**Background and Rationale:** When an ecosystem changes, either due to natural or anthropogenic causes, species within that system can only respond in three ways: they can escape the system, evolve or adapt to the changes, or go extinct (Holt, 1990). Since the first option is not possible if suitable habitat does not exist elsewhere, and the third is generally considered undesirable from a conservation standpoint, there is a pressing need to model and predict the demographic impact of adaptation to environmental change on populations (Merilä & Hendry, 2014). While quantitative genetic models have been developed on this subject (e.g. Lynch & Lande, 1993), genomic data, despite its substantial potential, has very rarely been used to improve predictions of population persistence during environmental change (Bay, Rose, Barrett, et al., 2017).

The most straightforward way to incorporate genomic data into quantitative models is via population additive genetic variance (), effective size (*Ne*), and heritability (*h2*), all of which can be estimated from genomic and phenotypic data (Meuwissen et al., 2001; Pérez & de los Campos, 2014; Waples et al., 2016). These parameters are critical to population persistence in the face of rapid environmental change. For example, according to Lynch and Lande (1993), in sexually reproducing populations that experience a constant rate of environmental change, , *Ne*, and *h*2 collectively determine the critical rate of change (*nc*) beyond which a species cannot adapt under a given selection scenario. Bürger & Lynch (1995), hereafter BL, expanded this model somewhat to include demographic stochasticity, and found that the expected time to extinction also depends principally on the same variables, as did a more recent model by Chevin et al. (2010). However, recent attempts to test or parameterize these models using *Drosophila* *birchii* or *Parus major*, respectively, did not attempt to incorporate genomic data and instead estimated , *Ne*, and *h2* purely via demographic and phenotypic data (Gienapp et al., 2013; Willi & Hoffmann, 2009).

While estimating , *N­e*, and *h2* with genomic data should improve quantitative genetics-based predictions of population persistence, these summary statistics still neglect a great deal of genomic data. Incorporating genotypes explicitly should further improve predictions. For example, Bay, Rose, Logan, et al. (2017) used candidate adaptive loci in *Acropora hyacinthus* populations to infer the probability of individual survival based on their genotypes under different temperature regimes, and then simulated populations forward in time with random mating to predict future demographics and estimate the probability of population extinction. While this method used more information than the three summary stats discussed above, it nonetheless assumed that all candidate loci have an equal effect on phenotype, which is certainly inaccurate. Incorporating information on the actual distribution of individual loci effect sizes, their linkage disequilibrium (LD) structure, and their interactions should therefore make for more accurate predictions.

Here, we propose to use modern population and quantitative genomic methods to directly estimate and incorporate individual marker effect sizes or individual breeding values (BVs, the sum genetic contribution of an individual to their phenotype) into recombination-explicit demographic/genomic selection simulations in order to better predict outcomes for populations that are exposed to environmental change. Specifically, we plan to use genome-wide association study (GWAS), genomic prediction (GP), and machine learning (ML) methods to estimate and predict marker effect sizes and individual BVs. Briefly, GWAS leverages LD genome-wide 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). GWAS requires very high sequencing resolution to be effective, particularly in large populations with lower levels of LD and may have difficulty identifying causative loci with small effects on phenotype (Risch & Merikangas, 1996; Visscher et al., 2017). GP methods generally rely on penalized or Bayesian methods to generalize or “shrink” estimates and enable regression models to be fit to genomic data, which typically has dramatically fewer individuals (*n*) than genotypes per individual (*p*) (de los Campos et al., 2013; Meuwissen et al., 2001). This enables the BV of individuals to be predicted via summing estimated effect sizes within individuals, although some methods bypass this and estimate BVs directly from genome-inferred pedigrees (de los Campos et al., 2013). GP works well in situations with many loci of minor effect, although it is still poor at detecting complex gene-gene interactions which effect phenotype (de los Campos et al., 2013; Meuwissen et al., 2001). Lastly, ML methods are non-parametric and model independent, and instead focus on algorithmically defining rules that allow for accurate prediction based on training data (Schrider & Kern, 2018). As such, they make few assumptions about the underlying mechanics which relate genotype to phenotype and are thus better able to incorporate many weakly influential or interacting loci to predict BV (Schwarz et al., 2010).

