October 28, 2022

Chris Setzke
University of Washington
School of Aquatic and Fishery Sciences
1122 NE Boat St
Box 355020
Seattle, WA 98195

Dear Chris Setzke,

We spoke earlier this morning in-person to discuss your research project on the genetics of kidney disease in rainbow trout populations. The motivation for this research is to breed disease-resistant rainbow trout in human-managed hatcheries. In this letter, I aim to document my notes from our meeting and provide some answers and recommendations to the questions and dilemmas you posed. I hope this letter is useful to you, your collaborators, and any other (bio)statistician you may consult for statistical and experimental design advice.

Meeting Notes:

- Client: Chris Setzke is a 2nd year PhD student in the SAFS program at UW. This project is a part of Chris' 50% RA appointment in which he works 20 hours a week.
    - Chris comes from a population genetics background.
    - Chris can code in R, write Bash scripts, and run programs on the terminal.
    - Chris has access to a cluster where he can run high-memory, parallelized jobs.
    - Chris has successfully run statistical genetics programs like Plink, Beagle, and vcftools.
    - Other team members have expertise in fishery sciences, genetics, and computational biology. They do not have a formal (bio)statistician on the team.
    - This project is in collaboration with the US Department of Agriculture.
- Project Goals:
    - Identify rainbow trout carrying disease resistant quantitative trait loci (QTL) to breed.
    - Determine haplotypes in designated QTL that help characterize disease resistance.
        - Study crosses between: strains November and May; phenotypes days and mortality; phenotype days and strains
        - For example, a QTL could contain 78 markers. They want to identify haplotypes of 3 markers in that market set that explain any signal.
    - Decide which sequencing technique and experimental design to move forward with to study the causal mechanisms of disease resistance. This task is ***time-sensitive*** and requires feedback in 1-2 months.
        - Animal genetics studies have less funding than human genetics studies.

- Dataset: variant calls from Affymetrix SNP array platforms for rainbow trout samples, and phenotype data on mortality and time-to-event after exposure to a pathogen.
    - There are two breeding lines, denoted as November 2018 and May 2019.
    - There are two generations, a parental level and an offspring level.
    - One line has ~1800 children and the other ~1700 children.
    - There are ~100 families in each line.
    - Both lines are bred in a common environment and separated based on genotyping at the time of the experiment (applying the exposure and monitoring mortality).
    - The offspring generation is formed via a circular mating scheme (link). This breeding scheme generates many siblings and half-siblings. Chris can explain these details.
    - Only the offspring were exposed to the pathogen and monitored for mortality. The monitoring occurred for 20-30 days, and the date of deaths were recorded.
    - The November line experienced 80% mortality. The May line experienced 90% mortality. However, the May line also experienced an additional secondary infection from an independent source.
        - While mortality seems high in both settings, Chris showed me some figures suggesting that mortality is lower in some families and higher in other families. This observation indicates that there may be some genetic mechanisms in play.
    - There were initially ~57k single nucleotide polymorphisms in the variant callset. After filtering steps, there were 30-40k SNPs. One filter may have been a 20% MAF threshold, which I consider to be a high threshold.
    - Chris reports that the markers are reported well with high accuracy. There are few Mendelian errors from Plink outputs.
    - Analyses to date, including various GWAS analyses, have been performed unphased.
    - They do not have a rainbow trout reference panel. Reference panels improve phasing.
    - Missing genotype calls are uncommon. Phasing programs can impute these. They intend to only impute the missing genotype calls because they do not have a reference.
    - The variant callsets are different for the different lines. Did you all use the same Affymetrix array platform? You could have different markers in the different callsets because of rare variants that do not appear in the other line.
    - They do not have MAP files? Some email correspondence indicates that there may be MAP files. Chris mentioned that rainbow trout may have a consensus MAP file, but it may be for an older chromosome build.
- Existing Analyses:
    - Prior work by the group and details of their analysis plans are detailed here (link).
    - They applied the methods wssGBLUP (weighted single-step genomic best linear unbiased prediction) method and ssBMR-BayesB (single-step Bayesian multiple regression BayesB) method to perform GWAS analyses.

