Genomic Comparisons among Varieties of a Primrose Species Complex in the Great Basin

by Austin Koontz

Introduction

Populations, and the genetic differences between them, are the causes of, and precursors to, speciation. Understanding the structure and origins of these differences is essential for making sense of species' evolutionary history and is of even greater importance for species of conservation concern. In a conservation context, population differences can carry implications for genetic robustness, establishment of source populations, and other management decisions. The goal of this study is to make such a population comparison between a plant endemic to Utah, Maguire primrose, and its sister varieties.

Maguire primrose (*Primula cusickiana* var. *maguirei*) is an herbaceous, perennial plant located exclusively in Logan Canyon, Utah, USA. The plant was listed as threatened on August 21st, 1985 (Fish and Wildlife Service, 1985). Its status as a rare species with a limited range motivated the need for an increased understanding of its genetic variation, given the greater potential for loss of that variation in small populations. Research in 1997 comparing allozyme marker genes revealed a significant degree of genetic structure between the relatively proximate (about 10 km) lower and upper Logan canyon populations of *P. cusickiana* var. *maguirei* (Wolf and Sinclair, 1997), with the two population groups being nearly fixed for different alleles. This structure was confirmed in 2008 through analysis of the same populations using 165 amplified fragment length polymorphism (AFLP) loci (Bjerregaard and Wolf, 2008); additionally, similar levels of polymorphism were found in most upper and lower canyon populations, suggesting this structure is not the result of a past bottleneck event. The same 2008 study compared crosses between and within upper and lower canyon populations and found that interpopulation crosses generated a higher seed set than crosses within a population (although resulting seeds were not tested for viability). This finding of possible inbreeding depression suggests environmental factors may be driving the population divergence in *maguirei*. Evidence of a fairly small degree of phenological overlap in the blooming period between upper and lower canyon populations (likely caused by significant temperature differences) supports this idea (Bjerregaard and Wolf, 2008, Davidson and Wolf, 2011). The population dynamics of *maguirei* are further affected by heterostyly, in which plants express one of two floral morphologies: "pins", with extended styles and anthers near the base of the corolla, and "thrums", with recessed styles and anthers near the mouth of the corolla. This self-incompatibility system promotes outcrossing between pin and thrum morphs (termed legitimate xenogamy) via insect pollination (Darwin, 1897). Because *maguirei* populations have been shown to be largely distylous, having a pin:thrum morphology ratio of about 1:1 (Davidson and Wolf, 2011), any individual can mate successfully with only half of the entire population, in scenarios of legitimate xenogamy. These reproductive limitations, along with *maguirei* 's patchy population distributions, likely serve as barriers to otherwise advantageous out crossing.

The taxonomic classification of Maguire's primrose has shifted over time. First described as a distinct species in 1936 (Williams, 1936), *P. cusickiana* var. *maguirei* lies within the section *parryi* of the *Primula* genus, along with the other three complex members of what is now the *P. cusickiana* complex: varieties *cusickiana*, *domensis*, and *nevadensis*). Varieties *cusickiana*, *maguirei*, and *nevadensis* were initially differentiated into species based on their different ecological traits (occupying different soil habitats) and geographic ranges, rather than their morphology (which is largely similar; Holmgren and Kelso (2001), and see Figure 3.1). The discovery and publication of *P. domensis* in 1985 (Kass and Welsh, 1985), along with the continued collection of the other varieties, began to cast doubt on the species distinction for each complex member. In 2001, a review of species within the section *parryi* concluded that the morphological differences between *P. maguirei*, *P. cusickiana*, *P. domensis*, and *P. nevadensis* were insufficient for distinguishing each as its own species, and established that these groups instead be distinguished as different varieties of the same species (Holmgren and Kelso, 2001). However, no genetic data was available to justify this taxonomic shift.

