Relatedness of subpopulations of an alpine butterfly in Grand Teton National Park
Audrey McCombs
EEOB 563, Spring 2021
April 27, 2021

Abstract

Grand Teton National Park supports subpopulations of *Parnassius clodius* butterflies in approximately 40 meadow habitat sites. Because many of the subpopulations are small and meadow sites tend to be isolated, park managers are concerned about loss of genetic diversity in subpopulations due to genetic drift. This project uses molecular phylogenetic techniques to investigate the genetic relatedness of subpopulations in the park, and provides support for the hypothesis that the larger park population is panmictic, and that smaller subpopulations breed freely with each other despite the seeming isolation of their meadow habitats.

Introduction

Parnassius clodius is a high-altitude butterfly species found in the western United States and Canada. Grand Teton National Park (GTNP) in Wyoming, USA, supports a population of *P. clodius* butterflies in approximately 40 meadow sites, which tend to be surrounded and isolated by forested areas, lakes, and anthropogenic development, and which differ in elevation by as much as 500m, affecting phenology. Meadow habitat for adult *P. clodius* in GTNP tends to support small population sizes: surveys in 2013 found 11 sites with ~25 butterflies observed per survey, and 21 sites with ~3 butterflies observed per survey (unpublished data). Because habitat areas can be isolated, and because some sub-populations seem to be small, loss of genetic diversity in smaller subpopulations is a concern for this species. As a result, information about genetic relatedness among subpopulations is of interest to conservation biologists.

This project uses molecular phylogenetic techniques applied to single nucleotide polymorphism (SNP) data to investigate the relatedness of 23 subpopulations of *P. clodius* in GTNP (Fig. 1: Site map). Because habitat areas are contained within a national park, they are highly managed and not subject to the rapid environmental change experienced by many other areas in the United States over the past century. It seems reasonable to assume, therefore, that genetic change in this population is due primarily to neutral forces such as mutation and drift. The most well-established distance measure for populations was developed by Nei (1972), and accounts for both mutation and drift:

$$D_{Nei}(X,Y) = -\ln\left(\frac{\sum_l \sum_u X_u Y_u}{\sqrt{(\sum_u X_u^2)(\sum_u Y_u^2)}}\right)$$

where X_u and Y_u represent the u^{th} allele frequency at the l^{th} locus in sub-populations X and Y respectively. An alternative model for genetic change was developed by Reynolds (1983) based on the coancestry coefficient θ: the probability that a random pair of genes at the same locus

