Genome-wide association studies in natural populations: managing expectations and avoiding error.

Susan E. Johnston1 and Lewis G. Spurgin2

1 Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3FL, UK.

2 School of Biological Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

# Abstract

Identifying the specific genomic regions underlying phenotypic variation is a central aim of evolutionary biology. High-throughput genomic technologies allow large numbers of genetic markers to be generated in almost any natural system, making trait mapping more tractable. The most common approach to trait mapping is a genome-wide association study (GWAS), a method that examines the association between individual genetic markers (e.g. single nucleotide polymorphisms, hereafter SNPs) and traits of interest. The efficacy of GWAS relies on linkage disequilibrium (LD) between genotyped SNPs and causal mutations contributing to the phenotype, and has been used with some success in mapping loci underlying both discrete and quantitative traits in several natural populations. There are numerous characteristics of genomic datasets obtained from natural populations that can impede the ability to conduct of trait mapping, leading to false positive (Type I error) and/or false negative results (Type II error). Yet we have limited understanding of how type I and II error in GWAS is affected by the interactive effects population and pedigree characteristics, particularly in the context of the marker densities and sample sizes typically available to ecologists and evolutionary biologists. We use simulation approaches to assess the degree to which Type I and Type II errors affect GWAS in natural populations. We use population and pedigree level simulations to generate genome-wide SNP datasets under different demographic scenarios (effective population size, cryptic population structure) and family structures, and highlight the major issues that are likely to affect GWAS. It is our aim that this study will allow researchers to have realistic expectations of the power of their datasets to detect trait loci, and to help researchers and readers be critical in interpreting GWAS results obtained from natural populations. Finally we present some (partial) solutions to the issues raised from our simulations and discuss some ways in which error can be minimised.

# Introduction.

**The fundamentals of GWAS.** What can it tell us, and why do we want to do it in natural populations? [Also – using genomic approaches to estimating heritability?]

* Helps us to understand underlying molecular mechanisms e.g. how many loci, their cumulative effects, if they are coding or regulatory (want to get to functional basis – Storz & Wheat 2011, Barrett & Hoekstra 2011)
* Help us to understand how traits are evolving – why they remain heritable – how genetic variants are associated with individual trait variation and fitness – solve evolutionary pizzles

**Background.**

* Mendelian traits obviously easy to map. QGen frameworks useful, based on infinitesimal (& additive?) model (Fisher 1918). Fishers geometric model of adaptation…  
  More and more studies starting to find large effect loci.
* On opposite side, truly quantitative traits hard to map e.g. human height… missing heritability.
* Rockman 2011 Evolution: GWAS doesn’t rely on QTL from inbred/outbred crossers (where LD is generated) but rather at the population level where “LD is generated by population history”.
* NGS etc. opens new perspectives to study genetic basis of traits…. Two approaches – 1. Identify VA (i.e. adaptive potential…) and 2. Identify genes involved in adaptation and test them for sigs of selection or correlations with fitness
* Distn of gene effect sizes (nice summary in Gagnaire & Gaggioti (2016) Curr Zool

**The false positive:** why it’s a problem. The Beavis effect, population structure, family structure, population bottlenecks, heteroscedasticity between categories, drift, sampling effects at rare alleles, temporal and spatial variation in allele frequencies/phenotypes, no accounting for other sources of variation (environmental effects).

**The false negative:** depends on LD between markers and trait loci. Not enough markers, strong sample sizes, small effect sizes, number of loci underlying the trait, the heritability. If not enough genetic variation is captured, then underestimate heritability. Markers are typed, not causal variants. Often common variation is on chips; rare variants are untyped or removed through QC.

[For both false positive and negatives, further problems may arise when investigating differences between the sexes, genetic correlations, GxE interactions, selection, bottlenecks, or categories with differences in variance, rare effect alleles, effect size distribution, variation in heritability, ascertainment bias.]

**Why these are issues in natural populations.**

Experimental setups – they would have done crosses between inbred lines/outbred populations to maximise LD between causal loci and markers… Natural populations could allow more fine mapping as lots of recombination breaks down linkage blocks, but at the same time makes things very difficult to detect.

Mapping approximate regions can be just as informative than QTN…

**How do we approach and fix these problems?** [see below for more detail] Generating enough marker data, sampling enough individuals, using hybrid or experimental populations, quantitative genetic approaches (standard animal model, chromosome partitioning, regional heritability.

**Why are we writing this review?** We want readers to have realistic ideas of what they can achieve and how to accurately interpret their findings. Also, we don’t want to put people off – rather make people aware of the problem. We offer solutions and help here. NB. As this is a review, we present case studies and emphasise that we cannot simulate new scenarios, but at least point out where problems can occur. Starting point for people wanting to investigate the extent of these issues in their own datasets, or in answering their own questions.