Each of these models allows genetic information to be used to directly estimate the BV of individuals given their genome and an initial dataset that relates genotype to phenotype. Incorporation of this data has the potential to radically improve our ability to forecast demographic outcomes for species adapting to environmental changes if employed in predictive simulation or model frameworks such as those used by BL or Bay, Rose, Logan, et al.(2017).

**Specific Objectives:** *Here we propose to incorporate results from GWAS, GP, and ML methods into a demographic simulation framework in order to improve on existing quantitative genetic models and better predict the persistence or extinction of populations that must adapt to environmental change*. If these methods improve demographic prediction accuracy, then simulations which incorporate estimates derived from them will better match simulations conducted using complete, accurate loci effect size, LD, and interaction information than do predictions derived from the incorporation of , *N­e*, and *h2* estimates into quantitative genetic models such as that derived by Bürger & Lynch (1995).

**Procedures:** In order to assess the ability of GWAS, GP, ML, and classic quantitative genetic models to predict population persistence, we will simulate genomes and loci effect sizes for populations under a range of different biological settings (hereafter Real Data, or RD) and run demographic simulations on those datasets. We will then compare these results to identical simulations run using BVs predicted from RD phenotypes and genotypes by GWAS, GP, and ML methods (hereafter Predicted Data, or PD), and to predictions of persistence generated by the BL model. Each RD or PD dataset will consist of individual genotypes and phenotypes derived from either real or predicted BVs.

We will generate RD data by first simulating genomes to generate genetic variance, assigning causal effects to genetic markers under a range of scenarios, and then calculating the phenotypes of each individual. We will simulate realistic genomes for two entire populations consisting of 1,000 and 10,000 individuals under several scenarios using the scrm coalescent simulator (Staab et al., 2015). The properties of the simulation and output genomes are given in Table 1. Note that we plan to simulate genomes for two different *Ne* values because the precision of GP varies depending on *N­e* (Lee et al., 2017; Meuwissen et al., 2001). We will check the resulting LD patterns for realism by calculating r2 for each pair of genomic loci using the snpR package (Hemstrom et al. in prep). Since the accuracy of the GWAS, GP, and ML results are all influenced by the distribution of loci effect sizes on phenotype (Meuwissen et al., 2001; Molinaro et al., 2011; Visscher et al., 2012), we will generate RD datasets using several different loci effect size distributions, as described in Table 2. Individual phenotypes will be calculated using the formula , where , *cij* is the count of the causal allele *j* in individual *i*, *ej* is effect of causal alleles at locus *j*, respectively, and *Ei* is a random environmental effect. *Ej* will be drawn from a normal distribution with mean 0 and standard deviation , where is additive genetic variance in the first generation and *h2* is heritability. For each individual simulation run, we will determine phenotypes twice, once each for *h2* 0.5 and 1.

PD datasets will be created by running random samples of 500 individuals from RD datasets through several different GWAS, GP, and ML methods in order to estimate loci effect sizes or BVs, as described in Table 3. Phenotypes will be estimated using the formula . will either be equal to estimated BV or, if BV is not directly estimated, calculated using , where is the estimated effect of the allele at locus *j*. *Ei* will be calculated as above, albeit with standard deviation , where is the estimated heritability. will be calculated by the GWAS, GP, or ML model if possible or, if not, via Genome-enabled Best Linear Unbiased Prediction, a GP method, using the JWAS software package (Cheng et al., 2018). In addition, since sequencing resolution also changes the accuracy of GWAS, GP, and ML results, we will create PD datasets under all parameter combinations using all single nucleotide polymorphisms (SNPs), 1,000,000 SNPs per genome, and 100,000 SNPs per genome.

