**Rationale:**

* Describe the contemporary neutral genetic structure along the Fundulus cline
* Test the different recent demographic scenarios that explain contemporary patterns of diversity (e.g. use *dadi, fastsimcoal* or *moments* to model different demographics scenarios)
* Use EAA approach to identify the proportion of genomic variation explained by spatial/demographic versus putatively adaptive variation (e.g use dbRDA to parse spatial structure from environmental structure)

**Minor hypotheses/ideas**

* Does larval juvenile transport actually occur and if it does might it help explain the maintenance of the steep genetic cline centered on New Jersey
  + It is true that most fundulus remain close to the natal estuaries, but for those that leave, are they recruited nearby or swept into the ocean leading to large dispersal distances for a small subset of individuals? If migration according to the latter scenario occurs at a high enough rate, ocean currents may predict gene flow to a greater extent than linear distance along the coast. This is a readily testable hypothesis: (a) test for migrants within sampled populations (b) model the genetic variation explained by linear distance compared to connectivity by current mediated dispersal and secondary contact of divergent clades

**ANGSD Lit Review**

(Bay *et al.* 2019)

Single nucleotide polymorphisms (SNPs) were identified using ANGSD (Korneliussen et al. 2014) with initial parameters: minor allele frequency > 0.01, SNP p -value < 1e -6, base quality Q>30, minimum individuals > 18. This individual coverage filter was used only for initial SNP identification and more stringent per -population requirements were used for analysis of introgression

We conducted principal components analysis of SNP genotypes using the PCAngsd (Meisner & Albrechtsen 2018), and used covariance matrices to calculate eigenvectors and generate PCA plots in R version 3.5.0 (R Core Team 2014). Additionally, we used NGSadmix (Skotte et al. 2013) to estimate ancestry proportions for each individual. We ran a range of K values (1 -6) with 20 replicate runs each and determined the optimal value using Accepted Article This article is protected by copyright. All rights reserved. the Evanno method as implemented in CLUMPAK (Kopelman et al. 2015). We also used CLUMPAK to combine runs and identify alternative solutions for each K value. We next identified SNPs fixed between westslope cutthroat and rainbow trout species, first removing individuals from “pure” populations that had admixed ancestry in our admixture analysis. After removing these individuals, we used genotype likelihoods to estimate allele frequencies for each admixed population (Sullivan Creek, AB and Gold Creek, BC) as well as each parental species using ANGSD. Measuring allele frequencies directly from genotype likelihoods has recently been shown to lead to more accurate inference than first estimating individual genotypes (Warmuth & Ellegren 2019).

We next identified SNPs fixed between westslope cutthroat and rainbow trout species, first removing individuals from “pure” populations that had admixed ancestry in our admixture analysis. After removing these individuals, we used genotype likelihoods to estimate allele frequencies for each admixed population (Sullivan Creek, AB and Gold Creek, BC) as well as each parental species using ANGSD. Measuring allele frequencies directly from genotype likelihoods has recently been shown to lead to more accurate inference than first estimating individual genotypes (Warmuth & Ellegren 2019).

(Murray *et al.* 2019)

While many analyses require knowledge of exact genotypes for each sample, some methods(e.g. ANGSD; Korneliussen et al., 2014) are able to represent uncertainty in individual genotypesthrough subsequent analyses. Given our low sequencing coverage, individual genotypes may havehigher error than we desire, particularly in detecting heterozygosity. To address these concerns,we used ANGSD (Korneliussen et al., 2014) to detect putative variants, and to calculate genotypelikelihoods at each variable site. ANGSD considered loci only if there were >10 reads at a SNP(summeds across all samples), considered reads only if they had a mapping quality >30, consideredbases within reads only if they had a base quality score >10, and removed variants with a minorallele frequency <2%, with fewer than three reads supporting the alternate allele, or if the p-valueof the likelihood-ratio test of non-zero minor allele frequency (i.e. test of polymorphism) was>10−3. Indel and block-substitution variation is not considered by ANGSD. We used a regionparallel approach similar to that used in variant calling to accelerate this computation. In total,ANGSD detected 55 million polymorphisms across our samples.From ANGSD likelihoods, we calculated several population genetic statistics. A two-dimensionalsite-frequency spectrum (SFS) between all E. albens and E. sideroxylon was calculated with realSFS(Nielsen et al., 2012), then estimated genome-wide FST between E. albens and E. sideroxylon usingthis two-dimensional SFS as a prior (see Supplementary fig. 15). Using ngsDist (Fumagalli etal., 2014), we calculated inter-sample genetic distances for all non-outlier samples. We estimatedinter-sample covariance using PCAngsd (Meisner and Albrechtsen, 2018). We calculated Euclideandistances from PCAngsd covariances using the Gower transformation (Di j = Cii + Cj j − 2Ci j;Gower, 1985).

