Title

Identifying Population Histories, Adaptive Genes, and Genetic Duplication from Population-Scale Next Generation Sequencing

Author

LInderoth, Tyler Philip

Publication Date

2018

Peer reviewed|Thesis/dissertation

Identifying Population Histories, Adaptive Genes, and Genetic Duplication from Population-Scale Next Generation Sequencing

by

Tyler Linderoth

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rasmus Nielsen, Chair
Professor Montgomery Slatkin
Professor Erica Bree Rosenblum

Spring 2018

Abstract

Identifying Population Histories, Adaptive Genes, and Genetic Duplication from Population-Scale Next Generation Sequencing

by

Tyler Linderoth

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Rasmus Nielsen, Chair

The arrival of next-generation sequencing (NGS) technologies in the mid 2000s opened the floodgates to a massive amount of genetic data. Not only does NGS permit relatively easy access to the genome of nearly any species, it also enables sequencing highly degraded DNA characteristic of ancient samples and museum specimens. The representation of genomic data across the tree of life has been spreading rapidly over the past decade owing to the emergence of numerous methods for inexpensively sequencing entire genomes and reduced representations of genomes based on NGS. However, without any high-quality preexisting genomic resources, species with large, highly paralogous genomes pose a major obstacle for NGS because accurately assembling short read data becomes extremely challenging. Furthermore, reads derived from paralogs will likely map to the same locus, which can inflate apparent levels of diversity, obscuring accurate population genetic inference and scans for adaptive loci. These problems can also effect population genetic studies using historic DNA from museum specimens, which often face the additional challenges of high sampling variability across space and time, and DNA degradation. The research presented in this thesis aims at overcoming these challenges using a combination of pioneering experimental and computational approaches. First, I present a method for identifying paralogy from NGS data, ngsParalog, that jointly leverages information from read proportions within and across individuals and sequencing coverage in a probabilistic framework. Combining information in this manner achieves superior power for identifying paralogy at lower false positive rates than using paralogy signatures separately as other current methods do. It also is widely applicable to both single and paired-end data ranging from low to high coverage. I use ngsParalog to detect paralogy in humans, chipmunks, and stick insects, representing a broad range of sequencing approaches. In the next chapter of the thesis I, along with colleagues, demonstrate how transcriptome-enabled exon capture applied to populations of century-old and modern *Tamias* chipmunks comprising multiple species, in conjunction with a new Approximate Bayesian Computation approach for fitting joint site frequency spectra between time periods can be used to infer recent population histories. Knowing these population histories allowed for disentangling the genetic signature of demographic changes from selection, which led to identifying a gene that may be helping chipmunk populations rapidly adapt to climate-induced environmental change. In the fourth chapter, I, along with other colleagues, employed the same exon capture technique and ngsParalog to overcome the challenge of mapping color and pattern genes in the $\sim$12 gigabase, highly paralogous genome of the mimic poison frog, *Ranitomeya imitator*. I applied statistical divergence and admixture mapping methods to different*R. imitator* color morphs in order to identify seven out of 13,086 examined genes that showed compelling evidence of influencing color and/or pattern in *R. imitator*. These candidate genes will likely be valuable for gaining insight into the *R. imitator* mimetic radiation. The combination of methods presented in this thesis advances the utility of NGS into taxa with genomes that previously precluded gene mapping and provides an analytical framework for identifying demographies and adaptive genes from museum collections.

Contents

Acknowledgments

First I would like to express my deepest thanks to my PhD advisors Rasmus Nielsen and Montgomery Slatkin. I always considered it a privilege to get to work in the Nielsen/Slatkin lab and I will miss my time here (granted I formated this thing correctly and I'm able to leave). I'm very grateful for your mentorship and support for my research in the lab and on my future endeavors. I honestly felt like I learned something every day, which was easy to take for granted at the time, but it certainly made the lab a unique place and I do realize how lucky I was for the opportunity. Equally as great were the people who made up the lab, it was always an awesome time. I'm very grateful for the friendship from all of my lab mates past and present, they are an extraordinary bunch of people, and I really mean that.

Apart from those in the lab, I need to give particular thanks to Natalie Graham who endured proofreading this thing, and for keeping me cranked out on coffee through the process so that I could get it done. Also, family, I hope you already know but will make explicit my deepest gratitude because you never for a second did anything but support me in all of my exploits. You guys even let me experiment with the French horn, and I still have no idea why, and I wish you wouldn't have.

Lastly, I need to thank my undergraduate advisors from the University of Alaska, Dave Tallmon and Sherry Tamone, because they played a huge role in helping me get to this moment.

1. Introduction

Next generation sequencing (NGS) technologies allowing for rapid generation of genetic data via massively paralleled sequencing were introduced in 2005 [1]. Over the subsequent years a multitude of innovative approaches to whole genome DNA sequencing, reduced genome sequencing, and RNA sequencing have produced a wealth of data that highlight the broad range of questions and aims that vast amounts of genetic data can address. Though today huge amounts of NGS data can be generated reasonably fast and affordably, utilizing the data to discover adaptive loci through population genetic approaches can be impended by a lack of preexisting, high-quality genomic resources. This barrier is amplified owing to the features of genomes and analytical settings described below. The combination of new computational methods and experimental approaches demonstrated in this thesis are intended to address these challenges and extend the utility of NGS into realms that previously posed major obstacles to population genetic inference.

The problem of paralogy

A significant challenge when analyzing genomic data for any species lacking high-quality genomic resources is paralogy, that is, the presence of homologous loci derived through duplication within a species. A fundamental assumption for most genetic analyses is that each analyzed unit, be it a single base pair, an extended genomic region, or a structural variant (e.g. insertions and deletions) represents a unique region in the genome; there is a one-to-one mapping between the data and genetic locus. Paralogy can lead to the violation of this assumption, with grave consequences. Sequence similarity between paralogs is often high enough that reads derived from paralogs can be assembled together and map to the same locus. The more loci that get wrongly confused as a single locus the greater the chances of inflating estimates of genetic diversity, simply because you have multiple genomic regions worth of variation in one. Collapsing loci in this way also obscures differences between populations and species by increasing the apparent number of heterozygous individuals. As an extreme example, imagine comparing two populations at a genetic site belonging to a duplicated gene. Let's say that the site for population 1 at the first gene copy is fixed for the ancestral allele, A, and the second copy is fixed for a derived allele, a. In population 2, the site at copy 1 is instead fixed for the a allele and copy 2 for the A allele. If these paralogous regions are similar enough such that sequencing reads derived from them map to the same location, every individual in both populations would appear heterozygous, which would result in a measure of genetic differentiation, like F_{ST} for example, to be zero (no differentiation), when it should be 1 (most extreme differentiation)! If one was trying to map genes underlying a particular trait by comparing allele frequencies between populations as is common in genetic association studies, biologically these genes with fixed differences would be very strong candidates, however the association would never be found. This example with fixed opposite alleles between paralogs in population 1 versus population 2 is perhaps unlikely, yet clearly demonstrates how paralogy can inflate measures of genetic diversity, decrease population differentiation, and lower the power to map genes. It is quite feasible to have a derived allele that is either fixed or segregating in one gene copy in either of the populations, in which case paralogy would have the same effects, including obscuring the ability to map genes if the copy with the mutation is causal or linked to a causal locus.

Even for species with the most well-developed reference genomes paralogy and similar sequence complexity issues that confound assembly and cause low mappability is a problem, including in humans [2, 3]. These issues are exacerbated in non-model species lacking quality genomic resources. In the second chapter of the thesis I present a statistical method for detecting paralogy from NGS data that improves upon existing methods. While previous methods use either excess heterozygosity [4], or similarly, the proportion of reads bearing different alleles within and across individuals [5], or sequencing coverage [6, 7] to detect paralogy, my method is the only one to use all of this information jointly in a probabilistic framework, which achieves greater true positive to false positive rates than using any of these signals independently. Unlike methods that use paired-end read information to detect duplications [8, 9], my method can be used for both paired-end and single-end data, and so is useful for a broader range of experimental designs. I demonstrate the effectiveness of this method by identifying regions in the human genome that show evidence for being duplicated in addition to finding paralogs in NGS datasets for chipmunks and stick insects, which use different sequencing approaches commonly utilized for non-model species.

Museum population genetics

Natural history museums are a trove of information for assessing genetic, phenotypic, and demographic change through time, which provides a direct route to understanding how species and populations respond to environmental change [10]. Museum collections have and still do represent a challenge for genetic studies because historic specimens are not often collected with molecular genetic analyses in mind. The first implication of this is that much historic DNA is degraded making PCR amplification and accurate Sanger sequencing a challenge, with special care required to avoid PCR artifacts [11, 12]. Some of the first genetic data to come from sequencing museum specimens focused on PCR amplifying and sequencing < 500 bp of the mitochondrial genome [13, 14, 15], and nuclear microsatellites [16, 17, 18]. The problem of scaling up the number of sequenced loci from museum DNA was overcome by NGS approaches since DNA library preparation involves ligating general sequencing adapters onto the ends of the degraded fragments, allowing for amplification and sequencing of entire genomes. Accordingly, over the past decade, there have been an increasing number of studies focused on vastly larger regions of the genome or the entire genome from museum specimens [12, 19, 20, 21], affording much greater genetic resolution for comparing populations through time. The ability to sequence entire genomes from museum specimens does not however overcome all of the problems facing museum genetics. The data from historic DNA sequencing is often plagued by artifacts from chemical damage such as cytosine deamination [22], and variability in sample sizes and localities over time can make direct population comparisons a challenge, particularly when samples were not collected with the intention of these types of comparisons. Nevertheless, the ability to essentially hop into a time machine and get a precise genetic snapshot of populations collected before genetic molecular techniques even existed is exciting and invaluable for understanding how organisms have adapted to the rapid global climate change that has occurred over the past century. Accordingly, the third chapter of my thesis focuses on a pioneering, multi-species, comparison of century-old, historic to modern populations at a genomic-scale in order to understand demographic and adaptive response to environmental change. We used transcriptome enabled sequence capture to obtain exonic data for 303 *Tamias* chipmunks, representing six species, and developed an Approximate Bayesian Computation framework for fitting population histories to joint frequency spectra between time periods. This approach afforded us the resolution to characterize recent demographic changes for various chipmunk populations and delineate the genomic signatures of demography from selection in order to robustly identify potentially adaptive genes in relation to climate-induced environmental change.

Mapping genes in large, highly duplicated genomes

The emergence of NGS has ushered in an enormous amount of progress towards mapping genes relevant to human disease [23] and adaptation [24]. Prior to NGS, population genetic approaches for mapping the genetic basis of traits in non-model species relied heavily on candidate gene approaches [25]. Now, however, genome-wide scans for selection have become increasingly common and generated many insights into the genetic mechanisms of adaption and speciation. Some notable examples include stickleback and cichlid fishes [26], butterflies [27], and stick insects [28]. Even with the ease and relative low cost of genome sequencing afforded by NGS technologies, genomic features of some species still limit the taxonomic breadth across which genes can be efficiently mapped. One example of this are poison frogs (family Dendrobatidae) for which at least some species have exceptionally large genomes sizes due to extensive paralogy making accurate genome assembly virtually impossible using short read sequencing. Even with a high-quality reference genome, the large genome size precludes cost effective sequencing of enough individuals to attain the statistical power necessary to confidently map genes. The fourth chapter of my thesis is focused on overcoming these challenges in order to identify genes underlying color and pattern in the mimic poison frog, *Ranitomeya imitator*. I, along with colleagues, reduced the daunting *R. imitator* genome down to the exome using a custom sequence capture system, which represented a subset of the genome that could be assembled and that is functionally relevant. We sequenced exomes from 124 *R. imitator* representing various color morphs and used a combination of statistical divergence and admixture mapping approaches to identify seven out of over 13,086 surveyed genes that show evidence for influencing color and/or pattern in *R. imitator*.

Conclusion

In order to advance the applicability of next-generation sequencing for identifying adaptive genomic regions across a broader breadth of taxa and time series, I, along with colleagues, have developed a suite of methodologies and computational tools for overcoming genetic inference challenges that are particularly pronounced in non-model species lacking quality genomic resources. These methods enable the discovery of adaptive genes from museum collections and organisms having genomes that have previously precluded population genetic approaches. We used these approaches to discover a gene in alpine chipmunks that is potentially adaptive in the face of climate change and candidate genes for involvement in a poison frog mimetic radiation.

2. Widely applicable, probabilistic methods for paralogy identification from population-scale, next generation sequencing data

Tyler Linderoth, Rasmus Nielsen

2.1 Introduction

Next-generation sequencing (NGS) is a widely used approach for gathering genome-wide data, at relatively low cost, for both model and non-model species. This most often involves the alignment of millions to billions of short sequencing reads through the course of assembly and mapping, which means that any two or more genomic regions with substantially similar sequence can produce reads that will align. This can provide useful information about the genome if recognized, but can also introduce severe bias into population genetic inference if not addressed. On the useful side, genomic regions that produce similar reads can represent paralogs (homologous regions of the genome derived through duplication within a species), which have much biological significance. First, these types of duplications are medically relevant due to the role they play in diseases [29, 30], and are interesting from an evolutionary perspective because they represent the opportunity for sub and neo functionalization. In this regard paralogs are involved in everything from fish coloration [31] to dosage effects [32, 33]. On the negative side, if ignored in population genetic NGS datasets, paralogy can inflate estimates of genetic diversity, and so can obscure analyses of population divergence, estimates of mutation rates, and scans for selection for example. Therefore, identifying paralogous regions is desirable for both investigating biological questions related to genetic duplication as well as a necessity for quality control.

One commonly used method for identifying paralogy is to use paired-end read information, since the arrangement of two mapped reads supposedly derived from a single library fragment can be informative about duplication and other structural variation. Two other commonly used signatures in NGS data used to detect copy number variation (CNV) are sequencing depth and excess heterozygosity. Sequencing reads derived from duplicated regions are often similar enough that they will map to a single locus, causing the coverage within an individual to be proportional to the number of genomic copies. When allelic differences exist among sites between the different copies, collapsing sequencing reads like this also makes individuals appear heterozygous. There are many existing methods for detecting CNVs from sequencing depth [6, 7], excess heterozygosity [5, 4], and paired-end sequencing information [8, 9]. Limitations of these previously listed approaches are that they are either inapplicable to studies that use either single-end sequencing (such as is widely applied in reduced representation restriction digest studies) or low coverage data whereby genotyping is too inaccurate to use excess heterozygosity, or they simply do not fully utilize all of the information available [5]. Consequently, we developed a probabilistic method for identifying paralogy that can be used with any type of NGS data and that is highly accurate at all sequencing depths (including very low coverage). The method first leverages information from excess heterozygosity to calculate a likelihood ratio for whether or not a particular site is duplicated. Most other methods rely on genotype calling to detect excess heterozygosity, whereas we focus on estimating the probability that sequencing reads within an individual are derived from multiple copies. This effectively incorporates within-individual read proportion information that is lost when calling genotypes and is important because variability at a duplicate copy within an individual and/or biased mapping between paralogs can obscure genotype calling even at high sequencing coverage and decrease power to detect duplicates. The duplication likelihood ratios from this first part of our method can themselves be used to statistically identify duplicated sites. The second part of the method uses these likelihood ratios in combination with sequencing coverage information in a hidden Markov model (HMM) to infer regions of duplication that achieves high sensitivity with low false discovery.

We used our method, called ngsParalog, to identify regions in the human genome that show evidence for duplication that are missing from the 1000 Genomes accessibility files, which are intended to contain such regions. These unrecognized duplicated regions are relevant to any studies on human population genomics. We also demonstrate the utility of ngsParalog for paralog discovery and quality control in non model species genomics by applying it to chipmunk exon capture and stick insect single-end sequencing, genotyping-by-sequencing (GBS) data. Our method is implemented in a very user-friendly program that can be downloaded from `https://github.com/tplinderoth/ngsParalog`.

2.2 Methods

Likelihood of single site duplication

Sequencing reads derived from paralogs will often be similar enough that they will be assembled together forming a chimeric region and then also map to this region. If a mutation occurs in one of the copies then any individual for which the mutant allele is present will appear to be heterozygous at the site to which the reads from both copies map. This generates an excess of heterozygotes relative to that expected under Hardy-Weinberg equilibrium

(HWE). To determine the likelihood that a particular site in the genome is duplicated we assume a model of duplication in which there are potentially two copies of the locus, and for which copy 1 is fixed for the reference allele and copy 2 is either fixed for a different (alternate) allele or is biallelic. We refer to the genotype at copy 1 and 2 as G_1 and G_2, respectively. For nonduplicated sites, we assume that genotypes occur at frequencies expected under HWE, and so deviation from this would indicate duplication. Leveraging this signal of excess heterozygosity assumes accurately calling genotypes, which is not possible with low coverage NGS data because chromosomes are sampled with replacement a variable number of times and with appreciable levels of sequencing error. Therefore we use genotype likelihoods to model all of these sources of genotyping error. The parameter of interest that causes the appearance of excess heterozygosity at duplicated sites is the probability that a sequencing read sampled at a site comes from copy 2, m. For a SNP, by assuming that copy 1 is nonvariable and copy 2 is variable, if $m < 1$, then the collapsed site appears variable due to duplication, and if $m = 1$, all reads are derived from the variable locus and so the SNP actually represents a single, nonduplicated site. For a site, given the observed sequencing data D, the joint likelihood of m and the population alternate allele frequency at copy 2, f, is

$$P(D|m, f) = \prod_{j}^{n} \sum_{G_2} P(X_j|G_1 = 0, G_2, m)P(G_1 = 0, G_2|f)$$

where n is the number of individuals and X_j is the sequencing data for individual j. $G_2 \in \{0, 1, 2\}$, denoting the number of alternate alleles. Note also that we always assume that G_1 is fixed for the reference allele, and so has genotype 0. Letting r_j denote the number of reads for individual j for the site in question, the genotype likelihoods are

$$P(D|G_1, G_2, m) = \prod_{k=1}^{r_j} P(x_k^j|G_1, G_2, m)$$

The genotype likelihoods can be broken down in terms of m and the read quality scores, which specify the probability that the identity of the associated read is a sequencing error, $\mathcal{E}$. Assuming a biallelic site where the only actual alleles are the reference, w, and alternative, z, and uniform error rates among erroneous bases

$$\begin{aligned}
P(x_k^j = w|G_1 = 0, G_2 = 0) &= 1 - \mathcal{E}_k^j \\
P(x_k^j = w|G_1 = 0, G_2 = 1) &= (1-m)(1-\mathcal{E}_k^j) + \frac{m}{2} \\
P(x_k^j = w|G_1 = 0, G_2 = 2) &= (1-m)(1-\mathcal{E}_k^j) + m\mathcal{E}_k^j \\
P(x_k^j = z|G_1 = 0, G_2 = 0) &= \mathcal{E}_k^j \\
P(x_k^j = z|G_1 = 0, G_2 = 1) &= (1-m)\mathcal{E}_k^j + \frac{m}{2} \\
P(x_k^j = z|G_1 = 0, G_2 = 2) &= (1-m)\mathcal{E}_k^j + m(1-\mathcal{E}_k^j)
\end{aligned}$$

Assuming that G_1 is indepedent of f, the genotype priors are specified by Hardy-Weinberg expectations

$$\begin{aligned}
P(G_1 = 0, G_2 = 0|f) &= (1-f)^2 \\
P(G_1 = 0, G_2 = 1|f) &= 2f(1-f^2) \\
P(G_1 = 0, G_2 = 2|f) &= f^2
\end{aligned}$$

We use the BFGS-B algorithm to obtain maximum likelihood estimates of m and f for each site. In the case of no duplication, $m = 1$, and so is nested within the duplicated case, allowing us to specify a likelihood ratio for whether a site is duplicated

$$LR_{\text{dup}} = -2\log\left(\frac{P(D|m=1,f)}{P(D|m,f)}\right)$$

LR_{dup} is asymptotically distributed as a 50:50 mixture between a χ_0^2 and χ_1^2 (Figures 2.1 & 2.2). Using LR_{dup} one can then decide at a chosen confidence level whether a given SNP is paralogous.