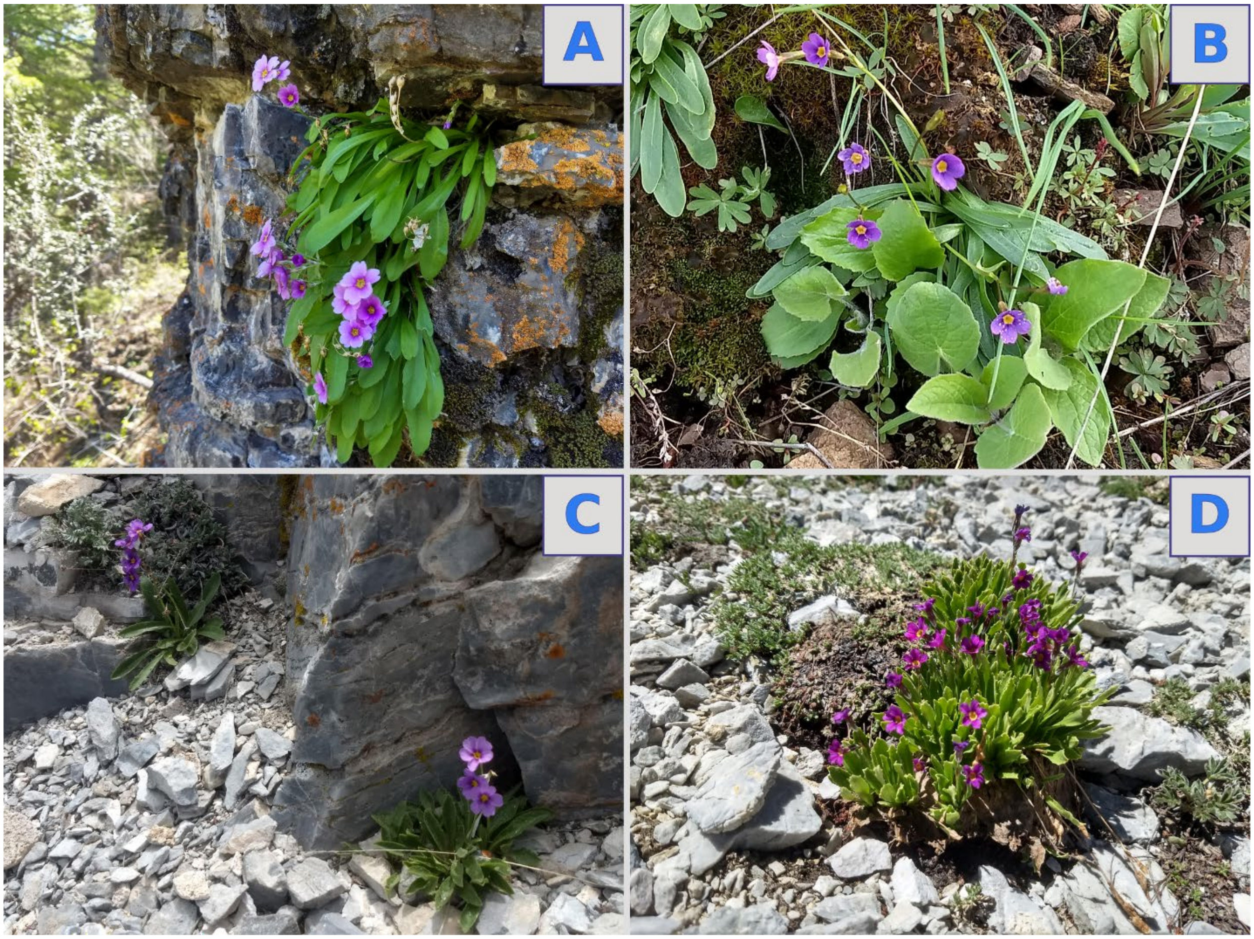

Figure 3.1: The four members (varieties) of the *Primula cusickiana* species complex: (A) *maguirei*, in Right Hand Fork of Logan Canyon, Utah; (B) *cusickiana*, near Cougar Point in Jarbidge, Nevada; (C) *domensis*, at Notch Peak in the House Range, Utah; (D) *nevadensis*, on Mount Washington in Great Basin National Park, Nevada.

A 2009 analysis using AFLPs and chloroplast DNA marker regions of section *parryi* of the *Primula* genus showed the *P. cusickiana* complex being more recently derived compared to other members of the *parryi* group, and supported the complex as monophyletic (Kelso *et al.*, 2009). Relationships within the complex were incongruent, however, with only weak support of a clade containing *nevadensis* and *domensis* being sister to a clade made up of *maguirei* and *cusickiana*. The position of *Primula capillaris*, the Ruby Mountain primrose, was similarly variable, being equally supported as nested within the complex and sister to it. In order to increase resolution and better determine the taxonomy of complex members, the authors suggested a wider analysis utilizing more populations. The restriction-site associated sequencing (RADseq) technologies available today, with their ability to generate reads over many sequence regions of closely related individuals, are well-suited to providing the data required for such an analysis.

We sought to clarify the relatedness of *P. cusickiana* complex members by genotyping all four varieties located at distinct populations scattered throughout the Intermountain West using a RADseq approach. In addition to taxonomic clarification, this analysis may help to explain the level of genetic structure existing between the upper and lower Logan Canyon populations of *P. cusickiana* var. *maguirei*. We hypothesized that one of the *maguirei* populations is more closely related to a population of another variety within the species complex than it is with the neighboring Logan Canyon population, given the genetic distance between the two. Comparing the genetic distances between the *maguirei* populations with those separating the populations of the other species complex members has the potential to provide insights into the biogeographic history of this species complex and could have important implications for conservation and management.

Methods

Sampling

All *Primula cusickiana* species complex samples were gathered in the field. *Primula parryi* samples were also collected in the field, to use as an outgroup. Populations and their respective flowering times were determined using herbarium specimens, and collection sites were selected to maximize the geographic distribution of each variety. At each population location, an individual plant was removed as completely as possible as a voucher specimen. For DNA samples, two leaves from each of five plants were removed and placed in labeled paper envelopes, which were stored on silica crystals to keep samples dry. Vouchers were deposited at the Intermountain Herbarium (UTC); *P. cusickiana* var. *nevadensis* voucher specimens collected from Mt. Washington were additionally deposited at the Great Basin National Park herbarium.

Because past research has shown variable relations between *P. capillaris* and the *P. cusickiana* species complex (Kelso *et al.*, 2009), we also tried to collect *P. capillaris* in the field. However, we were unable to locate any *P. capillaris* individuals in the Ruby Mountains: at one location suggested by past herbaria data, a population of *P. parryi* was found instead. To compensate, two *P. capillaris* samples were sourced from herbaria: one from the Intermountain Herbarium (catalog number UTC00138833), and one from the Arizona State University Vascular Plant Herbarium (catalog number ASU0020421).

Leaf tissue from 96 samples–87 field collections, 2 herbarium specimens of *P. capillaris*, and 7 replicates–were placed into 96 QIGAEN Collection Microtubes (catalog number 19560) and sent to University of Wisconsin-Madison Biotechnology Center, for DNA extraction, library prep, and DNA sequencing. Replicate samples were used to assess the quality of sequencing results, and were distributed across all *P. cusickiana* varieties, as well as *P. parryi*.

DNA Extraction

DNA was extracted using the QIAGEN DNeasy mericon 96 QIAcube HT Kit. DNA was quantified using the Quant-iT PicoGreenR © dsDNA kit (Life Technologies, Grand Island, NY).

Library Prep and Sequencing

Libraries were prepared following Elshire *et al.* (2011) Elshire *et al.* (2011). *ApekI* (New England Biolabs, Ipswich, MA) was used to digest 100 ng of DNA. Following digestion, Illumina adapter barcodes were ligated onto DNA fragments using T4 ligase (New England Biolabs, Ipswich, MA). Size selection was run on a PippinHT (Sage Science, Inc., Beverly, MA) to subset samples down to 300 – 450 bp fragments, after which samples were purified using a SPRI bead cleanu*P.* To generate quantities required for sequencing, adapter-ligated samples were pooled and then amplified, and a post-amplification SPRI bead cleanup step was run to remove adapter dimers. Final library qualities were assessed using the Agilent 2100 Bioanalyzer and High Sensitivity Chip (Agilent Technologies, Inc., Santa Clara, CA), and concentrations were determined using the Qubit © dsDNA HS Assay Kit (Life Technologies, Grand Island, NY). Libraries were sequenced on an Illumina NovaSeq 6000 2x150 S2.

Data Processing

Raw FASTQ data files were demultiplexed and processed using *ipyrad* version 0.9.31 (http://*ipyrad*.readthedocs.io/; Eaton and Overcast (2020)). For all downstream STRUCTURE analyses, SNPs recognized by *ipyrad* were used as the basis for variation between individuals. All *ipyrad* and STRUCTURE parameter files, as well as R scripts used for analysis and data visualization, are provided at https://github.com/akoontz11/*Primula*/tree/master/R.

Complex-Wide Genomic Survey

For our complex-wide genomic survey, we ran *ipyrad* twice: we used the results from our initial run to confirm sequencing consistency for replicate samples, and to identify samples with low coverage (generally, samples with less than 30 loci in the final assembly). For both our initial and secondary runs, demultiplexed sequences were paired and merged, and low quality bases, adapters, and primers were filtered prior to single nucleotide polymorphism (SNP) calling.

For our initial run, reads were clustered (*clust_threshold parameter*) at the default 85% threshold. We specified a minimum sequencing depth (*mindepth_statistical*) of 6, and a minimum number of samples per locus (*min_samples_locus*) of 35. Using the results from this run, we used the script vcf2Jaccard.py (https://github.com/27carol-rowe666/vcf2Jaccard) to compare samples with replicates by calculating the mean Jaccard similarity coefficients between all samples. We found that all replicates matched highly with their corresponding samples: see Figure 3.2.

