**Refining conservation boundaries of a sentinel stream-breeding frog (*Rana boylii*) using population genomics**

Ryan A. Peek 1,2

Sean O’Rourke 2

Michael R. Miller 2

1Center for Watershed Sciences, University of California, Davis, CA, 95616

2Department of Animal Science, University of California, Davis, CA, 95616

**ABSTRACT**

*Rana boylii* is an imperiled frog species native to California and Oregon, currently petitioned as candidate for Federal (USFWS) and State (CDFW) listing under Endangered Species Acts. As a lotic breeding amphibian, *R. boylii* is tied closely to the local hydrology within the watersheds it inhabits. It is particularly sensitive to alterations to flow regimes. Future conservation of this species requires several crucial components, including establishing higher resolution boundaries for populations across the species range, delineation of distinct population segments that can be utilized in conservation management, and quantification of genetic diversity and isolation within populations. We provide data on over 600 samples from 89 localities across the range of the species, identify six genomically distinct groups, as well as subdivision at local watershed scales. These data are congruent with additional molecular study of *R. boylii*, although we identify the Feather watershed as a distinctive group, at both population and subpopulation scales. Quantification and comparison of genomic variation across populations indicates populations in the southern coast, southern Sierra Nevada, and Northern Sierra/Feather basin in California should have high prioritization in conservation efforts due to patterns of bottlenecking and low genomic diversity.

**Keywords:** *genomics, frogs, rivers*

**INTRODUCTION**

The use of modern genomic sequencing technology has greatly advanced the ability for higher resolution analyses of both geographic and ecological patterns in populations (Hendricks 2018, Barbosa 2018, Nunziata 2017). Restriction site-associated DNA sequencing (RADSeq) (Miller et al. 2007) provides a powerful tool to address ecological genomics questions at scales that were previously impossible using traditional field methods. Furthermore, new methods such as RAD Capture (Rapture) (Ali et al. 2016) adapt RADSeq to target desired loci and allow highly efficient genotyping of thousands of individuals at once. As historical and future landscape use can influence species demography and migration patterns (CITE), these genomic tools will be invaluable for assessing critical factors for long-term persistence in sensitive populations or species.

Freshwater systems and biota are rapidly declining globally (Ricciardi, Rasmussen 1999), and conservation efforts will require assessment of the adaptive capacity of populations to rapid environmental change. For effective conservation management, it is important to establish clear geographic and genomic boundaries for distinct population groups across a species’ range. Delineation of distinct population segments can be utilized in conservation management, and furthermore, quantification of relative genomic health within populations groups can provide additional information as both a benchmark as well as a response metric to restoration actions. Thus, quantifying and linking landscape change with metrics of genetic health may provide an important benchmark to track how sensitive populations respond to future environmental change (through reduced adaptive potential) as well as evaluating whether restoration efforts are effective (i.e., increasing genetic connectivity, diversity, effective breeder/population size).

Amphibians are particularly sensitive to changes in the ecosystem due to their physiology and ontogeny (Davidson et al. 2002; Beebee and Griffiths 2005), thus the ability to utilize environmental variables as life history cues can be especially important. In highly dynamic riverine environments, organisms must constantly adapt to temporal and spatial changes. One such sentinel stream-breeding species is the Foothill yellow-legged frog (*Rana boylii*), a native to California and Oregon which historically occurred in lower elevation (0-1500m) streams and rivers from Southern Oregon to northern Baja California west of the Sierra-Cascade crest [(Stebbins 2003)](https://paperpile.com/c/95Dgjr/SFmh). As a lotic breeding amphibian, *R. boylii* is tied closely to the local hydrology in watersheds it inhabits, and therefore it is particularly sensitive to alterations to flow regimes (Kupferberg, Lind, Yarnell, etc).

As with many amphibians in California (Davidson 2004; Vredenburg et al. 2007; Kupferberg et al. 2012; Peek et al. 2014; Thomson et al. 2016), there have been significant population declines across the former range of this species, particularly in southern California and the Sierra Nevada where it has been extirpated from approximately 50 percent of its historical range [(Jennings and Hayes 1994; Davidson et al. 2002)](https://paperpile.com/c/95Dgjr/ysrB+y9KC). Thus while *Rana boylii* is currently only designated as a species of special concern (CDFW) in the state of CA, it has been petitioned as candidate for listing under the federal (USFWS) Endangered Species Act as well as the state (CDFW) Endangered Species Act.

Future conservation of this species requires several crucial components, including establishing higher resolution watershed-level boundaries for populations across the species range, delineation of distinct population segments that can be utilized in conservation management, and quantification of relative genomic health of these groups (i.e., genomic diversity, population connectivity). A recent study by McCartney-Melstad et al. (2018) identified five major clades in *R. boylii* with strong geographically structured genetic subdivision across its range in California and Oregon. Here we provide a comprehensive analysis of genetic variation across the range of a declining sentinel stream species that is currently a candidate for listing. We provide additional geographic and genetic resolution to McCartney-Melstad et al. (2018), as well as quantify genetic health across subpopulations and clades as both a reference and assessment of the potential for long-term persistence across this species range.