The method for calculating per site LR_{dup} is implemented as a freely available C++ program, ngsParalog, available for download at `https://github.com/tplinderoth/ngsParalog`. Specifically, the function calcLR is used for calculating LR_{dup} using pileup format data as input.

Hidden Markov Model for inferring duplicated regions

In order to use the likelihood that individiual sites are duplicated to infer entire regions of duplication we use an HMM that runs over LR_{dup}. We also use the HMM to incorporate sequencing coverage information because sites comprised of mismapped reads from copy number variants will tend to have coverage approximately proportional to the number of copies (Figure 2.3).

The HMM LR_{dup} emission probabilities under the nonduplicated state are determined from LR_{dup}'s asymptotic distribution, which is approximately a 50:50 mixture of a χ_0^2 and

χ^2_1, which we denote as ψ. The density for LR_{dup} under the duplicated state is a noncentral chi-square with 1 degree of freedom and noncentrality parameter Λ, $\Psi(lr_{dup};\Lambda) = \chi^2_{1,\Lambda}$.

A maximum likelihood estimate of Λ is found by jointly fitting it and p_{nd}, the probability that a randomly drawn site is not duplicated, to the density defined by

$$f(lr_{dup};p_{nd},\Lambda) = p_{nd}\psi(lr_{dup}) + (1 - p_{nd})\Psi(lr_{dup};\Lambda)$$

When the number of independent observations are large, even when the simpler model (no duplication) is true, the more complex model (duplication) is favored when the likelihoods are compared with likelihood ratios [34]. The Bayesian information criterion, BIC, for penalizing the LRs is based on asymptotic inflation of Bayes factors in favor of the more complex model, and so permits equal comparison of the models by more harshly penalizing the LRs with increasing sample size. When wanting to be more conservative, prior to estimating Λ we use BIC to penalize LR_{dup} since otherwise the additional parameter in the model involving duplication will inflate the likelihood ratio in favor of duplication. Accordingly, $k\ln(n)$ is subtracted from LR_{dup}, where k in our case is 1, and n is the number of individuals.

To incorporate coverage information into the HMM we focus on the average individual sequencing depth for a given site. The number of sequencing reads covering an individual, R, is Poisson distributed with mean parameter λ. We let $\delta(r;\lambda)$ denote the density for R. The average individual coverage at a particular site, R_n is defined as

$$R_n = \frac{1}{n}\sum_{i=1}^{n} r_i$$

λ will vary to some degree between sites based on features such as GC content, and the position of the site along the reference sequence. Under the central limit theorem, if λ is the same for all sites, then the average individual coverage among all sites, should be approximately normally distributed with mean λ and variance $\frac{\lambda}{n}$, where n is the number of individuals. The variation in λ among sites increases the variance of this normal distribution. Therefore, we consider the density for average individual coverage, $\phi(r_n;\mu,\sigma^2,a,b)$, to be that of a truncated normal distribution with mean μ, variance σ^2, and minimum and maximum values for r_n of a and b, respectively. The normal distribution is truncated since the minimum value r_n can take is zero. The average individual coverage for duplicated sites, assuming two copies, is the sum of two approximately normally distributed random variables, and so will also be approximately normally distributed with mean 2μ. As previously mentioned, since variation in λ among sites increases the variance among the average individual coverages, we obtain maximum likelihood estimates for the parameters of $\phi(r_n;\mu,\sigma^2,a,b)$ for nonduplicated and duplicated sites by fitting the set of observed r_n to

$$\begin{aligned} f(r_n;p_{nd},\mu_{nd},\sigma^2_{nd},\sigma^2_{dup},a_{dup}) &= p_{nd}\phi(r_n;\mu_{nd},\sigma^2_{nd},0,\infty) \\ &\quad + (1 - p_{nd})\phi(r_n;2\mu_{nd},\sigma^2_{dup},a_{dup},\infty) \end{aligned}$$

The left truncated normal density corresponds to nonduplicated sites (subscript nd stands for nonduplicated), while the right-hand density is for duplicated sites (subscript dup indicates duplicated).

The HMM observations are binned lr_{dup} and r_n pairs, $\{lr_{dup,[u,w]}, r_{n,[y,z]}\}$, with respective bin intervals of $[u, w]$ and $[y, z]$. The two possible states comprising the state sequence, Q, are $\{\text{nonduplicated}, \text{duplicated}\}$. We treat LR_{dup} and R_n as conditionally independent such that when site i is in state k, the probability of observing $\{lr_{dup,[m,n]}, r_{n,[r,s]}\}$ is

$$e_{nd}(lr_{dup,[u,w]}, r_{n,[y,z]}) = \begin{cases} \big(0.5 + 0.5 \int_u^w \chi_1^2(lr_{dup})\, dlr_{dup}\big) * \int_y^z \phi\big(r_n; \mu_{nd}, \sigma_{nd}^2, 0, \infty\big)\, dr_n, & m = 0 \\ 0.5 \int_u^w \chi_1^2(lr_{dup})\, dlr_{dup} * \int_y^z \phi\big(r_n; \mu_{nd}, \sigma_{nd}^2, 0, \infty\big)\, dr_n, & m > 0 \end{cases}$$

when the site is not duplicated, and

$$e_{dup}(lr_{dup,[u,w]}, r_{n,[y,z]}) = \int_u^w \Psi(lr_{dup}; \Lambda)\, dlr_{dup} \;*\; \int_y^z \phi(r_n; 2\mu_{nd}, \sigma_{dup}^2, a_{dup}, \infty)\, dr_n$$

when the site is duplicated. The emission probabilities are normalized to sum to one for each state

$$e_{nd}(v_i) = e_{nd}(v_i) / \sum_{j=1} e_{nd}(v_j)$$

$$e_{dup}(v_i) = e_{dup}(v_i) / \sum_{j=1} e_{dup}(v_j)$$

where V is the set of all possible observations.

With the emission probabilities calculated as outlined above, the Baum-Welch algorithm is used to estimate the transition probability matrix, A, for the observed sequence, x, as

$$A_{hl} = \frac{1}{P(x)} \sum_t f_h(t) a_{hl} e_l(x_{t+1}) b_l(t+1)$$

and initial state distribution, π, for each $h \in \{\text{nonduplicated}, \text{duplicated}\}$, as

$$\pi_h = \sum_h f_h(0) a_{hl} e_l(x_1) b_l(1)$$

where $f(t)$ and $b(t)$ are the standard forward and backward variables for site t, respectively.

The Viterbi algorithm is used with A, π, and the emission probabilities estimated as outlined above to infer the most probable states among the SNPs.

The duplication HMM method is implemented as an executable R script, dupHMM, within the ngsParalog suite of tools (`https://github.com/tplinderoth/ngsParalog`). It uses the output of ngsParalog calcLR and average individual coverage for each site as input and outputs a position file of regions inferred to be duplicated.

Simulated genomic regions

We simulated NGS data in pileup format using the program ngsSimPileup `https://github.com/tplinderoth/ngsSimPileup`. The average individiual coverage for each locus, λ, was drawn from a normal distribution and then the number of sequencing reads for each individual was drawn from a Poisson(λ). λ was treated as a random variable in order to mimic variablity in coverage due to factors such as GC content that generate site-wise differences in total sequencing depth. Reads for heterozygous individuals were determined with uniform probability of sampling each allele. Given a specified allele frequency, the genotype counts at each locus were those expected under Hardy-Weinberg equilibrium. Genotype counts were altered to account for differences in relative fitness, which allowed for simulating overdominance. We assumed no inbreeding. Each sequencing read was assigned a Phred quality score, which were distributed according to a beta distribution with alpha specified as a function of the average sequencing error rate, 0.00015, and beta=0.3 for all simulations. Sequencing errors were then introduced according to these Phred-scaled error probabilities. In the case of a sequencing error, the allelic identity of the read was uniformly changed to any of the three other alleles.

For simulating data from paralogs that get collapsed into a single site, each copy was simulated as described above using the same λ for each copy. Mapping bias between paralogous loci is possible based on the amount of sequence divergence between the copies. Therefore, the reads from each copy were combined based on their respective, and potentially unequal probability of mapping. ~250,700 bp long genomic regions consisting of nonduplicated and duplicated sites were simulated to evaluate the HMM. Specifically, sequences of SNP positions were generated based on a specified probability that a randomly drawn site is variable, θ, and a state assignment of duplicated or nonduplicated was assigned to each SNP based on a an initial distribution and transition matrix. Then sequencing data for each position was simulated with ngsSimPileup as outlined above.

One of the primary advantages of the the ngsParalog method is that it is effective at very low sequencing depths for even moderate sample sizes, and so we analyzed simulated datasets for 20, 50 and 80 individuals, each sequenced to an average depth of 2X and 4X. For all simulations the MAF for nonduplicated sites varied uniformly between 5-50%, while for duplicated sites copy 1 was always fixed for the reference allele, while copy 2 had an alternate allele frequency uniformly distributed between 5-100%. Mapping bias between duplicate copies was also simulated so that both copies either contributed reads to a site equally (the reads covering a duplicated sites were comprised 50/50 between the two copies), or in skewed proportions of 60/40 and 75/25 between copy 1 and copy 2, respectively. For copies with equal contribution of reads, since all reads from both copies map to a single site, the coverage for that site will be approximately double that of nonduplicated sites. This inflation of coverage decreases with greater mapping bias since only a fraction of reads from copy 2 are mapping.

Since both paralogy and overdominance generate excess heterozygotes relative to HWE expectations, LR_{dup} could potentially confound regions of heterozygote advantage for duplication. There is, however, no reason that such regions under selection should exhibit inflated sequencing coverage and so dupHMM, since it uses information from the duplication LRs and coverage jointly, should be able to distinguish regions of overdominance from duplication. In order to test this we introduced three ∼2kb long windows with selection coefficient equal to 0.1 in favor of heterozygotes into our simulated regions. Strength of selection against both homozygous types was the same. We then analyzed these simulated datasets the same way as for regions without selection.

Duplication detection with *Tamias* chipmunk exon capture data

Virtually all NGS analyses that rely on population genetic data assume a one to one mapping between a particular site and the data derived from that site (e.g. sequencing reads). This can be violated when paralogy is involved as the data from genomic copies can be collapsed together during assembly and mapping, which inflates apparent levels of genetic diversity, and, ultimately, biases inference. This problem is most acute for studies involving non model organisms with lower quality reference genomes where duplication has to be identified *de novo*. When samples are sequenced to a high enough depth to accurately call genotypes, duplicated SNPs can be identified by testing for a significant excess of heterozygotes relative to HWE expectations. With current analytical methods, the accuracy and power of studies that rely on population genetic sampling often increases with more individuals even at the cost of reduced coverage [35, 36]. That is, a desirable experimental design is often to sample more individuals at sequencing coverage that prevents calling genotypes. The calculation of LR_{dup} is based on genotype likelihoods and so is able to effectively identify duplicated SNPs at low sequencing depths where other methods would lack power. To demonstrate the effectiveness of ngsParalog at identifying duplicated SNPs in a nonmodel species we applied it to an alpine chipmunk, *Tamias alpinus*, exon capture dataset from Bi *et al.*[19]. This study sequenced 20 historic museum specimens and 20 modern samples, so we limited the current duplication analysis to only the contemporary samples. The *T. alpinus* reference sequence was a *de novo* assembly of the exon capture data and so comprised many small contigs (N50 = 705 bp, longest contig is 5,025 bp) representing exons and portions of introns. Since the reference lacked large scaffolds most suited for dupHMM we limited our analysis to the per-site duplication LRs. Using the BAM files generated in the original study, we identified 12,154 sites with a p-value of being variable ≤ 0.05 using ANGSD [37]. The average individual sequencing depth among the identified SNPs was 15.5X. We calculated LR_{dup} at all of the SNPs with ngsParalog calcLR, considering only bases with a Phred-scaled quality score ≥ 20. We identified duplicated SNPs at a 0.05 sigificance level after a Bonferroni correction for multiple testing by comparing LR_{dup} to a 50:50 χ^2_0 and χ^2_1 distribution. We then generated the site frequency spectrum for the 20 modern individuals with ANGSD for all sites, and for a subset of the sites that had the duplicated positions filtered out in order to test whether removing the putatively duplicated sites would control the inflation of the 50% allele frequency category in the SFS caused by paralogs.

Duplication detection in *Timema* stick insects using GBS

In order to demonstrate how ngsParalog and dupHMM can be used to identify duplicated sites and regions in a typical population genetic study involving a nonmodel species with a more complete reference genome, we analyzed a low-coverage, restriction digest dataset for stick insects, *Timema cristinae,* generated by Lindtke *et al.*[28]. Using the aligned data for the N1 population from Lindtke *et al.*[28], we analyzed only biallelic SNP positions contained in the VCFs from their study found at `https://datadryad.org//resource/doi:10.5061/dryad.jt644`. We analyzed all 435 N1 individuals, which had an average individual sequencing depth of 4.9X for the analyzed SNPs. First, ngsParalog was used to calculate LR_{dup} for all SNPs, discarding any reads with Phred-scaled base quality < 20. SAMtools [38] was used to obtain individual sample coverages at each site with no minimum mapping or base quality for retaining reads, and then the average individual coverage, R_n, for each SNP was calculated. The HMM LR_{dup} and R_n emission density parameters were estimated from all SNPs across all scaffolds using dupHMM, with the largest 2% of LR_{dup} excluded to avoid fitting to extreme values. Using these estimates for the emission density parameters, we then ran dupHMM on each scaffold separately, first without a minimum R_n cutoff for duplicated sites, and also with a semi-conservative and conservative R_n lower bound of 6X and 7.3X, which corresponds to 1 and 2 standard deviations above the mean average individual coverage for sites with $LR_{dup} = 0$. Note that these lower bounds were imposed when estimating the coverage density parameters. Running dupHMM with the coverage lower bound is for conservative identification of duplicated regions. All duplicated region coordinates correspond to the MSSY03 *timema* assembly `https://www.ncbi.nlm.nih.gov/nuccore/MSSY00000000`.

Human genome analysis

We also looked for duplication in the human genome, except for the Y chromosome, using all 2,504 unrelated individuals for the autosomes and 1,271 unrelated females for the X chromosome represented in the Phase 3 1000 Genomes dataset. For the analyzed sites, the average individual coverage is ∼7.29X across all chromosomes, and ranges from 6.8X for chromosome 19 (5.3X for chromosome X with males and females) to 7.4X for chromosome 16. There has been extensive effort to identify and handle structural variation and low complexity regions with low mappability in this dataset already, including using decoy sequences during mapping, which is outlined in the supplement of the 1000 Genomes Consortium [39].

We analyzed all of the biallelic SNPs released in the official Phase 3 VCF files found at `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`. SAMtools was used to generate pileup input for ngsParalog from BAM files for the analyzed SNPs. LR_{dup} were calculated with ngsParalog calcLR, considering only reads with a minimum Phred quality score of 20 and requiring at least 100 individuals to have data for a site to be analyzed. Average individual coverage was calculated from the VCF files and this, along with the duplication LR_{dup} were used as input to dupHMM, which was run over each chromosome.

Coverage for chromosome 12 was calculated directly from the BAM files with Samtools because it was not specified for all analyzed SNPs in the VCF file. All of our reported inferred regions correspond to the hg19 human genome assembly.

2.3 Results

Duplication detection from simulated data

We simulated NGS data for 5K SNPs spanning a $\sim$ 250,700 bp long region for 20, 50, and 80 individuals at an average sequencing depth of 2X and 4X. Inflated sequencing depth is commonly used to reveal duplications, and so we used receiver operator curves (ROC) to compare the true positive to false positive rates for identifying duplicated sites based on average sequencing depth at a site, to LR_{dup}, and then also to dupHMM which combines information from both the duplication LRs and sequencing depth (Figure 2.4). In situations where both sequencing coverage and sample sizes are small, 2X and 20 individuals, respectively, coverage and LR_{dup} perform similarly for false positive rates under $\sim$5% (Figure 2.4). As allele frequencies are more accurately estimated with larger sample sizes, LR_{dup} outperforms sequencing coverage. Under even the most extreme circumstances of mapping bias between paralogs, with the exception of one case where it matched LR_{dup}, dupHMM achieved the highest, true positive to false positive rate ratio, TPR/FPR, with maximum false discovery rates, FDRs, under 0.05%. When the sequencing data is most sparse (2X, n=20), dupHMM has 81-90% power for accurately identifying duplicated sites, while this increases to 92-95% with slightly more than double the sample size, n=40.

When sequencing coverage is still low, but is increased from 2X to 4X, except for one case where the FDR is 0.04%, dupHMM, had all FDRs of 0% and power that ranges from 96% when there was extreme mapping bias and only 20 individuals to 100% when paralogs map equally and with 80 sequenced individuals (Figure 2.4). When few individuals are sequenced (n=20), a positive relationship between mapping bias and the TPR/FPR of LR_{dup} results in coverage performing better than LR_{dup} when mapping is equal and vice versa when mapping bias becomes extreme. This minimal sample size scenario is when dupHMM shows the greatest performance advantage over coverage and LR_{dup} under any degree of mapping bias. When sample sizes increase to 50 and 80 individuals, LR_{dup} primarily outperforms coverage in all mapping scenarios, while dupHMM still has the best performance between either LR_{dup} or coverage.

We introduced three $\sim$2 kb long regions under selection in favor of heterozygotes into our simulations to determine whether dupHMM could distinguish overdominance regions apart from duplication. dupHMM TPR/FPR for regions involving selection and without were primarily the same (Figure 2.5), with selection not increasing the duplication FDR. Therefore, dupHMM was able to accurately distinguish between regions of overdominance and duplication.

Tamias chipmunk exon capture analysis

We estimated the SFS from exon sequencing data for 20 *T. alpinus* chipmunks and found an excess of of heterozygotes (Figure 2.6), which is the telltale sign of paralogs. Under neutrality, the number of sites belonging to allele frequency category x should be proportional to $1/x$, meaning that in this case the 50% category is expected to have approximately 0.05 as many sites as the singleton category (baring the small increase in singletons from folding the SFS). However, there are nearly 0.25 times as many 50% frequency sites as singletons, that is, nearly 5x the neutral expectation. This could bias the many population genetic summary statistics and analyses based on the SFS. We applied ngsParalog to this dataset to determine whether it could be used to identify these problematic, likely paralogous SNPs, and improve the SFS. From the Bonferroni-corrected LR_{dup} p-values, we found 1,707 SNPs to likely be duplicated at a 0.05 significance level. After removing these SNPs from the dataset we estimated the SFS in the same way as before and found that there was no longer an excess of sites with an allele frequency of 50% (Figure 2.6), and that now there were 0.052 as many 50% frequency sites as singletons, which agrees exactly with the neutral expectation.

