}$). We then calculate the proportion of variance explained by each gene (PVE_g) as $\sigma^2_{PB,g}/\sigma^2_{PB}$

$$PVE_g = \frac{\sigma^2_{PB,g}}{\sigma^2_{PB}}. \quad (2)$$

Note that PVE_g does not provide a measure of local adaptation, merely a measure of how much phenotypic variation between populations can be explained by a a particular gene.

Analysis of simulation data

We added neutral mutations to each simulated tree sequence at a rate of 1×10^{-8} using PySLiM (version). While this gave us a population scaled mutation rate of $4N_e d\mu = 0.00078$, and resulted in an average of 30 SNPs per gene that passed a minor allele frequency filter of 0.05.

For each SNP, we calculated Kendall's τ and recorded the p-value. We chose Kendall's τ over Spearman's ρ as an uncorrected GEA statistic as it can handle ties in data,

Application to data from Lodgepole pine

We re-analysed a population genomic dataset collected for lodgepole pine distributed across the North West of North America. The data were initially generated and described by Yeaman et al (2016). Initially, the top-candidate test was applied to this data. We calculated Z_W scores for the same genes analysed by Yeaman et al (2016). Data were accessed from the Dryad respository associated with Yeaman et al (2016) (DRYADLINK)

Data Availability

The simulation configuration files and code to perform the analysis of simulated data and generate the associated plots are available at github/TBooker/GEA. Tree-sequence files for the simulated populations are available at Dryad and all processed GEA files are available on (https://doi.org/10.5061/dryad.0t407).

Results

Analysis of an island model

To assess the statistical properties of the WZA and the top-candidate test, we first performed GEA analyses on neutrally evolving populations structured according to an island model. While highly unrealistic, analysing this model allowed us to determine the statistical properties of the WZA and the top-candidate test without the need to correct for the confounding effects of population structure.

The distribution of Z_W scores obtained for 5,000 neutrally evolving genes was very close to the expectation of the standard normal distribution. The mean Z_W was 0.00X and the variance was 1.XXX. Figure **??**A shows the distribution of Z_W scores obtained when analysing a sample consisting of 50 individuals from 40 demes (2,000 total).

Imposing selection on the island model resulted in

Discussion

Genomic regions of low recombination exhibit greater variance in population genetic summary statistics than do more highly recombining ones , complicating statistical inference. We simulated genomes that experience uniform recombination rates, though that is highly unrealistic. We did so for statistical convenience. When analysing real datasets, researchers should be mindful that analysis windows of a constant physical size may

Acknowledgements

Thanks to Tongli Wang for help with BC climate data and to Simon Kapitza for help with wrangling raster files.

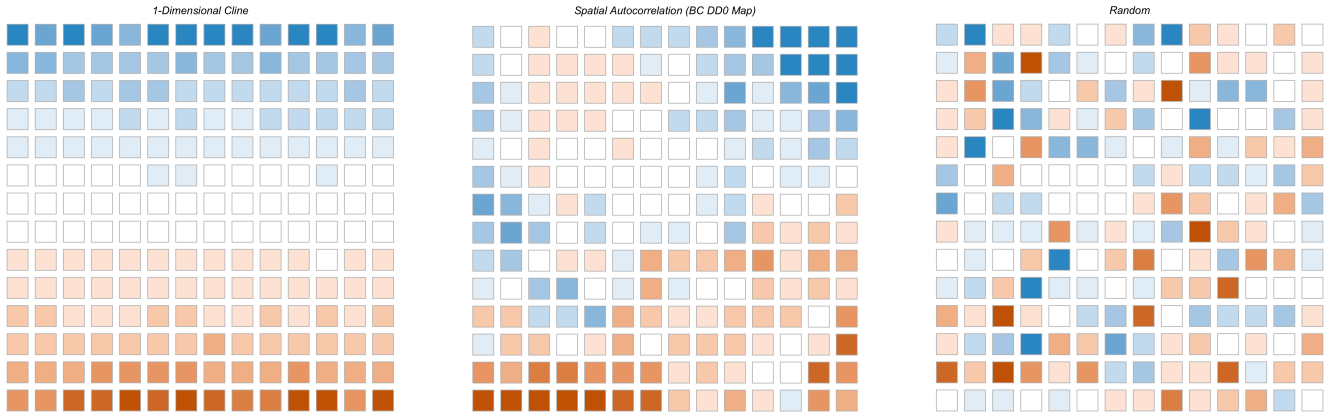

Figure 1 Three models of population structure used to simulated varying degrees of spatial autocorrelation in the environment. A) A highly discretized map of degree-days above 0 (DD0) in South-Western British Columbia, capturing realistic spatial autocorrelation in an environmental variable species may respond to. We refer to the map in A as the BC map. B) A 1-dimensional cline in phenotypic optimum, we refer to this as the cline map. C) A heterogenous distribution of phenotyic optima. The distribution of phenotypic optima in the cline and random maps