# MATERIALS AND METHODS

## Sampling and DNA extraction

A total of 1103 individual tadpole tail clips, buccal swabs, or tissue samples were compiled, collected between 1992 and 2016 across the range of *R. boylii.* Field sampling was conducted following methods in Heyer et al. (1994) under CDFW SCP Permit #0006881, with IACUC protocol #19327. Individual post-metamorphic frogs were buccal-swabbed following established protocols (Broquet, Berset-Braendli, Emaresi, & Fumagalli, 2007; Goldberg, Kaplan, & Schwalbe, 2003; Pidancier, Miquel, & Miaud, 2003). Each post-metamorphic individual was comprehensively swabbed underneath tongue and cheek for approximately one minute. Swabs were air dried for approximately five minutes and placed in 1.5 mL microcentrifuge tubes while in the field. Samples were stored in the laboratory at -80°C until DNA extraction. Where possible, tail clips from tadpole larvae were collected, and tadpoles greater than 15 mm total length were targeted (Parris et al., 2010; Wilbur & Semlitsch, 1990). One clip was taken per individual tadpole and dried on Whatman filter paper (grade 1) and stored at room temperature. Some older tissue samples consisted of toe clips placed in 100% ethanol for storage, and DNA extraction used Qiagen DNeasy kits following manufacturer protocols. Buccal swabs and tail clip DNA were extracted using an Ampure magnetic bead-based protocol [(Ali et al. 2016)](https://paperpile.com/c/5S37iG/wUFjJ) and stored at -20°C.

## Generating high-quality sequencing data

To produce a high-quality genomic resource for frog species with large genome sizes, we interrogated a significant fraction of the *boylii* genome using a *Sbf1* restriction enzyme and high-density RAD sequencing on an Illumina HiSeq [(Miller et al. 2007; Baird et al. 2008)](https://paperpile.com/c/5S37iG/AhOo0+1HfN). Paired-end sequence data were generated using 24 *R. boylii* individuals (Table S2). RAD libraries were constructed following the protocol described in [(Ali et al. 2016)](https://paperpile.com/c/5S37iG/wUFjJ). *De novo* loci discovery and contig extension were carried out via custom PERL scripts [(Miller et al. 2012)](https://paperpile.com/c/5S37iG/TVJnh), the alignment program Novoalign and the genome assembler PRICE [(Ruby et al. 2013)](https://paperpile.com/c/5S37iG/9I48m). This pipeline resulted in a set of 77,544 RAD contigs ranging from 300 to 800 bp which served as a de novo partial genome reference for all subsequent downstream analyses (File S3: de novo). Using these data, we filtered data to loci with 4 or fewer SNPs, and randomly selected 10,000 loci from this subset. Using these RADSeq data, 8,533 RAD capture baits (120bp) were designed by Arbor Biosciences from the de novo alignment (File S4: Baits). The number of polymorphic loci identified across all *R. boylii* study samples was 44,406. RAPTURE was then used to identify putative high quality SNPs.

Three different sequencing runs on an Illumina HiSeq were merged together, filtered, and duplicates were removed using ANGSD and Samtools [(Li et al. 2009)](https://paperpile.com/c/5S37iG/0VV5). Sampled individuals were aligned against the de novo partial genome reference using the BWA-MEM algorithm [(Li & Durbin 2010; Li 2013)](https://paperpile.com/c/5S37iG/p57MY+cnSB2) and saved to BAM format (*S1-the denovo fa*). To generate SNP (segregating site) data, a probabilistic framework was used for all population genetic analyses as it does not require calling genotypes and is suitable for low-coverage sequencing data [(Korneliussen et al. 2013; Fumagalli et al. 2013)](https://paperpile.com/c/5S37iG/2jPzp+tRndv). Estimates of per site minor allele frequencies (MAF), genotype probabilities and SNP discovery were conducted using ANGSD [(Korneliussen et al. 2014)](https://paperpile.com/c/5S37iG/4U5tA). Genomic sites were designated as polymorphic only if MAFs were greater than 0.05 and the probability of the site not being polymorphic was less than 10-12. ANGSD analyses were conducted following methods from Prince et al (2017), with a minimum mapping quality score (minMapQ) of 10, a minimum base quality score (minQ) of 20, the genotype likelihood model (GL 1), and only sites represented in at least 50% of the included samples (minInd) were used (Li, 2011).

**Quantifying genetic structure**

To characterize and quantify genetic population structure within and among watersheds, we conducted principal component analysis (PCA) using data subsampled to different alignment thresholds (e.g., all individuals had a minimum of 100,000 alignments) to determine the best genomic resolution for population analyses. For downstream analysis, we selected individuals that had greater than 100,000 alignments. To assess population structure and coancestry, ANGSD was used to generate PCA and NGSadmix was used to calculate admixture. Settings used in ANGSD for PCA to identify polymorphic sites included a SNP\_pval of 1e-6, inferring major and minor alleles (doMajorMinor 1), estimating allele frequencies (doMaf 2) (Kim et al., 2011), retaining SNPs with a minor allele frequency of at least 0.05 (minMaf), estimation of genotype posterior probabilities using a uniform prior (doPost 2), specifying the RAPTURE bait locations using the -sites flag, calculate the PCA matrix with the -doIBS 1 and -doCov 1 options, and limiting to higher quality alignment data (-minMapQ 10, -minQ 20). Principal components (PC) summarizing population structure were derived from classic eigenvalue decomposition and were visualized using the ggplot2 package in R (R Core Team, 2017). To assess admixture in *R. boylii*, genotype likelihood data (-GL 2) was generated in ANGSD with the same settings as above, in addition to retaining only SNPs that were shared in at least of 50% of the samples, -doPost 2, -doGLF 2, and limiting to higher quality alignment data (-minMapQ 10, -minQ 20). We then used NGSadmix [(Skotte et al. 2013)](https://paperpile.com/c/5S37iG/amOL) to infer ancestry proportions in *R.* *boylii* individuals. NGSadmix is a robust admixture method that can be applied to low-depth NGS data, and does not require called genotypes, thus reducing error associated with potential ascertainment and uncertainty in the data [(Skotte et al. 2013)](https://paperpile.com/c/5S37iG/amOL).