Identical demographic simulations will be run on both the RD and PD datasets. 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 determined by Gaussian stabilizing selection around the optimum phenotype in that generation (), as described in BL equation 1. For each generation, will be calculated according to the equation , where *k* is the constant rate of environmental change and *E* is a random stochastic effect with mean 0 and variance . Genomes for generation *t+1* will be generated by randomly assigning sex to surviving individuals in generation *t*, a mother and father to each of *Nt+1* individuals, and then one paternal and one maternal copy of each chromosome to each offspring. Recombination will be simulated by randomly positioning *n* recombination events along each chromosome, where *n* is drawn from a Poisson distribution with λ = 1 (equivalent to the recombination rate ρ used to simulate the genome). All simulations will be run until either population extinction or for 1,000 generations. All growth, selection, and stochastic parameters will be tuned via small test runs for each set of RD datasets in order to cause extinction in RD datasets at *t* ≈ 500. For PD datasets, sampled individuals will first be randomly mated up to *Ne* in order to simulate demographic scenarios for the entire population.

For each possible combination of parameters, we will simulate 1,000 PD datasets and their corresponding demographic outcomes. Likewise, each 1,000 copies of the RD dataset for each combination of relevant parameters will be created. All the variables that define these groups of simulations are listed in Table 4. This equates to 24 RD dataset groups, each of which corresponds to 15 PD datasets, and thus 384,000 (24 RD + 24\*15 PD x 1,000) total simulations. The mean time to extinction within each group of PD datasets will be compared that of the corresponding RD dataset to determine the viability of each prediction method.

In addition, the time till the start of strong population decline (*t1*) will be calculated using BL equations for each of the RD datasets by estimating *Ne* according to Waples et al. (2016) and and *h2* using JWAS G-BLUP (Cheng et al., 2018). For comparison, we will also use the BL equations in stepwise stochastic simulations where *Ne* and are sequentially calculated generation to generation using the BL equations according to Gienapp et al. (2013). These results will also be compared to the simulation results from the RD and PD datasets.

**Results to Date:** Simulated genomes have been created as described above and much of the core R code needed for the project has been written. Specifically, the code needed to simulate population growth, selection, random mating, and genetic recombination under a range of user-defined models and parameters based on loci effect sizes has been completed. Revisions to this code to allow for BV estimation from ranger, JWAS, and BGLR models are ongoing. Code to estimate marker effect sizes and generate models to predict BVs is complete for ranger and BGLR, and is in progress for JWAS. Code can be found at https://github.com/hemstrow/genomic\_prediction\_accuracy.

**Tables and Figures:**

|  |  |
| --- | --- |
| Table 1: Parameters and their values for simulated genomes. | |
| Parameter | Value |
| Mutation Rate (μ) | 1x10-8 |
| Recombination Rate (ρ) | 1x10-7 |
| Genome Size | 100mb |
| Chromosome Length | 10mb |
| Chromosome Count | 10 |
| Effective Population Size (Ne) | 1,000 and 10,000 |

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2: The names, number of causal loci, number of genes that contribute to each individual effect on phenotype, and the percentage of effects that are epistatic. | | | |
| Name | # Causal Loci | Max # Interacting Genes | % Epistasis |
| Major Effect | 1 | 1 | 0 |
| Slightly Polygenic | 20 | 1 | 0 |
| Slightly Polygenic with Epistasis | 20 | 2 | 20 |
| Polygenic | 200 | 1 | 0 |
| Polygenic with Epistasis | 200 | 2 | 20 |
| Highly Polygenic | 1000 | 1 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| Table 3: Method name and citation, category (GP, GWAS, or ML), model name, and output variable for each method used to generate predicted BVs for demographic simulations. | | | |
| Name | Category | Model | Output |
| JWAS (Cheng et al., 2018) | GP | G-BLUP | BVs |
| BGLR (Pérez & de los Campos, 2014) | GP | BayesB | BVs |
| PLINK (Purcell et al., 2007) | GWAS | Quantitative Trait Association | Marker Effect Sizes |
| TASSEL (Bradbury et al., 2007) | GWAS | GLM | Marker Effect Sizes |
| Ranger (Wright & Ziegler, 2017) | ML | Random Jungle | BVs |

