UC SANTA CRUZ BIOMOLECULAR ENGINEERING AND BIOINFORMATICS TRACK PH.D. STATEMENT OF PURPOSE

JARED G. GALLOWAY

Biomolecular engineering and bioinformatics research has opened new horizons for human health and understanding of the forces driving evolution. With recent advances in technology, quantitative biology research, and access to large datasets, the field has become more equipped to solve complex problems that will play a large role in current and future generations to come. I am applying to graduate school because I believe it is a necessary challenge to nourish my biological skill set and interest in quantitative problem-solving. This training will present an opportunity for me to leverage my computational background and to help find creative solutions to problems facing the field at an exciting time for biology. Moving forward, I would like to help progress the field and participate in the discussions that define our understanding of complex systems. My professional goals are to build tools (software) that biologists can use to help solve a wide variety of problems in genetics research.

Over the course of the past two years I have completed two major projects and am currently working as a full-time scientific programmer in the Institute of Ecology and Evolution at the University of Oregon. Advised by Dr. Peter Ralph and Dr. William Cresko, the first project I took the lead on was studying local polygenic adaptation of stickleback fish populations in Alaska. This northern hemisphere-wide metapopulation includes both marine populations and a large number of smaller freshwater populations that have repeatedly adapted to freshwater conditions, often by using standing genetic variation. For this project we wanted to know what range of introgression between marine and freshwater populations was required to maintain the transportation of freshwater alleles. We were also interested in the genetic signals we can expect to see in real data and how gene flow impacts the those signals. Using SLiM, I wrote code to run large evolutionary simulations which emulated the geography and evolutionary history of stickleback populations in Alaska. We then varied the gene flow by changing migration rates to observe the impact of selection on standing genetic variation. From this we found that rapid, repeated adaptation using alleles maintained at low frequency by migration-selection balance occurs over a realistic range of intermediate rates of gene flow. We outlined the rates of gene flow which allowed us to see causal loci from F_{st} scans across the genome as well as the rates at which migration load prevents adaptation. Lastly, we traced back to the origin of all alleles which came to high frequency in the introduced populations after adaptation had occurred, and found the majority were pre-existing in the first generation as opposed to being carried in by subsequent migration. As a first author, the manuscript for this paper is in the final stages of editing and will be posted to the BioArxiv soon, then submitted

to Genetics. The most recent draft can be found at `https://github.com/jgallowa07/SticklebackPaper/blob/master/Stickleback_Paper.pdf`.

Working with Dr. Ben Haller and Dr. Peter Ralph on the implementation and profiling of *Tree Sequence Recording* in SLiM 3.0, was a large part of my undergraduate thesis. Genealogical tree sequence recording is a strategy for efficiently recording the genealogical history from forward-moving simulations. This history is represented by the forest of trees relating all sampled individuals to each other over every genomic interval. *TreeSeq* uses a collection of tabular data structures to encode this history which was introduced for use in the coalescent simulator msprime. Using TreeSeq, simulations in SLiM can avoid the cost of tracking and propagating neutral mutations as a by-product of obtaining the origins of all sampled genotypes. For this project we utilized a variety of software engineering tools including C/C++, cmake, xcode and an agile workflow. After successfully implementing *TreeSeq* with rigorous testing, we found simulations where individuals had realistic size genomes experienced a speedup of over 2 orders of magnitude. The paper describing the applications of this strategy, titled *Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes*, was published by Molecular Ecology Resources and can be found online at `https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12968?af=R`.

My current project working with Dr. Andrew Kern involves using deep learning to infer population genetics parameters from sampled data. For this project, we want to know if the recent advances in deep learning architecture can learn complex patters in genotype matrices resulting in accurate predictions of recombination and mutation rates. Using Tensor Flow and other data analysis packages for python, I have set up a pipeline used for: simulating and storing large datasets efficiently, concurrently prepping data batches while training on the previously generated batch, and testing the performance of trained neural networks. Using this pipeline, we have found architectures and data prepping heuristics which have resulted in predictions of recombination rates that consistently outperform industry standards such as LD Hat.

The projects described above have exposed me to a wide variety of research environment practices including organization of experiments, scientific writing, seeking and approaching others when I need help, and effectively communicating my work. I have gained a familiarity with many concepts in genomics, population genetics, and applied computer science skills such as simulation and machine learning. Through graduate training in biomolecular engineering and bioinformatics, I aspire to be truly interdisciplinary. Lying at the intersection of computer science and biology, I strive to effectively communicate with experts on both sides so as to bridge the fields. The current advances produced by biology experts along with the flood of data we anticipate has shaped an exciting future for genetics that I am eager to be a part of. The wide range of study that PBSE offers ranging from paleogenomics to precision medicine, would allow me to further explore where my skill set intersects with my interests.

If shortlisted as a candidate, I would like the opportunity to speak with David Haussler, Benedict Paten, and Beth Shapiro

Jared G. Galloway