**Genetic differentiation and diversity estimates**

FYLF are cryptic, and often occur in low densities within the study area. Thus, we retained a minimum of three individuals per site for estimates of diversity and FST. With genomic data, population genetic parameters can be accurately estimated from even low sample numbers (Hotaling et al., 2018), and genomic analyses in non-model organism often use fewer loci (Narum et al., 2013).To quantify genetic variation and differentiation, pairwise population differentiation (FST) was calculated and scaled [mean FST / (1-mean FST)] to quantify genetic differentiation between populations (Rousset, 1997; Wright, 1943, Weir & Cockerham, 1984). FST was estimated by first calculating a folded site frequency spectrum (SFS) for each population from site allele frequencies (SAF) in ANGSD (doSaf 1, fold 1, minMapQ 10, minQ 20, GL 2) and specifying the Rapture bait locations using the -sites flag (Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012). The two-dimensional SFS and global FST between each population pair were then estimated from folded SAF.idx files using a maxIter of 100 with realSFS (Korneliussen et al., 2014). FST statistics were then calculated from two-dimensional SFS (2DSFS) for each possible pairwise combination of unique collection locations using an estimator preferable for small sample sizes implemented in ANGSD (-whichFst 1). These values were plotted in R (R Core Team, 2017).

We summarized patterns of genetic variation using two estimators of θ (4N𝛍): Tajima’s θ (θ𝜋) is based on the average number of pairwise differences (Tajima, 1983), and Watterson’s θ (θS) is based on the number of segregating sites (Watterson, 1975) (see methods). These estimators are influenced by the demographic history of a population and provide information on the trajectory of changes in genetic diversity. When genetic diversity has been stable, these estimates are generally equal; but when genetic diversity has been increasing, θ𝜋 > θS; and when genetic diversity has been decreasing, θS > θ𝜋.

To calculate θ statistics from Rapture data, we used folded SFS in ANGSD with -GL 2, -doThetas 1, -doSaf 1 -fold 1, and -pest. Outputs were used to calculate each statistic for each allelic site using thetaStat with make\_bed and then do\_stat. These data were averaged over every variable allelic site to obtain a single “genome-wide” value for each statistic for each population/locality (Korneliussen et al., 2013).

# RESULTS

A total of 1,103 individual samples were sequenced using Rapture (see Methods). After selecting samples that had greater than 100,000 alignments, 480 individuals from 89 distinct localities across the range of the species were used in PCA and admixture analyses (Figure 1, Table 1). These localities overlap many of the localities used in McCartney-Melstad et al. (2018), with a few notable differences. There were more individuals available for analyses at most of the localities (Table 1), there was higher resolution sampling in certain areas (i.e., the northern coast of California, the Feather watershed), and a few additional localities fall outside of the clades delineated by McCartney-Melstad et al. (2018) (e.g., Locality 1 in the SF American basin in El Dorado County, Locality 4 in the Honey-Eagle Lakes basin in Lassen County).