Timema stick insect genome analysis using GBS data

We estimated duplication levels in the stick insect genome by applying ngsParalog and dupHMM to GBS data from 435 *T. cristinae* that were sequenced and mapped to the draft reference genome by Lindtke *et al.*[28]. Without imposing any coverage requirements for a site to be considered duplicated, on average 3.8% of the analyzed sites among all linkage groups (LGs) appear to be duplicated (4% duplication including scaffolds not assigned to LGs) (Table 2.1). Duplication was highest at 6.15% for unassigned scaffolds, followed by LG1 with 5.78% of sites inferred to be duplicated. LG9 had the least duplication of 2.41%. The number of duplicated regions among the 13 LGs ranged from 89 on LG12, comprised of 64 scaffolds to 660 for LG3 with 166 scaffolds. In total we found 3,602 potentially duplicated regions among the 1,312 LG scaffolds and 1,622 possibly duplicated regions among the 536 analyzed scaffolds not assigned to any LG. The largest duplicated region was on scaffold 1157, belonging to LG1, between positions 15,980 and 91,104 (Figure 2.7).

We also analyzed the *Timema* data with dupHMM whereby we required duplicated sites to have average individual coverage that was at least 1 (semi-conservative) and 2 (conservative) standard deviations above the average among sites with $LR_{dup} = 0$, that is, sites that are likely not duplicated. This provides more conservative estimates of duplication levels and identifies high-confidence regions of duplication that should be masked from analyses that can be confounded by paralogy. The average percentage of duplication among LGs for the semi-conservative and conservative analyses were 2.3% (range 1.3-3.6%) and 1.9% (range 1.3-3.2%), respectively (Tables 2.2 & 2.3). When being most conservative we found 1,161 duplicated regions total among all of the LGs (36-220 regions per LG), which increases to 1,702 regions (43-325 regions per LG) with the semi-consevative coverage threshold. The largest potential duplicated region when using the less stringent coverage threshold is a 74,792 bp long region on scaffold 1157 nested within the largest region identified when imposing no minimum coverage threshold for duplicated sites. Under the most strict coverage requirements for duplicated sites, the largest duplicated region changes to scaffold 1326 spanning positions 181,369 to 204,012. The longest duplicated region among scaffolds unassigned to LGs stays consistent across the three coverage stringency levels, being a ~54 kb long region on scaffold 1017. Levels of duplication among unassigned scaffolds under the semi-conservative and conservative coverage requirement was 3.6% and 3.2% respectively.

Duplication detection in humans from the Phase 3 1000 Genomes dataset

To find duplication in the human genome we analyzed all unrelated individuals represented in the 1000 Genomes Phase 3 dataset for the autosomes and all unrelated females for the X chromosome. ngsParalog was used to calculate LR_{dup} for all of the biallelic SNPs in the officially released Phase 3 VCFs. We ran dupHMM using LR_{dup} and average individual sequencing coverage in a conservative manner that identifies sites as duplicated only when there is high evidence for duplication from inflated coverage and excessive heterozygosity, which would be best suited for identifying duplications for quality control purposes. Since the 1000 Genomes dataset has already undergone extensive filtering for regions that are inaccessible to next generation sequencing for reasons such as duplication, we analyzed the human genome without any previous masking by the 1000 Genomes Project, then with the pilot mask (intended to represent regions that are essentially unaccessible to short reads and so should not include any duplicated sites), and then with the strict mask (representing the most unique regions of the genome). Without any masking we found on average 0.37% of each chromosome to be duplicated (Table 2.4), which ranged from 0.13% for chromosome 20 to 1.9% for chromosome 21. The largested duplicated regions for all chromosomes spanned approximately 1 kb (chromosome 20) to 21.06 mb (chromosome 1, encompassing the centromere), with a median maximum length of ~11.5 kb.

The large region on chromosome 1 was identified by the 1000 Genomes pilot mask. With this mask applied we find on average 0.1% (range 0.05-0.4%) of each chromosome to be duplicated such that the pilot mask removes approximately two thirds of duplicated regions identified by dupHMM (Table2.5). The largest remaining regions missed by the pilot mask are on average 269 bp long, encompassing only a few SNPs. Coverage and LR_{dup} distributions for SNPs inferred to be duplicated by the HMM and that were missed by the pilot filter show minimal overlap and are right-shifted compared to nonduplicated SNPs (Figures 2.8 & 2.9) suggesting that the duplicated SNPs do in fact belong to small duplicated regions.

The largest remaining conservatively-identified duplicated region after the pilot filter is a 1.1 kb long region on the X chromosome spanning positions 452,904-454,014 (Table 2.5), which exhibits higher-than-average sequencing depth and SNPs with excess heterozygosity (Figure 2.10). This region contains multiple subregions of repeated elements according to RepeatMasker [40] and shows some enrichment of the H3K27Me3 histone mark (UCSC

Genome Browser [41], hg19). When dupHMM is run with less stringent coverage requirements (more suitable for discovery versus data quality control), the largest duplicated region in the pilot-masked human genome is a 1.8 kb long region on the X chromosome between positions 1,522,669-1,524,475 (2.7). The conservative run of dupHMM also provides evidence that this region, which spans cytokine receptor and acetylserotonin O-methyltransferase-like genes (UCSC Genome Browser [41], hg19), is duplicated (Figure 2.11).

The 1000 Genomes strict accessibility mask is intended to represent the most unique regions of the human genome. After applying this filter we find only 0.008% of each chromosome on average to be duplicated (Table 2.6), implying that the strict mask removes 98% of the potentially paralogous sites present in the unmasked data. The largest duplicated region remaining after strict masking is 240 bp long and is located on chromosome 2, while the average length for the largest duplicated regions across all chromosomes is 77 bp. This means that the majority of putatively duplicated regions missed by the strict mask are very small, mostly containing a single SNP. The coverage and LR_{dup} for duplicated sites are clearly greater than for nonduplicated sites after the strict mask (Figures 2.12 & 2.13), supporting that these missed sites are duplicated.

2.4 Discussion

Evaluation of duplication detection from simulations

Our probabilistic method for duplication detection involves first estimating the likelihood that reads at a SNP are derived from a secondary locus. This leads to a likelihood ratio test based on LR_{dup} for whether or not a site is duplicated that achieves high TPR/FPR even at low sequencing depth, e.g. 2-4X, where genotyping error would normally be too high to reliably use deviation from HWE to detect paralogs. We find based on simulations that this likelihood ratio test is superior to using sequencing depth when sample sizes are large enough to accurately estimate allele frequencies ($>$ 20 individuals). However, when there is mapping bias between paralogs LR_{dup} outperforms coverage alone even in these low sample size cases. The second feature of our method is that we use both excess heterozygosity via LR_{dup} and coverage information in an HMM to infer regions of duplication. We find from simulation that dupHMM achieves better TPR/FPR than using coverage or LR_{dup} alone under nearly all combinations of sample sizes, sequencing depth, and mapping bias between paralogs, and shows the clearest advantage over LR_{dup} and coverage under the worst sequencing scenarios, 20 individuals at 2-4X. In addition, dupHMM's almost 0% FDR is uneffected by fairly strong overdominance (s=0.1)that can mimic the signal of duplication.

Duplication detection in nonmodel species

We showed how our method for identifying paralogs can be used to discover copy number variation and obtain more reliable population genetic inference in studies using experimental approaches that are more typical for nonmodel species lacking well developed reference genomes. When there is not a draft genome but instead only a *de novo* assembly comprised of many short contigs like for the *Tamias* chipmunk exon capture dataset, based on LR_{dup} at each SNP, we can accurately statistically identify which contigs may represent paralogs. By removing nearly 4 fold of the sites from the 50% allele frequency category of the SFS that were identified as duplicated, we achieved a nearly perfect fit to theoretical expectations in the portion of the spectrum that had been distorted by paralogs. Population genetic inference using the SFS before being fixed with ngsParalog would results in increased estimates of diversity and obscure inferences of population differentiation and scans for balancing selection.

Mapped NGS data is all that is required to estimate LR_{dup}, and so our method can also be used on restriction digest datasets in the same way that it was applied to the exon capture data. In the case of studies like that of Lindtke *et al.* [28] where they were able to map GBS data to a draft reference genome comprised of linkage groups and unassigned scaffolds, dupHMM can also be used. This may be desirable as our simulations demonstrate that dupHMM nearly always achieves improved power to false discovery rates compared to LR_{dup}. By analyzing the Lindtke *et al.*[28] data we find that duplication levels in the *Timema* stick insect genome is conservatively around 1.9% and could very possibly be as high as 3.8%. Inferred levels of duplicaiton were almost always highest among scaffolds unassigned to linkage groups. It could be that some of these scaffolds are unassignable because too much of their sequence is chimeric, representing multiple regions of the genome or low complexity regions, and so detecting the highest amount of duplication among them makes sense. A number of recent papers [42, 28, 43, 44] make it clear that stick insects are becoming a very interesting species in which to address evolutionary questions, particularly now that there is a draft reference genome. Reliable inference from *Timema* genomic data will often necessitate masking potential regions of duplication, and so we provide masking position files at `https://github.com/tplinderoth/ngsParalog_masks`, which can be used for this purpose.

The regions of duplication inferred with dupHMM for data generated using restriction enzymes and then mapped to a longer contiguous reference should be considered more approximate than for true whole genome sequencing data because identifying where the start and stop coordinates of duplication occur must coincide with SNP locations proximal to restriction sites. Long regions of duplication may be inferred only because there are few restriction sites between the start and end positions. Therefore, for GBS or RAD data mapped to a longer contiguous sequence, it should be acknowledged that a section of duplication may end sooner than the inferred stop position, which becomes increasingly likely when restriction sites are sparse within the inferred region. Certainly the region proximal to a restriction site spanned by sequencing reads and corresponding to a start position for an inferred duplicated region is likely to be duplicated and so specific restriction loci can be masked.

Duplication detection in the human genome

Our analysis of the 1000 Genomes Phase 3 data geared towards quality control finds that ∼0.37% of the human genome is potentially paralogous. This estimate is primarily limited to regions in the human reference genome that are chimeric from the assembly of paralogous regions. Most notable is that we found 58,488 regions that are not included in the 1000 Genomes pilot mask and that show strong evidence of being duplicated. Inclusion of these regions in analyses could falsely inflate estimates of diversity, thereby potentially increasing mutation rate estimates, biasing inferrences based on the site frequency spectrum, and generating false positives in selection scans. Therefore, it is advisable that these regions, which primarily consist of one or a few SNPs should be removed for any analyses that would normally apply the 1000 Genomes pilot mask. The 1000 Genomes strict filter appears to do a thorough job of masking putatively duplicated sites. Nevetheless, we do find 3,119 regions (mostly consisting of single SNPs), that show signs of duplication in terms of inflated coverage and excessive heterozygosity, which are not in the strict mask. Based on simulations, when using the combination of coverage and excess heterozygosity with dupHMM, the false discovery rate is extremely low ($< 0.05\%$), and so it is likely that these low-coverage regions still may represent duplications. We think it advisable for any analyses that do intend to use only the most unique regions of the genome, to mask these regions from the inference. All of the regions that we found showing evidence for being duplicated and that are missing from the 1000 Genomes masks are contained in region files at `https://github.com/tplinderoth/ngsParalog_masks`.

2.5 Acknowledgements

We would like to thank Dorothea Lindtke, Patrik Nosil, and Zach Gompert who were extremeley helpful with sharing their *Timema* sequencing data in order to aid in the development of our method. TL was supported by NIH training grant 1T32LM1241-7.

2.6 Tables

Table 2.1. Paralogy in the *Timema cristinae* genome identified with dupHMM using no minimum coverage requirement for duplicated sites. The '% paralogy' is the percentage of SNPs analyzed by dupHMM that were inferred to be duplicated. 'Max region' gives the positions on the respective scaffold that the largest duplicated region spans. The average length and corresponding standard deviation for all of the duplicated regions found for each scaffold are shown in the 'average length' and 'sd length' columns, respectively.

linkage group	scaffold	% paralogy	number of regions	max length (bp)	max region	average length (bp)	sd length
1	1157	5.782	384	75,125	15,980 - 91,104	501	4,297.09
2	95	4.694	173	10,306	30,679 - 40,984	111	862.62
3	60	3.030	660	28,602	289,588 - 318,189	226	2,004.81
4	312	3.196	544	62,523	300,500 - 363,022	248	2,863.95
5	1020	4.277	167	51,832	168,582 - 220,413	409	4,105.77
6	251	2.649	274	61,550	33,598 - 95,147	237	3,717.68
7	104	4.783	202	22,427	464,381 - 486,807	278	2,101.18
8	707	3.892	298	36,096	292,969 - 329,064	181	2,152.90
9	688	2.413	189	12,766	98,845 - 111,610	91	940.31
10	91	2.762	202	6,723	659,027 - 665,749	78	621.47
11	469	3.004	184	15,879	431,879 - 447,757	185	1,635.04
12	793	5.416	89	6,169	93,342 - 99,510	117	689.88
13	2387	3.370	236	41,675	38,748 - 80,422	309	2,878.07
NA	1071	6.153	1,622	54,059	3,316 - 57,374	494	3,198.95

Table 2.2. Paralogy in the *Timema cristinae* genome identified using dupHMM with the requirement that duplicated sites have average coverage $\geq$ 1 standard deviation above that for sites with $LR_{dup} = 0$. The '% paralogy' is the percentage of SNPs analyzed by dupHMM that were inferred to be duplicated. 'Max region' gives the positions on the respective scaffold that the largest duplicated region spans. The average length and corresponding standard deviation for all of the duplicated regions found for each scaffold are shown in the 'average length' and 'sd length' columns, respectively.

linkage group	scaffold	% paralogy	number of regions	max length (bp)	max region	average length (bp)	sd length
1	1157	3.573	201	74,792	16,313 - 91,104	615.11	5601.99
2	95	3.192	97	1,237	30,679 - 31,915	30.59	126.96
3	257	1.769	325	20,940	594,894 - 615,833	104.09	1179.88
4	75	1.867	248	12,068	928,173 - 940,240	126.96	928.23
5	1020	2.359	84	11,597	49,474 - 61,070	150.42	1264.15
6	10	1.618	128	108	522,335 - 522,442	17.70	24.74
7	2084	2.869	96	18,911	69,497 - 88,407	243.68	1944.53
8	565	2.264	123	848	5,148 - 5,995	30.73	90.50
9	140	1.254	85	73	917,182 - 917,254	10.87	18.47
10	91	1.647	87	6,723	659,027 - 665,749	103.20	723.36
11	469	1.869	90	15,879	431,879 - 447,757	319.36	2036.62
12	793	3.589	43	6,169	93,342 - 99,510	182.40	938.25
13	2387	1.780	95	18,213	62,210 - 80,422	271.89	1919.39
NA	1071	3.565	757	54,029	3,316 - 57,344	496.01	3300.54

Table 2.3. Paralogy in the *Timema cristinae* genome identified using dupHMM with the requirement that duplicated sites have average coverage ≥ 2 standard deviations above that for sites with $LR_{dup} = 0$. The '% paralogy' is the percentage of SNPs analyzed by dupHMM that were inferred to be duplicated. 'Max region' gives the positions on the respective scaffold that the largest duplicated region spans. The average length and corresponding standard deviation for all of the duplicated regions found for each scaffold are shown in the 'average length' and 'sd length' columns, respectively.

linkage group	scaffold	% paralogy	number of regions	max length (bp)	max region	average length (bp)	sd length
1	1326	2.738	139	22,644	181,369 - 204,012	301.96	2,266.84
2	95	2.707	60	1,237	30,679 - 31,915	42.33	159.76
3	257	1.392	220	20,940	594,894 - 615,833	144.45	1,431.88
4	75	1.475	167	12,068	928,173 - 940,240	149.75	1,057.91
5	1020	2.169	53	11,597	49,474 - 61,070	234.96	1,590.89
6	10	1.485	91	108	522,335 - 522,442	22.98	26.99
7	2084	2.673	68	18,911	69,497 - 88,407	337.15	2,308.79
8	565	1.640	76	848	5,148 - 5,995	36.91	105.85
9	140	1.132	62	73	917,182 - 917,254	14.50	20.52
10	91	1.227	57	6,723	659,027 - 665,749	150.32	892.61
11	469	1.544	60	15,879	431,879 - 447,757	474.55	2,486.59
12	1262	3.231	36	389	94,038 - 94,426	33.19	66.09
13	2387	1.618	72	18,213	62,210 - 80,422	297.96	2,155.42
NA	1071	3.171	549	54,029	3,316 - 57,344	539.47	3,660.87

Table 2.4. Paralogy in the human genome conservatively identified from the Phase 3 1000 Genomes data (no mask) using dupHMM. LR_{dup} were BIC-penalized and the average individual coverage lower bound for duplicated sites was set two standard deviations above the inferred nonduplicated coverage distribution mean. The '% paralogy' is the percentage of SNPs analyzed by dupHMM that were inferred to be duplicated. 'Max region' gives the positions on the respective chromosome that the largest duplicated region spans. The average length and corresponding standard deviation for all of the duplicated regions found for each chromosome are shown in the 'average length' and 'sd length' columns, respectively.

		number	**max**		**average**	
	%	**of**	**length**		**length**	**sd**
chr	**paralogy**	**regions**	**(bp)**	**max region**	**(bp)**	**length**
1	0.5832	9,505	21,055,606	121,484,721 - 142,540,326	2,338.88	215,968.70
2	0.2479	6,976	25,458	91,796,971 - 91,822,428	61.57	575.47
3	0.1442	4,438	1,934	195,435,052 - 195,436,985	24.56	94.87
4	0.2413	5,294	55,978	9,270,201 - 9,326,178	79.21	972.70
5	0.1506	4,273	16,527	34,178,620 - 34,195,146	42.49	416.08
6	0.2372	5,289	10,120	161,044,540 - 161,054,659	35.25	219.71
7	0.2771	6,880	7,075	61,187,579 - 61,194,653	30.62	192.27
8	0.2732	4,775	3,010,095	43,835,896 - 46,845,990	737.54	43,645.12
9	0.3344	5,549	9,150	66,476,878 - 66,486,027	53.79	307.14
10	0.3463	3,963	9,740	42,535,553 - 42,545,292	51.87	251.19
11	0.1450	3,321	7,050	51,585,328 - 51,592,377	21.88	180.42
12	0.1320	2,527	51,215	95,179 - 146,393	51.06	1,028.89
13	0.1484	1,602	2,889	112,962,725 - 112,965,613	28.56	120.40
14	0.4005	4,944	5,901	19,838,213 - 19,844,113	45.52	214.37
15	0.5551	3,857	11,506	20,483,492 - 20,494,997	91.04	471.94
16	0.7333	10,661	23,605	46,386,567 - 46,410,171	29.43	304.26
17	0.3117	2,998	13,929	36,358,538 - 36,372,466	64.23	509.30
18	0.1417	1,614	10,514	44,546,370 - 44,556,883	33.89	296.98
19	0.3181	2,902	26,446	36,773,503 - 36,799,948	66.94	801.39
20	0.1280	1,371	953	20,337,722 - 20,338,674	20.03	68.41
21	1.8829	3,536	51,788	9,774,635 - 9,826,422	190.70	1,241.98
22	0.3917	2,494	8,037	20,630,977 - 20,639,013	44.30	214.00
X	0.2794	4,285	29,098	114,968,488 - 114,997,585	52.38	529.50

Table 2.5. Paralogy in the human genome conservatively identified from the pilot-masked Phase 3 1000 Genomes data using dupHMM. The 1000 Genomes Pilot mask is intended to filter out regions of the human genome that are not effectively accessible to short read sequencing. LR_{dup} were BIC-penalized and the average individual coverage lower bound for duplicated sites was set two standard deviations above the inferred nonduplicated coverage distribution mean. The '% paralogy' is the percentage of SNPs analyzed by dupHMM that were inferred to be duplicated. 'Max region' gives the positions on the respective chromosome that the largest duplicated region spans. The average length and corresponding standard deviation for all of the duplicated regions found for each chromosome are shown in the 'average length' and 'sd length' columns, respectively.