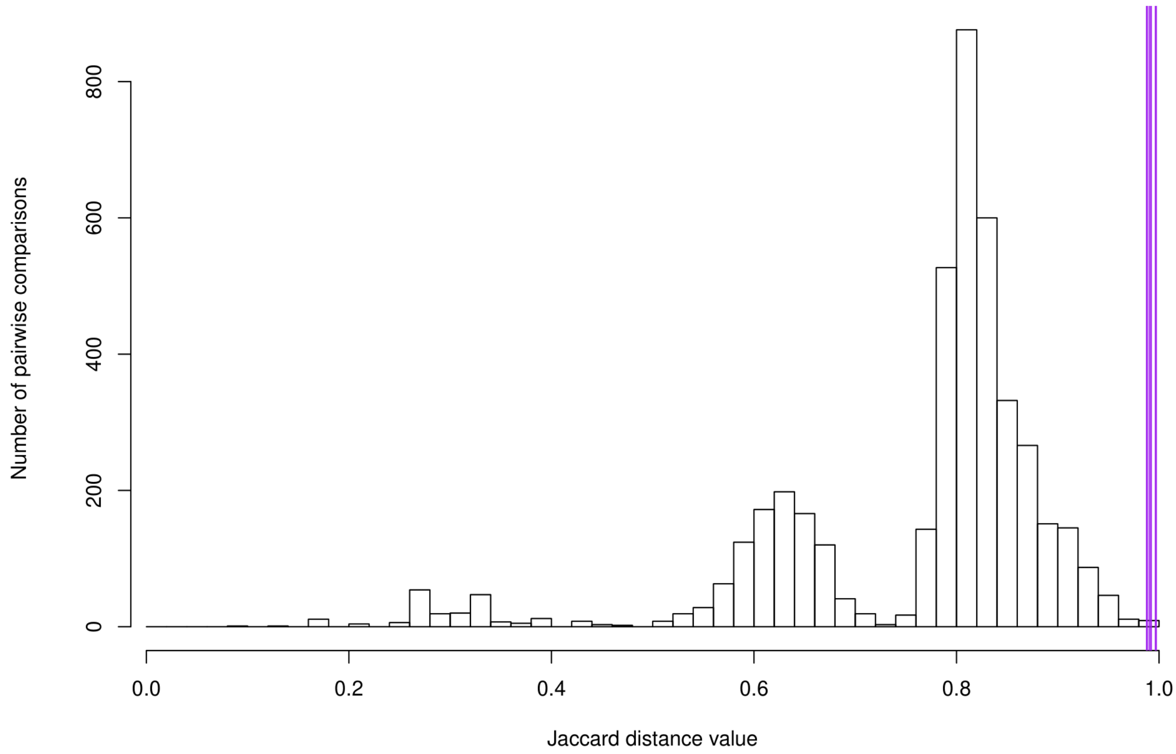

Figure 3.2: Distribution of Jaccard similarity values between samples analyzed. Jaccard similarity values for replicates are shown as purple lines. Similarity coefficient values were calculated based off the SNP matrix of samples generated by our initial *ipyrad* run.

After merging replicates and removing low coverage samples from the dataset, 82 samples remained for our complex-wide analysis. We reran *ipyrad* using these 82 samples to select for loci specific to this subset. For this second run, reads were again clustered at the default 85% threshold; we used a *mindepth_statistical* parameter of 6 and a *min_samples_locus* parameter of 32.

Variety Specific Clustering

In addition to our complex-wide survey, we were interested in exploring within certain varieties, in order to resolve relations at too fine a scale to be captured at the overall complex level. We first analyzed only variety *maguirei* by running *ipyrad* on just the 18 *maguirei* samples used in our complex-wide survey. Reads were again clustered at the default 85% threshold; we used a mindepth_statistical parameter of 6, and a *min_samples_locus* parameter of 5.

We were also interested in relationships in variety *cusickiana*, which has the greatest geographic range of all the species complex members. We ran *ipyrad* on the 24 *cusickiana* samples sourced from 4 populations across the Snake River Plain in Idaho (*cusickiana* samples from Nevada and Oregon were not included). For this run, we used a *clust_threshold* value of 85%, a *mindepth_statistical* parameter of 6, and a *min_samples_locus* parameter of 7.

Population Analyses

To visualize relations between complex members across their geographic range, we used the program STRUCTURE version 2.3 (Pritchard *et al.*, 2000). STRUCTURE uses Bayesian clustering analysis to probabilistically assign individuals to one or more of K populations, where the loci within each population are assumed to be at Hardy-Weinberg equilibrium and linkage equilibrium. For all STRUCTURE runs, we used a burnin length of 50,000, and 100,000 MCMC reps after burnin. For our complex-wide survey, we ran STRUCTURE for K values of 2 – 16, with 50 replicates per K value. For both our *maguirei* and Snake River Plain *cusickiana* analyses, we ran STRUCTURE for K values of 2 – 6, with 50 replicates per K value. We used the CLUMPAK server (http://clumpak.tau.ac.il/; Kopelman *et al.* (2015)) to summarize results across replicates for each K value, and for building STRUCTURE plots.

In addition to STRUCTURE, we analyzed the results of our complex-wide survey using discriminant analysis of principal components (DAPC; Jombart *et al.* (2010)). DAPC is a statistical technique capable of differentiating within-group variation from between-group variation that is designed to accommodate the size of genomic datasets. SNP data is first transformed using a principal components analysis (PCA), and then k-means clustering is run to generate models and likelihoods corresponding to each number of population clusters. Bayesian information criterion (BIC) is then used to determine the best supported number of population clusters. We compared our STRUCTURE results with our results using DAPC and used our DAPC output to visualize population clusters in a PCA format.

Results

P. cusickiana species complex relations

We retrieved, on average, 2.04 x 10 6 reads per sample, and our complex-wide *ipyrad* run identified 1,277 loci that were used in our STRUCTURE analysis. Using the Evanno method for finding the optimal K value (Evanno *et al.*, 2005) found K = 5 to yield the greatest ∆K value; using the method described in the STRUCTURE manual (Pritchard *et al.*, 2000), which identifies the K value with the greatest likelihood, generated an optimal K value of K = 14. We visualized the STRUCTURE results for K values ranging from K = 2 - 16 (Figures 3.3, 3.4, and 3.5). Based on the output of these STRUCTURE plots, we determined K = 7 to be the most biologically relevant number of clusters. At this K value, varieties *domensis* and *maguirei* are clearly delineated, whereas variety *cusickiana* is split into three distinct groups made up of populations from Idaho, Nevada, and Oregon, as are populations of variety *nevadensis*. Higher values of K further emphasized these divisions and did not reveal any more information regarding overall relations within the complex generally.