the above pipeline have been implemented as a generic, modular Snakemake workflow.In particular, the region-parallelisation of variant calling is handled specifically in Snakemake,allowing abstraction of the execution environment. Project and cluster specific configuration ofthis pipeline is separate to pipeline code, allowing easy adaptation to other systems and datasets. Infact, this pipeline has subsequently been used in at least three additional projects (wheat, tomato,and potato population genomics). This pipeline and associated scripts are open source, andavailable online at https://github.com/kdmurray91/euc-dp14-workspace.

(Oh *et al.* 2019)

Long and detailed, go to ms for methods. In general they use AFS estimated in ANGSD as prior to run analyses in NGStools

(Krohn *et al.* 2019)

Same, but a summary is below

Genotype calling: filter first based on coverage and missigness, then run ANGSD

Structure:

* + PCA – NGScovar using genotype posterior probabilities from ANGSD as input
  + IBD – pairwise genetic distance matrix using ngsDist, then plot against geographic distance

EAA: used mantel tests on SNPs called in ANGSD

Differentiation:

* FST: used ANGSD and its built‐in function realSFS to calculate FST. For each population trio, we calculated FST based on the site frequency spectrum of every pairwise combination of populations (three combinations per population trio).

Demographic

* Dadi: called SNPs identically as in our Mantel analyses, but using –doGeno 2 in ANGSD to output allele calls. We converted the allele calls into the ∂a∂i input format using custom Pythonscripts (available at <https://github.com/alexkrohn/LavaFlowLizards/>). Using ∂a∂i, created joint site frequency spectra for each pair of populations

(Hamala *et al.* 2018)

Great supplemental methods that explain all commands used and filtering techniques

(kerth)

Thesis with very pedagogical methods section, helpful for understanding rationale underlying methods. Also has a well maintained git repo for the project with rendered markdown docs! <https://github.com/claudiuskerth/PhDthesis>

Uses ANGSD to construct GLs and SFS, NGScovar for PCA, FST from the realSFS from ANGSD, then dadi using GLs

(Mischa Matz)

Has github repo with for ANGSD and moments: <https://github.com/z0on/AFS-analysis-with-moments/blob/master/momens_scripts_README.txt>

**Tools that use GLs**

Genotype likelihoods / SFS:

ANGSD: has output for dadi and moments from realSFS! See link <https://github.com/ANGSD/angsd/issues/101>

Neutral Structure:

PCA:

* + PCAngsd – might be preferable to other options becomes is more robust to variable coverage than ngscovar (Meisner & Albrechtsen 2018)
  + NGScovar

Admixture:

Demographic modeling:

* Dadi: can use SFS from ANGSD as input
* Fastsimcoal2
* moments

**ANGSD workflow**

Matz example (for input into moments):

1. collect maf information to do filtering and do initial filtering: ANGSD -gl doMajorMinor 1 -doMaf 1 -dosnpstat 1 -dogeno 3 -doPost 2 -minMapQ 30 -minQ 35 -minInd 36 -doHWE 1 -sb\_pval 1e-2 -hetbias\_pval 1e-2 -skipTriallelic 1"
2. extract and index list of sites to make SFS from
3. estimate site frequency likelihoods for each population: to dos: -doSaf 1 -anc $GENOME\_REF -ref $GENOME\_REF
4. generate per-population SFS: realSFS pop0.saf.idx > pop0.sfs
5. convert to input for dadi / moments

**Dadi/Moments lit review**

*Here we wish to understand how these tools have been used in the literature to make sure I understand their use and get familiar with the conventions and limitations. For each paper using dadi/moments to model demography of a non-model organism I will describe the a priori knowledge about the demography and how that leads to the models being tested, a brief review of the methods and how the model fits lead to evolutionary and demographic conclusions.*

*(Le Moan et al. 2016)*

*A priori demography and questions*

* High degree of spatial structure for a pelagic broadcast spawner
* Phenotypic and genetic similarity among coastal populations suggests an coastal-oceanic ecotype pair, these pairs are repeated along the range
* May be evidence of reproductive divergence among ecotypes (different spawning zones)
* What are the past and present patterns of gene flow between the ecotype pairs?
  + Is the present pattern of parallelism driven by repeated primary divergence (i.e. actual parallel evolution) or has there been secondary divergence (spread of locally adapted alleles to other habitats after divergence either by secondary contact or spread of adapted alleles via gene flow across the range)

