Arun Durvasula

Elucidating local archaic ancestry using machine learning

Background: The past decade of genomics research has demonstrated that introgression, the passing of DNA from one species to another, is a powerful force in shaping and maintaining patterns of genetic variation. In particular, we have learned that the human evolutionary tree is rife with archaic hominins whose genetic legacy lives on modern humans. For example, Malick et al (2016) cataloged the human diversity across the globe and quantified the amount of Neanderthal and Denisovan present in their genomes[1]. Nearly every population outside of Africa, they showed, contains some amount of archaic ancestry. A key question is why we are the only hominin species alive today? Discovering these regions of archaic ancestry allows us to begin to answer this question.

However, the state-of-the-art methods for pinpointing these regions of archaic ancestry are incomplete. They make use of an archaic reference genome derived from preserved remains to boost their power and accuracy. While these methods excel at discovering introgressed DNA for species we have sequenced, they have low power when applied to species without reference information. In particular, there is evidence of archaic introgression from a population that diverged about 700 thousand years ago into African populations, but archaic remains are unlikely to be sequenced due to poor environmental conditions for preservation[2]. This motivates the creation of a reference sequence-free method to detect archaic ancestry. Previous population genetic methods for this reference-free inference are powerful in certain circumstances, but no single method retains power across a wide range of parameter space.

Machine learning has increasingly been shown to hold great promise in prediction problems because it is able to combine many weakly informative statistics into a powerful inference framework. These methods take high dimensional observations and probabilistically assign them to classes. Research in applying machine learning methods to problems in evolutionary genomics has shown some recent success[3]. Here, I propose to use coalescent simulations to train a neural network to distinguish between archaic and non-archaic DNA.

Aim 1: Detecting introgression. Applying machine learning methods to inference problems in evolutionary genetics requires training data where the ground truth is known. In the area of introgression, there are very few examples where this is the case. Thus, I propose to use simulations from the coalescent model to train a neural network, labeling simulations as the result of archaic introgression or not. The coalescent has been shown to match real genetic data in important ways and is thus a suitable source of training data.

The resulting simulated data must then be summarized before being used as input for inference with a neural network. I propose to use several statistics which are informative of archaic admixture including the minimum distance to a reference non-admixed population, the distribution of distances in the focal population, and S*, which is a statistic designed to be sensitive to archaic introgression.

With these statistics and class labels, I will train the neural network using stochastic gradient descent. I will then use the inference method on real genomic data, first in European populations to detect Neanderthal introgression. I can compare the neural network results to the results of previous methods that have access to the Neanderthal reference genome. This important step will establish the ability of the method to infer archaic introgression and provides an empirical test. Preliminary results suggest that the method has good performance in simulated test data and is able to recover previously discovered introgression regions in Europeans at the BNC2 locus. After a more complete performance evaluation, I can look at African populations from the 1000 Genomes project and Simons Diversity Genome Project and produce a map of local archaic introgression.

Aim 2: Functional impact of introgression. A natural question to ask once local introgression maps are made is, to what extent has selection shaped the assortment of introgressed DNA? Results from Neanderthal and Denisovan introgression suggest that there have been several instances of adaptive archaic introgression[4]. These results rely on haplotype patterns around the selected site and the observation of increased allele frequency of the archaic haplotype.

I propose to develop a machine learning framework to detect instances of adaptive introgression using a similar strategy proposed in Aim 1. I will simulate adaptive introgression from an archaic population using a modified version of the coalescent that allows for positive selection and create input features from the resulting data. Here, I can calculate summary statistics that have previously been used to characterize positive selection such as Extended Haplotype Homozygosity, locus specific population branching statistics, F_{ST}, and measures of derived allele frequency. These statistics all have power in certain circumstances, but are weak in many other scenarios. By combining them in a machine learning framework, I can take advantage of their signals across a wide range of scenarios and increase the overall power. Previous instances of archaic adaptive introgression have found results in immune related genes and genes related to environmental conditions, which I hypothesize may be the case here as well.

Future extensions: The method proposed here is focused on detecting archaic human populations, but can easily be extended to many other species where introgression has contributed to genetic variation (for example, fruit flies, bonobos, and *Heliconius* butterflies). Because there is often less information in these species, I will need to modify the method to take into account extra uncertainty in the demographic parameters and sequencing data.

Broader impacts: As an undergraduate, I was fortunate to receive extensive mentoring from graduate students, postdocs, and principle investigators (see Personal Statement). I believe that this mentorship directly drew me into a career in science and I am eager to pass it on. The proposed project touches on a broad area of interest, spanning genetics, computer science, and machine learning, which is conducive to recruiting students. UCLA is a particularly strong institution for this, with on campus programs such as Bruins in Genomics (BIG) connecting undergraduate students with on-campus researchers.

I have already begun mentoring a student in computer science, who is leading a project focusing on methods to automatically create features for a neural network. We hold weekly meetings to discuss the project and troubleshoot, and I assign papers for him to read to round out his knowledge of biology. This project will lead to authorship on a paper and help propel him into a career in science, should he choose to pursue one. As the project progresses, I will recruit more students from diverse backgrounds to tackle the interdisciplinary problems this research focuses on, including biology students who I can mentor in computer programming.

In addition, I am eager to engage the public with this research. Because it provides clues about our own species' origin, it will be of broad interest to the general public. I will reach out in venues such as my personal blog (~1,000 readers/month), and social media sites like Twitter and Reddit where science communication has proven very successful. This study will benefit society by providing us with empirical evidence of the contribution of other archaic hominins, deepening our understanding of where we come from as a species. Additionally, this study could provide answers to the questions of why we are the only hominin species alive today and what our close relatives were like. **References:** [1]Malick et al 2016 *Nature*; [2]Hammer et al 2011 *PNAS*; [3]Sheehan and Song 2016 *PLoS Comp Bio*; [4]Racimo et al 2015 *Nat Rev Gen*