Homework #1 (Population Genomics)

Homework assignment #1 – Due by Monday October 07, 2024

We've now completed our first module of coding sessions. Let's review all you've learned!

- *Understanding the format of raw sequencing reads (fastq files) and interpreting their quality (Q-scores)*
- *How to work with variant call format (VCF) files, read them into R and filter them based on different quality metrics*
- *Estimating population genomic diversity in terms of expected heterozygosity within subpopulations (Hs) and across the entire sample (Ht) and estimating levels of genetic differentiation between a-priori defined groups (Fst)*
- *Estimating and visualizing population structure in the form of genetic PCA and Admixture analyses*
- *Testing for selection on specific genomic loci using outlier analyses*

You've accomplished a lot! :)

Along the way, we've learned and discussed how sequencing quality affects various inferences in population genomics, and started to get a picture of the genomic diversity, structure, and selection that may be acting on these *Centaurea* populations in their native range and where they've been introduced.

An important question among biologists studying introduced species is whether hybridization/admixture among populations increases genetic variation in the introduced range and allows those populations to respond to novel selection pressures. The hypothesis that admixture may increase introduction success has been discussed and tested across many different types of taxa (classically reviewed by Ellstrand & Schirenbeck, 2000 and more recently by Pfenning et al., 2016), and tested by many studies, including some older ones by your professor :) (Keller et al., 2010; Keller et al., 2014), and genomics tools are increasingly applied in this space (McGaughran et al., 2024) These studies often find evidence of genetic mixing in PCA/Admixture analyses, increased levels of genome-wide diversity, and evidence of selection on introduced populations, all consistent with the hypothesis of admixture increasing the evolutionary potential of introduced populations. Is this true for *Centaurea* as well?

One important issue is **individual-level missingness** (i.e., for each individual, what % of SNP loci are they missing data?), since this has been shown to affect inference of population structure in PCA (Yi and Latch, 2022), and may affect other aspects of diversity as well. For example, we saw differences among regions in average Hs and the proportion of loci with non-zero Hs. Is this difference in diversity real or an artifact of too much missing data? Similarly, we saw some genetic clustering by region in the PCA and Admixture analyses, but also many individuals near the center of the PCA or with mixed ancestry in the Admixture analysis. Is this real evidence of gene flow through hybridization or artifacts from missingness?

Your first homework assignment is to assess the sensitivity of population diversity and structure to invididual-level missingness.

The basic approach:

1. Go back to the filtering stage on our original VCF file. Keeping all other filters constant, vary the level of individual missingness to evaluate its effects. Recall we originally allowed individuals with 75% missing data. **Choose 2 other thresholds to evaluate, and generate 2 new filtered vcf files** (again, keeping all other filters the same as before).

2. For each new VCF file representing different levels of individual missingness, **estimate** the following:

a. The number of individuals and SNP loci after filtering
b. Genome-wide diversity for each region (Hs) and its standard deviation (SD)
c. The number of loci with Hs=0 vs. Hs>0 for each region
d. Genetic structure using **either** PCA **or** Admixture analysis

3. **Compare** the effect that the 3 different levels of individual missingness (the original 75% threshold, and the 2 new thresholds you choose) have on the above measures of diversity and structure in this *Centaurea* dataset.

4. **Interpret** your findings in light of the hypothesis that hybridization/admixture is increasing diversity in the introduced range of *Centaurea*.

Guidelines and expectations:

- The main text of the write-up should be 2 pages (max) single spaced. Tables, figures, and references can be on separate pages.
- You may collaborate with each other to discuss details and share notes, but your write-up should be done independently and represent your own work.
- Approach the writing as a technical report based on the work you've done to date. That is, write for a scientific audience using appropriate technical language and narrative style, with citations used when referring to methods or making factual assertions.
- Your write-up should include the following essential elements:

Background (1-2 paragraphs):

- A brief description providing context and motivation of the problem we're trying to address with these data, including a clear statement of your objective for this analysis.
- Brief background on the study species, biological samples, library prep, and sequencing strategy (look through the first tutorial for info, plus your notes from class)

Bioinformatics Pipeline (~2 paragraphs):

- Description of the various steps you used for the analysis of the sequencing data. Take it from the initial VCF file you were given up to estimation of population diversity and structure. No need to include things we did in class that aren't used explicitly in your homework analysis unless you think it's relevant. Write this section at a level of detail similar to what we've seen in papers we've read for Discussion. That is, write with enough detail that a capable person with training similar to yours could reproduce your work. No need to include minutia (e.g., "First, I logged into the VACC. Then I opened up RStudio, ...").
- This section should demonstrate both your technical knowledge of the flow of the different steps in the pipeline, and your level of proficiency in understanding why each step was done.