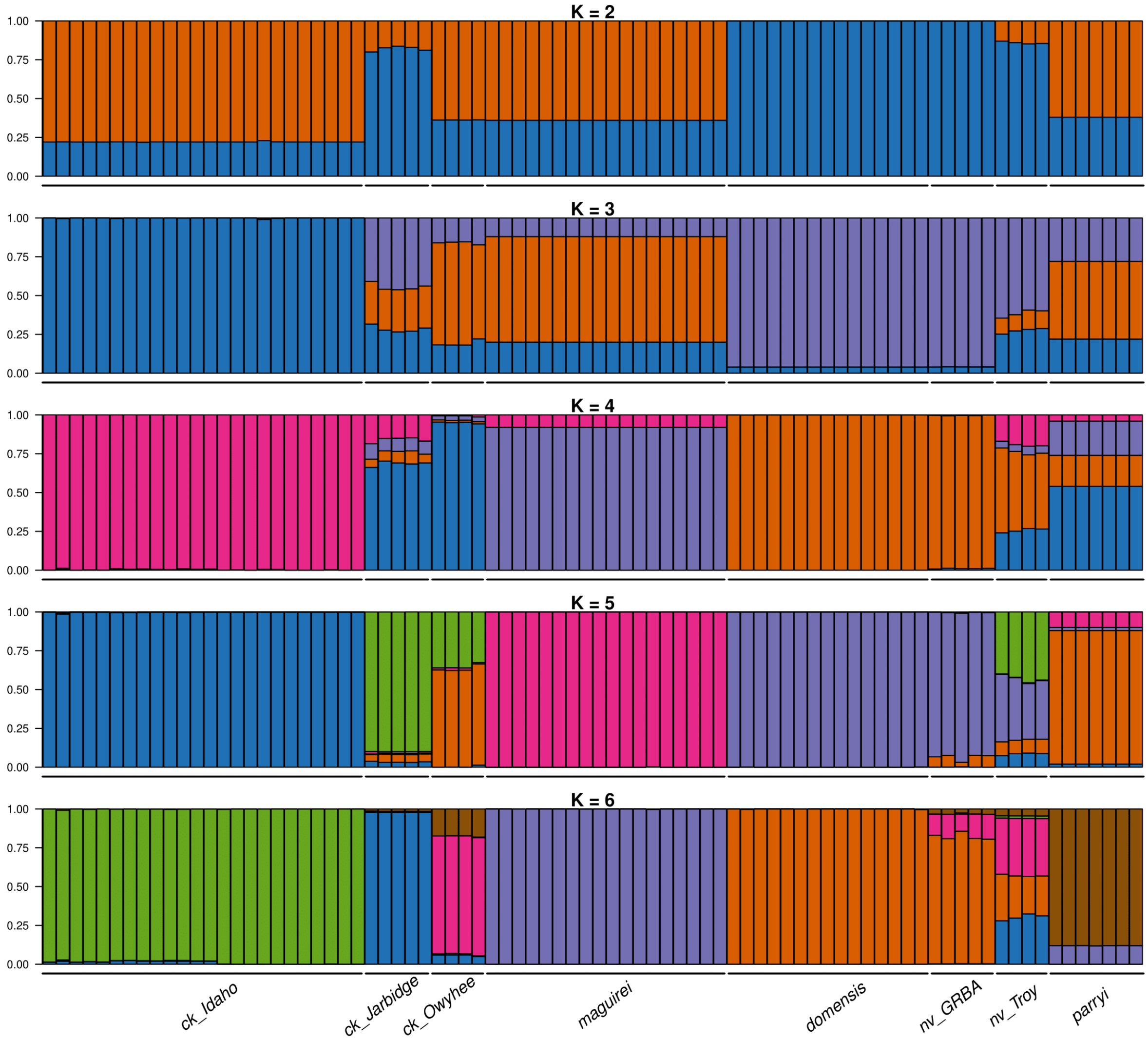

Figure 3.3: STRUCTURE plots for values of K=2 through K=6

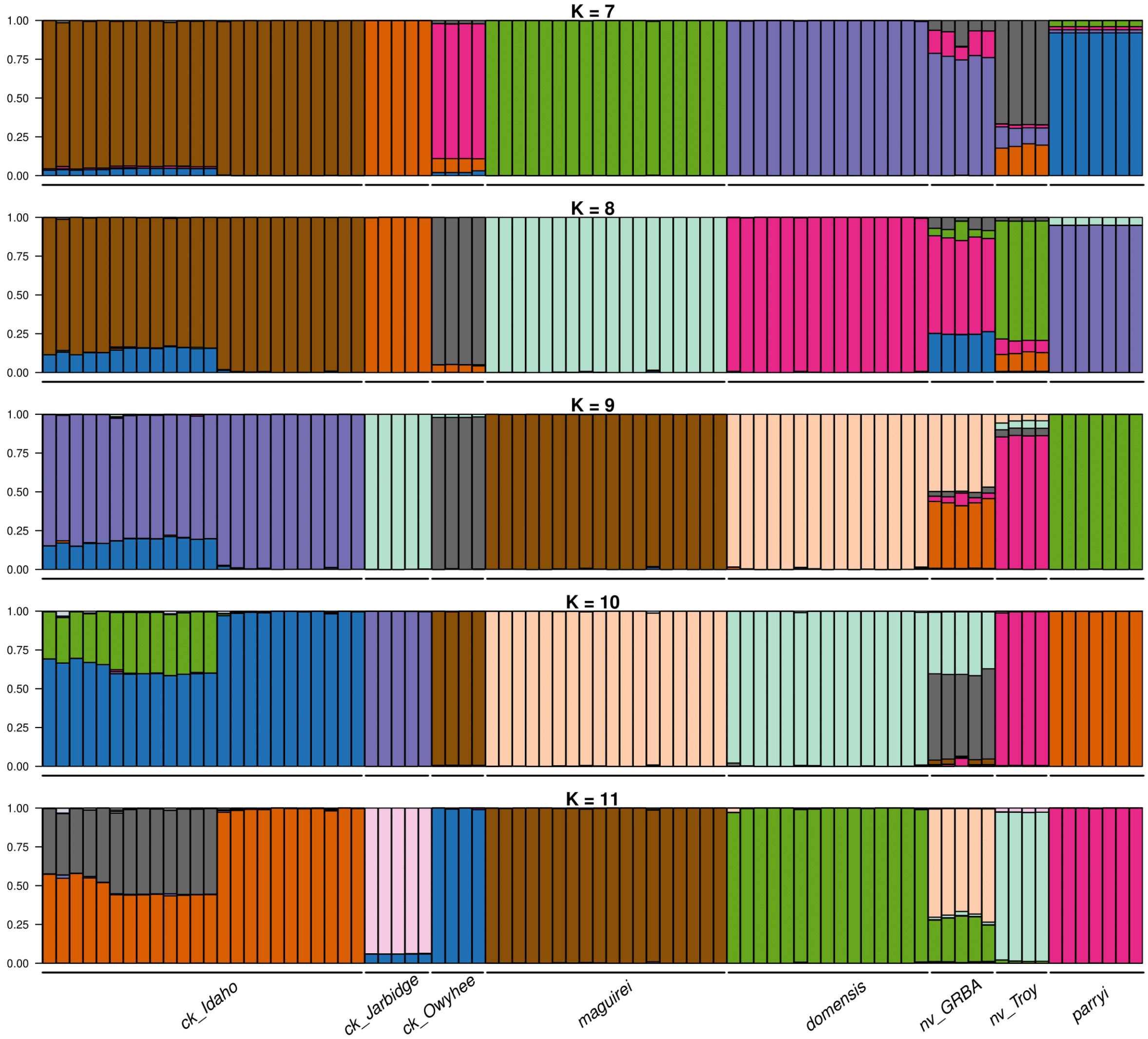

Figure 3.4: STRUCTURE plots for values of K=7 through K=11

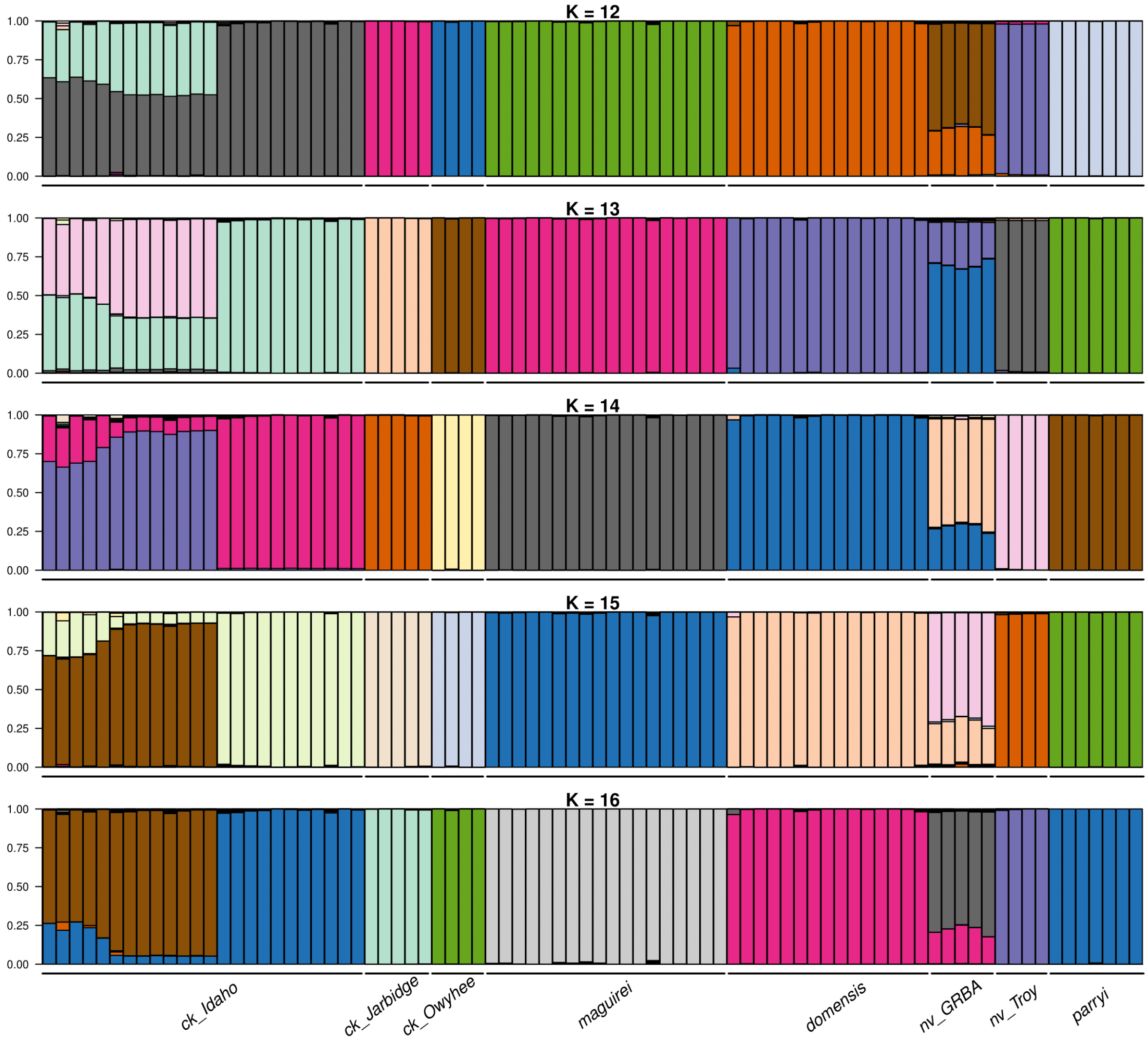

Figure 3.5: STRUCTURE plots for values of K=12 through K=16

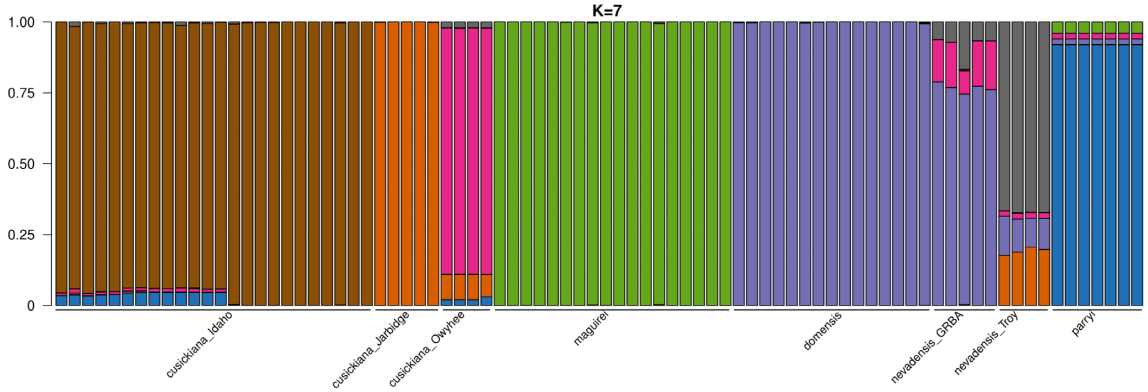

Figure 3.6: STRUCTURE plot at K = 7, with labels representing the populations from which samples were sourced

Our DAPC reported the number of clusters with the lowest BIC value (i.e. the greatest supported number of clusters) was 11 (data not shown). However, several of these clusters were quite small (consisting of only one or two samples), and groupings at 11 clusters poorly reflected on the biology of the species complex. Therefore, we ran our DAPC with a specification of 7 clusters, to examine the distances between groups at this level and compare results to our STRUCTURE output. DAPC results for 7 populations clusters are given in Figure 3.7, and illustrate the distinctiveness of varieties *cusickiana* and *maguirei* and the similarities between varieties *nevadensis* and *domensis*, along with *cusickiana* populations outside of the Snake River Plain. Figure 3.8 provides an image of sampled populations in their geographic contexts, with charts used to illustrate membership to different clusters according to our STRUCTURE analysis.