		number	**max**		**average**	
	%	**of**	**length**		**length**	**sd**
chr	**paralogy**	**regions**	**(bp)**	**max region**	**(bp)**	**length**
1	0.0901	3,908	311	144,925,427 - 144,925,737	10.92	22.40
2	0.0752	3,725	269	89,834,594 - 89,834,862	6.84	17.28
3	0.0581	2,272	230	195,447,513 - 195,447,742	8.57	19.72
4	0.0693	2,591	234	7,726,116 - 7,726,349	8.24	18.55
5	0.0672	2,478	244	251,917 - 252,160	7.79	18.92
6	0.1037	3,236	276	57,243,461 - 57,243,736	11.54	26.04
7	0.1309	4,113	291	158,870,000 - 158,870,290	7.63	18.67
8	0.0923	2,757	303	143,337,214 - 143,337,516	9.25	21.27
9	0.1030	2,635	215	39,059,787 - 39,060,001	7.92	18.26
10	0.1202	2,902	217	38,567,575 - 38,567,791	8.39	18.37
11	0.0678	1,972	201	11,268,184 - 11,268,384	5.78	14.30
12	0.0544	1,470	238	126,664,908 - 126,665,145	8.24	20.66
13	0.0578	937	147	114,453,782 - 114,453,928	7.61	16.38
14	0.1795	3,251	187	19,090,957 - 19,091,143	8.59	19.76
15	0.1566	2,349	300	22,488,047 - 22,488,346	14.61	29.40
16	0.4041	7,501	271	32,649,869 - 32,650,139	8.66	20.81
17	0.1313	2,127	243	34,456,117 - 34,456,359	8.82	21.18
18	0.0678	978	157	74,473,810 - 74,473,966	8.04	18.23
19	0.1566	1,825	212	16,180,151 - 16,180,362	7.60	16.77
20	0.0749	985	208	47,059,931 - 47,060,138	7.81	20.55
21	0.2053	1,532	155	9,583,843 - 9,583,997	9.31	17.66
22	0.1447	1,161	162	23,772,319 - 23,772,480	7.06	16.11
X	0.1161	1,783	1,111	452,904 - 454,014	15.04	54.32

Table 2.6. Paralogy in the human genome conservatively identified from the strict-masked Phase 3 1000 Genomes data using dupHMM. The 1000 Genomes strict mask is intended to identify the most unique regions of the human genome that should be most accessible to short read sequencing. LR_{dup} were BIC-penalized and the average individual coverage lower bound for duplicated sites was set two standard deviations above the inferred nonduplicated coverage distribution mean. The '% paralogy' is the percentage of SNPs analyzed by dupHMM that were inferred to be duplicated. 'Max region' gives the positions on the respective chromosome that the largest duplicated region spans. The average length and corresponding standard deviation for all of the duplicated regions found for each chromosome are shown in the 'average length' and 'sd length' columns, respectively.

chr	% paralogy	number of regions	max length (bp)	max region	average length (bp)	sd length
1	0.0016	62	91	17,211,757 - 17,211,847	4.47	12.19
2	0.0045	184	240	113,046,093 - 113,046,332	4.31	18.53
3	0.0031	105	64	195,729,979 - 195,730,042	3.93	8.08
4	0.0017	64	36	1,634,750 - 1,634,785	2.27	5.91
5	0.0045	148	105	757,975 - 758,079	3.90	10.74
6	0.0051	162	82	275,071 - 275,152	3.50	9.36
7	0.0055	137	78	54,183 - 54,260	4.13	10.36
8	0.0045	130	133	23,011,545 - 23,011,677	4.57	14.40
9	0.0026	48	46	137,873,834 - 137,873,879	3.29	7.28
10	0.0066	149	49	39,111,408 - 39,111,456	3.45	6.34
11	0.0055	143	67	18,953,270 - 18,953,336	3.38	8.76
12	0.0041	93	42	38,114,822 - 38,114,863	3.04	6.34
13	0.0030	47	38	113,222,296 - 113,222,333	4.51	7.69
14	0.0480	698	151	106,927,706 - 106,927,856	7.91	18.38
15	0.0000	2	11	28,705,645 - 28,705,655	8.50	3.54
16	0.0166	249	122	72,103,288 - 72,103,409	5.12	12.58
17	0.0243	297	115	34,472,963 - 34,473,077	6.39	16.27
18	0.0042	60	54	76,515,070 - 76,515,123	3.48	8.87
19	0.0177	137	41	21,755,570 - 21,755,610	3.40	6.81
20	0.0032	36	37	59,748,311 - 59,748,347	2.92	6.62
21	0.0010	7	8	45,631,883 - 45,631,890	2.14	2.61
22	0.0095	54	75	23,772,319 - 23,772,393	5.19	12.33
X	0.0059	107	79	1,578,618 - 1,578,696	5.52	11.77

Table 2.7. Paralogy in the human genome identified from the pilot-masked Phase 3 1000 Genomes data using dupHMM ran in a manner most suitable for discovery. LR_{dup} were not penalized and the average individual coverage lower bound for duplicated sites was set slightly more than three standard deviations below the inferred nonduplicated coverage distribution mean. The '% paralogy' is the percentage of SNPs analyzed by dupHMM that were inferred to be duplicated. 'Max region' gives the positions on the respective chromosome that the largest duplicated region spans. The average length and corresponding standard deviation for all of the duplicated regions found for each chromosome are shown in the 'average length' and 'sd length' columns, respectively.

		number	**max**		**average**	
	%	**of**	**length**		**length**	**sd**
chr	**paralogy**	**regions**	**(bp)**	**max region**	**(bp)**	**length**
1	1.1482	54,191	396	144,937,890 - 144,938,285	6.40	17.45
2	0.8600	45,592	500	133,117,550 - 133,118,049	5.53	15.82
3	0.8206	35,634	452	195,453,591 - 195,454,042	5.74	17.25
4	1.0635	45,207	278	125,536,276 - 125,536,553	5.36	14.62
5	0.8375	33,021	391	975,677 - 976,067	5.66	16.93
6	1.0138	34,913	495	57,277,858 - 57,278,352	7.08	21.47
7	1.1195	38,159	379	61,886,490 - 61,886,868	6.13	17.20
8	0.8537	28,196	393	2,224,643 - 2,225,035	5.96	17.14
9	1.1506	30,621	445	66,520,472 - 66,520,916	5.97	16.69
10	1.2148	34,715	355	46,960,620 - 46,960,974	5.92	16.59
11	0.9503	29,135	317	11,268,184 - 11,268,500	5.12	14.50
12	1.3440	38,752	425	132,964,460 - 132,964,884	5.59	17.10
13	1.0872	22,615	640	112,629,117 - 112,629,756	5.65	17.05
14	1.2696	23,219	394	19,052,769 - 19,053,162	7.06	19.73
15	1.3034	22,311	391	22,301,953 - 22,302,343	7.91	21.11
16	1.4487	27,494	338	15,065,313 - 15,065,650	7.76	20.61
17	1.7109	29,711	360	75,480,941 - 75,481,300	5.75	17.49
18	1.1991	20,148	321	37,381,549 - 37,381,869	5.58	16.46
19	2.1490	27,972	383	24,531,071 - 24,531,453	5.63	15.89
20	0.7946	10,929	356	29,506,031 - 29,506,386	5.64	17.52
21	1.1917	8,884	283	10,961,722 - 10,962,004	7.36	18.32
22	1.3463	10,959	319	45,964,384 - 45,964,702	5.73	15.91
X	1.1783	27,198	1,807	1,522,669 - 1,524,475	6.70	29.09

2.7 Figures

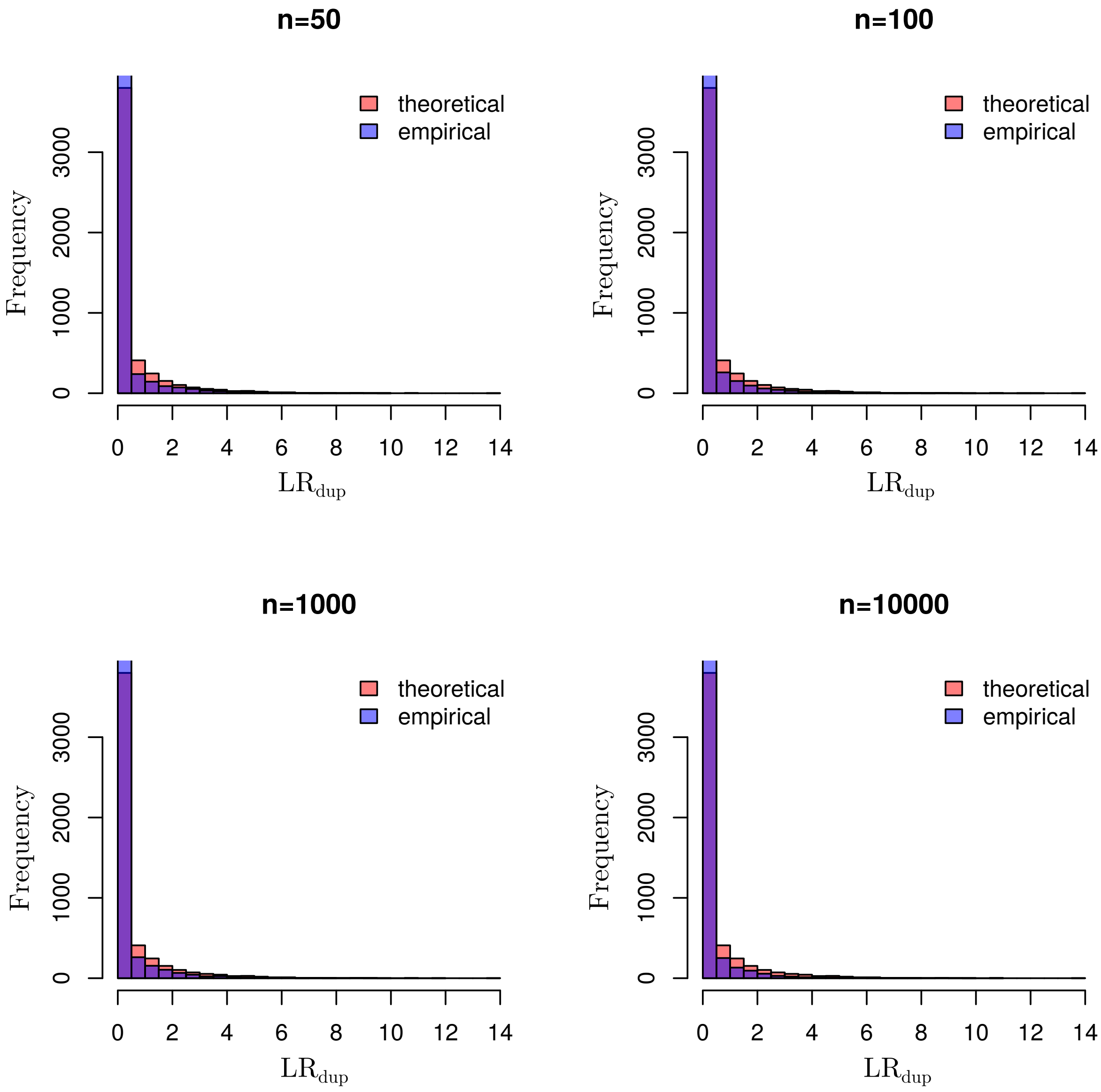


Figure 2.1. Asymptotic distribution of LR_{dup}. Empirical LR_{dup} were calculated from simulated NGS data for nonduplicated SNPs at sample sizes of 50, 100, 1000, and 10000 individuals at 8X average depth.

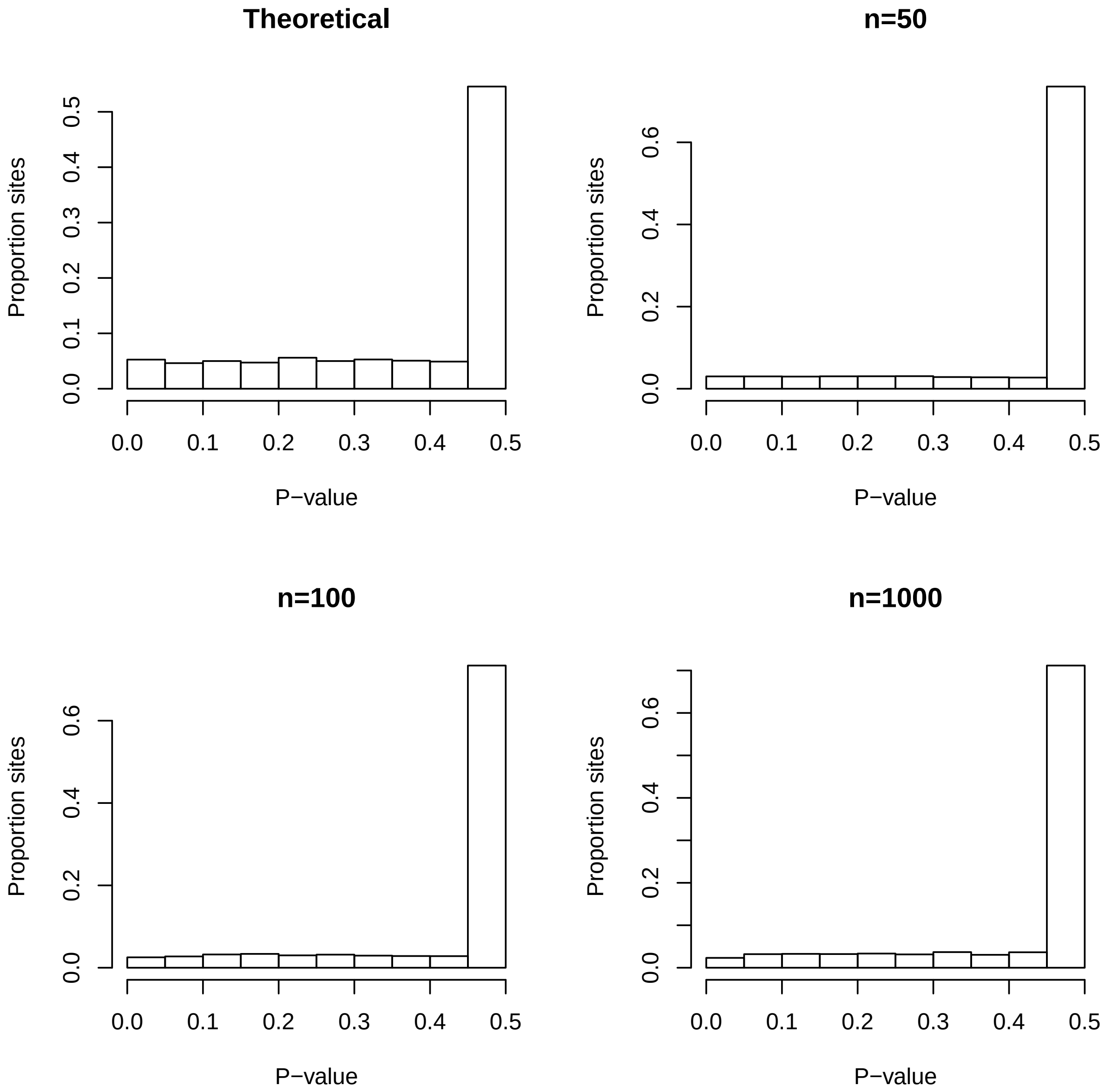


Figure 2.2. Distribution of p-values for LR_{dup} under its theoretical distribution. P-values were obtained from data simulated under the mixed 50/50 χ_1^2 and χ_0^2 distribution (theoretical) and from LR_{dup} calculated from simulated NGS data for 50, 100, and 1000, individuals respectively. The p-values are uniformly distributed over the continuous part of the distribution with a discrete mass at 0.5, corresponding to the point mass at zero.

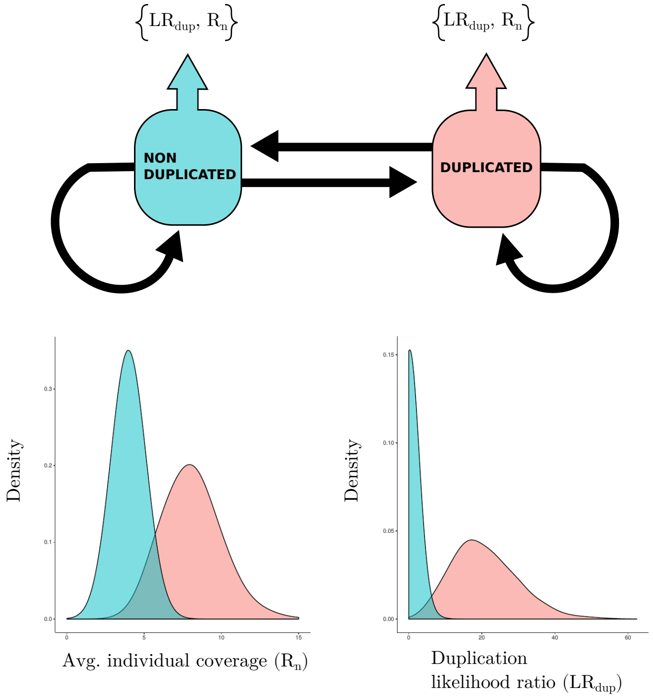


Figure 2.3. Depiction of dupHMM. The HMM is comprised of two states, nonduplicated and duplicated, and uses likelihood ratios of duplication, LR_{dup}, which are based on the probability that the collection of sequencing reads for individuals are derived from multiple regions in the genome, jointly with average individual sequencing coverage, R_n, as emissions. The two bottom plots exemplify the expected difference in the distribution for LR_{dup} and average coverage between non-duplicated (aqua), and duplicated (pink) sites.