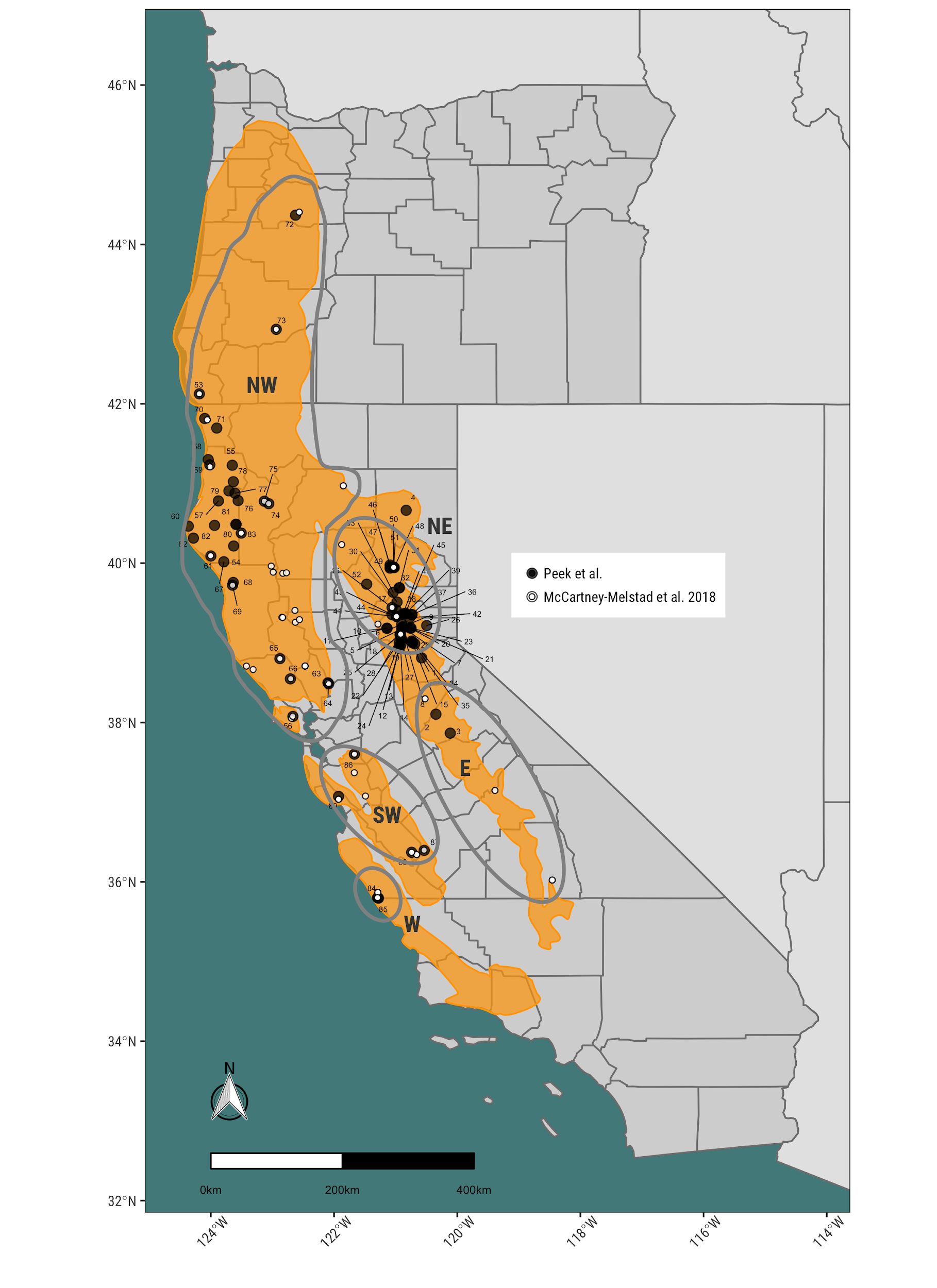


Figure 1. Map comparing sampling localities between recent *R. boylii* study (McCartney-Melstad et al. 2018) and the current study with a simplified historical range of *R. boylii* (adapted from USFS 2011, Thomson et al 2016). Map is annotated with the clades identified by McCartney-Melstad et al. (2018).

## PCA and Admixture shows strong separation between California ecoregions

To compare and assess population structure patterns across the range of *R. boylii*, principal components analysis (PCA) were used to provide a dimensionless comparison of genetic variation. Strong differentiation was primarily observed across four major groups, the Sierra Nevada (comprised of North-East [NE] and East [E] clades, McCartney-Melstad et al 2018); the North Coast (comprised of the North-West [NW] clade); the Central/South Coast (comprised of the South-West [SW] and West [W] clades) and the Feather Basin (previously part of the NE clade) (Figure 2). Additional subdivision was evident in the Central/Southern coastal region (two groups which match SW and W clades), and the Sierra Nevada (two groups, matching the NE and E clades, as well as a smaller group from the Feather Basin), yielding seven total distinct structured groups. Interestingly, a small cluster of individuals from the North Fork (NF) Feather River consistently clustered along intermediate axes between the North Coast and Northern Sierra/Southern Sierra groups. These individuals were consistently intergrades between the larger Feather group and the Northern Coastal group, regardless of the PC axes.

A single sampling locality from Lassen National Forest near Eagle Lake (Locality 4, Table 1) clustered with the Northern Sierra group, suggesting the geographic boundary for this genetic group should be extended (Figure 2). Furthermore, individuals from the South Fork (SF) American basin clustered with the Southern Sierra (East) group, which would extend the boundary delineated by McCartney-Melstad et al. (2018). To assess how strongly the Feather and Southern/Central coastal samples may have affected PCA structure, we ran several additional separate *post-hoc* PCA analyses where all Feather samples were excluded, and all Southern/Central coastal samples were excluded. Significant genetic structure was evident, and patterns remained consistent with those observed when all samples were used, regardless of which samples were excluded or retained. We conclude there is significant congruence with the McCartney-Melstad et al. (2018) genetic clades, but an additional group appears distinct from the original five delineated, the Feather watershed, and boundaries should be expanded to include the SF American basin in the Southern Sierra, and the Honey-Eagle Lakes basin in the Northern Sierra.