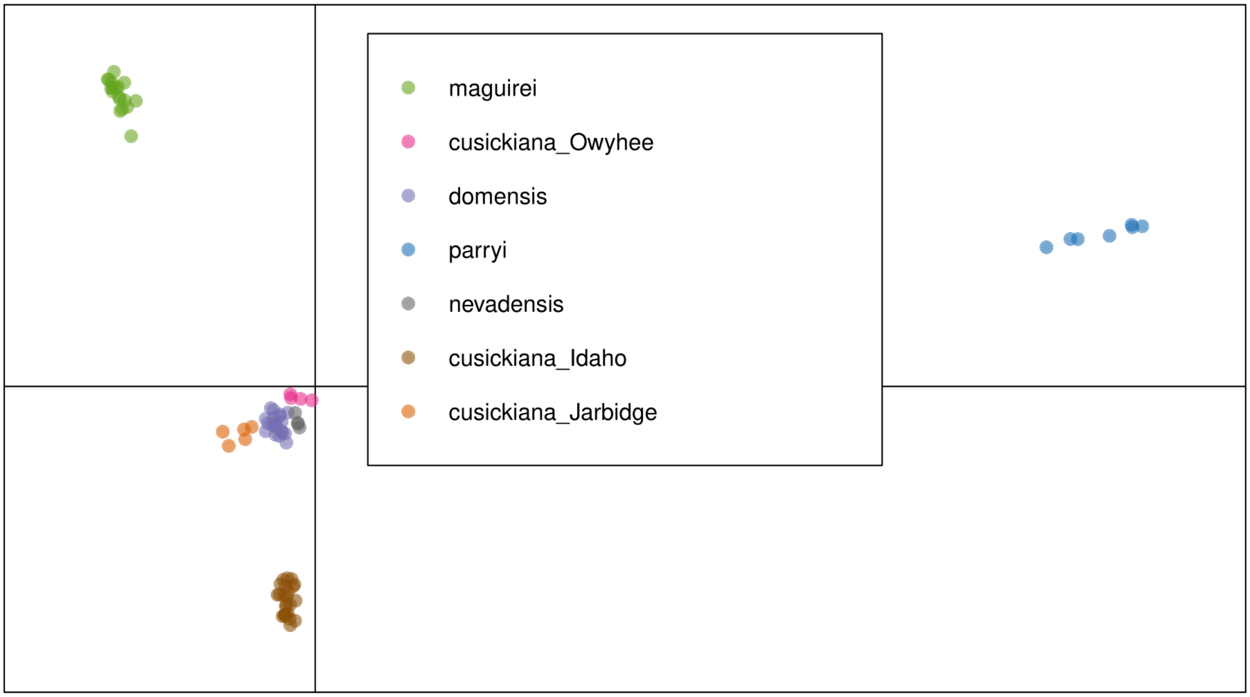

Figure 3.7: Discriminant Analysis of Principal Components (DAPC) illustrating ordination of 7 population clusters. Coloration matches STRUCTURE plot at K=7 and Sample Map. Populations of *domensis*, *nevadensis*, and *cusickiana* populations from Nevada and Oregon are shown to group together using this clustering analysis, while *maguirei* remains distinct.

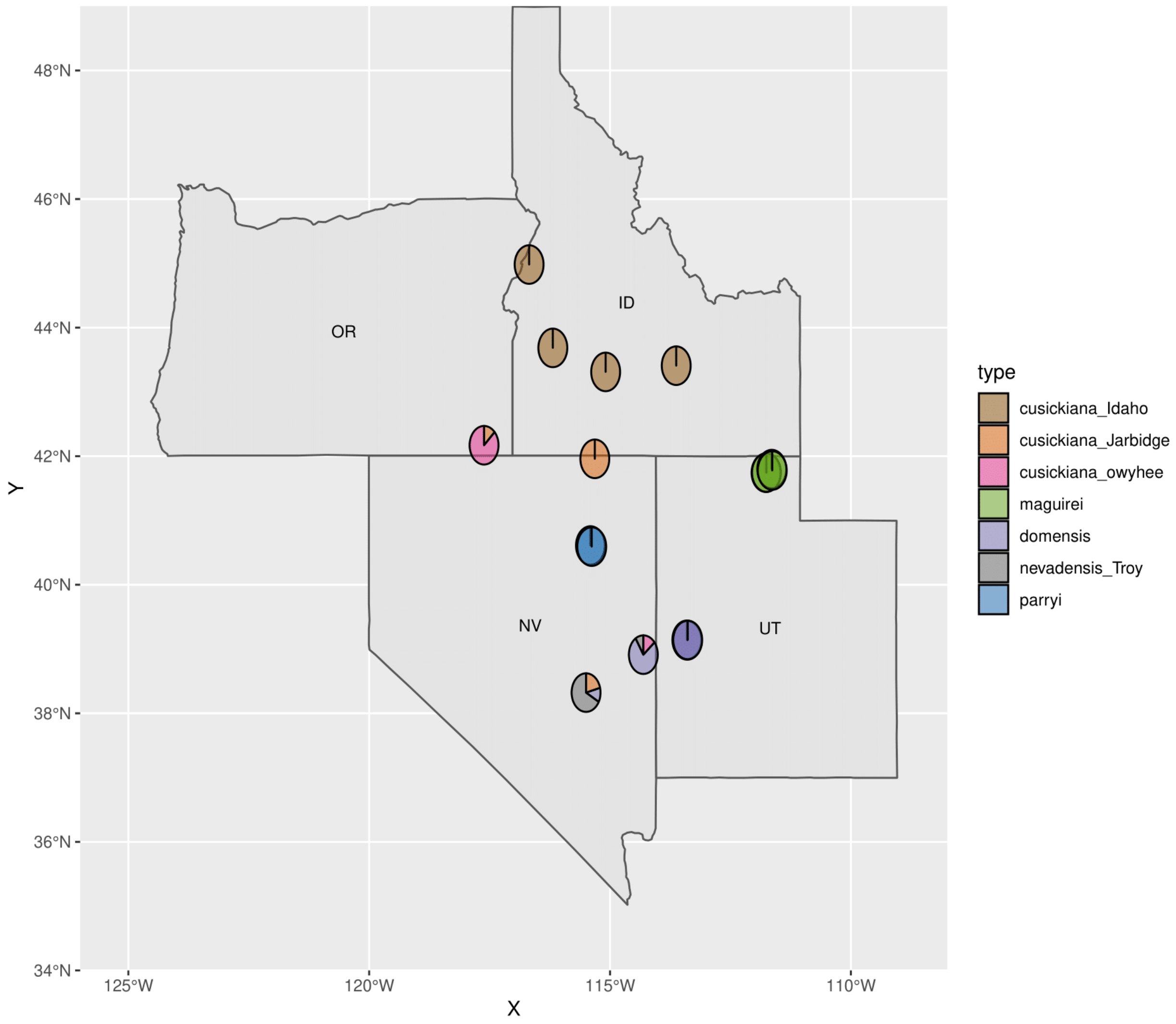

Figure 3.8: Map of sample locations, with pie charts used to represent sample membership to STRUCTURE clusters at K = 7. Coloration matches STRUCTURE plot at K=7 and DAPC.

