Background:

There is a pressing need to model and predict demographic impact on populations as species adapt to environmental changes (Merilä & Hendry, 2014). In conservation genetics and genomics, the effect of genetic diversity on population persistence and demographics is well studied (Allendorf & Luikart, 2009), but the use of genomic data to predict population demography during selection, particularly when traits are highly polygenic, is rare and challenging (Bay, Rose, Barrett, et al., 2017).

In one recent example, Bay, Rose, Logan, et al. (2017) sequenced *Acropora hyacinthus* coral adapted to different thermal environments at many single nucleotide polymorphisms (SNPs), and SNPs for which FST (a measure of population differentiation) was significantly larger than others were considered candidate loci for adaptation and considered to have an identical effect on thermal tolerance. The authors then simulated the coral populations forward in time to predict future demographics and estimate the probability of population extinction. This specific approach may not work in most cases, since example populations such as those used by Bay, Rose, Logan, et al. (2017) may not always exist, and all causative loci rarely have identical effect sizes. More importantly, genome-scan methods for detecting causative alleles like those employed above, are biased towards the detection of major effect alleles (Jensen et al., 2016), but many adaptive traits, particularly those which may confer critical adaptive capacity regarding climate change, are likely influenced by many genes (Bay, Rose, Barrett, et al., 2017). Genome-wide association studies (GWAS) and genomic prediction (GP) offer alternative but generally unexplored routes, since both allow direct discovery and quantification of allelic effects as long as phenotypic variation exits in a population. Additionally, GP methods in particular may be well suited to reducing bias introduced by highly polygenic traits (Visscher et al., 2017). It is important, however, to quantify error in demographic simulations driven by estimations from GWAS or GP.

*Overview of GWAS and GP*

GWAS has been used for over ten years to discover loci which influence a given phenotype and estimate the size of their effects (Visscher et al., 2017). In essence, GWAS involves leveraging linkage disequilibrium (LD) to locate regions of the genome which contain causative loci, typically via associating phenotypes with allele frequencies at nearby (likely neutral) loci (Visscher et al., 2012). Since LD degrades rapidly with genomic distance, GWAS requires high sequencing resolution—ideally thousands to millions of markers depending on genome size—to be effective (Risch & Merikangas, 1996). GWAS also has difficulty identifying causative loci in highly polygenic traits or when causative variants are rare unless sample sizes are extremely high (Visscher et al., 2017).

The reliance of GWAS on LD may also create problems when a trait is controlled by a single major effect locus. For example, the time of year during which Chinook and steelhead salmon return to river systems from the ocean to breed is likely controlled by a single ancestrally derived causative allele (Prince et al., 2017). In their original study, Prince et al. found that a single SNP in the GREB1L gene was approximately X% predictive of early (pre-mature) migration in steelhead. This SNP and the next most highly predictive SNP were separated by a gap of approximately 50kb, which presumably contains the actual causative variant. The imperfect association between the sequenced SNP and phenotype could be taken to imply that some individuals without the causative pre-mature allele nonetheless return as pre-mature migrants, and, as a result, that any future loss of the pre-mature migration allele may not correspondingly result in the loss of the pre-mature migration phenotype. However, a recent study that sequenced the GREB1L gene at a much higher resolution identified a SNP which was perfectly associated with pre-mature migration (cite Tasha’s paper). The difference in association with phenotype between the two SNPs is due solely to imperfect LD between those SNPs and the causal variant.

In this case and others, demographic predictions based on GWAS are likely to be biased by imperfect LD. This effect is likely to be more severe when sequencing resolution is lower, since the genomic distance between causative and sequenced loci will increase. Regardless of sequencing resolution, traits should also generally appear more polygenic than in reality, since the effects of single variants should be spread out somewhat amongst partially associated nearby SNPs. As Prince et al. (2017) illustrates, such errors can create profoundly different predictions for future demographic trajectories by causing overestimates of population evolvability.

An alternative to traditional GWAS, GP methods estimate the effect of all SNPs on phenotype jointly (typically using a linear model framework) in order to reduce the bias created by imperfect linkage and highly polygenic traits (see for example Cheng et al., 2018; Meuwissen et al., 2001; Yang, Lee, et al., 2011; Yang, Manolio, et al., 2011). These methods are typically employed principally in the human health or agricultural fields, since they rely on high sequencing resolution and large sample sizes or detailed pedigree information, and have been only occasionally used in natural systems. For example, genome-wide complex trait analysis (GCTA) was pioneered on a study of 294,831 SNPs in 3,925 unrelated individual humans (Yang et al., 2010), and GP in general likely requires large sample sizes and intensive sequencing to render accurate predictions. 3000+ samples sequenced at 50,000SNPs may be required to reach ~15% prediction accuracy if the effective population (Ne) is large, and 12,0000+ individuals may be required to reach ~30% accuracy (Lee et al., 2017). When Ne is smaller (~1000), ~45% accuracy may be reached with 3000+ samples, and >80% accuracy can be achieved with similar sample numbers when Ne is very small (<100) (Lee et al., 2017). Sample numbers this high may not be achievable in natural populations, however, so it is critical that the efficacy of GP with smaller samples be evaluated. This issue may fade somewhat as the cost of sequencing drops, although not in situations where acquiring many samples is more difficult than funding their sequencing. Best linear unbiased prediction (BLUP), on the other hand, often employs pedigree information alongside genotypes (Henderson, 1975) which may not be available in natural populations without long-term study (see Kruuk, 2004).

