Exploring genealogical relatedness, migration, and demography using population genomics

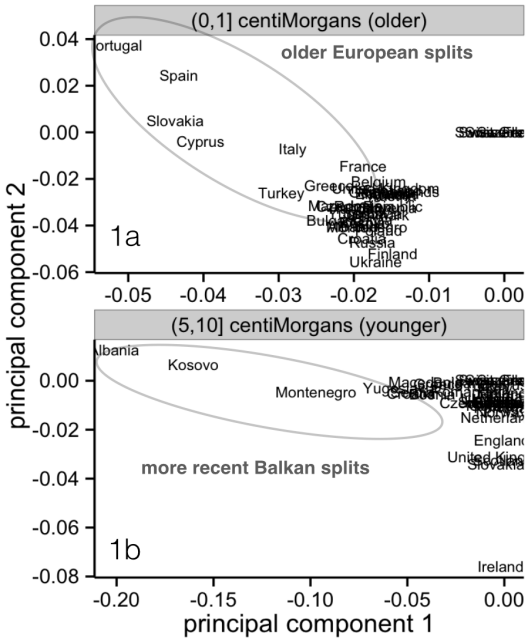
Background: Most species live in spatially structured populations, connected through a complex demographic history of population movements and splits, migrations, expansions, and declines. The demography of populations both reflects and impacts processes on evolutionary time scales (e.g. speciation, local adaptation, gene flow) and ecological timescales (e.g. competition, disease spread, predation). Aided by the recent abundance of genomic datasets and modern population genetics methods, researchers have begun to infer detailed demographic histories of numerous species including humans (Li and Durbin, 2011), dogs (Freedman et al., 2014), and agriculturally significant domesticated crops (Wright et al., 2005). Genetic data also captures information on the geographic distribution of individuals: those more distant from each other are less likely to share a recent common ancestor than those close to one another. This has allowed results like those of Novembre et al. (2008), which show that the geography of Europe can be recovered from European individuals' genetic data projected on two dimensions using Principal Component Analysis (PCA). These population genetics methods extract different summaries of the same underlying signal: the distribution of time to most recent common ancestors (or coalescent time) of the sample.

However, the current methods have two limitations. First, these methods are usually unable to reconstruct the recent demographic and migratory events, over the past tens to hundreds of generations. Such resolution is vital if we are to move population genomics towards ecological timescales and monitor populations' response to climate change. Second, existing methods are not well positioned to fully exploit the hundreds to thousands of individuals that researchers are already sequencing and genotyping. As datasets of this scale become the norm, we need population genomics methods that can take advantage of these large samples, which supply us with the power to study population history at greater resolutions. **Goal: I intend to develop an open source, novel population genomics inference framework that allows us to better reconstruct and visualize the demographic and geographic history of large samples of individuals over a variety of timescales.**

Method Aim 1: Extract genealogical histories. The time-depth of genetic relatedness (the coalescent time) differs along the chromosome for each pair of samples; this information records rich information about their shared ancestry. I will develop an approach to extract the distributions of these pairwise coalescent times between all individuals in the sample from full genome, genotyping (e.g. SNP chip or RAD-tag) data by extending existing approaches (e.g. Li and Durbin's, 2011). From this I will calculate the number of common ancestors that a pair of individuals share during different time periods (e.g. 0-10, 11-30 generations, etc). Over all pairs of individuals we obtain a matrix of the number of shared common ancestors for each time period, which captures individuals' relatedness in particular time slices. However, each matrix is likely to be sparse (contain many zero entries), due to noise in coalescent time estimation. To remedy this, I will develop a smoothing approach that exploits the concordance of coalescent probabilities in adjacent time slices (e.g. Li and Durbin, 2011; Ralph and Coop 2013).

Method Aim 2: Visualization of ancestral relationships through time. I will then summarize the information in these common ancestry matrices to learn about the geographic and demographic history of the sample through time. To do this I will utilize a low-rank approximation the original full-data matrix (similar to principal component analysis). These low-rank matrix approximations will allow us to visualize major axes of relatedness in different time intervals. Previous applications of principal

components to genetic data have revealed high resolution geographic information (Novembre et al., 2008's map of Europe) but these average over patterns of co-ancestry across all time periods. By separating out different time strata we will be able to track demographic and geographic signals back through time.



Hypotheses Tested with this Novel Method: After developing this method, I will apply it to two systems. First, I will apply my method to the POPRES dataset (2,257 individuals from varying European countries, genotyped at 500,000 SNPs) during the development process, as European demography is relatively well known and serves as a good benchmark. **Hypothesis 1:** my method's smoothing and low-rank projection techniques will recover a clearer picture of the various recent large scale migration events (e.g. the Balkan expansions). Preliminary versions of this method (Figure 1a), show that my low-rank approximations of older coalescent times (as captured by short shared genomic regions) capture older European population history, while low-rank approximations (Figure 1b) of more recent coalescent times (very long shared genomic regions) capture more recent history between individuals from Albania and Kosovo.

Second, the recent sequencing of 544 poplar trees (Evans et al, 2014) has provided extensive evidence for an adaptive response to climate change in *Populus trichocarpa*. Such adaptations are often geographically structured (Savolainen, et al. 2007), and changing climates can shift ranges and lead to recent demographic changes. Current methods that infer geographic and demographic history capture history tens of thousands of generations ago, which would not perceive these recent subtle shifts. **Hypothesis 2:** the more recent coalescent times my model utilizes will reveal recent range shifts at latitudes and elevations predicted to be most affected by climate change. Consequently, this model will capture much more recent demographic and geographic shifts in individuals due to exogenous pressures such as climate change.

Broader Impacts: I think that the application of this method to poplar data will be a terrific joint project for a group of driven undergraduates from backgrounds historically underrepresented in STEM research recruited through UC Davis's Biology Undergraduate Scholars Program (http://www.busp.ucdavis.edu/). This project will lead to undergraduate authorship(s) on a paper and help propel these individuals into science. Additionally, the growing popularity of ancestry and personal genomics websites indicate the broader public is interested in learning more about how human populations are related. Given that the Coop lab blogs about outreach topics (e.g. http://bit.ly/shared-ancestry) and discusses their research on large internet community sites such as Reddit (e.g. http://bit.ly/coopama), I can use these outreach channels to share results from my method (especially the visualization component) to increase public engagement and scientific literacy about how closely related related all humans are. Finally, I can use ancestry examples in my outreach blog for women interested in science (discussed in personal statement).

References: Evans, et al. *Nat. Genet.* (2014); Freedman, et al. (2014) *PLoS Genet.*; Li and Durbin (2011) *Nature*; Ralph and Coop (2013) *PLoS biology*; Savolainen, et al., (2007) *Annu. Rev. Ecol. Evol. Syst.*; Wright, et al. (2005) *Science*.