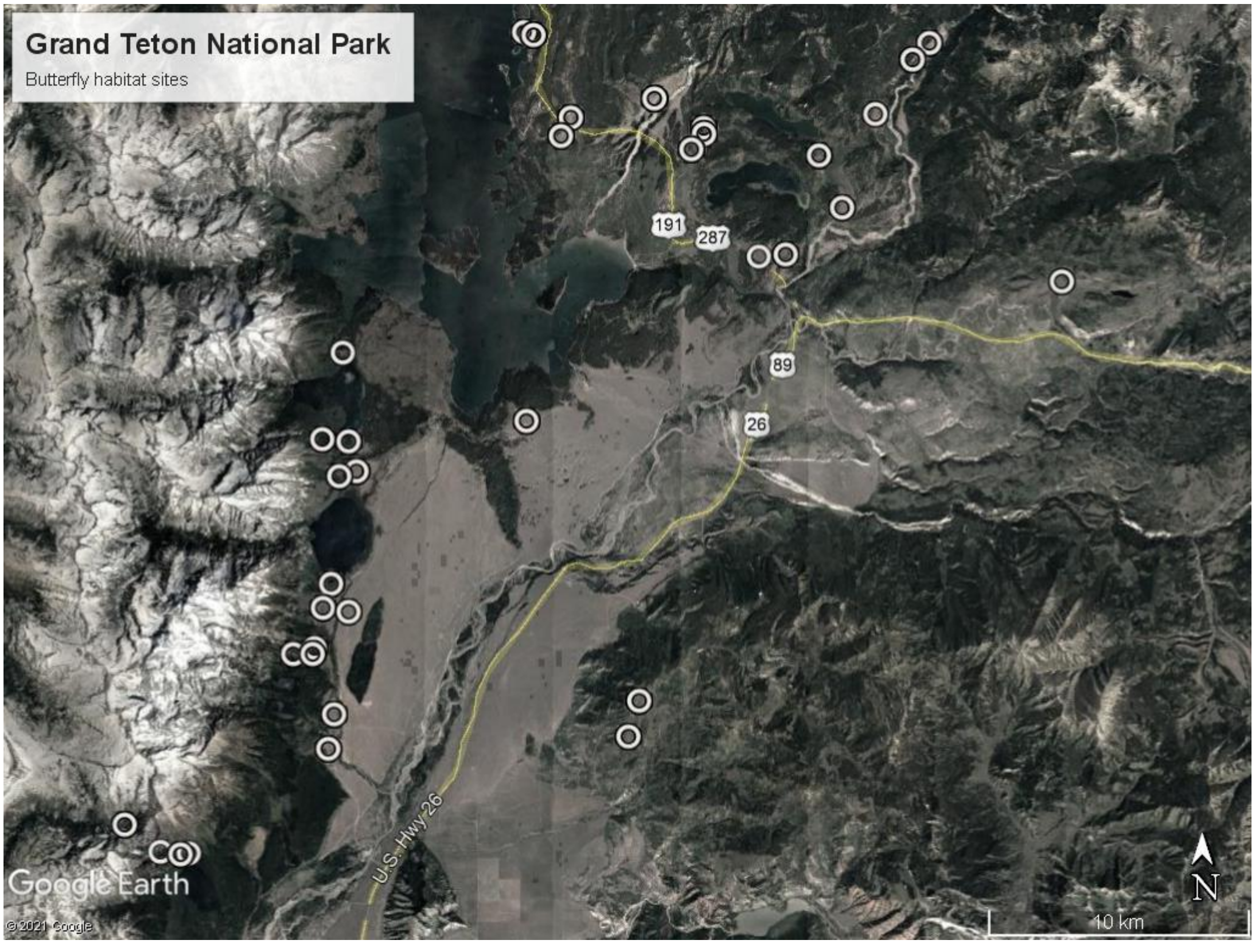


Figure 1: Map of sample sites at Grand Teton National Park

within a randomly chosen population are identical by descent. Reynolds' distance measure based on θ accounts for drift but not mutation:

$$D_{Reynolds}(X,Y) = \sqrt{\frac{\sum_l \sum_u (X_u - Y_u)^2}{2\sum_l (1 - \sum_u X_u Y_u)}}$$

Algorithms for building phylogenetic trees, such as the neighbor-joining and UPGMA algorithms, are essentially dependent on the distance matrices from which they are built; getting the distance matrix right is therefore critical to phylogenetic analysis. The two distance measures described above represent two different models for genetic change in populations—this project will compare trees built from these two distance matrices.

Several hypotheses regarding the relatedness of the 23 *P. clodius* populations present themselves: 1) Panmixia, in which all subpopulations randomly mate with all other subpopulations and all subpopulations are equally related to all others. 2) An island-mainland model, in which one large population serves as a source for smaller sink populations. This model is plausible, as there is one population in GTNP—the Pilgrim Creek population—that seems much larger than the others (observed butterfly counts typically number in the hundreds). The Pilgrim Creek population may therefore serve as a mainland source to the rest of the smaller island sink populations. 3) A stepping stone model, in which subpopulations that are closer together in space are more closely related genetically. And, 4) an isolated-population model, in which most subpopulations interbreed, but a small number of subpopulations are isolated by landscape features or phenology and are therefore more distant genetically from the larger group of interbreeding populations.