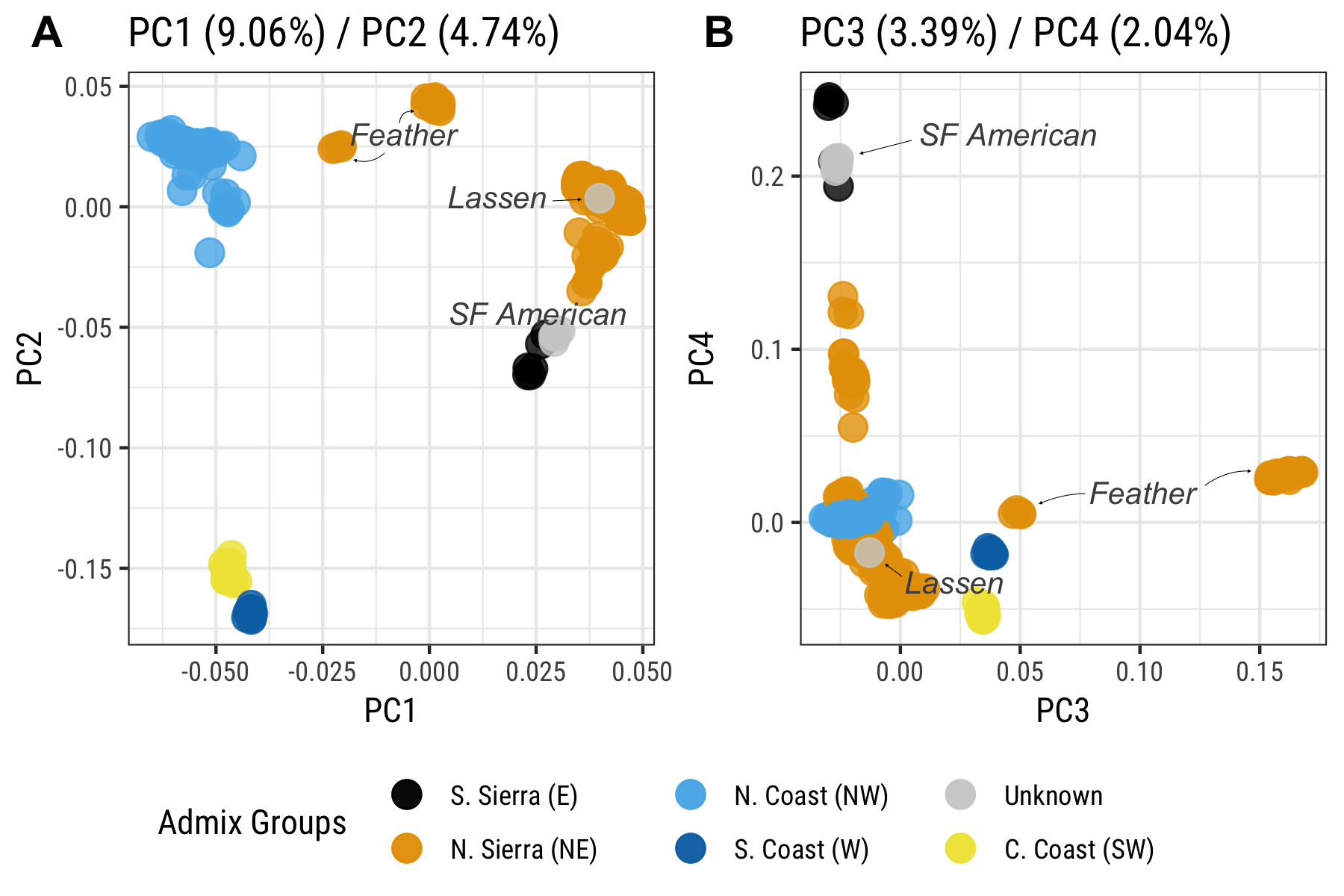
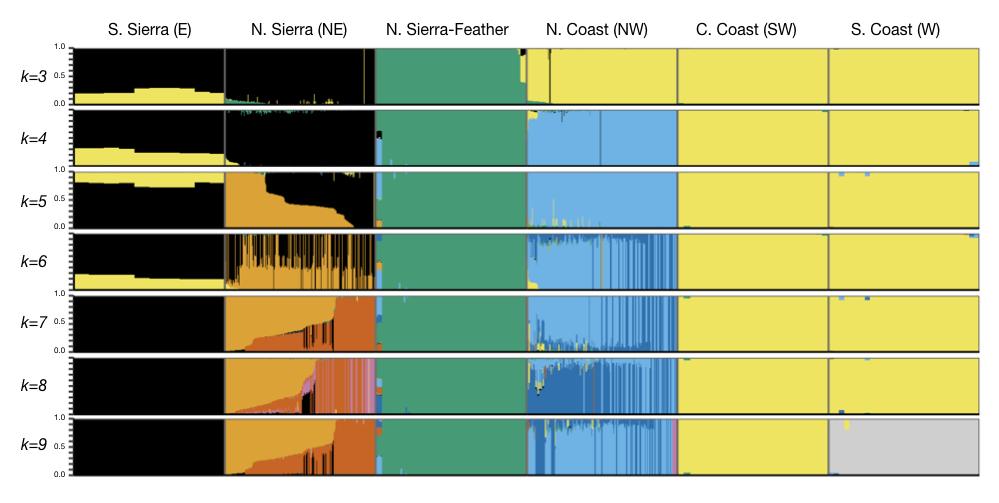


Figure 2. Principal component analysis of Rapture data colored by genetic groups (in parentheses) from McCartney-Melstad et al. (2018). A) PC1 vs. PC2; B) PC3 vs. PC4. Note, samples from South Fork American River (gray) had not previously been studied, and thus had no assigned genetic grouping. *R. boylii* in Feather basin were assigned as part of the Northern Sierra (North-East) clade by McCartney-Melstad et al. 2018.

To further evaluate population structure across the range of *R. boylii*, we used NGSAdmix estimate the proportion of coancestry among individuals (admixture) from genome-wide SNPs [(Skotte et al. 2013)](https://paperpile.com/c/5S37iG/amOL). We used a range of k-values to evaluate the number of potential groups based on genetic ancestry and found strong patterns of divergence which largely matched the patterns observed in the PCA. However, there were several notable differences. While the northern and central coastal (SW and W) groups showed strong divergence from all other groups, they did not separate into distinct groups until k=9 (Figure 3). At k=3 through k=8, the N. Sierra and N. Coast groups continued to show patterns of subdivision, while other groups like the N. Sierra-Feather remained strongly distinct with little or no admixture observed. At k=9, the Central Coast and South Coast split into distinct groups, with additional patterns of substructure and admixture observed in the N. Sierrae (NE) and the N. Coast (NW) groups (Figure 3). While the Central/Southern Coastal populations are clearly distinct from all other clades, our NGSadmix analysis of our data shows less support for delineating Central (SW) and Southern (W) groups as distinct. Figure 3. NGSAdmixture for k=3 through k=9 for *R. boylii* across major geographic groupings (adapted from McCartney-Melstad et al. 2018). The two Central/Southern Coastal groups did not differentiate until a k=9.

Sampling localities were then updated with genetic groups identified from admixture and PCA analyses to delineate geographic boundaries for these clades, building on McCartney-Melstad et al. (2018) (Figure 4). In summary, we identified *R. boylii* from the Feather Basin as a unique genetic group, the Southern Sierra (East) clade should extend to include the South Fork American basin in El Dorado County (Locality 1), and the Northern Sierra (North-East) clade should be expanded to include Honey-Eagle Lakes basin in Lassen County (Locality 4, see Figure 1 and Figure 4).