*Methods*

* SNP filtering: HWE 1%, exclude singletons, one snp per rad locus
* Removed admixed individuals
* Demo models – done separately for population pairs in different regions
  + Strict isolations (SI) – pure parapatry
  + divergence with gene flow – homogeneous across genome (i.e. only provided one model of gene flow for the whole genome)
    - Isolation with migration - always some gene flow
    - Ancient migration – gene flow restricted to period before split, then total isolation
    - Secondary contact – no ancient gene flow, but contemporary gene flow
  + divergence with gene flow – heterogeneous rates of gene flow for different classes of loci to account for neutral and selected loci – same models fit as other three divergence with gene flow models
* dadi run: masked singletons, 20 independent runs of each model, then best fit among models was compared using the AIC

*Results and conclusions*

* The best demo fit was secondary contact with heterogeneous gene flow
* Estimated the proportion of the genome with restricted gene flow (i.e. due to selection against hybrids) vs neutral loci
* Puts a time estimate on secondary contact

(Filatov *et al.* 2016)

*A priori knowledge and questions*

* Altitudinal cline with a stable hybrid zone
* Discordant clines for neutral loci and phenotypic traits, suggesting selection maintains phenotypic variation
* Previous demo inference points to high gene flow and recent divergence but may be due to samping of highly admixed indivuiduals
* Is the incipient species pair due to allopatric speciation and secondary contact, or parapatry?

*Methods*

* Calculated folded 2d sfs of “pure” samples
* Conducted an EDA with a complex model then removed parameters if they did not improve the model using likelihood ratio tests, then did the same for a model that varied gene flow over time to model a secondary contact scenario

*Results*

* best model fit did not include a variable migration rate over time since split, suggesting parapatric speciation and the formation of an in situ hybrid zone
* timing of the split is coincident with emerge of the volcano that created the altitudinal cline

(Ortego *et al.* 2018)

This one gets a brief summary. They use a neutral dataset, filtering outlier loci (bayescan) for the SFS calculation and removed entire populations that demonstrated admixture

**dadi / moments questions**

* How to divide the populations?
  + as two (north and south) excluding NJ
  + as two (north and south), including NJ, but excluding recent hybrids
* how to account for selection
  + use a neutral dataset – this has some issues because so much of the clinal structure is coincident with the pattern expected under selection
  + set two different migration rates for different sets of loci – scripts are here https://github.com/crougeux/Dadi and here http://methodspopgen.com/2017/11/23/scripts-for-incorporating-heterogeneous-gene-flow-in-models-of-divergence-for-dadi-and-abc/

Bay RA, Taylor EB, Schluter D (2019) Parallel introgression and selection on introduced alleles in a native species. *Mol Ecol* **28**, 2802-2813.

Filatov DA, Osborne OG, Papadopulos AS (2016) Demographic history of speciation in a senecio altitudinal hybrid zone on mt. Etna. *Mol Ecol* **25**, 2467-2481.

Hamala T, Mattila TM, Savolainen O (2018) Local adaptation and ecological differentiation under selection, migration, and drift in arabidopsis lyrata. *Evolution*.

kerth <thesis.Pdf>.

Krohn AR, Diepeveen ET, Bi K, Rosenblum EB (2019) Local adaptation does not lead to genome‐wide differentiation in lava flow lizards. *Ecology and Evolution*.

Le Moan A, Gagnaire PA, Bonhomme F (2016) Parallel genetic divergence among coastal-marine ecotype pairs of european anchovy explained by differential introgression after secondary contact. *Mol Ecol* **25**, 3187-3202.

Meisner J, Albrechtsen A (2018) Inferring population structure and admixture proportions in low-depth ngs data. *Genetics* **210**, 719-731.

Murray K, Janes J, Bothwell H*, et al.* (2019).

Oh KP, Aldridge CL, Forbey JS, Dadabay CY, Oyler-McCance SJ (2019) Conservation genomics in the sagebrush sea: Population divergence, demographic history, and local adaptation in sage-grouse (centrocercus spp.). *Genome Biol Evol*.

Ortego J, Gugger PF, Sork VL (2018) Genomic data reveal cryptic lineage diversification and introgression in californian golden cup oaks (section protobalanus). *New Phytol* **218**, 804-818.