Methods

Sample collection, DNA preparation, and genotyping-by-sequencing
In June and July of 2016, we collected 220 samples from 39 unique meadow sites using aerial nets and leg removal (NPS permit GTRE-2016-SCI-0020). DNA was isolated with the Qiagen DNeasy Blood & Tissue Kit (Qiagen, Valencia CA). At least 50ng genomic DNA for each sample was used for library preparation. DNA was digested with the restriction enzyme ApeKI, Illumina adapters and unique barcode adapters were ligated to the digested fragments, and fragments were pooled across individuals and amplified with the Illumina Solexa PCR protocol. Samples were divided among three 96-well plates and sequenced as 100 bp single-end reads, each in one lane of an Illumina HiSeq 2000 (Illumina, San Diego, CA) sequencing system at the University of Wisconsin-Madison Biotechnology Center.

SNP identification
Single-end reads were demultiplexed according to each unique barcode adapter using the process_radtags script from STACKS v1.44 (Catchen et al. 2013). Single nucleotide polymorphisms (SNPs) were filtered in STACKS based on three criteria: 1) loci needed to have at least 8x sequencing depth, 2) loci had to be represented in 70% of all populations, and 3) loci were represented in 30% or more individuals per population. SNP genotypes were output to a VCF file prior to running rxstacks data filtering, resulting in 42,215 SNPs in all 217 individuals. These data were processed in LinkImputeR (Money et al. 2015), which imputed missing genotype data after filtering for minor allele frequency (>0.05), percent missing data at loci (no greater than 70%), and Hardy-Weinberg deviations (p=0.01), producing a complete dataset with >95% accuracy in genotype calls. The complete dataset includes 1,001 SNP loci in 146 individuals.

Phylogenetic analysis
All analyses were conducted in R version 4.0.2. Final pipeline files in .vcf format were imported into R and converted first to a `genind` object, then to a `genpop` object with population annotations using the package **adegenet** (Jombart 2008). Distance measures were calculated using the `dist.genpop` function in the package **adegenet** with options as appropriate for Nei's and Reynolds' distance measures. To test if different distances measures ranked pairwise site distances similarly, I conducted a Kendall concordance test using the `kendall.global` function in **vegan** (Oksanen et al 2019), where the two distance measures were the judges ranking average pairwise distances for each site. The null hypothesis for a Kendall's concordance test is that the ranking of sites by the two judges is no more concordant than would be expected by random chance (Legendre 2009).

Geographic distances between sites were calculated based on latitude and longitude recorded using a Trimble© Geo 7x handheld GPS unit during field work. Geodesic distances were calculated using the `geodist` function in the package **geodist** (Padgham 2021). To test if the distance matrices were correlated with each other, I used a Mantel test from the R package **vegan**. Several papers in recent years have argued that the Mantel test is not appropriate for assessing correlation between two distance matrices (Legendre et al 2015, Guillot & Rousset 2013), however this is still standard practice and therefore I use it here. The null hypothesis for a Mantel test is that the correlation between two matrices is zero (Mantel 1967).

Trees were built using the `nj` function in the **ape** package for unrooted neighbor-joining trees, and `hclust` from **stats** and `as.phylo` from **ape** for UPGMA trees (Paradis & Schliep 2019). Cophenetic distances were calculated using `cophenetic` from **stats**. Finally, bootstrap analyses were conducted with 1,000 replicates on unrooted trees using `boot.phylo` in the **ape** package.

Results

Distance measures

A comparison of pairwise distances between populations using Nei's distance and Reynolds's distance formulas suggest several similarities and some differences. Both distance measures indicate that the populations with the largest pairwise distances are Bearpaw Lake Intersection, Lozier Road, and Buffalo Fork (Fig. 2: Heatmap). Kendall's coefficient of concordance indicates that the ranking of sites based on average pairwise distances is more concordant than would be expected by random chance (W stat = 0.092, p-value = 0.0001 on 9,999 permutations). While the two distance measures rank the sites similarly, the difference between the largest distances and the smallest distances is larger for the Nei's distance calculations than for the Reynolds'. We expect trees based on the Nei's distance matrix, therefore, to differentiate between sites more clearly.