More importantly, most GP methods by design are optimized to accurately predict breeding values (the net contribution of an individual’s genotypes to phenotype) rather than identify and quantify the effects of specific loci. The effect of this on demographic simulations based on sampled genotypes is not established. Furthermore, like with GWAS, imperfect association may still cause effect sizes of major effect alleles to be spread across nearby genes, which should subsequently influence estimates of population evolvability.

Despite these issues, information on causal loci inferred via GWAS and/or GP offer a straightforward way to identify and quantify the effect of causal loci which can then be used to simulate population level responses to changes in selection optima even when comparative data like that employed by Bay, Rose, Logan, et al. (2017) are not available. However, it is critical to first estimate and account for error during demographic simulations due to any bias or error in the effect size estimates derived from GWAS and GP under a range of sampling scenarios and genomic architectures.

Proposed Work:

Here, we propose to assess the accuracy of demographic simulations based on results from GWAS and GP. To do so, we propose to simulate whole genome sequences for a sample of individuals, assign effect sizes to random loci, and simulate the effect of a shifting selection optima on the population demographics. We will then use one GWAS and two GP tools to estimate allele effect sizes from phenotypes derived from the simulated data, conduct similar demographic simulations based on those estimates, and compare the results to those from the “real” allele effect sizes. In order to assess the effect of various parameters on the accuracy of demographic simulations, will also run 500 simulations using combinations of three marker effect size distributions, two heritabilities (h2), three sequencing resolutions, and two effective population sizes (see Figure 1), for a total of 60,000 simulations (54,000 based on estimated effect sizes plus 6,000 based on actual effect sizes).

*Genome, phenotype, and effect size simulation and estimation*

We will use coalescent simulations to generate realistic simulated genomes. Specifically, we will generate two sets of 1000 chromosome copies for ten 10mb long chromosomes with recombination rate rho = 40,000 and mutation rates theta = 4,000 and 400. This is equivalent to two samples of 500 diploid individuals from populations with effective sizes of 10,000 and 1,000, respectively, belonging to a species with a 100mb long genome with mutation and recombination rates of 1x10-8­ and 1x10-7 per base. In order to generate samples with realistic LD patterns and allelic distributions, we will use the scrm coalescent simulator (Staab et al., 2015) and check the resulting LD patterns by calculating r2 for each pair of loci using the snpR package (Hemstrom et al. in prep).

Since results from GP may be influenced by traits that are either controlled by many or few genes, we will then generate three different effect sizes for each SNP in the resulting dataset. First, we will simulate a highly polygenic trait by randomly drawing SNP effects from a distribution where each SNP has a one in 1000 chance of having an effect, and SNP effects are drawn from a normal distribution with mean = 0 and standard deviation = 0.5. Second, we will simulate a somewhat polygenic trait by randomly drawing SNP effects from a distribution where each SNP has a one in 10000 chance of having an effect, and SNP effects are drawn from a normal distribution with mean = 0 and standard deviation = 0.5. Lastly, we will simulate a trait controlled by a single major effect allele by randomly assigning a single SNP an effect size of 0.5 (such that homozygous individuals for the causal allele have a BV of 1 and heterozygotes have a BV of 0.5). For all cases, we will run the analysis with both h2 of 0.5 and 1.

The resulting genotypes and phenotypes will be run through several GWAS and GP software tools to estimate marker effect sizes. Specifically, we will perform GWAS using the PLINK software (Purcell et al., 2007), which is arguably the most widely used GWAS utility available. To conduct GP, we will use both the BGLR (Pérez & de los Campos, 2014) and JWAS software packages (Cheng et al., 2018). To simulate different sequencing resolutions, we will perform GWAS and prediction on the full dataset, a set of 100,000 random SNPs, and a set of 10,000 random SNPs.

*Demographic simulations*:

Using the simulated genomes and simulated or estimated allele effect sizes, we will simulate the effect of a gradual shift in phenotypic optimum on population demographics. Specifically, we will model population demographics for a hypothetical semelparous, randomly mating population with no mutation that grows using a logistic model where *K* is the carrying capacity, *r* is the growth rate, *Nt*is the population size in the current generation, and *Nt+1* is the population size in the next generation. *Nt* in each generation is calculated by summing the number of individuals who survive until mating, where the probability of survival is drawn from a scaled normal distribution with standard deviation , where is the phenotypic variance in the first generation, and individuals with the optimum phenotype have a survival probability of 0.7. Individual phenotypes are given by and , where *eij* is the count of the causal allele *j* in individual *i* and *aj* is effect of causal alleles at locus *j*, respectively, and *Ei* is a random environmental effect drawn from a normal distribution with mean 0 and standard deviation , where is the additive genetic variance in the first generation and *h2* is heritability. Each generation, the optimum phenotype will increase by a constant arbitrary value. Genomes for generation *t+1* will be generated by randomly assigning sex to surviving individuals in generation *t* by drawing from a binomial distribution, randomly assigning a mother and father to each of *Nt+1* individuals, and then randomly assigning one paternal and one maternal copy of each chromosome to those individuals. Recombination will be simulated by randomly positioning *n* recombination events along each chromosome, where *n* is drawn from a Poisson distribution with λ = 1 (as given by a recombination rate of 1x10^-7 per base pair and a chromosome length of 10mb). At each recombination event, alleles will switch source from one maternal or paternal chromosome copy to the other. All simulations will be run until either population extinction or for 1,000 generations.

Predicted results:

After all simulations are complete, those run based on allele effect sizes estimated by GWAS or GP will be compared to those run with identical parameters (effect size distribution, h2, and Ne). Specifically, we will calculate covariance between population size and genetic diversity changes over time as well as average time to extinction between simulations based on actual and estimated effect sizes. Principally, we hope the accuracy of each method of effect size estimation under each of the conditions tested. Ideally, we then hope to suggest corrections to compensate for any systematic biases, such as overestimation of polygenicity, and quantify error rates otherwise. In doing so, we aim to show the utility of effect size estimation via GWAS and GP for directly predicting the effects of environmental change on populations.

Effect sizes:

1/1000

1/10000

1 SNP

h2:

.5

1

nSNPs:

all

100000

10000

Ne:

10000

1000

Programs:

PLINK

JWAS

BRLP

real

500 per, so 60,000 sims. Thus the small genome. Should I skip Ne?

Allendorf, F. W., & Luikart, G. (2009). *Conservation and the genetics of populations*: John Wiley & Sons.

Bay, R. A., Rose, N., Barrett, R., Bernatchez, L., Ghalambor, C. K., Lasky, J. R., Brem, R. B., Palumbi, S. R., & Ralph, P. (2017). Predicting Responses to Contemporary Environmental Change Using Evolutionary Response Architectures. *The American Naturalist, 189*(5), 463-473. doi:10.1086/691233

Bay, R. A., Rose, N. H., Logan, C. A., & Palumbi, S. R. (2017). Genomic models predict successful coral adaptation if future ocean warming rates are reduced. *Science Advances, 3*(11). doi:10.1126/sciadv.1701413

Cheng, H., Kizilkaya, K., Zeng, J., Garrick, D., & Fernando, R. (2018). Genomic Prediction from Multiple-Trait Bayesian Regression Methods Using Mixture Priors. *Genetics, 209*(1), 89-103. doi:10.1534/genetics.118.300650

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics, 31*(2), 423-447. doi:10.2307/2529430

Jensen, J. D., Foll, M., & Bernatchez, L. (2016). The past, present and future of genomic scans for selection. *Molecular ecology, 25*(1), 1-4. doi:doi:10.1111/mec.13493

Kruuk, L. E. B. (2004). Estimating genetic parameters in natural populations using the ‘animal model’. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359*(1446), 873-890. doi:10.1098/rstb.2003.1437

Lee, S. H., Clark, S., & van der Werf, J. H. J. (2017). Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. *PLoS ONE, 12*(12), e0189775. doi:10.1371/journal.pone.0189775

Merilä, J., & Hendry, A. P. (2014). Climate change, adaptation, and phenotypic plasticity: the problem and the evidence. *Evolutionary Applications, 7*(1), 1-14. doi:doi:10.1111/eva.12137

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics, 157*(4), 1819-1829.

Pérez, P., & de los Campos, G. (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics, 198*(2), 483-495. doi:10.1534/genetics.114.164442

Prince, D. J., O’Rourke, S. M., Thompson, T. Q., Ali, O. A., Lyman, H. S., Saglam, I. K., Hotaling, T. J., Spidle, A. P., & Miller, M. R. (2017). The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. *Science Advances, 3*(8). doi:10.1126/sciadv.1603198

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., & Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*(3), 559-575.

Risch, N., & Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. *Science, 273*(5281), 1516-1517. doi:10.1126/science.273.5281.1516

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics, 31*(10), 1680-1682. doi:10.1093/bioinformatics/btu861

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics, 90*(1), 7-24.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics, 101*(1), 5-22.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet, 42*, 565. doi:10.1038/ng.608

<https://www.nature.com/articles/ng.608#supplementary-information>

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics, 88*(1), 76-82.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., Hill, W. G., Landi, M. T., Alonso, A., Lettre, G., Lin, P., Ling, H., Lowe, W., Mathias, R. A., Melbye, M., Pugh, E., Cornelis, M. C., Weir, B. S., Goddard, M. E., & Visscher, P. M. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet, 43*, 519. doi:10.1038/ng.823

<https://www.nature.com/articles/ng.823#supplementary-information>