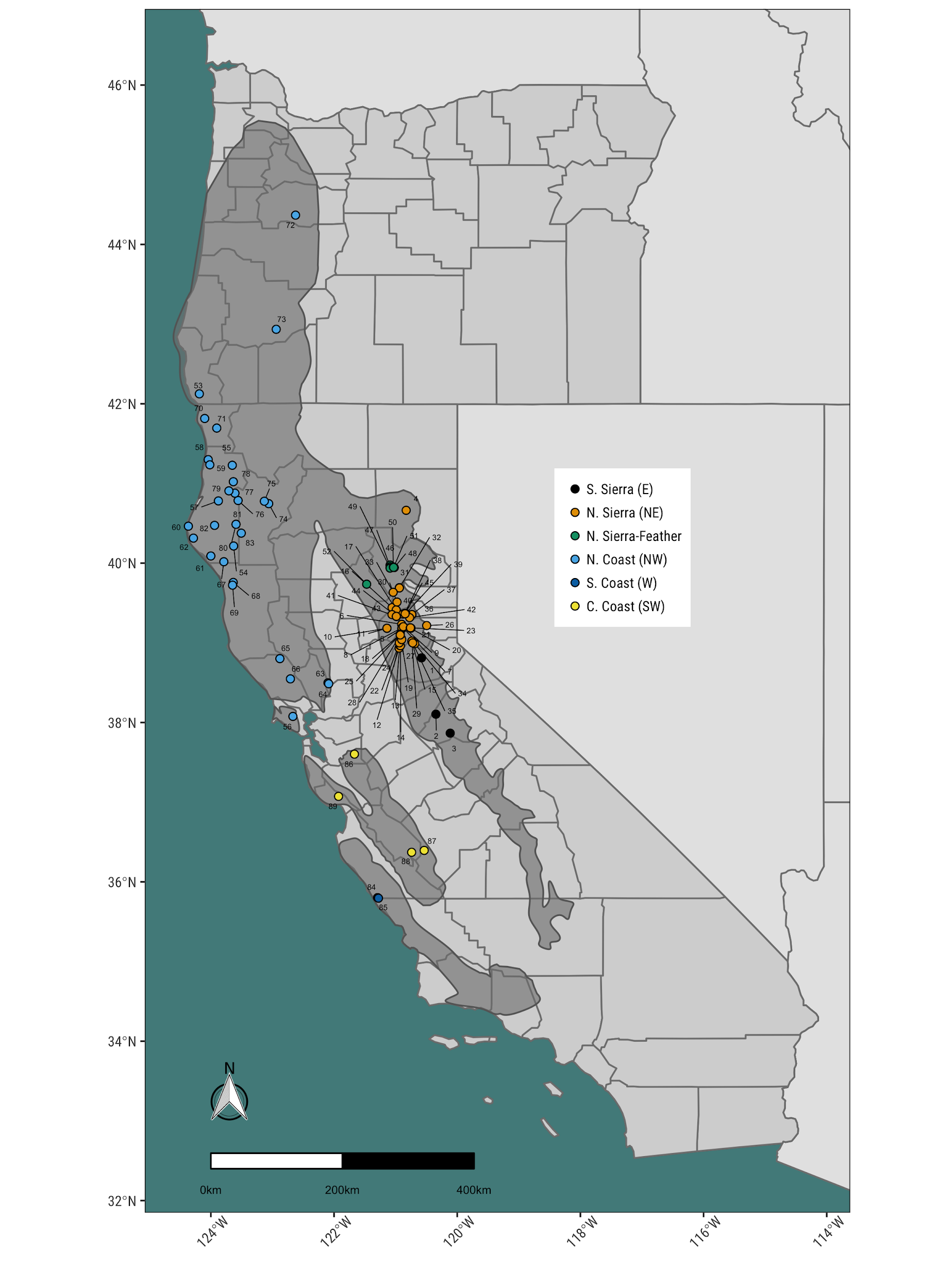


Figure 4. Map of localities colored by genetic groupings. The Northern Sierra-Feather group is adds an additional genetically distinct group, the Feather Basin to previous clades; Locality 1 extends the boundary for the Southern Sierra (E) genetic clade up to the SF American basin in El Dorado County, and Locality 4 extends the Northern Sierra (NE) clade into Lassen County.

## FST shows strong divergence across regions

To assess how distinct *R. boylii* populations were across the range of the species, we selected localities with at least three individuals, and randomly selected no more than ten individuals from localities with greater than ten samples. A total of 630 individuals from 63 localities were selected for analysis (Table 2). Pairwise FST was calculated between each population pair (total combinations = 1,953), values ranged from 0 to 0.646. Extremely high FST values were observed between Central and South Coastal (W/SW) localities and all localities from other regions, indicating these regions are strongly divergent and unique from *R. boylii* in other parts of the species’ range (Figure 5A). Comparison of pairwise FST values against geographic distance showed higher FST values between clades than within, and largest geographic distances between locality pairs (i.e., between Southern Coastal localities and Northern Coastal localities in Oregon) correlated with higher FST values. However, the highest FST values were not associated with the greatest geographic distances, the Northern Sierra (NE) and Southern Coast (SW) comparisons had the greatest pairwise FST values (> 0.6) of any pairwise comparison (Figure 5B).