Neither Nei's distance nor Reynolds' distance were associated with geodesic distance between sites, however both measures of genetic distance were closely associated (Fig. 3: Distance comparisons). Mantel test results comparing 3 distance matrices as follows:

- Nei to Reynolds: p-value = 0.0001
- Nei to geodesic: p-value 0.445
- Reynolds to geodesic: p-value = 0.808

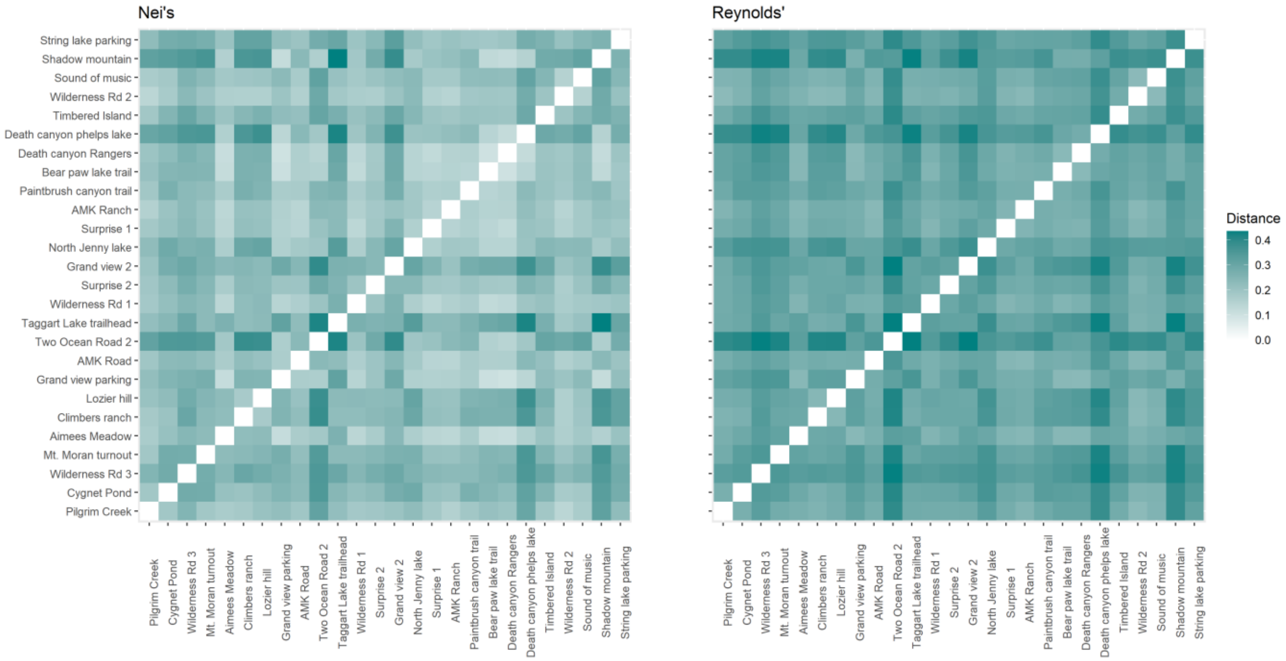


Figure 2: Heatmap of pairwise distances between sites. Left panel: Nei's distance measure. Right panel: Reaynolds' distance measure based on coancestry coefficient.