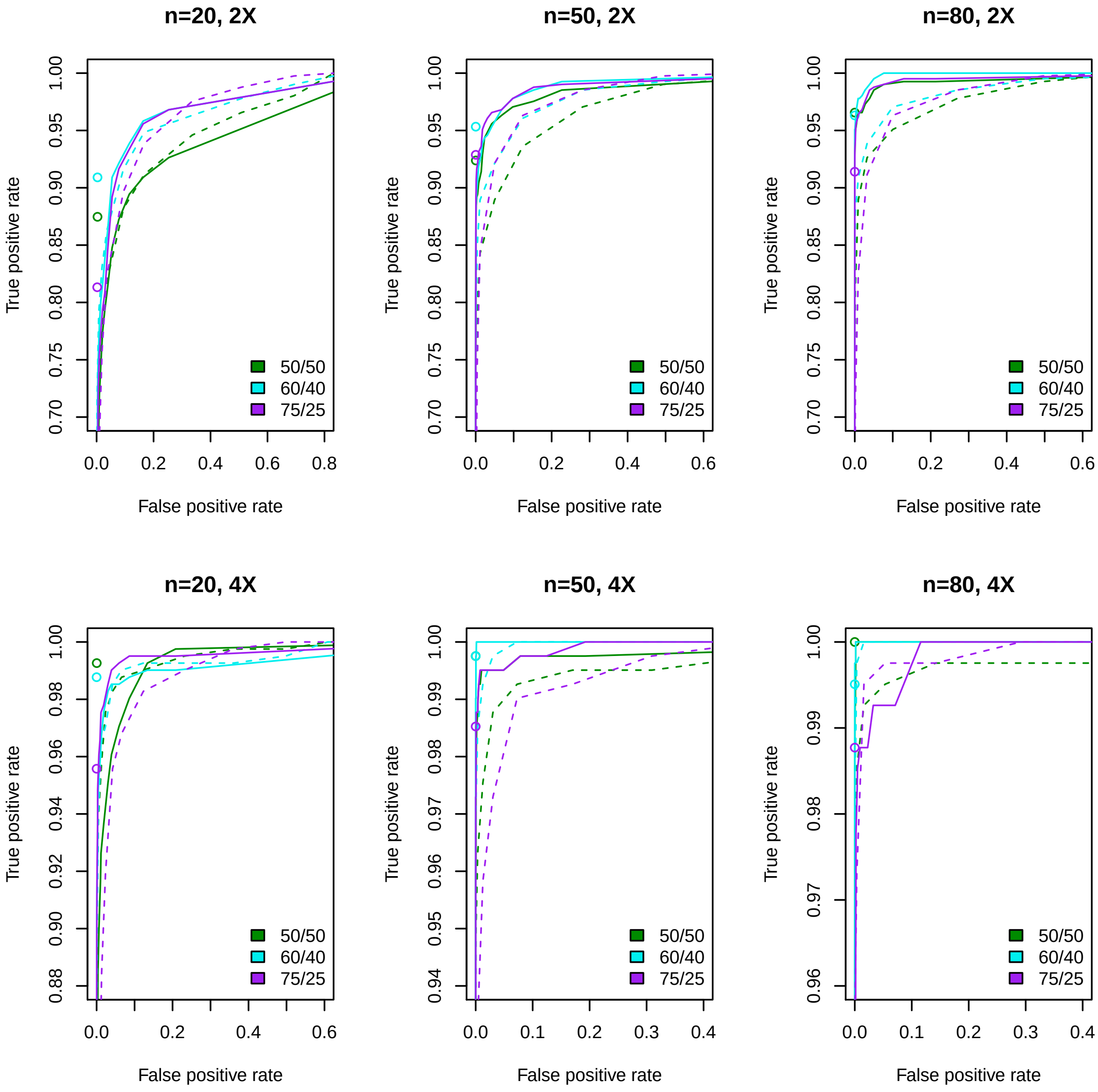


Figure 2.4. Comparison of power and false discovery from using average individual coverage (dashed line), LR_{dup} (solid line), and dupHMM (points) for identifying duplicated sites. The top row of panels is for data simulated at an average sequencing depth of 2X for 20, 50, and 80 individuals, while the bottom row shows results for data simulated at an average sequencing depth of 4X for the same sample sizes. The different colors indicate different levels of mapping bias between paralogs, where 60/40 , for example, indicates that in the pool of reads covering a duplicated site, 60% represent copy 1 and 40% represent copy 2 because fewer reads from copy 2 mapped.

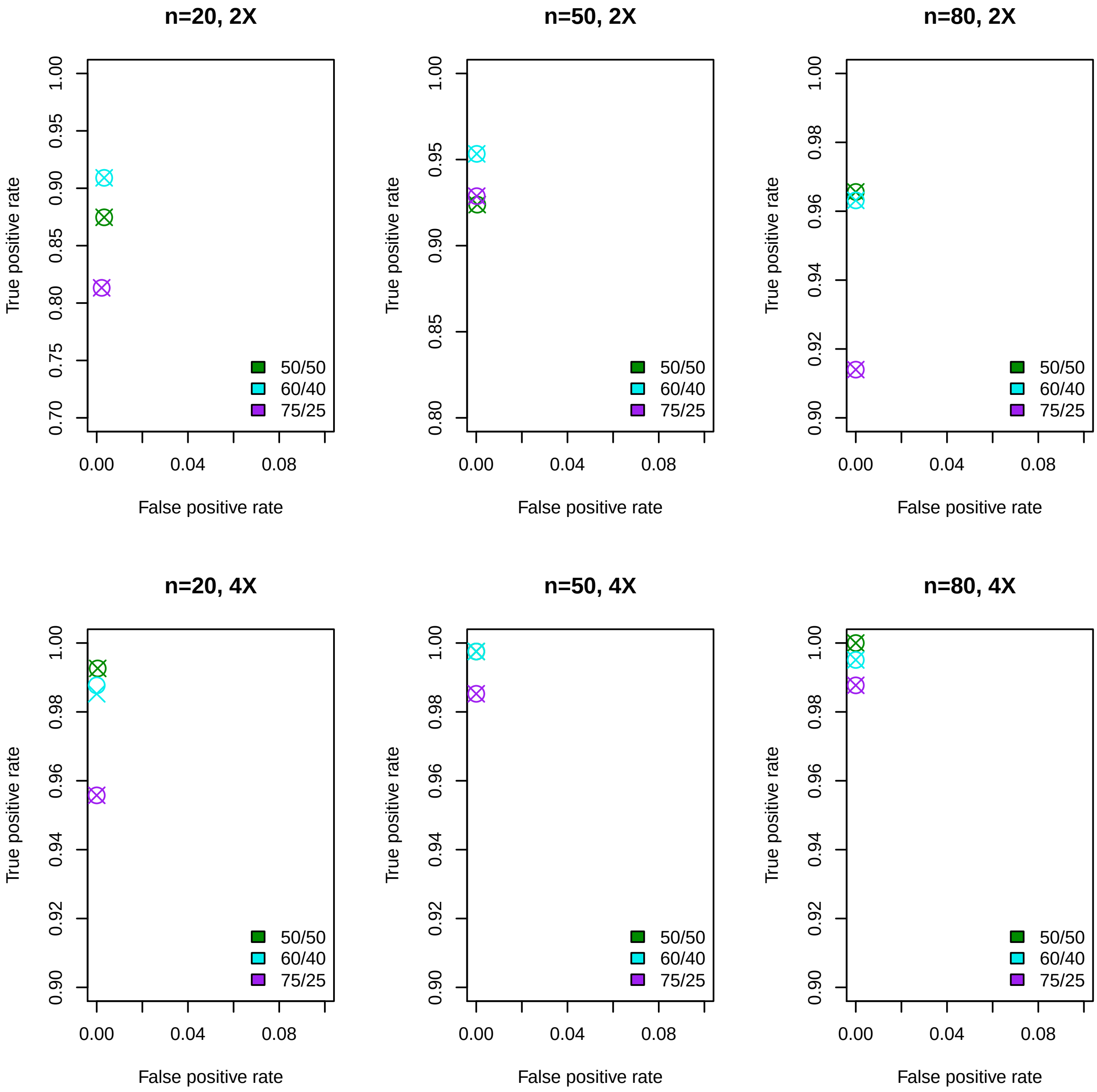


Figure 2.5. Comparison of dupHMM power and error rates for selectively neutral regions (circles) and the same regions with selection for heterozygotes introduced (X). The top row of panels is for data simulated at an average sequencing depth of 2X for 20, 50, and 80 individuals, while the bottom row shows results for data simulated at an average sequencing depth of 4X for the same sample sizes. Different mapping biases between paralogs are represented by the colors where 60/40, for example, means that in the pool of reads covering a duplicated site, 60% of reads are derived from copy 1 and 40% are from copy 2.

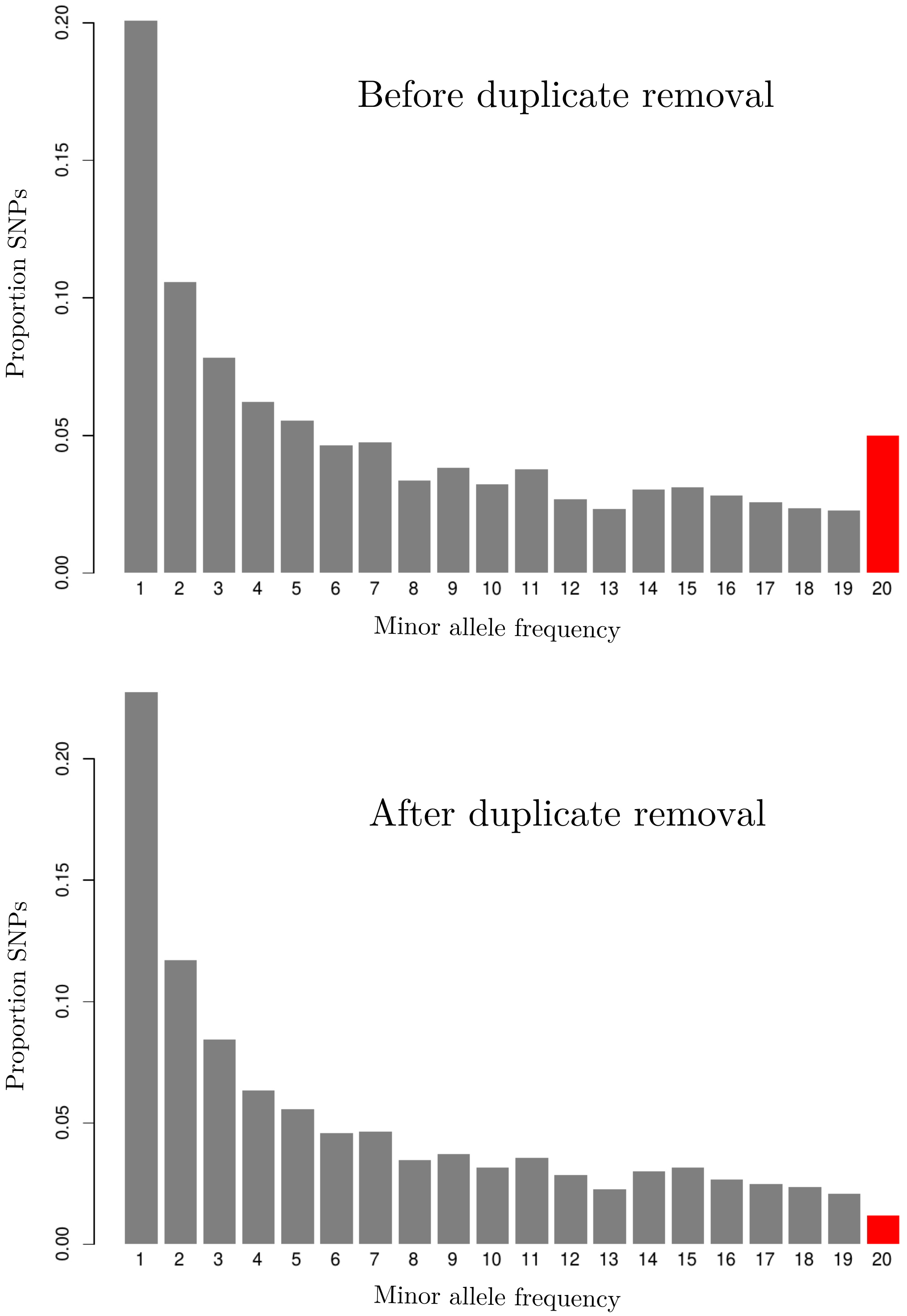


Figure 2.6. The folded site frequency spectrum from exon capture data for 20 *Tamias alpinus* chipmunks before and after sites inferred to be duplicated with ngsParalog are removed. Paralogy effects the 50% allele frequency category, which is denoted in red.

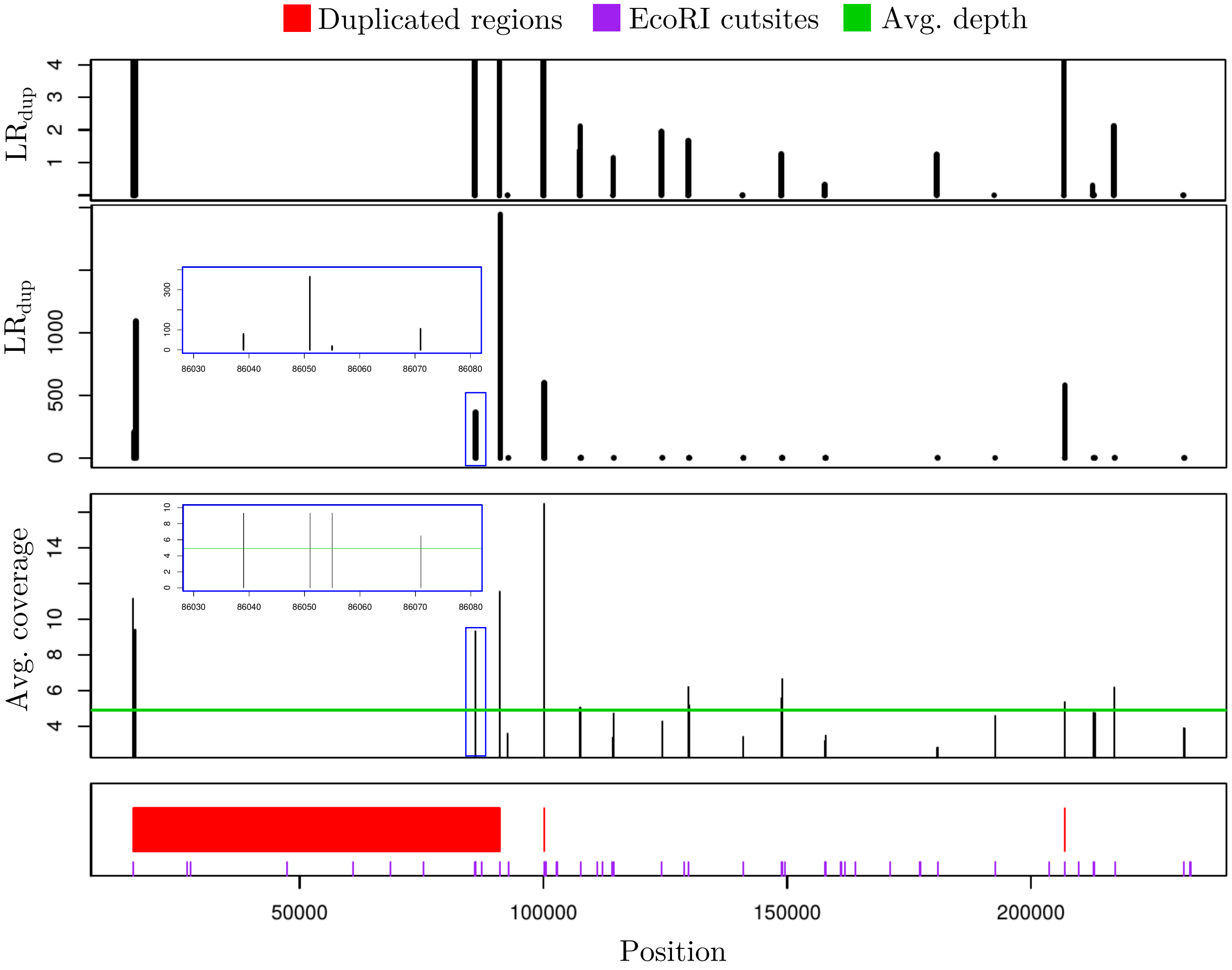


Figure 2.7. Trace of the largest paralogous region in *Timema cristinae* (deplicated by red bars), which is located on linkage group 1 scaffold 1157. The top LR_{dup} plot is a zoomed-in version along the y-axis of the one below it to offer better resolution of the LR_dup values for sites that were being overwhelmed by those with extremely large values. Since the data represented by this plot is a reduced representation of the genome using the restriction enzyme *Eco*RI, all SNPs fall within 100 bp of the restriction sites. Accordingly, the blue outlined insets are a zoomed-in view of the blue outlined areas, offering better resolution of the LR_dup and coverage values at one restriction site. The average individual sequencing coverage for this dataset is depicted by the horizontal green line.

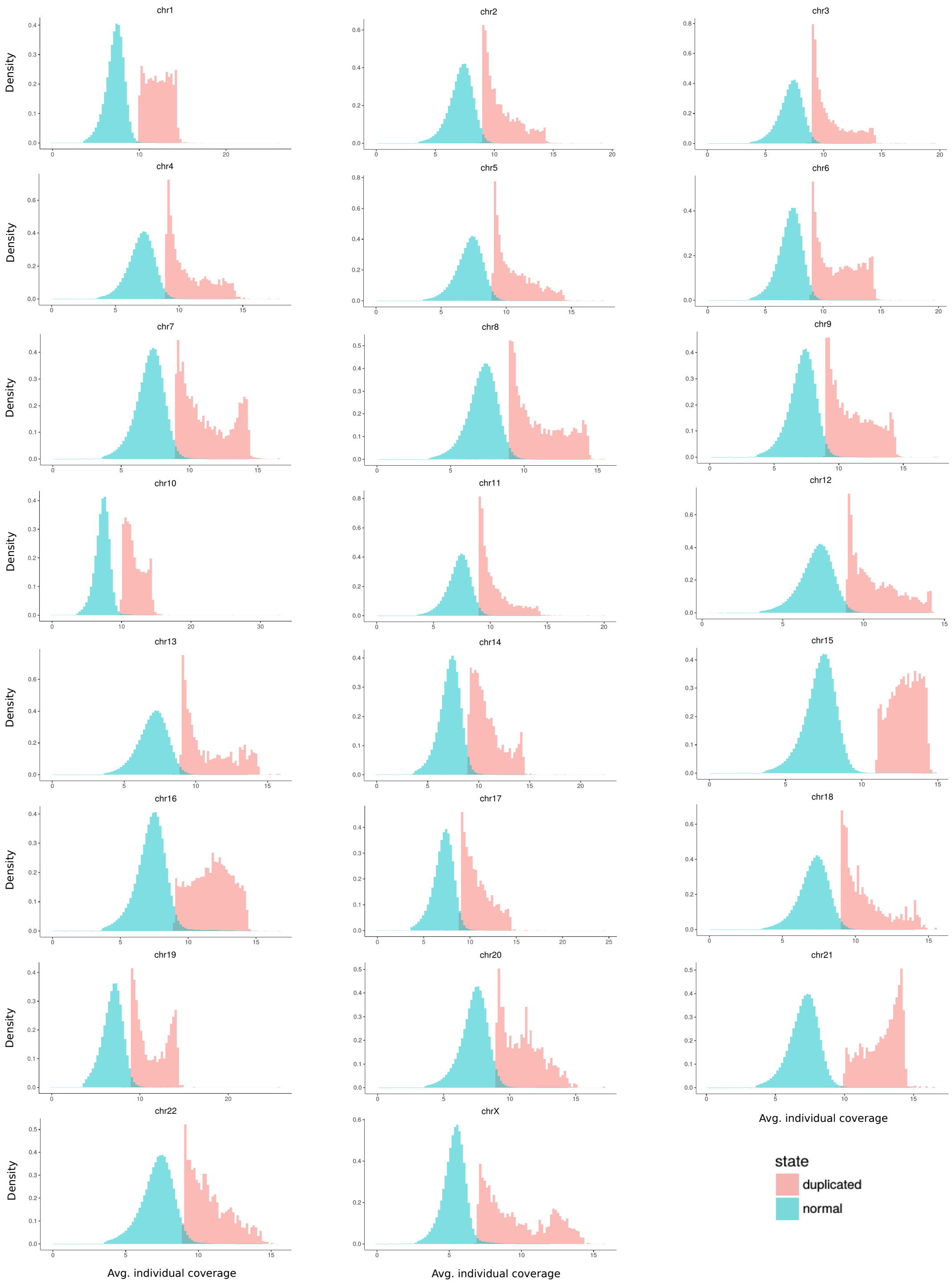


Figure 2.8. Comparison between the distributions of average individual sequencing coverage for sites conservatively identified as nonduplicated (cyan) and duplicated (pink) by dupHMM after applying the 1000 Genomes pilot accessibility filter. dupHMM was run using BIC-penalized LR_{dup} and required duplicated sites to have average individual coverage of at least ∼9X (7X for chromosome X).