    - They defined 1 mega basepair (Mbp) sliding windows as quantitative trait loci. If adjacent sliding windows show a signal, they may merge them into a single QTL. This decision of sliding window size and merging appears to be ad-hoc.
    - In their current work, they focus on these QTL because they have found marginal effect sizes in SNPs.
- Question/Answer
    - Should Chris phase entire chromosomes, QTLs, or just 3 markers?
        - Best practice is to always phase using as many markers as possible. For example, the Beagle program, which is based on a localized haplotype cluster model, requires a good many markers to warm up.
    - Is it best to phase with known pedigree information?
        - Best practice is to use pedigree information as this is very informative. However, many popular phasing programs are not built for pedigree applications but instead population-based GWAS studies. On one hand, these phasing programs can improve with the inclusion of related individuals, even if the program does not know about the relatedness (Figure 4; Browning and Browning, 2013). On the other hand, Beagle 4.0 can use pedigree information (link). As well, Beagle 4.0 and Beagle 4.1 (link) can use genotype likelihoods, which are commonly preferred in fish studies where coverage is low. Another option is the pedigree phasing PULSAR of Blackburn et al. (2020), where in Tables 1-2 Beagle has poor switch error rate without a reference panel. However, these are small sample sizes, much smaller than the 1-3k sample size, and population-based phasers like Beagle usually improve with increasing sample size. The downsides of PULSAR are that heterozygous sites are dropped and computation time is quadratic in the # of meiosis, which would be high for this dataset with so many sibships.
    - Are there methods to identify haplotypes associated with a trait?
        - Yes, there are methods for haplotype-based association testing designed for array data. See the paper Browning (2006), the Beagle 3.3 program (link), and the documentation that describes the advanced analyses using cluster2haps.jar program. You can perform this association testing after phasing with a more recent Beagle program. Important outputs for this are the *.pval and the *.dag.gz files. Caution should be exercised in interpreting the *.pval file without first consulting a statistician. There are two types of Fisher exact tests conducted here: a 2 x M omnibus test and a 2 x 2 edge against all the others test, where M is the number of edges. Beagle version 3- takes the input file form *.bgl. I can share a Python script to translate a VCF file to a BGL file.
    - Which experimental design to take for the sequencing follow-up study?

        - I recommend you reach out UW STAT and BIOST consulting service (link). They could help design some simulation studies to answer this question. Nobu Masaki in the Browning group is in the consulting course this autumn quarter and is still looking for a final project. It is possible Nobu could also help you interpret the haplotype-based association testing. Otherwise, this consulting service seldom has expertise in statistical genetics.
- Other References:

    - Blackburn, August N., Lucy Blondell, Mark Z. Kos, Nicholas B. Blackburn, Juan M. Peralta, Peter T. Stevens, Donna M. Lehman, John Blangero, and Harald H. H. Göring. 2020. “Genotype Phasing in Pedigrees Using Whole-Genome Sequence Data.” European Journal of Human Genetics: EJHG 28 (6): 790–803.
    - Browning, Sharon R., and Brian L. Browning. 2011. “Haplotype Phasing: Existing Methods and New Developments.” Nature Reviews Genetics. Nature Publishing Group. https://doi.org/10.1038/nrg3054.
    - Browning, Sharon R. 2006. “Multilocus Association Mapping Using Variable-Length Markov Chains.” American Journal of Human Genetics 78 (6): 903–13.
    - Browning, Brian L., Xiaowen Tian, Ying Zhou, and Sharon R. Browning. 2021. “Fast Two-Stage Phasing of Large-Scale Sequence Data.” American Journal of Human Genetics 108 (10): 1880–90.
    - Browning, Sharon R., and Brian L. Browning. 2007. “Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering.” American Journal of Human Genetics 81 (5): 1084–97.
    - Loh, Po-Ru, Pier Francesco Palamara, and Alkes L. Price. 2016. “Fast and Accurate Long-Range Phasing in a UK Biobank Cohort.” Nature Genetics 48 (7): 811–16.
    - Delaneau, Olivier, Cédric Coulonges, and Jean-François Zagury. 2008. “Shape-IT: New Rapid and Accurate Algorithm for Haplotype Inference.” BMC Bioinformatics 9 (December): 540.

Again, I hope these notes and suggestions are helpful. You are welcome to share this document with your collaborators. If consenting, I can alert Tamre Cardoso and Nobu Masaki in statistical consulting about your project.

Sincerely,

Seth D. Temple
UW Statistics PhD Student
NDSEG Fellow
NIH Trainee in Statistical Genetics

UNIVERSITY of WASHINGTON
Department of Statistics