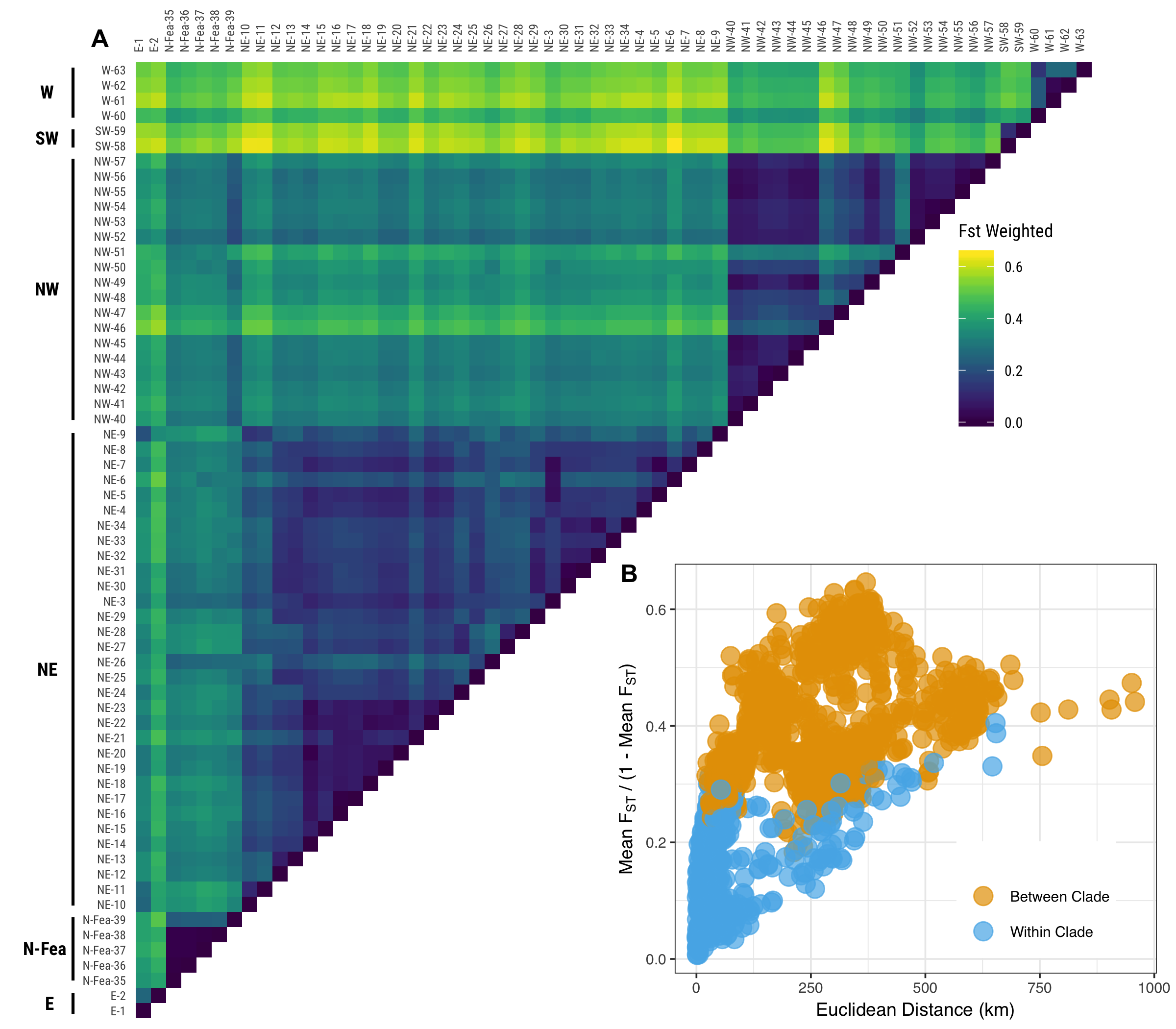


Figure 5. Plots of FST for all localities sampled, A) comparison of pairwise FST for all localities grouped by *R. boylii* clades; and B) pairwise FST comparisons between and within clades vs. geographic distance between localities.

## Genetic variation is highest in the North-West and lowest in the East/ South-West clades

Genetic diversity estimates for each region showed *R. boylii* from the North Coast group had the greatest range in genetic variation, with both the most diverse and least diverse θ estimates (Figure 6). Populations from the Southern Sierra (E) and Central Coast (SW) had the lowest mean Tajima’s θ𝜋 (0.0064) and Watterson’s θW (0.0059 = E, 0.0060 = SW), but both these groups had the lowest number of sampling localities. The Northern Sierra (NE) had the greatest number of sampling localities, yet the Δθ showed limited evidence of increasing genetic diversity (e.g., a θ𝜋 - θS  that was less than zero). Tajima’s D was negative across all groups, with the North Coast exhibiting the lowest value and the Central Coast (SW) and Southern Sierra (E) estimates being highest. Both coastal groups (SW and W) had very wide confidence intervals around Tajima’s D estimates. With the exception of the North Coast which contains the greatest genetic variation among all the clades, diversity data for *R. boylii* show a pattern of low genetic diversity, with many populations exhibiting trajectories of recent genetic variation loss (Figure 6D).

Table 3. Tajima’s D by group

|  |  |  |
| --- | --- | --- |
| Genetic Group | Tajima’s D | 95% CI |
| Southern Sierra (E) | -0.0417 | -0.0217 |
| Northern Sierra (NE) | -0.0479 | -0.0052 |
| N Sierra - Feather | -0.0630 | -0.0146 |
| North Coast (NW) | -0.0922 | -0.0191 |
| Central Coast (SW) | -0.0410 | -0.0766 |
| South Coast (W) | -0.0496 | -0.0514 |