**What we will not talk about:** marker selection, …

Examples:

Mendelian: Colosimo 2005, Linnen et al 2009, Johnston et al 2013, Barson et al 2015, Bosse? Price? Mothes?

Quantitative traits: Hancock et al 2011, Arnegard et al 2014, Berenos et al 2015….

# Illustrating the problem: a simulation approach

We will illustrate how Type I and II error arise in GWAS of wild populations using a combination of population and pedigree level simulations. We simulated four “scenarios”, representing i) a large panmictic population, ii) a large population with local family structure, iii) a structured metapopulation and iv) a small population with family structure. These scenarios were chosen to capture the diversity of study systems used in GWAS of wild populations, from outbreed marine and continental species (e.g. FISH/FLYCATCHER/GREAT TIT REFS), to isolated island populations (SHEEP/FINCH REFS). We stress here that our aim is not to provide a systematic analysis of the factors that affect error rate in GWAS, but rather to provide an illustrative account of how GWAS performs in simulations resembling some of the different types of study system typically used by ecologists and evolutionary biologists.

We used the software QMSim (REF) for all simulations. QMSim allows the user to first generate equilibrium levels of LD and standing variation using population-level simulations, and then to use this to simulate recent populations with family structure. Although designed for livestock studies, the flexible nature of QMSim makes it highly suitable for simulating natural population scenarios.

For all scenarios, we first simulated genomes in a historical population for 1000 generations. Each individual genome was divided into 10 chromosomes, each of 100 cM in length and each containing 50,000 markers and a single QTL, giving 500k markers and 10 QTL in total. Markers and QTL were randomly distributed along chromosomes, and QTL effects followed a gamma distribution with a shape parameter of 5. The combined QTL effects were simulated to contribute to a continuous trait with heritability of 0.4.

In scenarios involving family structure (scenarios 2-4), we used pedigree simulations, with the historical population as the founder population. In all pedigree simulations we simulated equal sex ratios. Offspring production of each pair varied between 1 and 10, with a mean value of approximately two. Survival was simulated as a function of age, with a 50% replacement ratio; thus, generations were overlapping, and population size was approximately constant over time. All other parameters varied among our simulated populations, and are outlined in the specific scenarios below. The code to recreate all simulations is available from XXX.

We performed standard GWAS analyses on our simulated datasets using …

**Scenario 1: Large panmictic population**

We first simulated a large panmictic population, at a population size of 105, for 1000 generations, using the population-level parameters described above. Because this approach was designed to mimic very large outbred populations with no family structure likely to occur in samples (e.g. atlantic herring, some salmonids), we did not perform pedigree simulations, and instead sampled 1000 individuals from the historical population.

These simulations revealed that with 500k markers and 1000 individuals…

Reducing the sample size revealed that…

**Scenario 2: Large panmictic population with local family structure**

Many study systems in ecology and evolution are highly outbred with local family structure (e.g. many bird systems sampled from nest boxes, some fish REFS). To simulate such a system we first simulated a historical population of 105 individuals identical to that described in scenario 1. From this, we simulated 20 overlapping generations using a pedigree simulation with the parameter settings described above, starting with a founder population of 500 individuals. We sampled XX individuals from the final 5 generations of the pedigree… ADD DETAILS OF HOW SAMPLED.

With a sample size of 500K markers and 100 individuals etc etc

**Scenario 3: Structured meta-population with family structure**

We simulated a structured meta-population as follows. First, we simulated a single, historical population with a population size of 105 for 1000 generations. From these we simulated four pedigree populations, each with a founder population size of 50, using the parameter settings defined above. Samples were randomly drawn from across the four subpopulations from the final five generations of the pedigree for GWAS analysis.

**Scenario 4: Small population with family structure**

To test how GWAS performs in a small, isolated population, with family structure we simulated a historical population of 300 individuals, for 100 generations. A pedigree simulation was then performed in the same way as for Scenatio 2, expect that the founder population size was 300. Individuals were sampled from the last five generations for analysis.

# Limitations of our analysis

Comment about the fact that variation in recombination rate is not really accounted for and why

# Recommendations, solutions and future directions

**Designing the study:** Marker density, selecting suitable individuals? Case/controls from same populations, or multiple populations… bleh blah

**Checking the data for structure and accounting for it:** PCA, fitting relationship matrices...

**Checking and fixing results for inflation:** lambda, what kind of correction to use (genomic control, eigenstrat/PCs, kinship matrix

**Power analysis?**

Stop focusing on the QTN - Distn of effects across genes, rather than individual effects e.g. Bayes-R

Regional heritability, partitioning, quant gen approaches

Near isogenic lines may pick up