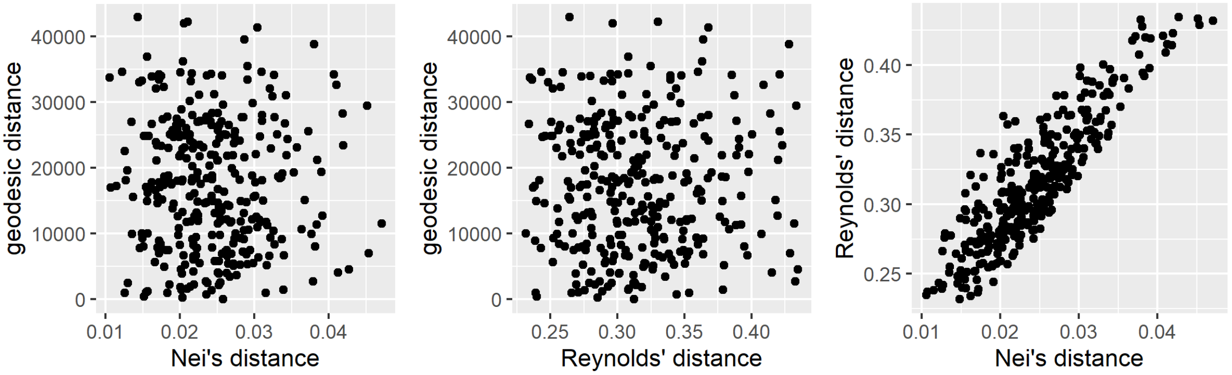


Figure 3: Pairwise plots for Nei's genetic distance, Reynolds' genetic distance, and geodesic distance.

Tree algorithms

Cophenetic plots depict the population distance as calculated in the distance matrix against the tree distance produced by the final dendrogram. Cophenetic plots for the neighbor-joining tree and the UPGMA tree for both Nei's distance and Reynolds' distance indicate that the neighbor-joining trees better capture the distances in the population distance matrices (Fig. 4: Cophenetic plots). The variability in the spread of the plots is slightly larger in the Nei's NJ tree versus the Reynolds' NJ tree, but both UPGMA trees assign similar tree distances to very different population distances. The inability of the UPGMA trees to capture population distances increases as the population distances increase.

This result is not surprising, for two reasons. The UPGMA tree is a rooted tree, and the algorithm is therefore more constrained than the neighbor-joining algorithm that doesn't assume a particular root. The UPGMA procedure also produces a tree in which every leaf is equidistant from the root. Because the NJ tree is not constrained in these ways, the resulting tree has more flexibility to capture the distances in the original distance matrix.

Phylogenetic trees

Because the UPGMA trees were not able to capture the population distances very well, I report only the results of the neighbor-joining trees. Bootstrap support for most nodes on both trees was very weak. For the tree constructed using Nei's distance measure, of the 27 nodes in the dendrogram fully 17 had zero bootstrap support. On 1000 permutations, support for other nodes ranged from 5 to 38. For the tree constructed using Reynolds' distance measure, 13 nodes (of 27) had bootstrap support of zero while support for other nodes ranged from 1 to 38.

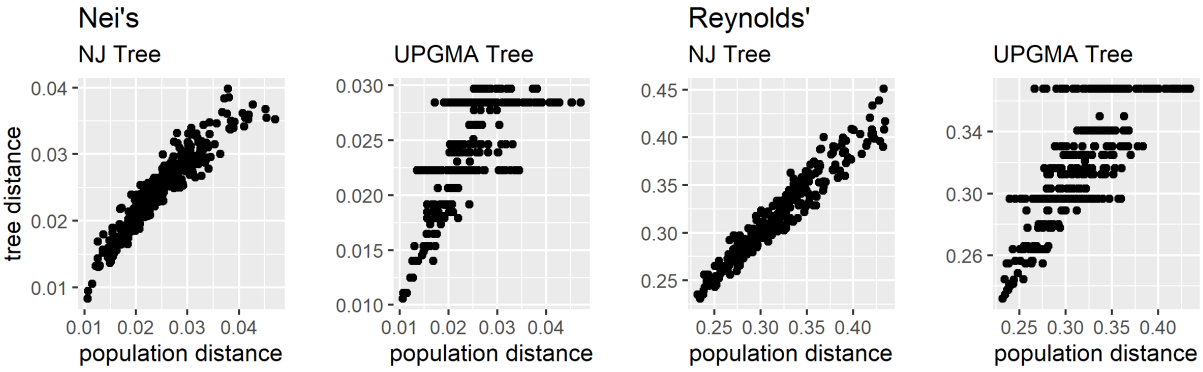


Figure 4: Cophenetic plots of population distance by tree distance.