|  |  |  |  |
| --- | --- | --- | --- |
| Table 4: List of variables that define each distinct group of datasets, their possible options, and whether they vary within real data (RD) and predicted data (PD) datasets. | | | |
| Name | Possible Options | Varies within RD Groups | Varies within PD Groups |
| Effective Population Size (Ne) | 1,000 or 10,000 | Yes | Yes |
| Causal Locus Effect Size Distribution | See Table 2 | Yes | Yes |
| Heritability (h2) | 0.5 or 1 | Yes | Yes |
| Sequencing Resolution | All, 10,000, or 100,000 SNPs | No | Yes |
| BV Prediction Method | See Table 3 | No | Yes |

**References:**

Bay, R. A., Rose, N., Barrett, R., Bernatchez, L., Ghalambor, C. K., Lasky, J. R., … Ralph, P. (2017). Predicting Responses to Contemporary Environmental Change Using Evolutionary Response Architectures. *The American Naturalist*, *189*(5), 463–473. https://doi.org/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). https://doi.org/10.1126/sciadv.1701413

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*. https://doi.org/10.1093/bioinformatics/btm308

Bürger, R., & Lynch, M. (1995). Evolution and Extinction in a Changing Environment: a Quantitative-Genetic Analysis. *Evolution*, *49*(1), 151–163. https://doi.org/10.1111/j.1558-5646.1995.tb05967.x

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. https://doi.org/10.1534/genetics.118.300650

Chevin, L.-M., Lande, R., & Mace, G. M. (2010). Adaptation, Plasticity, and Extinction in a Changing Environment: Towards a Predictive Theory. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1000357

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, *193*(2), 327–345. https://doi.org/10.1534/genetics.112.143313

Gienapp, P., Lof, M., Reed, T. E., McNamara, J., Verhulst, S., & Visser, M. E. (2013). Predicting demographically sustainable rates of adaptation: can great tit breeding time keep pace with climate change? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1610), 20120289. https://doi.org/10.1098/rstb.2012.0289

Holt, R. D. (1990). The microevolutionary consequences of climate change. *Trends in Ecology & Evolution*, *5*(9), 311–315. https://doi.org/https://doi.org/10.1016/0169-5347(90)90088-U

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), 1–22. https://doi.org/10.1371/journal.pone.0189775

Lynch, M., & Lande, R. (1993). Evolution and extinction in response to environmental change. In R. Kareiva, PM and Kingsolver, JG and Huey (Ed.), *Biotic Interactions and Global Change* (pp. 234–250). SUNDERLAND: SINAUER ASSOCIATES.

Merilä, J., & Hendry, A. P. (2014). Climate change, adaptation, and phenotypic plasticity: the problem and the evidence. *Evolutionary Applications*, *7*(1), 1–14. https://doi.org/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 LP-1829. Retrieved from http://www.genetics.org/content/157/4/1819.abstract

Molinaro, A. M., Carriero, N., Bjornson, R., Hartge, P., Rothman, N., & Chatterjee, N. (2011). Power of data mining methods to detect genetic associations and interactions. *Human Heredity*, *72*(2), 85–97. https://doi.org/10.1159/000330579

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

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/https://doi.org/10.1086/519795

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

Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, *34*(4), 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Schwarz, D. F., König, I. R., & Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, *26*(14), 1752–1758. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btq257

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. Retrieved from http://dx.doi.org/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*, 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

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. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Waples, R. K., Larson, W. A., & Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity*, *117*, 233. Retrieved from https://doi.org/10.1038/hdy.2016.60

Willi, Y., & Hoffmann, A. A. (2009). Demographic factors and genetic variation influence population persistence under environmental change. *Journal of Evolutionary Biology*, *22*(1), 124–133. https://doi.org/10.1111/j.1420-9101.2008.01631.x

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software; Vol 1, Issue 1 (2017)*. https://doi.org/10.18637/jss.v077.i01