Overall, our complex-wide analyses found support for several distinct groupings within variety *cusickiana*, notably for populations located in the Owyhee High Desert in Oregon and in Jarbidge, Nevada. Even at very low values for K (i.e. K = 2 - 3), both populations emerge as distinct from the remaining *cusickiana* populations located primarily along the Snake River Plain in Idaho. This motivated our analysis into solely Snake River Plain populations (see *P. cusickiana* var. *cusickiana* section below). Additionally, we found support for the closeness of varieties *domensis* and *nevadensis*, with the latter characterized by differing levels of admixture between the sampled populations. Finally, all of our analyses point to the uniqueness of variety *maguirei* within the species complex.

P. cusickiana var. *maguirei*

In our complex-wide analysis, variety *maguirei* was grouped as a single population cluster, distinct from all other populations of all other varieties. Even at values of K = 16, the upper and lower Logan Canyon populations of *maguirei* were not resolved. Thus, we reject our hypothesis that either Logan Canyon *maguirei* population is more closely related to another population of a different variety.

However, reducing our sample set to only *maguirei* samples allowed us to retain loci informative to this variety but unshared with other complex member populations. Our *maguirei*-only *ipyrad* run generated a STRUCTURE file with 68,492 loci, indicating a large number of loci specific to *maguirei* and not shared with the wider species complex. To speed up processing times, we ran STRUCTURE on a 17,988 loci subset of *maguirei*-specific markers. Using the CLUMPAK server, we found optimal K values of K = 4 (using the Evanno method) and K = 3 (using the likelihood method described in the STRUCTURE manual). Figure 3.9 shows the STRUCTURE plot at K = 3, which resolves similar groupings of *maguirei* populations supported in Bjerregaard and Wolf (2008), and the distinctions between upper and lower canyon populations.

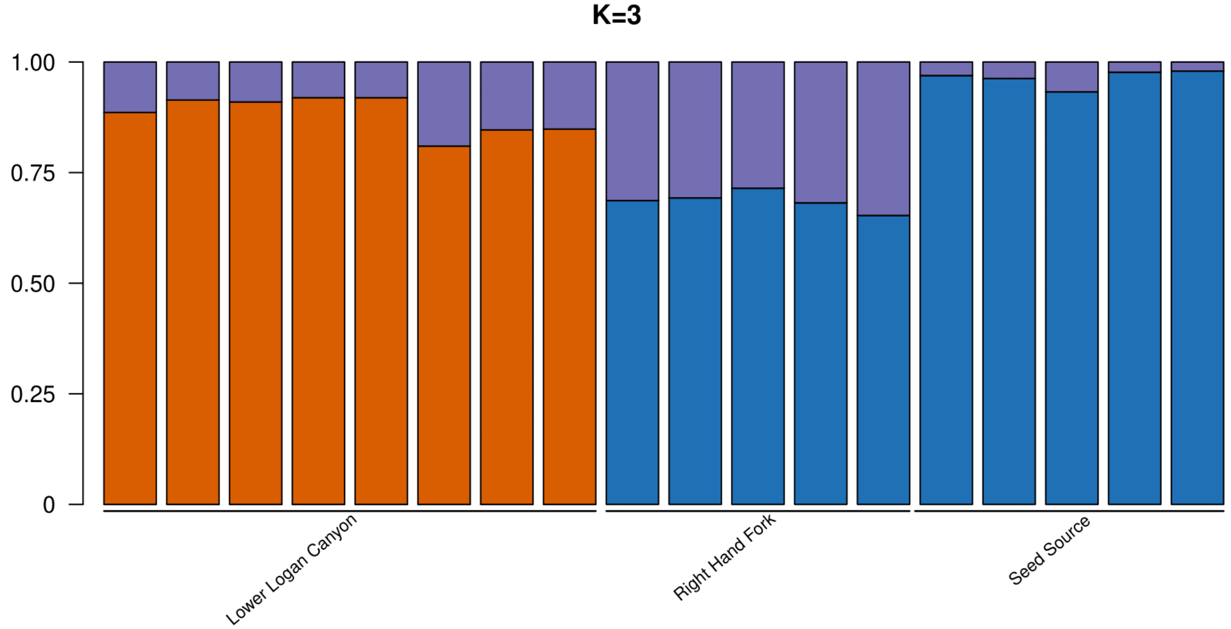

Figure 3.9: STRUCTURE plot for only *maguirei* samples at K = 3.

P. cusickiana var. *cusickiana*

Populations of variety *cusickiana* from Nevada and Oregon clustered separately from *cusickiana* populations along the Snake River Plain in Idaho at relatively low values of K (i.e. K = 4; see Figure 3.3). At higher levels of K, distinctions began to appear between the Snake

River Plain *cusickiana* individuals. We ran *ipyrad* and STRUCTUCRE solely using these *cusickiana* populations to better resolve relations in this group. Our *ipyrad* run reported 38,751 informative loci between Snake River Plain *cusickiana* individuals, again indicating a sizable number of loci not shared with other members of the species complex. We ran STRUCTURE on a subset of 4410 loci for K values ranging from 2 to 6; K = 4 was best supported using the Evanno method, while K = 3 was best supported using the likelihood technique described in the STRUCTURE manual. We visualized K = 3 STRUCTURE plots for these populations (see Figure 3.10) as we felt this level of clustering best represented the biology in this group.