Keeping in mind that the bootstrap support was very weak for all clades in both trees, for those clades with non-zero bootstrap support five were identified in both trees (eleven total populations), one was identified in the Reynolds's tree but not the Nei's tree (2 populations), and six populations were involved in clades that differed topographically between the trees (Fig. 5: Phylogenetic trees).

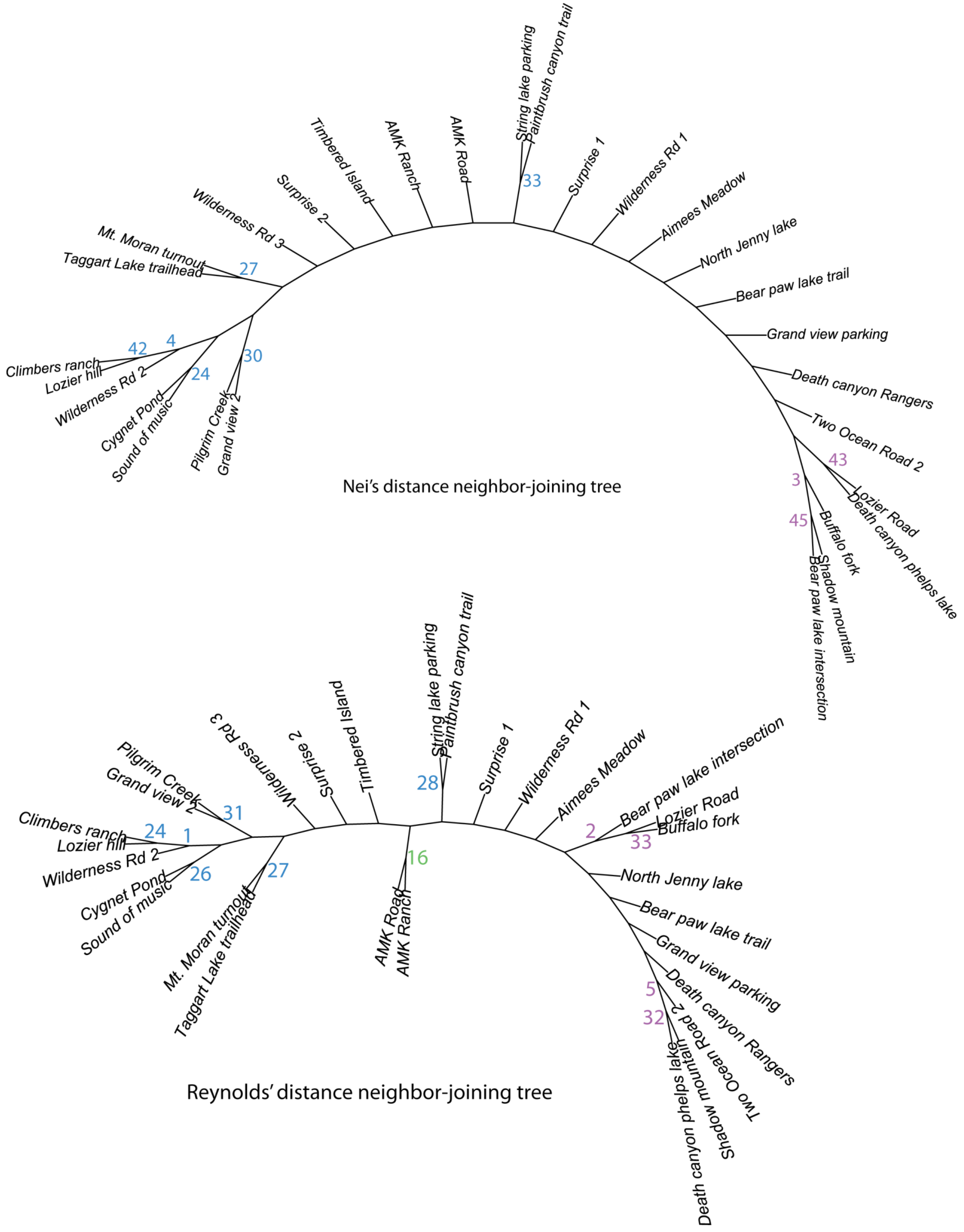
Discussion

The question of interest in this study involves the relatedness of the sub-populations of *P. clodius* in Grand Teton National Park. Of primary concern to NPS managers is the possible loss of genetic diversity that might result from small populations becoming genetically isolated. Several hypotheses capture different possible scenarios for genetic mixing in the park: 1) panmixia, 2) island-mainland, 3) stepping stone, and 4) one or a few large interbreeding population(s) with a few small isolated populations. Because the primary question of interest concerns relatedness but not ancestry, unrooted trees are presented in this analysis, avoiding the need to make a possibly incorrect choice of root for the trees.

The phylogenetic analysis in this study supports the first hypothesis, with perhaps some small support for the fourth. The very weak bootstrap support for all clades suggests that the sub-populations in GTNP are approximately equally related. Both distance matrices identified three populations as being genetically more distant from the others: Bearpaw Lake Intersection, Buffalo Fork, and Lozier Road. That the first two sub-populations are identified as more distant seems reasonable—both of these populations are spatially isolated from the main park area, although there are other populations more isolated (e.g., Death Canyon, Sound of Music, and Shadow Mountain). Lozier Hill, however, is centrally located among the meadow habitats, and the larger genetic distance attributed to this population is more surprising. It is possible that the populations at Bearpaw Lake Intersection and Buffalo Fork are diverging from the main population group, but I would need stronger evidence to support a conclusion that the Lozier Hill population is similarly diverging.

