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

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). False negatives can arise in cases where there is low LD between typed SNPs and causal variants (e.g. due to large population size and/or low marker densities), where the effect sizes of causal variants are low (e.g. due to low heritabilities and/or polygenic architectures), or where causal loci interact with other genetic and/or environmental variables. False positives pose a significant problem, and can arise through generation of LD between loci and phenotypic variation due to cryptic family or population structure, population history (e.g. bottlenecks). Type I error can also arise due to small sample sizes, which are subject to over-estimation of effect sizes (Beavis effect) or spurious associations with phenotype due to chance, particularly at rare alleles. Whilst there are some statistical approaches to ameliorate false positives from model systems (e.g. genomic control, accounting for relatedness), they may be overly conservative or ineffective in natural datasets. Furthermore, 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.

The aim of our paper is to use simulation approaches to assess the degree to which Type I and Type II errors affect GWAS in natural populations. Our objectives are as follows:

1. Population level GWAS: We will use coalescent approaches to simulate genome-wide SNP datasets under different demographic scenarios (effective population size, cryptic population structure, bottlenecks) and genetic architectures (e.g. variation in heritability, number of loci, locus effect sizes), incorporating information on chromosome size, recombination landscape and marker density. This will allow us to assess the relative effects of each parameter in generating Type I and/or Type II errors in GWAS under different population histories.
2. Pedigree level GWAS: We will extend coalescent approaches from (1) to generate founder population haplotypes and use a gene-dropping approach accounting for linkage and recombination to assess levels of Type I and/or Type II error in pedigree data (i.e. datasets with family structure). Our aim is to generate pedigrees reflective of published or in progress GWAS studies where demographic histories are well characterised, including Seychelles Warbler (Ne of magnitude 101), Soay sheep (Ne = 102), Atlantic salmon (104) and great tits (106).

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. We will conclude by present potential solutions to the issues raised from our simulations. For example, we will explore whether multi-locus approaches, such as regional heritability, offer a viable solution to provide the required data for conducting evolutionary genetic studies.