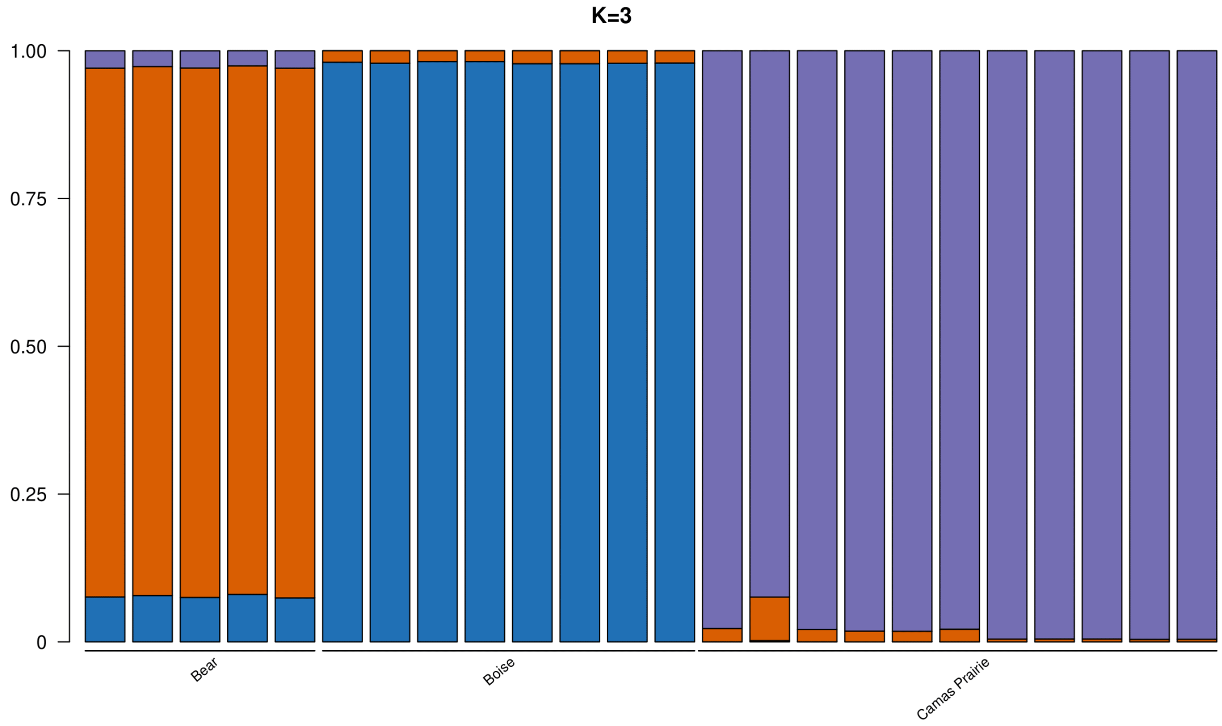

Figure 3.10: STRUCTURE plot for only *cusickiana* samples sourced from the Snake River Plain in Idaho, at K = 3.

Discussion

Primula cusickiana var. *maguirei*

Results from our complex-wide DAPC and STRUCTURE analyses strongly support variety *maguirei* as its own unique variety, separate from others within the species complex. Simultaneously, our *maguirei* -only analyses support previous research revealing notable genetic distance between the upper and lower Logan Canyon populations of this threatened endemic. Taken together, these findings reflect the overall biogeographic trends of this plant. Members of the *P. cusickiana* species complex are generally adapted to the cool, moist habitats prevalent during the Pleistocene. Due to warming across the Great Basin throughout the Holocene, ranges have become restricted to narrow, disjunct pockets, largely in high elevations and on limestone substrates. Research into the population structure of *P. cusickiana* var.

maguirei was originally motivated by its threatened status and identifying potential source populations. Our results suggest that it will be important for future management of *maguirei* to consider and protect both upper and lower canyon populations, given the genetic diversity between them. This is further emphasized by past research showing that inter-population crosses generated a significantly higher seed set than intra-population crosses (Bjerregaard and Wolf (2008))—but more work needs to be done on *maguirei* 's breeding dynamics to confirm these results.

Genetically distinct populations within var. *Cusickiana*

Previous morphological analysis of 76 individuals from four different populations from Southwest Idaho has posited support for dividing then species *cusickiana* within Idaho into three unique species: *P. cusickiana*, P. wilcoxiana, and P. broadheadae (Mansfield (1993). Prior analyses using restriction fragment length polymorphisms (RFLPs) of chloroplast DNA from different *cusickiana* morphotypes from Owyhee and southwestern Idaho populations found insignificant correlation between genetic and geographic distances between morphotypes, but the authors note that this was possibly a product of the genetic markers used (Owen *et al.* (2003)). Our results support the differentiation of Owyhee populations of *cusickiana*, which has been variably classified in the past as *P. wilcoxiana* by Mansfield and others.

Jarbidge populations of *cusickiana* were first discovered in 1995 (New York Botanical Gardens, catalog #801526), and to our knowledge, ours is the first study making any type of genomic comparison including these Jarbidge populations. At all values above K = 5, our results separate these two populations from the rest of *cusickiana* and from each other, although our DAPC results suggest close relations. Future morphological analyses and breeding surveys, along with more precise genetic data and estimates of divergence times, will help to clarify whether these populations stand to be classified as separate varieties or species from the rest of the complex.

Admixture in var. *Nevadensis*

Varieties *nevadensis* and *domensis* are the most recent additions to the *P. cusickiana* species complex, being described in 1967 (Holmgren (1967)) and 1985 (Kass and Welsh (1985)), respectively. The 2009 phylogenetic analysis by Kelso *et al.* of the *Parryi* section of *Primula* found support for grouping these two varieties together, suggesting they may even be considered a single taxon. Our analyses support the closeness of these two varieties, revealing *nevadensis* populations from Mt. Washington, in the Snake Range, to be hybrids of *domensis*, in the House Range to east, and *nevadensis* populations in the Grant Range, to the south. At K = 6, Grant Range populations of *nevadensis* are also characterized by hybridization, with segments coming from *domensis*, Jarbidge populations of *cusickiana*, and Owyhee populations of *cusickiana*, whereas at K = 7, Grant Range populations are less admixed. Minor clusters at each of these K values (as seen in CLUMPAK; not shown) similarly suggest possible admixture with Jarbidge and Owyhee *cusickiana* populations for both *nevadensis* populations. Clearly, more detail is needed to determine the extent and nature of the admixture within *nevadensis* populations.

Primula capillaris

We originally considered *P. capillaris* as an outgroup species candidate with which to compare *P. cusickiana* complex members. However, Kelso *et al.*'s 2009 review described the relation of capillaris to the complex members as “notably dissonant”, with capillaris variably being shown as nested within and sister to the complex varieties. This motivated the inclusion of *P. parryi* in our genomic survey as an outgroup. Unfortunately, our study was unable to clarify the relation of *capillaris* with the rest of the species complex: we were unable to locate any wild populations within the Ruby Mountains, and the two obtained herbarium specimens showed highly variable relations, possibly due to their age (both were sourced from 1965). Future research is needed to determine how *capillaris* fits in with populations of the *P. cusickiana* complex broadly, and to determine potential threats to this narrow endemic.

Conclusion

Our results underline the biogeographic history of the *P. cusickiana* complex and the *Parryi* section of the *Primula* genus generally. The vicariance exhibited by these populations is almost certainly a product of their range contraction since the end of the Pleistocene 2.5 mya, justifying the relatively large genetic distances separating them. Such processes continue to this day, and several of the populations considered in this survey are likely threatened by a continued loss of suitable habitat given future climatic warming. Narrow range endemics—including *P. cusickiana* var. *maguirei*, but also *nevadensis*, *domensis*, and *P. capillaris*—warrant concern of extinction, and more work needs to be done to better understand the breeding limitations faced by each of these taxa and what can be done to ensure their survival in an increasingly more arid Great Basin.