Both models of genetic change—mutation and drift captured by Nei's distance measure, and drift only captured by Reynold's measure—closely agree in the clades they identify. Based on my familiarity with the landscape of GTNP and the meadows that provide habitat for adult *P. clodius* butterflies, some of the identified clades make sense and some do not. For example, the String Lake parking lot is located at the mouth of the canyon up which the Paintbrush Canyon Trail travels. Canyon sides are steep and likely funnel butterflies from the valley up to higher elevations. Climbers Ranch and Lozier Hill, however, are located at the southwest and northeast areas of the park, respectively, and are separated not only by distance but by landscape features such as lakes and forested areas. Analysis results that place these two populations together in a clade are surprising, and warrant further investigation.

In conclusion, this study supports the hypothesis that the population of *P. clodius* butterflies in Grand Teton National Park is not subdivided into smaller sub-populations with limited mating among the subpopulations. Rather, random mixing occurs among adult butterflies that feed and mate in different meadows, and these different meadow populations are approximately equally related to each other.



Nei's distance neighbor-joining tree

Reynolds' distance neighbor-joining tree

Figure 5: Phylogenetic trees of 23 subpopulations of *P. clodius* butterflies in Grand Teton National Park. Numbers at nodes indicate bootstrap support on 1,000 iterations; unlabeled nodes had a bootstrap support of zero. Blue: nodes identified in both trees; green: nodes identified in one tree; purple: nodes identified in both trees but with different topology.

References

Guillot, G. and Rousset, F. 2013. Dismantling the Mantel tests. *Methods in Ecology and Evolution* 4(4), pp.336-344.

Jombart, T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24: 1403-1405.

Legendre, P., Fortin, M.J. and Borcard, D. 2015. Should the Mantel test be used in spatial analysis? *Methods in Ecology and Evolution* 6(11), pp.1239-1247.

Legendre, P. 2005. Species associations: the Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics* 10: 226-245.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. *Cancer research* 27(2 Part 1), pp.209-220.

Nei, M. 1972. Genetic distances between populations. *American Naturalist* 106, 283--292.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. 2019. vegan: Community Ecology Package. R package version 2.5-6.

Padgham, M. and Sumner, M. D. 2021. geodist: Fast, Dependency-Free Geodesic Distance Calculations. R package version 0.0.7.

Paradis E. and Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526-528.

Reynolds, J. B., Weir, B. S., and Cockerham, C. C. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105, 767--779.