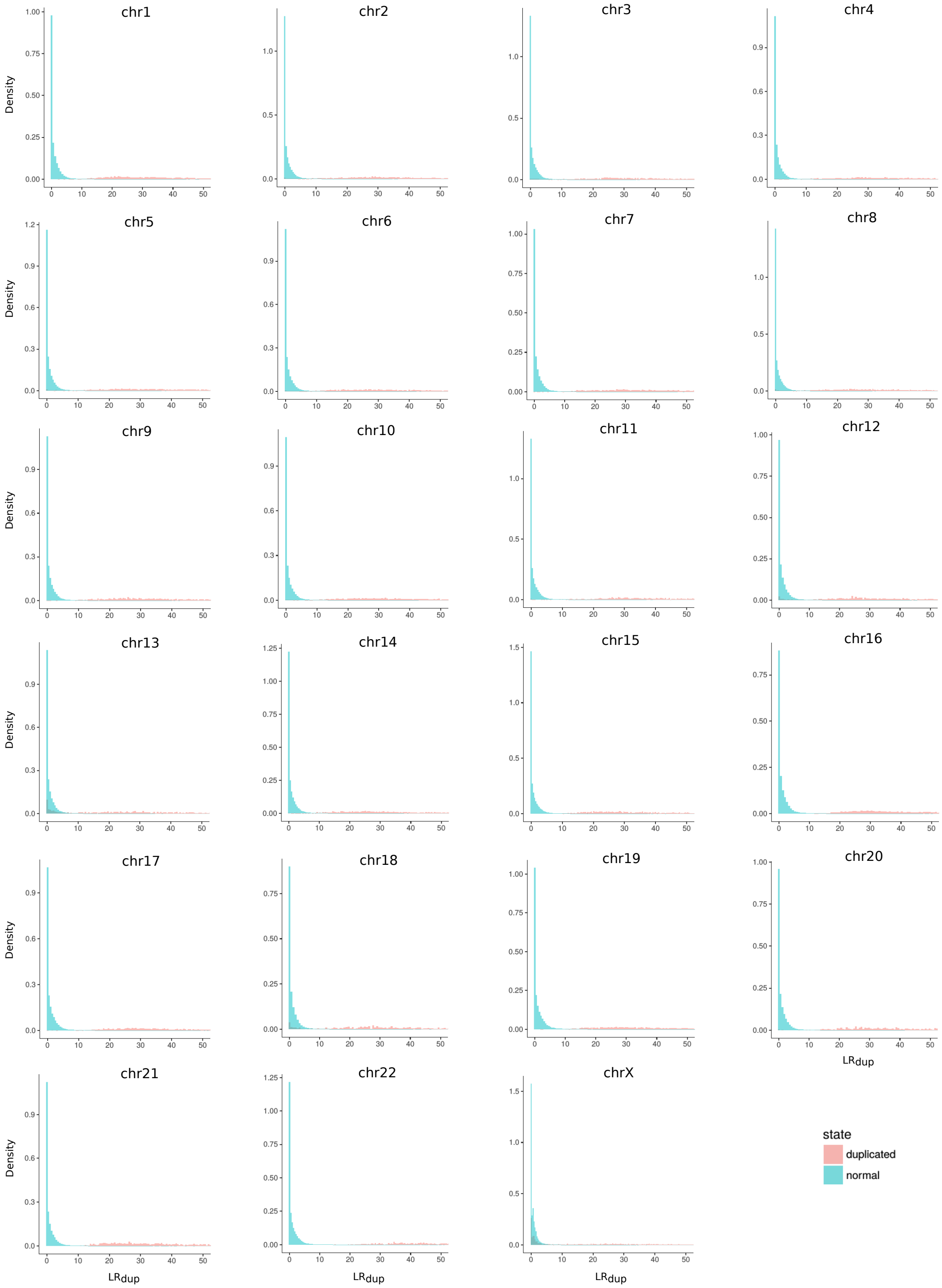


Figure 2.9. Comparison between the distributions of LR_{dup} for sites conservatively identified as nonduplicated (cyan) and duplicated (pink) by dupHMM after applying the 1000 Genomes pilot accessibility filter. dupHMM was run using BIC-penalized LR_{dup} and required duplicated sites to have average individual coverage of at least ∼9X (7X for chromosome X).

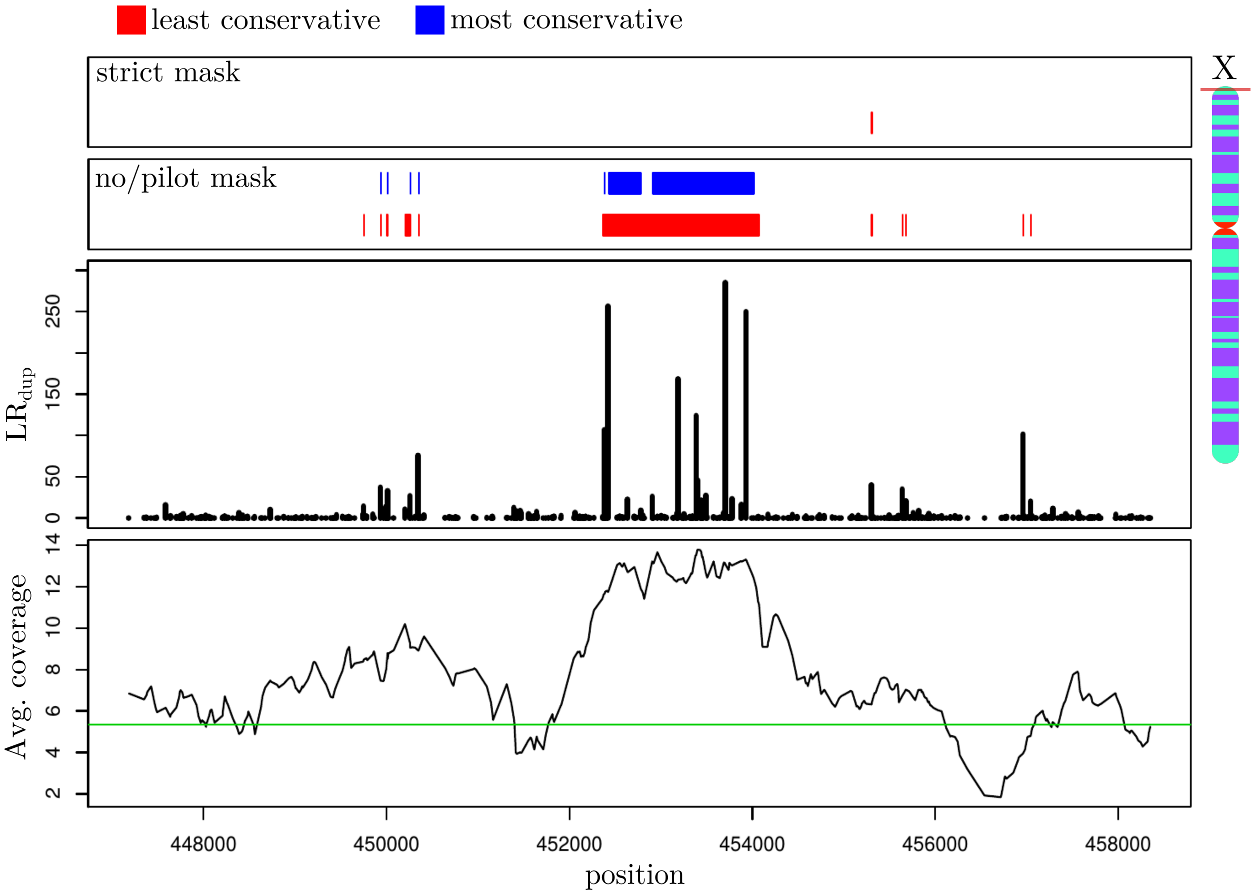


Figure 2.10. Putatively duplicated region in the human genome that was missed by the 1000 Genomes pilot accessibility filter. The ∼1.1 kb long region spanning position 452,361-454,014 on the X chromosome was identified as duplicated with dupHMM run in a conservative manner most appropriate for filtering data (blue bars) and in a way more suited for discovery (red bars, Table 2.7). The conservative run penalized LR_{dup} using BIC and set an average individual coverage lower bound for duplicated regions at 7X, while the more exploratory run did not penalize LR_{dup} and set a lower coverage bound for duplicated regions at 2.5X. The top two panels show the duplicated regions identified without any accessibility filter and after applying the strict and pilot filters (note that results were the same for no masking and after applying the pilot mask). The bottom two panels show traces of LR_{dup} calculated with ngsParalog and the average individual coverage in and around the region identified as duplicated. The horizontal green line on the coverage trace indicates the average individual coverage for chromosome X ($\sim$ 5.3X).

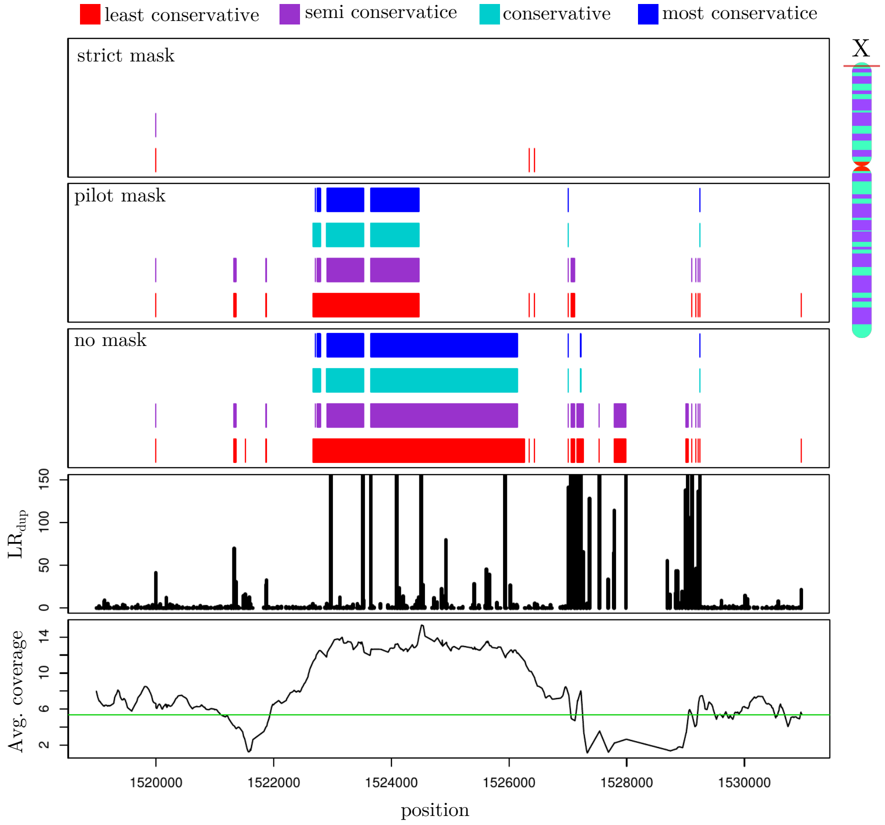


Figure 2.11. Largest region in the human genome that is likely duplicated and missing from the 1000 Genomes accessibility filters. This putatively duplicated 1.8 kb long region between positions 1,522,669 and 1,524,475 of the X chromosome span cytokine receptor and acetylserotonin O-methyltransferase-like genes. The top three panels show regions identified as duplicated with dupHMM run in four different ways. The blue regions represent the most conservative way of running dupHMM, where LR_{dup} are penalized using BIC and duplicated sites must have at least $\sim$ 7X average individuals coverage. The cyan region used the same coverage threshold but without penalizing LR_{dup}. The purple region used BIC-penalized LR_{dup} and an average individual coverage lower bound of 2.5X for duplicated sites, which is slightly more than three standard deviations below the overall average sequencing depth among all sites. The red region represents dupHMM run in the most liberal way, best suitable for discovery, in which LR_{dup} are unpenalized and the allowed minimum average individual coverage for duplicated sites is 2.5X (Table 2.7). The top three panels show regions identified as duplicated before applying any accessibility filters and after applying the pilot and strict masks. The bottom two panels are traces of LR_{dup} and the average individual sequencing coverage. The green horizontal bar in the coverage trace indicates the average coverage for chromosome X ($\sim$5.3X).

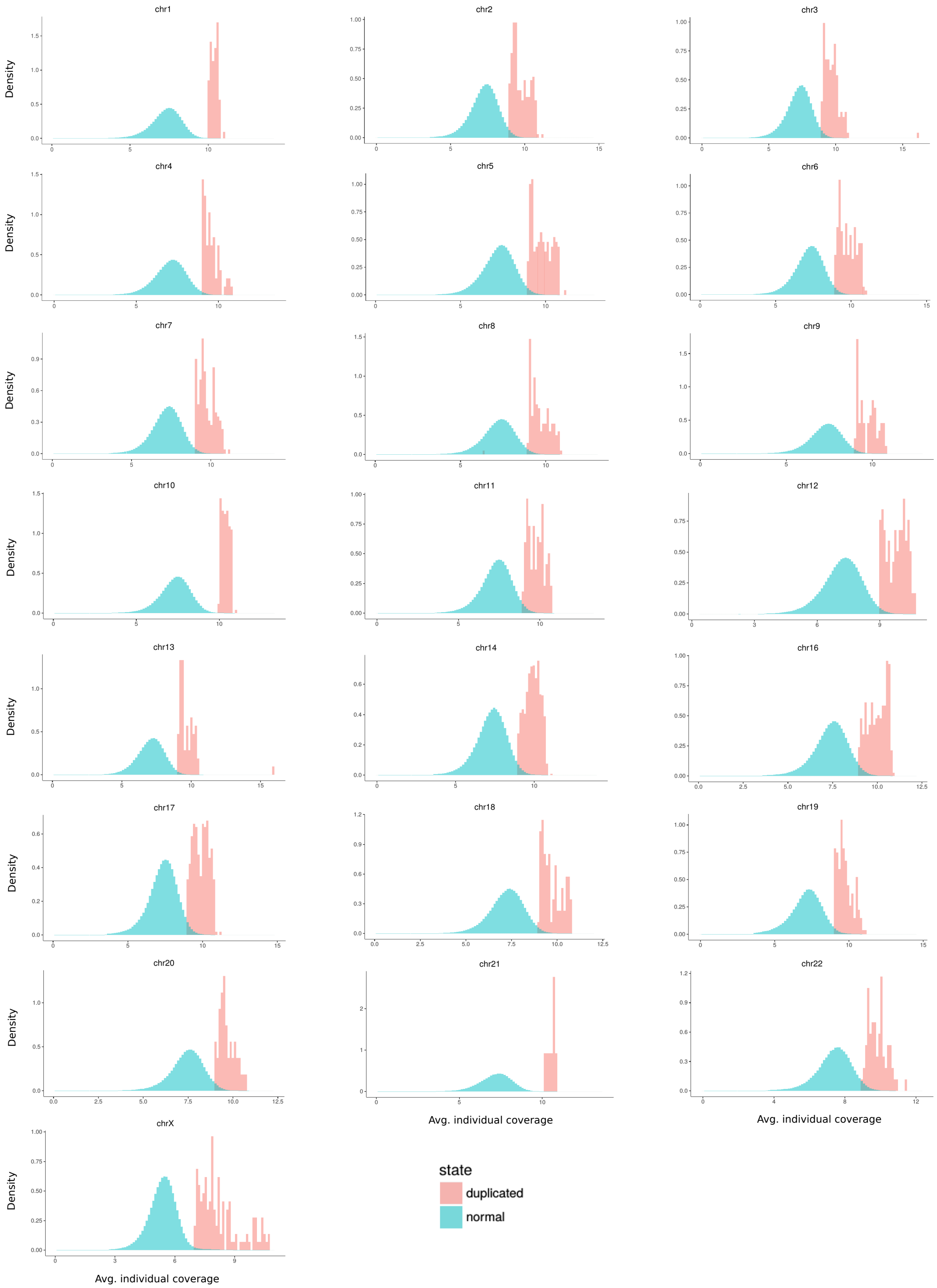


Figure 2.12. Comparison between the distributions of average individual sequencing coverage for sites conservatively identified as nonduplicated (cyan) and duplicated (pink) by dupHMM after applying the 1000 Genomes strict accessibility filter. DupHMM was run using BIC-penalized LR_{dup} and required duplicated sites to have average individual coverage of at least ∼9X (7X for chromosome X).

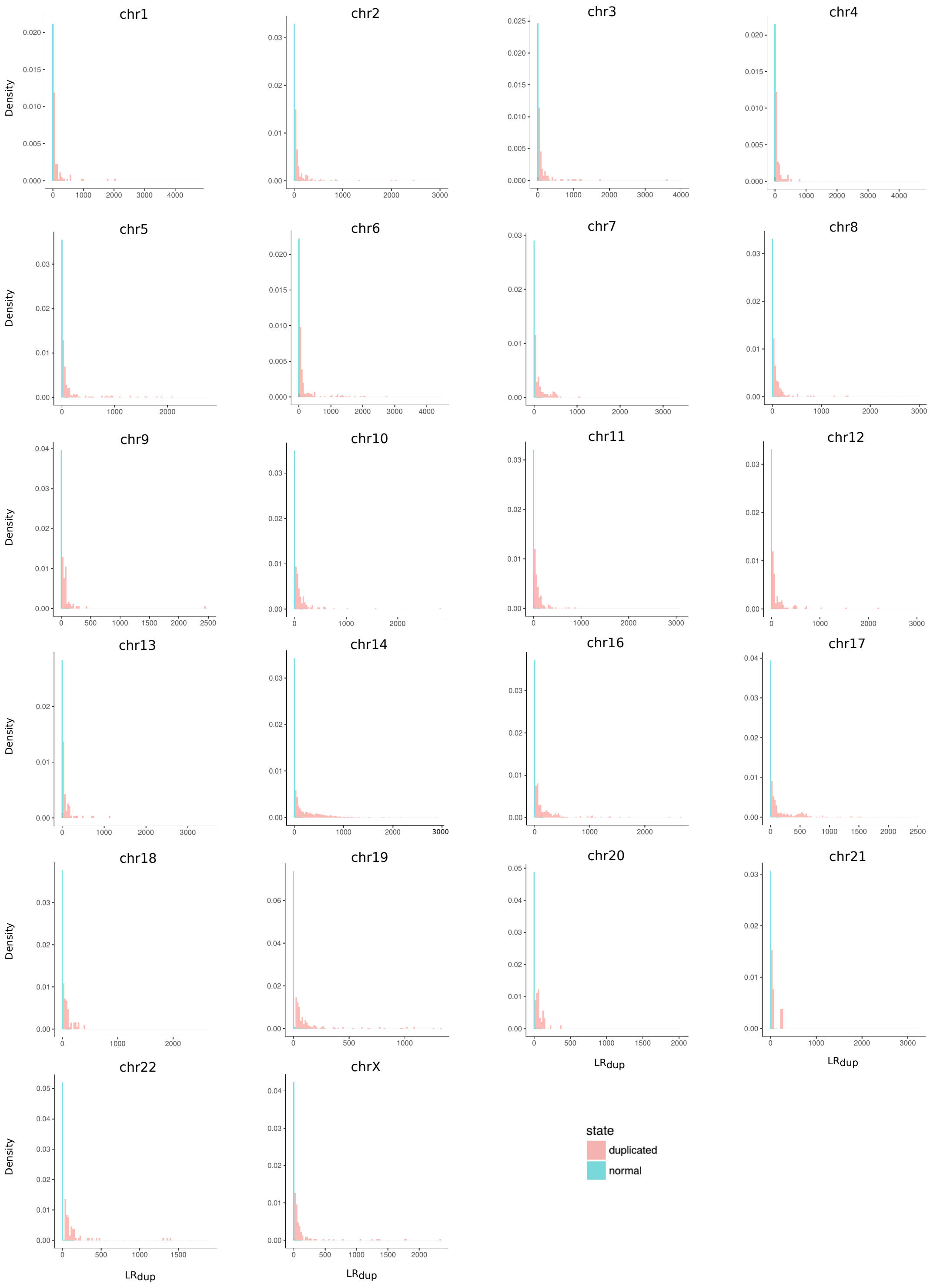


Figure 2.13. Comparison between the distributions of LR_{dup} for sites conservatively identified as nonduplicated (cyan) and duplicated (pink) by dupHMM after applying the 1000 Genomes strict accessibility filter. DupHMM was run using BIC-penalized LR_{dup} and required duplicated sites to have average individual coverage of at least ∼9X (7X for chromosome X).