Results (1-2 paragraphs)

- Report your findings from the different analysis steps. Use a combination of reporting results in-line in your text and summarizing more detailed information in tables and/or figures.
- You may use a **max of 3 tables/figures total** (not counted towards the page limit). These should be placed at the end of your document.
- Be sure each table/figure has a title, and a very brief legend describing its contents.

Conclusion (1-2 paragraphs)

- Give your conclusion so far from the data: How did your levels individual-level missingness affect your results? What can we conclude at this point about evidence for/against the admixture hypothesis and invasive species?
- Discuss any caveats or uncertainties that should be considered when interpreting your results.
- Discuss any methodological challenges encountered along the way that are relevant to your results and their interpretation.
- Discuss opportunities for future directions.

References (listed on a separate page)

- You need to cite peer-reviewed literature in your homework. Somewhere between 3-5 references is expected, not including citations of R packages.
- In addition to your 3-5 references, be sure to cite any R packages you use in your analysis (citations provided at the bottom of this document)
- Only include references to papers you cite in your text!
- Cite papers in APA format. Example:

Lachmuth, S., Molofsky, J., Milbrath, L., Suda, J., & Keller, S. R. (2019). Associations between genomic ancestry, genome size and capitula morphology in the invasive meadow knapweed hybrid complex (Centaurea× moncktonii) in eastern North America. AoB Plants, 11(5), plz055.

Github: Have your github lab notebook and scripts up to date.

- For your homework, combine your analyses into a single script named `homework1_yourlastname.r` and save it to your `docs/` folder on github.
- Also be sure your in-class scripts are available in your github `docs/` folder inside your `population_genomics/` folder.
- Your lab notebook should also be up to date

Submission and deadlines:

- Due by Monday Oct 07, 2024 (end of day)

- Upload your homework as a Word doc or pdf to Brightspace

Extra Credit! (10 pts possible; NTE 100 pts total)

- Include a selection analysis using `PCAdapt` in your evaluation of the effect of individual missingness.
- For this, you'd want to run your selection analysis as we did in class and determine the number of significant outliers on each PC axis tested
- Significance can be determined from the `pvalues` slot in the results object that `PCAdapt` returns. Best practice would be to adjust those p-values for multiple testing using the "p.adjust" function in `PCAdapt` and base the number of significant loci on the adjusted p-values. Like so:

```
padj = p.adjust(pcadapt.pca$pvalues, method="BH")
outliers = which(padj < 0.05)
length(outliers)
```

Reach out via email, office hours, or schedule an appointment if you get stuck, and don't forget to have fun with it!

R package citations:

`vcfR` – *For reading/writing VCF files and calculating genetic diversity and differentiation*

Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. Molecular ecology resources, 17(1), 44-53.

`SNPfiltR` – *For filtering and thinning VCF files*

DeRaad, D. A. (2022). SNPfiltR: an R package for interactive and reproducible SNP filtering. Molecular Ecology Resources, 22(6), 2443-2453.

`LEA` – *for doing PCA and Admixture ('snmf') analyses*

Gain, C., & François, O. (2021). LEA 3: Factor models in population genetics and ecological genomics with R. Molecular Ecology Resources, 21(8), 2738-2748.

`PCAdapt` *(only if you choose to do selection analysis for extra credit)*

Privé, F., Luu, K., Vilhjálmsson, B. J., & Blum, M. G. (2020). Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. Molecular Biology and Evolution, 37(7), 2153-2154.

`tidyverse` (which includes `ggplot2`) – *for general data wrangling/plotting*

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. Journal of open source software, 4(43), 1686.