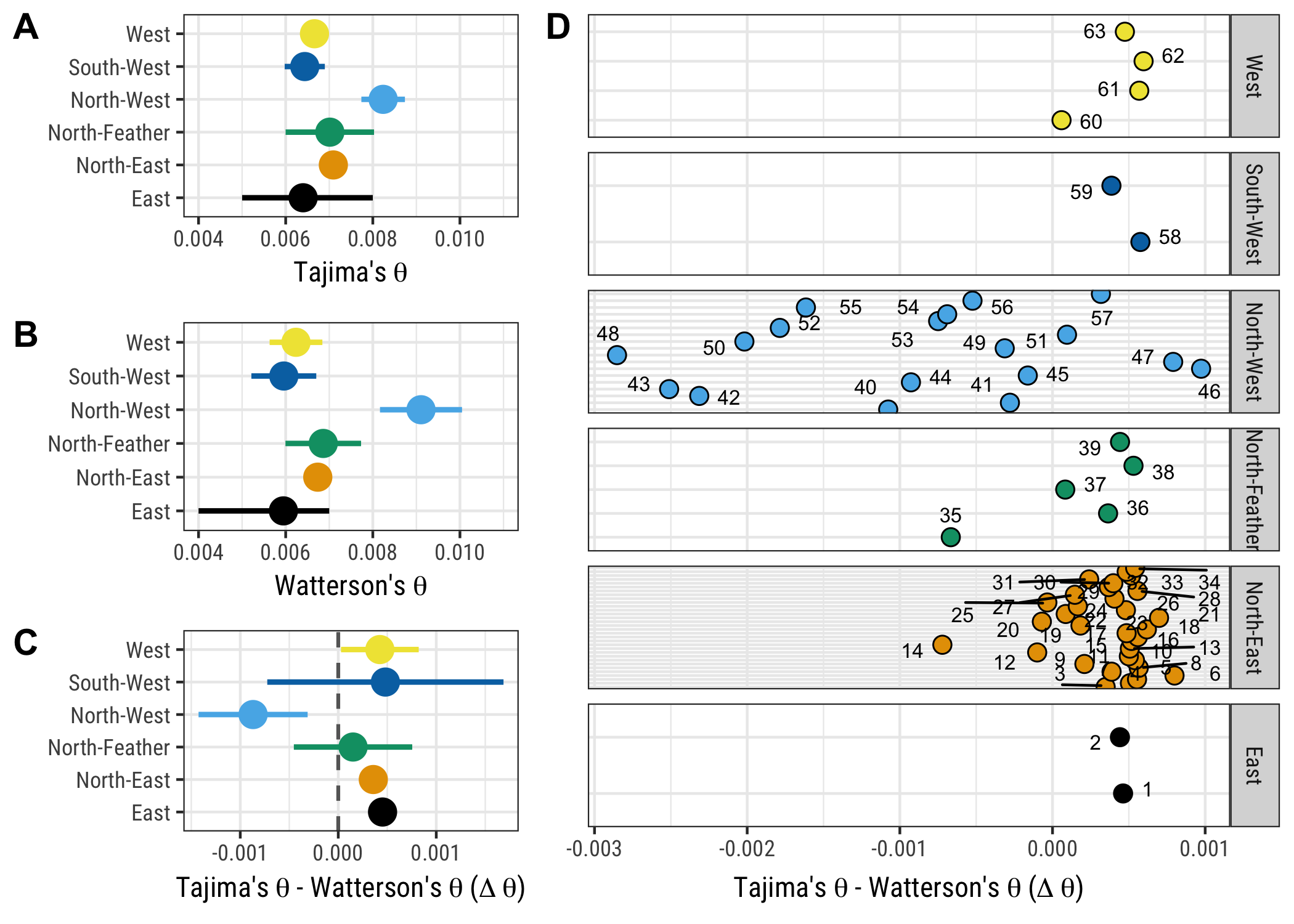


Figure 6. Tajima’s θ𝜋 and Watterson’ θS diversity estimates for *R. boylii* by clade groups (adapted from McCartney-Melstad et al. 2018) with 95% CI. A) Tajima’s θ𝜋 by clade; B) Watterson’ θS by clade; and θ𝜋 - θS  aggregated by clade C) and for each locality within clades D).

# DISCUSSION

As a species with a broad geographic range, *Rana boylii* utilizes a wide array of different landscapes throughout California and Oregon, however our study indicates there are strong patterns of geographic structure and diversity loss across a large part of the species’ range. This is important for conservation for several reasons. First, the higher resolution sampling provided by this data set compliments a recent genomic study of *R. boylii* (McCartney et al)*,* and largely supports the same clades identified using a different RADSeq method. Moreover, admixture and PCA results are congruent with the five major clades identified in McCartney-Melstad et al, with the exception of an additional novel group, populations from the Feather Basin. This congruence indicates these clades should be important units which conservation actions (i.e., such as translocation, recolonization, and habitat restoration) should consider carefully in planning future management of the species.

Second, we show genetic data can provide resolution to prioritize populations and regions at multiple scales. Evaluating the efficacy of potential restoration actions can be difficult, particularly with rare and cryptic species. Genetics provides an tool that can more effectively track changes in population connectivity or genetic trajectories (i.e., low diversity, high FST) that correlate with the probability for long-term population persistence. Populations in the Northern Coastal clade contained the greatest amount of genetic variation, yet pairwise comparisons with populations from Central and Southern Coast localities showed extremely high FST values, indicating these groups are genetically very distinct.

Recent sampling has largely occurred in the North Coast, Northern Sierra, and Northern Sierra-Feather groups. The limited number of sampling localities from the southern extent of the range of *R. boylii* correlates with research documenting population declines (Adams et al. 2017, Davidson 2004). Thus conservation actions should prioritize sites in the Central/Southern Coast and Southern Sierra clades, as these represent both the most genetically divergent and distinct *R. boylii*, and simultaneously the most at-risk populations. Extremely high FST values between populations from the same species as well as strong phylogeographic structuring support patterns of limited gene flow from biogeographic barriers (O’Connell et al. 2017, Spinks et al. 2010). This suggests conservation management should consider these clades as distinct, and future actions that include translocation or captive breeding should utilize genetic data to assess potential impacts such as outbreeding depression in locally adapted populations.

*R. boylii* is an important sentinel stream species—not only as link between both aquatic and riparian ecosystems—but also as a sensitive gage of the natural hydrology in riverine environments. The impacts of regulated hydrology on *R. boylii* populations have been well documented (Lind, Yarnell, Peek, Kupferberg). For populations in regulated streams, genetic monitoring may be a crucial component for assessing flow management actions on the probability of long-term persistence (via maintenance of genetic diversity), as well as tracking population responses (i.e., using metrics like FST and θ) to restoration actions.

*Probably need a wrap up paragraph here...*

## Acknowledgements

Many thanks to all who helped collect/prepare samples: Corey Luna, Rick Wachs, and Sarah Mussulman.

# CITATIONS