3. Temporal genomic contrasts reveal rapid evolutionary responses to climate change

Tyler Linderoth*, Ke Bi*, Sonal Singhal, Dan Vanderpool, James L. Patton, Rasmus Nielsen, Craig Moritz, Jeffrey M. Good

*These authors contributed equally to this work.

3.1 Introduction

Rapid environmental change threatens global biodiversity and has already led to population decline or extirpation of many species [45, 10, 46]. Although phenotypic plasticity may enable populations to rapidly track changing climates, evolutionary adaptation will be essential for the long-term persistence of many species[47]. Disentangling plasticity from evolutionary responses ultimately requires resolving the genetic basis of adaptation. However, it remains challenging to differentiate recent or ongoing positive selection from stochastic genetic changes in populations that are also undergoing extreme demographic changes [48, 49]. Natural history museum collections may hold the key to overcoming many of these difficulties by providing crucial temporal information on species distributions, phenotypes, and population genetic variation spanning periods of recent environmental change [50, 19]. Temporal genomic contrasts have yielded powerful insights into human evolution [51, 52], but have yet to be fully leveraged to understand more recent evolutionary responses in species impacted by rapid anthropogenic climate change.

Using contrasts between early 20th century and modern museum surveys, Moritz and colleagues [10] showed that the ranges of many high elevation small mammal species in Yosemite National Park (YNP), California, USA, have retracted upward over the past century of climate change. This landmark study demonstrated the potential of using museum archives to understand community level ecological responses to climate change. Here we focus on two chipmunk species within the YNP montane mammal community. The alpine chipmunk (*Tamias alpinus*) has undergone severe elevational range retraction [10, 53] combined with pronounced shifts in diet and cranial morphology [54] across the alpine zone of YNP and elsewhere in Sierra Nevada mountains. For YNP *T. alpinus*, modern samples show increased phenotypic integration relative to historic samples across a suite of skull characters consistent with strong directional selection [55]. In contrast, the range, diet, and morphology of the partially overlapping lodgepole chipmunk (*T. speciosus*) has remained stable within YNP [53, 54]. It remains unclear why *T. alpinus* has contracted with 20th Century climate change whereas its congener has not. A previous temporal survey of eight microsatellite markers revealed increased subdivision and stable heterozygosity but declining allelic diversity in YNP *T. alpinus*, but no significant genetic changes in *T. speciosus* over the same interval [18]. Extending these basic descriptions of phenotypic and genetic variation to a detailed understanding of demographic and evolutionary responses in these species requires genomic data.

Here we build on these previous works by generating targeted genome-wide sequence data from over 300 contemporary and archived chipmunk specimens spanning a century of climate change. We then develop a novel analytical framework that allows us to both characterize general demographic responses in these species and to localize positive selection on standing genetic variation at specific genes. In addition to providing important insights into climate-induced evolutionary responses in this system, this genomic time-series approach should be broadly applicable to detecting biological responses to recent climate change given the wealth of archived specimens contained within natural history museums.

3.2 Methods

Biological samples

Tamias speciosus and *T. alpinus* surveyed in this study were collected from montane transects in Yosemite National Park (YNP) and the Southern Sierras (SS). Historic samples were collected by Joseph Grinnell and his colleagues from 1911 to 1916, and are preserved as dried skins in the Museum of Vertebrate Zoology (MVZ), at the University of California, Berkeley. Modern samples were collected from the same sites by the 'Grinnell Resurvey' team led by MVZ researchers and collaborators from 2003 to 2012 (Figure 3.1). We examined 100 YNP *T. speciosus* (52 historic, 48 modern), 104 YNP *T. alpinus* (56 historic, 48 modern), and 90 SS *T. alpinus* (52 historic, 38 modern) from each transect. We also sampled six *T. minimus* (the least chipmunk) immediately east of YNP, which were used to control for sample misidentification and potential signals of recurrent hybridization between *T. alpinus* and *T. minimus* [56]. Furthermore, we included one sample each of three outgroup species (*T. striatus*, *T. ruficaudus*, and *T. amoenus*) in order to polarize SNPs identified in our focal populations. Historic DNA was extracted from toe pad tissue (∼3 x 3 mm) in a separate dedicated laboratory using a previously described protocol [19]. DNA was extracted from modern samples using Qiagen DNeasy Blood and Tissue kits following the manufacturer's protocol. Genomic libraries for all samples were constructed following Meyer and Kircher

[57] with slight modifications [19].

Exome capture and sequencing

We previously developed Agilent SureSelect custom 1M-feature microarrays to target 11,975 exons in chipmunks [19, 58, 59] identified from multi-tissue transcriptome sequencing. We enriched and sequenced (Illumina HiSeq 2000, 100 bp paired-end) this target in 40 *T. alpinus* YNP samples. The resulting data was used to *de novo* assemble targeted exons and their flanking sequences, yielding 9,774 contiguous regions from 8,053 genes (6.9 Mb, including flanking introns and intergenic regions). In addition to these published assemblies [19], we also extracted a broad set of candidate genes from the AmiGO and NCBI mouse (*Mus musculus*) protein databases with associated functional annotations that were potentially relevant to environmental stress responses (e.g., HSP/HSF, hemoglobin, cytokines, apoptosis, immunity, oxidative stress, oxidative phosphorylation, metabolism, c-reactive protein, MHC, pyruvate, citrate cycle, T-cell signaling, glucocorticoids). We then located 2,054 orthologous transcripts (2.4 Mb) from the *Tamias* transcriptome using a BLASTx search against the mouse candidate genes. These *Tamias* transcripts were included as target sequences in our capture. Following our previous designs [19, 58], we also targeted the *T. alpinus* complete mitochondrial genome (16.5 Kb) to assess empirical error rates and potential sample contamination and five previously sequenced nuclear genes [60, 61] used as positive controls in post capture qPCR assays of global enrichment efficiency. The total target size was 9.6 Mb. In-solution capture probes were designed and manufactured by NimbleGen (SeqCap EZ Developer kits) using soft masking relative to the targeted sequences as well as the genome for the thirteen-lined ground squirrel, *Ictidomys tridecemlineatus* (Ensembl v2.68).

Barcoded genomic libraries were pooled together and hybridized in seven independent reactions with Tamias Cot-1 DNA (prepared following [62]) and barcode-specific blocking oligonucleotides. Six hybridization experiments were used for the focal species (one per time point for each of the three temporal contrasts) and one additional capture was performed on pooled libraries from six *T. minimus* and three outgroup samples (*T. striatus*, *T. ruficaudus*, and *T. amoenus*). After hybridization, each of the enriched genomic libraries were amplified using PCR and sequenced using one lane of Illumina HiSeq2000 per capture with 100-bp paired-end reads.

De novo assembly and mapping

The bioinformatic workflow that we employed for processing the *de novo* exon capture data was previously outlined [19, 63]. Briefly, raw fastq sequencing reads were treated to remove adapters, exact duplicates, low complexity, and reads sourced from bacteria and human contamination. Overlapping paired reads were merged to avoid inflating estimates of coverage and biasing downstream genotype likelihoods. The cleaned paired-end and unpaired reads for the 48 contemporary YNP *T. alpinus* genomic libraries were assembled together all at once using ABySS [64] at kmer sizes of 21, 31, 41, 51, 61, and 71. The resulting assemblies were then merged using Blat [41], CD-HIT [65], and CAP3 [66] to remove redundancies. This merged assembly was then compared to the original set of targeted sequences to obtain the assembly subset associated with targets. This in-target reference was then error-corrected using the method suggested by Bi *et al.* [19]. We aligned cleaned reads from *T. alpinus*, *T. speciosus*, and *T. minimus* individuals to the *T. alpinus* reference using NovoAlign (`http://www.novocraft.com/products/novoalign`) and retained only uniquely mapped reads. The resulting SAM format alignments were initially analyzed using SAMtools [38] and BCFtools to produce data quality control information in VCF format. To generate an outgroup reference, we aligned cleaned reads from *T. striatus*, *T. ruficaudus*, and *T. amoenus* to the error-corrected *T. alpinus* reference using NovoAlign and then used BCFtools on the alignments to generate a multi-species VCF followed by 'vcfutils.pl vcf2fq' implemented in SAMtools. We only retained sites that had data for all three focal species (*T. alpinus*, *T. speciosus*, and *T. minimus*), passed quality filters, and were monomorphic among the three outgroup samples (*T. striatus*, *T. ruficaudus*, and *T. amoenus*).

Post mapping data quality control

Quality control was applied hierarchically down from the individual level, contig level, and then to the site level (3.1). We previously described the quality control filtering in detail [19], which was carried out using the script snpCleaner (`https://github.com/tplinderoth/ngsQC/tree/master/snpCleaner`). We restricted downstream analyses to sites that passed all three levels of quality control.

Empirical error rates were measured as the percentage of mismatched bases out of the total number of aligned bases in the mitochondrial genome. On average, the empirical error rate was almost fourfold higher in historic (0.16%) than in contemporary (0.04%) samples. The observed error rates were consistent with previous findings based on data generated from array-based exon capture of museum skin DNA [19]. There were no concerning biases in empirical error rates or sequencing coverage of individual samples compared to population averages. All samples passed individual-level quality control. Within populations at the contig level, we removed contigs showing signatures of paralogy based on excessively high coverage (99th percentile) and sites that strongly deviated from Hardy-Weinberg equilibrium (HWE) proportions ($p<0.0001$). For each transect, we retained the intersection of remaining contigs between historic and contemporary populations. As a result, 2,569, 2,451, and 2,738 contigs (11.6 - 13% of the total) were eliminated from YNP *T. speciosus*, YNP *T. alpinus*, and SS *T. alpinus* datasets, respectively. At the site level, we removed sites showing excessive strand bias, end-distance bias, base quality bias, and map quality bias. We also filtered out sites with extensive missing data among samples within each population.

Errors associated with long-term DNA degradation were of particular interest in our study. DNA derived from archaeological and historic samples is usually characterized by postmortem nucleotide damage from hydrolytic deamination. This causes conversion from cytosine (C) to uracil (U) residues resulting in mis-incorporation of thymine (T) during PCR amplification [67, 68, 69]. We found elevated frequencies of C-to-T and G-to-A substitutions at the 5'- and 3'-most positions, respectively, and compared to other changes, their frequencies remained elevated throughout the sequence (Figure 3.2). These results are in strong agreement with patterns of damage accumulation characteristic of museum [19, 70] and ancient samples [68]. To mitigate the effects of misincorporation-bias on population genetic and demographic analyses [67], we took a rigorous filtering approach and eliminated all C-to-T and G-to-A SNPs (relative to the reference, not to the outgroup). In total, 9.0, 9.3, and 8.5 Mb of data from YNP *T. speciosus*, YNP *T. alpinus*, and SS *T. alpinus* passed all quality controls and was used in downstream analyses.

Population genetic analyses

For the SNPs that passed quality control we used probabilistic methods for variant discovery and allele frequency estimation as implemented in the software ANGSD [37]. These approaches account for genotyping uncertainty associated with low-medium coverage NGS data [71]. This entailed using a population-specific SFS estimated from allele frequency likelihoods as a prior to obtain allele frequency posterior probabilities. We then called SNPs using a 95% probability cutoff of being variable. We used ANGSD to estimate levels of diversity in terms of the number of segregating sites (S), Watterson's theta (θ_W), and Tajima's theta (π) in the historic and modern *T. alpinus* and *T. speciosus* populations. We calculated Tajima's D to evaluate skew in the different SFS [72]. Population differentiation within and between the modern and historic populations of *T. speciosus* and *T. alpinus* was determined using probabilistic methods for estimating F_{ST} [36] and individual covariance matrices for principle component analysis (PCA) implemented in ngsTools [73]. As an overall comparison between allele frequencies over time, we estimated the 2D-SFS between the pooled modern and pooled historic demes of each species and/or transect. SNPs identified in *T. speciosus* and *T. alpinus* individuals were polarized relative to the sampled outgroups. We further examined population genetic structure using NGSadmix [74], which estimates admixture proportions from genotype likelihoods. We ran 10 replicates for K (number of clusters) ranging from 1-10. Results across runs were summarized to determine the best K using the method of Evanno and colleagues [75]. To investigate possible hybridization between sampled *T. alpinus* and *T. minimus* samples, we used the program Admixture [76], which estimates individual ancestries from multilocus SNP data. We randomly sampled one SNP per contiguous sequence for all admixture analyses to minimize the effect of linkage disequilibrium.

Demographic inference using Approximate Bayesian Computation

Studies utilizing museum specimens often require sampling schemes that are suboptimal for population genetic analyses due to uneven sampling across time and space. ABC is particularly well suited for demographic inference under such circumstances. Simulation-based approaches allow modeling of serial sampling from populations that have experienced complex demographic histories. Further, simulations can be processed in the same way as observed data (e.g., removal of deamination SNP categories), permitting meaningful comparisons between expected and observed results. A major difficulty of ABC is the choice of statistics that sufficiently describe demographic parameters of interest. Multiple, jointly informative, summary statistics are often used to sufficiently estimate parameters while reducing the risk of any particular statistic biasing the results [77]. In many cases, the SFS is an optimal choice for fitting demographic histories as many commonly used summary statistics can be derived from it. In practice, high dimensionality and low count categories of joint site frequency spectra make them difficult to fit. Consequently, we developed an effective means of fitting binned 2D-SFS using an ABC framework that is particularly useful for inferring demographic histories from serially sampled populations or metapopulations. We applied this approach to the serially sampled chipmunk populations to test hypotheses about their demographic histories (see Figure 3.3 for a method overview).

For Yosemite populations, we modeled each sampling locality as its own deme in an island model with symmetric migration. SS *T. alpinus* was modeled in a similar manner except that, given fewer demes, we tried to fit specific pairwise migration rates according to a stepping stone model. We fitted between five and nine explicit demographic models (Figure 3.4) characterized by possible changes in migration and population size to each of the temporal contrasts. For YNP *T. alpinus* we tested models A, B, C, D, E, F, G, H, and N, for *T. speciosus* we tested A, B, D, E, G, H, J, and N, and for *T. alpinus* in the southern Sierras we tested models A, G, H, J, and N. Different subsets of our total model set were tested for different populations because it was apparent through the course of the analysis that fitting some of the nested models was redundant since the more flexible models converged on them. For each model we performed 25,000 simulations. Each simulation entailed first drawing demographic parameter values from uniform or log-uniform prior distributions and then simulating ∼20.2 Mb of sequence split evenly among 38 unlinked chromosomes for each individual under the specified history using the coalescent simulator fastsimcoal [78] assuming a mouse-based mutation rate of 2.2 x 10^{-9}, an empirically determined transition bias of 0.725, and no recombination. Lineages from the different demes were sampled at the present (modern sample) and 90 generations in the past (historic sample) according to the actual number of sampled individuals. Then all samples within a respective time period were pooled and the historic versus modern 2D-SFS was calculated. Diagonal and anti-diagonal bins of this joint SFS were then calculated using a bin width of 2. The bin width refers to the number of joint SFS categories on either side of the diagonal that are included in each bin (Figure 3.5). The joint SFS was binned in this way to reduce noise caused by trying to fit categories with no or few counts, reduce the dimensionality of the summary statistic vector, and to ensure that we fit the mass correctly everywhere throughout the spectrum. Binning in this manner fits the shape of the 2D-SFS, which should be a result of demography (barring selection).

Best fitting models where chosen as those with the highest posterior probabilities approximated with rejection sampling at a 0.8% tolerance-level after performing 1000 rounds of leave-one-out cross validation per tested model to determine if different models were distinguishable based on the 2D-SFS bins using the R [79] package ‘abc’ [80]. The same 0.8%

tolerance level was also used for the cross-validation. Under the rejection sampling framework, the posterior probability of a model is approximated by the proportion of accepted simulations determined to have come from that model when at least two models are compared. We additionally based model selection on the maximum likelihood (ML) history for each model, which is the set of parameter values that minimized the Euclidean distance between the observed and expected joint SFS bins. This allowed us to identify close-fitting histories when the rejection method posterior probability for the corresponding model may have been low as a result of fitting relatively more free parameters (a small proportion of parameter value sets may resemble the true history under a model that could potentially generate a greater array of histories at a fixed number of ABC simulations). ML parameter values for each model were used to perform 1,000 simulations with fastsimcoal to encapsulate variance due to randomness under the ML histories, and the distances between our observed and the set of simulated joint SFS bins, $D_{ML,obs}$, were calculated. This produced a distribution of $D_{ML,obs}$ for each of the models, which were compared using a Kolmogorov-Smirnov 2-sample test (KS test) to determine if models were significantly different. The model that was most probable to minimize $D_{ML,obs}$ was chosen as the most likely demographic scenario under this framework. The best models chosen using posterior probabilities versus maximum likelihood always agreed, except for one case involving model H as it was the most flexible (had the most free parameters) among the tested models, and even then, its ML history resembled the other best fitting models (see 'YNP T. alpinus demography').

Once we had chosen amongst competing models, we evaluated the fit of the selected model to our observed data by comparing the distribution of the Euclidean distance between the model's ML history joint SFS bins and the observed bins, $D_{ML,obs}$, to the distribution of the Euclidean distance between the ML history and itself, $D_{ML,pseudo}$. $D_{ML,pseudo}$ is calculated exactly like $D_{ML,obs}$ except that we use one set of joint SFS bins produced under the ML history as pseudo-observed values. To quantify the agreement between the $D_{ML,obs}$ and $D_{ML,pseudo}$ distributions we used Weitzman's coefficient of overlapping (OVL) [81], which given two probability density functions $f_1(x)$ and $f_2(x)$ defined on n-dimensional real numbers R_n is

$$OVL = \int_{R_n} \min\left\{f_1(x), f_2(x)\right\}dx$$

In our case the OVL can be interpreted as the probability of incorrectly exchanging the ML history for the true demography or vice versa as the population history producing $D_{ML,obs}$ and $D_{ML,pseudo}$. Greater OVL values indicate more similarity between the ML history and the true population history. The OVL ranges from zero to one, where an OVL value of one would indicate that the ML history is the same as the true history assuming that the demography is identifiable from the 2D-SFS bins. Lastly, we obtained posterior probability distributions for the demographic parameters under the chosen models using the standard rejection method by which we retained 8% of the parameter value sets that minimized the Euclidean distance between the simulated and observed 2D-SFS bins. The entire ABC procedure is implemented in scripts available at `https://github.com/tplinderoth/ABCutils`.

