# Summary Discussion on Landscape Genomics

Discussion Leaders: Amy Zyck and Jennifer Soukup

Papers Discussed

***Brauer, Hammer, & Beheregaray 2016.*** Riverscape genomics of a threatened fish across a hydroclimatically heterogeneous river basin. Molecular Ecology: 25: 5093-5113.

***Hancock et al. 2011.*** Adaptation to climate across the *Arabidopsis thaliana* genome. Science: 334: 83-86.

The class began the discussion by summarizing the paper by Brauer, Hammer, and Beheregaray. The authors completed a riverscape genomic analysis on the threatened southern pygmy perch (*Nannoperca australis*). They collected environmental data and high-quality filtered SNPs to determine if environmental factors contribute to adaptive genetic divergence of *N. australis* populations. The authors used a combination of FSToutlier tests and genotype-environment association (GEA) analyses, with FSToutlier tests defining neutral loci and GEA analyses identifying candidate adaptive loci. The authors gathered data for around 40 environmental variables, which were then broken down into categories, narrowed down using variance inflation factor (VIF) analysis, and then further narrowed down using principal components analysis (PCA). Linear mixed-models were used to identify associations between allele frequencies and environmental variables. The authors found evidence of strong population structure associated with the river network, which confirms that genetic drift is a major evolutionary process shaping the genetic diversity in these populations. However, they also found evidence of selection driven by environmental factors, with temperature and precipitation acting as the major factors influencing the allele frequencies of the adaptive loci. The authors also tested for the effects of human disturbances on variation, and found that it influenced adaptive variation, but for less adaptive loci and only at a local scale.

The group then discussed the following topics:

*Gene Ontology (GO) annotations –* A method used to compare the sequenced genes of your study organism to already annotated genes in a database. The authors in this paper had few genomic resources for this particular species, so using the GO database to determine the function of the loci they were examining could have been helpful. However, the enrichment analysis that was run did not find any GO terms that were significantly under- or over-represented in the candidate adaptive data set compared with all loci. Additional information on gene ontology can be found [here](http://www.cs.tau.ac.il/~rshamir/ge/09/scribe/lec14a.pdf) and in du Plessis, Skunca, & Dessimoz (2011) - link [here](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3220872/).

*Enrichment –* In Gene Ontology there are a variety of categories such as metabolism, cell cycles, homeostasis, etc. When you see that the majority of your SNPs are associated to a single category instead of being dispersed among all the categories, then your protein is enriched for that category. For example, if all your SNPs are associated with metabolism, then metabolism is enriched for that protein. More information can be found [here](http://www.cs.tau.ac.il/~rshamir/ge/09/scribe/lec14a.pdf).

*Admixture Plot -* The admixture plot in this paper was used to demonstrate the presence of spatial genetic structure and showed that the majority of the pygmy perch populations were isolated. There were a few populations that hybridized, most likely due to the fact that they inhabited river terminals that were just starting to diverge.

*PCAs –* Principal components analysis was used to further reduce the number of environmental variables to be used in the GEA analysis. Within each environmental category that was created, principal components (PCs) that could explain most of the total variance were determined. For the temperature category, winter and summer temperature were the two variables that contributed the most to each axis and these variables were retained for GEA analysis. In the precipitation category, only one component had an eigenvalue > 1, so all the individual precipitation variables rather than PCs were used for further analysis.

*Eigenvalue –* Eigenvectors (the principal components) are associated with an **eigenvalue**

which can be interpreted as the “length” or “magnitude” of the corresponding eigenvector. In this study, eigenvalues > 1 were retained for GEA analysis. More information on PCAs and eigenvalues can be found [here.](https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html)

*Log-Bayes factor scores to determine outliers -* The log-Bayes factor scores for all of the high-quality filtered SNPs were plotted and those with a log Bayes-factor >15 were highlighted in red. The highlighted points represent candidate loci that are associated with the environmental variable being examined in each plot. This technique helps to remove all neutral genes as the goal was to identify selection driven by environmental factors.

*Redundancy analysis (RDA) -* RDA is an approach used to test hypotheses about specific environmental factors. With RDA, one can build and test models of varying complexity. Partial RDA includes models that condition results based on neutral genetic structure or spatial effects. Partial RDA can be used to assess correlations between multivariate climate and multivariate genetic variation while controlling for spatial effects. Putatively adaptive SNPs can be identified by looking at the contribution of each SNP to the first RDA axis.

The RDA plot in this paper was particularly difficult to understand. The blue arrows represent the environmental variables that were determined to be significant drivers for the selection observed. The length of the vector represents the weight the particular variable had on the model. The direction of the arrowhead indicates high values. The yellow color-coded sites had the highest amount of rainfall as indicated by the rainfall vector. Whereas the red color-coded sites were at the opposite end of the plot, indicating they had the lowest rainfall. The fact that the Lower Murray-Darling Basin sites (light blue) are in the middle indicate that these areas did not face extremes and were relatively stable in comparison to the rest of the sites. More information on RDA can be found in Chapter 12 of *Analyzing Ecological Data* (Springer, 2007).

*Hard sweeps -* A hard selective sweep results in the rapid fixation of an advantageous mutation. This rapid fixation will cause linked alleles to be dragged along to fixation because they are near another gene that is undergoing a selective sweep and is on the same DNA chain (genetic hitchhiking). A hard sweep tends to eliminate diversity in the region around the advantageous allele, create an excess of low-frequency variants, and increase linkage disequilibrium and haplotype homozygosity. More information on hard sweeps can be found [here](http://www.genetics.org/content/210/4/1429).

*Soft sweeps -* A soft sweep is a selective event acting on standing variation where this standing variation becomes selectively favored and sweeps to fixation. A soft sweep can also be a selective event where multiple alleles at a locus are simultaneously favored and increase in frequency. Typically, soft sweeps result in a smaller reduction of genetic diversity compared to hard sweeps. More information on soft sweeps can be found [here](http://www.genetics.org/content/210/4/1429).

The class then briefly discussed the Hancock paper. The authors aimed to identify climate-adaptive loci in the plant *Arabidopsis thaliana*. The authors looked for enrichments of non-synonymous, synonymous, and intergenic SNPs in the 1% tail of the combined climate correlation distributions. SNPs in the 1% tail of the distribution represent the extremes of the distribution and are likely explained by selection. They found that non-synonymous SNPs were significantly enriched among the loci correlated with climate (overall and individual variables), indicating that they detected adaptive alleles. A common garden experiment was used to predict the relative fitness among a set of geographically diverse *A. thaliana* accessions. Finally, they found that hard selective sweeps were prevalent in this system. The rapid fixation of new mutations may limit the adaptability of this species to rapid climate change.

*Non-synonymous and synonymous SNPs* - Synonymous SNPs are SNPs that have different alleles that encode for the same amino acid. Non-synonymous SNPs are SNPs that have different alleles that encode different amino acids. Non-synonymous SNPs are associated with adaptation and synonymous SNPs are associated with neutral variation. More information on SNP classifications can be found [here](https://www.ncbi.nlm.nih.gov/books/NBK44488/).