Selection inference

We considered SNPs with large allele frequency shifts between the modern and historic time periods that could not be attributed to demography as evidence for positive selection. We used the program OutFLANK [82] to detect F_{ST} outlier SNPs. This approach empirically adjusts the degrees of freedom of χ^2-distributed F_{ST} values in order to control for how demography distorts this distribution. We considered SNPs to be outliers if the false detection rate (FDR) adjusted p-value (q) for the test of neutrality was less than 0.01. We then confirmed that outlier SNPs identified through OutFLANK could not have been a consequence of demography by comparing the observed F_{ST} values for the SNPs to null distributions of F_{ST} generated under the population histories inferred through ABC. Specifically, null exome-wide and per-site F_{ST} distributions were generated by performing 1,500 simulations under each of the best fitting ML histories for YNP *T. alpinus*. These null distributions were generated for the scenario involving all sampled demes and for when the calculation of F_{ST} was limited to demes providing samples in both time periods. Comparing our observed F_{ST} values to these distributions allowed us to determine the probability that demography alone could generate such extreme values. The outlier loci identified using OutFLANK were then annotated relative to the *Mus musculus* reference genome. When the outlier nucleotide positions did not match the mouse reference sequence their locations were inferred relative to annotated exon-intron boundaries. Finally, we used Latent Factor Mixed Models (LFMM) [83] to test if outlier loci were correlated with elevation when taking underlying population structures into account. LFMM uses a hierarchical Bayesian mixed model that controls for population structure via latent factors (K), which roughly correspond to the number of clusters identified by NGSadmix (historic K = 2; modern K = 6). We performed 50 LFMM runs, each with 500,000 iterations and a 10% burn-in for both the historic and modern populations, respectively. We considered the focal loci to be significantly correlated with elevation at an FDR $q < 0.05$.

3.3 Results and Discussion

Exome capture efficiency

We designed a custom targeted capture to enrich and sequence exons from 10,000 protein coding genes (9.4 Mb) in 294 *T. alpinus* and *T. speciosus* samples and nine samples from other chipmunk species (total n=303). We sampled modern and historic (~100 year-old) populations in Yosemite National Park (YNP) for both species as well as past and present populations of *T. alpinus* in the Southern Sierras (SS), where this species has also contracted [53], with an average of 49 individuals per population (Figure 3.1). This design allowed us to compare stable and retracting species within the same montane mammal community, as well as replication within the same range-retracted species across two different areas that span its latitudinal range.

Exome enrichment was highly specific (90-93% of reads on target) and sensitive (>92-93% of the target regions sequenced), resulting in high coverage of targeted regions (26-35X average individual coverage per population). While historic DNA samples are notorious for poor technical performance, all 303 individuals yielded moderate to high coverage data with similar capture success between modern and historic samples. Analysis of mitochondrial DNA indicated that empirical error rates were ~fourfold higher in historic (0.16%) versus modern samples (0.04%), due primarily to DNA damage typical of century-old museum samples [19, 70].

Population genetic characterization

After applying a series of quality filters [19] (Table 3.1), we identified 20,395, 10,395, and 10,954 high-quality single nucleotide polymorphisms (SNPs) in YNP *T. speciosus*, YNP *T. alpinus*, and SS *T. alpinus*, respectively. Consistent with prior microsatellite results [18], we observed only a slight reduction in heterozygosity in modern versus historic samples (Table 3.2). We also quantified the degree of population genetic structure within each species by estimating the fixation index (F_{ST}) for historic and modern populations. Within YNP, population structure was relatively low overall but increased considerably in modern *T. alpinus* ($F_{ST,historic} = 0.032$, $F_{ST,modern} = 0.058$). By contrast, population structure in *T. speciosus* was more stable over time ($F_{ST,historic} = 0.027$, $F_{ST,modern} = 0.030$). These patterns were supported by a maximum likelihood analysis of population structure [74] that revealed a substantial increase from two to six genetic clusters for YNP *T. alpinus* compared to stable temporal structure (K=2) within *T. speciosus* (Figure 3.6). Principle component analyses of these data also indicate less genetic similarity among modern YNP *T. alpinus* samples (3.6). In contrast, population structure appears to have remained relatively more stable for *T. alpinus* in the Southern Sierras ($F_{ST,historic} = 0.034$, $F_{ST,modern} = 0.044$; Figure 3.6). Less detectable increases in SS *T. alpinus* population structure could reflect differences in local responses to rapid environmental changes and/or more spatially clustered samples across both historic and modern periods in this region (Figure 3.1). Finally, we found no evidence for gene flow between *T. alpinus* and neighboring (lower elevation) populations of the closely related least chipmunk (*T. minimus*; Figure 3.7), indicating that recent admixture has not impacted modern genetic diversity in this high elevation endemic [56].

Tamias chipmunk population histories

While a previous analysis of a few microsatellite loci produced qualitative insights on overall patterns of genetic variation [56] that are reinforced and extended here, reconstructing detailed demographic changes and identifying loci under selection requires more comprehensive data and analyses. To fully leverage information afforded by our temporal genomic data, we developed an Approximate Bayesian Computation (ABC) framework designed to infer population histories from serially sampled metapopulations (Figure 3.3). We constructed a two-dimensional site frequency spectrum (2D-SFS) for each pairwise temporal contrast by pooling individuals across populations within a time period (Figure 3.8). Allele frequencies should be highly correlated over such short time scales but the overall shape of this joint spectrum will depend on various demographic processes that contribute to skews in the distribution of allele frequencies. We fitted the 2D-SFS to multiple demographic models (Figure 3.4) describing changes in population size (constant size, bottlenecks, expansions) and connectivity (migration) among subpopulations or demes.

YNP *Tamias alpinus* demography

The best fitting models for YNP *T. alpinus* were B, F, H, and N, which mostly converged on a population history characterized by a relatively small, constant population size and increased fragmentation occurring within 90 generations ago. Models B, F, and N had the highest posterior probabilities (Table 3.3), while the distribution of the distance between the observed and expected 2D-SFS bins under the ML history for model H suggested that it was also a relatively good fit (Figure 3.9, Table 3.4). There was an overall high cross validation misassignment rate among nested models because they converged on similar histories that produced similar joint frequency spectra (Table 3.5). It is worth noting that model H had a low posterior probability relative to B, F, and N because it was the most flexible model such that a large proportion of histories produced under it did not resemble the B, F, and N histories and were rejected. Consequently, in the case of model H, we focused primarily on its ML history for parameter inference. The ML histories under models B, F, H, and N all produced joint spectra similar to the true history as indicated by OVL values greater than 0.81, with model F having the highest value of 0.89 (Figure 3.10).

Posterior median parameter estimates for models B, F, and N (Figure 3.11) and the ML estimate for model H indicate modern deme effective sizes of around 1,350 individuals for YNP *T. alpinus* (Table 3.6). Estimates for the strength and timing of population size change for any models which allowed for it indicated no population size change. With the exception of N (which involves constant migration), the best fitting models specify at least a two order of magnitude decrease in migration from historic rates of approximately 7e-5 (based on model B and F median values and model H's ML value). Models B and H indicate that migration slows from 20-90 generations ago, while migration entirely stops under model F 90 generations in the past (Table 3.6). Model N had the worst fitting ML history among the other best fitting models (Figure 3.10), supporting a history involving increased fragmentation as specified under the better fitting ML histories of B, F, and H.

SS *Tamias alpinus* demography

The best fitting models for *T. alpinus* in the southern Sierras were A, G, and N, which had both the highest posterior probabilities (Table 3.3) and ML histories that produced joint frequency spectra most closely resembling the observed joint SFS (Figure 3.9). The population history inferred from these three models is one with fragmentation potentially recently increasing among demes and a constant population size that is likely around three times larger than YNP *T. alpinus.* There is, however, some evidence for a potential recent, weak, population bottleneck. While there was only a 29% chance of correctly differentiating between the tested models with cross validation due to the nested nature of the models, simulations under the ML histories indicate that models A, G, and N are a significantly better fit to the observed data than the other models (all KS test p-values $<$ 2.2e-16) (Figure 3.9, Table 3.7).

The parameter posterior medians for models A, G, and N (Table 3.6) indicate that SS *T. alpinus* has a modern effective deme size of around 4,600 individuals, which has likely remained constant through time. The histories based on the parameter posterior medians for models A and G suggest constant population size over the past 90 generations, which is supported by model N and the ML history for model G, which had the highest OVL of 0.93 (Figure 3.10). The 95% credible intervals for the intrinsic shrink rate and bottleneck time for models A and G do not however exclude the possibility for a population bottleneck (Table 3.6), but if one did occur, the parameter posterior distributions indicate that it was weak ($r <$ -3e-4) and likely within the past 100 years (Figure 3.12). The historic migration rates between adjacent demes {1,2} and {2,3} were around 3e-4 and 6e-5 respectively, while the migration rate was lower between the geographically most distant demes, {1,3}, ranging from possibly around 2e-9 (based on model G and N ML estimates) to 3.5e-7 (based on posterior median estimates). The posterior median values for models A and G indicate that somewhere between 29-90 generations ago migration rates between demes decreased to around 2e-5, 1e-5, and 2e-8 for demes {1,2}, {2,3}, and {1,3} respectively. While the ML history for model A supports this increased fragmentation among all demes, the ML history for model G has migration changing only 5 generations ago suggesting a history with effectively no change in migration, which is supported by model N. This conflicting result implies that it is uncertain whether population fragmentation has changed in the southern Sierras, but if it has, it is likely a subtle increase within the past 100 years. Any potential SS *T. alpinus* demographic changes being minor is also supported by the fact that the 2D-SFS bins produced under various models were hardly distinguishable from those produced under a history with constant population size and migration, which is contrary to the case for *T. alpinus* and *T. speciosus* in YNP (Tables 3.5 - 3.10).

YNP *Tamias speciosus* demography

Models D, H, and J had the best fit for YNP *T. speciosus* in terms of both their posterior probabilities (Table 3.3) and distance of their ML histories from the observed data (Figure 3.9). The fits for these models indicate that *T. specious* is characterized by a past population expansion and a modern effective size that is likely at least three times larger than *T. alpinus* in YNP. The model fits also provide some evidence to suggest that migration among *T. speciosus* demes has also decreased recently. Cross validation indicated only a 0.19 probability of being able to correctly distinguish among models due to their nested nature, however simulations under the ML histories showed a clear and significant difference (all KS test p-values $<$ 2.2e-16) in the fits between models D, H, and J and the other tested models

(Figure 3.9, Table 3.9).

Based on the parameter posterior medians for models D, H, and J (Figure 3.13, Table 3.6), *T. speciosus* demes were expanding at an intrinsic rate of around 8e-6 until 45 - 1,234 generations ago, at which point deme sizes have remained constant at around 4,560 individuals. The migration rate posterior medians for models D, H, and J decreased from around 3e-4 to rates of 0 - 1.47e-5 within 33 generations in the past. It should be noted that the ML estimate for model J, which produced a nearly perfect fit to the observed history according to an OVL value of 0.98 (Figure 3.10), specified a decrease in migration rates from 3.42e-3 to 1.17e-8 341 generations ago, but the parameter posterior distributions for all three best fitting models (Figure 3.13) suggest that if migration has decreased, it likely happened within the last 100 generations. The ML histories for models D and H indicate constant migration and so it is not entirely certain that fragmentation has increased in YNP *T. speciosus* but the majority of evidence would suggest it has.

Comparison of population histories among *Tamias* chipmunk populations

The best fitting population history for YNP *T. alpinus* was characterized by relatively small but constant deme effective sizes through time (∼1,350 individuals) and a strong decrease in migration within the past 90 years (Figure 3.14). In contrast, both SS *T. alpinus* and YNP *T. speciosus* were found to have much larger effective deme sizes (∼4,600 and ∼4,560 individuals respectively) and higher migration rates overall. Although SS *T. alpinus* showed some evidence for a possible, recent decline in effective size (Figure 3.14), YNP *T. speciosus* was the only population that showed a strong signature of size change, which involved a historic expansion followed by constant population size (Table 3.6).

We found that modern individuals tended overall to be less genetically similar to each other than did historic individuals in all three comparisons (Figure 3.6). Consistent with this observation, the fitted demographic histories for each comparison provide some evidence for a recent (< 90 years ago) decrease in migration among demes. Decreased migration is expected if climate change is broadly affecting species within this montane community, but strong genetic structuring is only currently apparent in YNP *T. alpinus* (3.6). Long-term changes in gene flow should ultimately impact the genetic composition of metapopulations across these heterogeneous landscapes. It is possible that higher overall migration rates and larger effective population sizes have so far buffered the population genetic effects in *T. speciosus* and SS *T. alpinus*, but that genetic structure could increase over time according to the inferred histories. This is significant in that it emphasizes the importance of high-resolution demographic inference from genomic data not only for reconstructing population histories, but also as a potentially powerful tool that could enable proactive conservation management. Our ABC framework is generalizable to other temporally sampled genetic datasets, allowing high-resolution inference into demographic histories over shallow evolutionary timescales that are relevant to recent anthropogenic climate change.

Adaption of *Tamias alpinus* to climate change

We investigated specific genetic changes that might underlie adaptive responses to climate change by directly comparing genetic differences between historic and modern populations. These temporal population pairs are very closely related, however, complex population histories can result in stochastic changes in allele frequencies that confound standard signatures of positive selection [84]. We tested for individual SNPs that had undergone large frequency shifts between historic and modern populations using an approach that accounts for the confounding influence of complex population histories on the genomic distribution of F_{ST} [85, 86, 82]. We found no significant allele frequency shifts in YNP *T. speciosus* or SS *T. alpinus*. In contrast, five SNPs in YNP *T. alpinus* populations appeared as very strong outliers (false discovery rate [FDR] q-value < 0.01) relative to the inferred null distribution of per-site, genome-wide F_{ST} between the very closely related temporal populations (temporal $F_{ST} = 0.012$; Figure 3.15). To further verify the influence of positive selection on these SNPs, we compared the observed F_{ST} values to null distributions simulated under the best ABC-fitted demographic history for YNP *T. alpinus* (Figure 3.14). Our simulated F_{ST} distributions were in close agreement with the observed F_{ST} values and thus it is very unlikely that demography alone could produce the extreme changes in allele frequencies that we observed at the outlier loci (p-value $<$ 3e-7; Figure 3.16).

The five significantly differentiated SNPs showed three-fold increases in derived allele frequencies between historic and modern samples (average frequencies of 0.22 versus 0.65; Figure 3.15B) and all were located in the protein-coding gene, Arachidonate 15-Lipoxygenase (Alox15)(Figure 3.15D). Alox15 is a broadly expressed lipoxygenase involved in inflammatory responses and is upregulated in response to hypoxia [87, 88] as part of the Hypoxia-inducible factor-1α (HIF-1α) regulation pathway [89]. Two of the variants represent synonymous changes in non-adjacent exons (positions a, b; Figure 3.15D) while the three other SNPs (positions c-e) were at non-coding positions within the same intron. All five positions were in strong linkage disequilibrium (historic $r^2 = 0.86$; modern $r^2 = 0.93$) in YNP *T. alpinus* but invariant in all other populations except for one site (b) that was at similar frequency across the SS *T. alpinus* temporal contrast (historic = 0.13, modern = 0.2; Figure 3.17).

Given the extreme loss of low elevation range in YNP *T. alpinus* over the last century [53] and the potential functional association between Alox15 and hypoxia [87, 88], we reasoned that standing genetic variation at this locus might be associated with elevation (or correlated environmental variables). To test this, we first examined the frequency of derived Alox15 alleles as a function of elevation (Figures 3.15C and 3.18). Given small sample sizes at some localities, individuals were pooled into discrete elevation bands to enable more precise allele frequency estimation. For historic samples we observed a positive correlation between derived allele frequencies and 100-meter elevation bands (adjusted R^2=0.34, p-value=0.0004), while modern samples showed a negative association (adjusted R^2=0.28, p-value=0.01). However, it is important to note that allele frequencies are not independent across this landscape due to shared ancestry and gene flow [90]. To account for this, we used Latent Factor Mixed Models (LFMM) [83] to test for correlations between historic and modern genetic variation and elevation while accounting for underlying population structure. We found significant associations (FDR q < 0.05) between elevation gradients and derived allele frequencies at all five outlier positions in modern YNP *T. alpinus*, and all but one outlier SNP (position a) in historic YNP *T. alpinus*.

These correlations reveal an apparent link between genetic variation at Alox15 and elevation and suggest that the strength of selection on this variation might differ across demes at different elevations. Given range retraction and an association between derived allele frequencies at Alox15 and elevation, it is possible that higher overall frequencies in modern populations could simply reflect non-sampling of extinct low elevation populations. This interpretation is inconsistent with the observation that the magnitude of derived allele frequency changes at Alox15 vary by elevation and have been much greater at lower elevations (Figure 3.15C and 3.18). To examine this further, we repeated our temporal F_{ST} contrasts excluding low elevation sampling localities present only in the historic YNP *T. alpinus* transect. All five positions remained strong outliers in these comparisons (OutFLANK FDR q < 0.05, ABC-fitted F_{ST} distribution p-value 4e-7). Thus, evolutionary responses at Alox15 are consistent with *in situ* evolutionary change driven by selective pressures that are strongest among remnant demes at the lower bound of the modern YNP *T. alpinus* range (<3000 meters elevation). The strong association of the derived alleles with lower elevations may partially explain why we did not detect selection at Alox15 in SS *T. alpinus* where populations currently do not exist below 3200 meters and all but one of the outlier YNP SNPs were fixed for ancestral alleles.

3.4 Conclusions

By integrating high throughput sequencing, cost and time-effective targeted enrichment technologies, and sophisticated inference methods we provide powerful insights into demographic and evolutionary responses of an alpine species threatened by rapid climate change. Our unique time-series approach demonstrates how historical archives of biological specimens can unlock the potential of genomics to transform the study of climate change [50, 19]. Temporal genomic data can provide a means to understand the current state of populations and their potential evolutionary trajectories, providing powerful tools to inform the conservation of populations experiencing changing environments.

This framework also enables the detection of specific evolutionary responses in adaptive genetic variation over timescales that are usually refractory to population genomic inference. Detailed functional understanding of Alox15, and other such targets of positive selection identified using these methods, will continue to be challenging in species of conservation concern. Even in the absence of specific links to phenotypes or fitness, the identification of evolutionary responses at specific genes will inform and improve future on-ground studies focused on identifying the proximate causes of warming-related population declines across the range of this montane species. However, we suggest that the true power of this approach lies in its potential to extend across species. Though the occurrence of museum records tend to be highly punctuated through space and time for a given species, historic collection efforts, such as those led by Grinnell and other early naturalists, usually surveyed many co-distributed species. These invaluable genetic resources now enable comparative community level insights into the impacts of and evolutionary responses to rapidly changing environments.

3.5 Acknowledgements

The authors would like to thank Chris Conroy, Eileen Lacey, Karen Rowe, Kevin Rowe and Jack Sullivan for providing access to samples, A. Assis, L. Chow, P. Moore, and C. Patton for and field assistance, Sara Keeble for lab support at the University of Montana, Tali Hammond for identification of candidate genes and Maria Santos for assistance with maps. KB wants to give a special thanks to the Evolutionary Genetics Lab (EGL) and its manager, Lydia Smith, for the consistent help and support throughout the work. This research was supported by an NSERC postdoctoral fellowship (KB), University of Montana research funds (JG), and University of California Berkeley VCR-BiGCB and the Gordon and Betty Moore Foundation (RN & CM). TL was supported in part by the NIH Genomics Training Grant (Grant T32HG000047-13). Fieldwork was supported by grants from the Yosemite Conservancy, the National Park Service Inventory and Monitoring Program, the National Geographic Society (Grant 8190-07), and the National Science Foundation (DEB 0640859), permits provided by YNP and Sequoia-Kings Canyon national parks (SEKI). Additional instrumentation, laboratory, and computational support was provided by the University of Montana Genomics Core, supported by a grant from the M.J. Murdock Charitable Trust, the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303, and the Texas Advanced Computing Center (TACC) at the University of Texas at Austin. A grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01HD073439; JG) supported the development of experimental protocols for exome capture.

3.6 Tables

Table 3.1. Hierarchical data quality control protocol.

filtering at individual level
(a) Remove individuals with extremely low or high coverage (<1/5 or >5 x the average coverage across all individuals).
(b) Remove individuals with excessively high sequencing error rates measured as the percentage of mismatched bases out of the total number of aligned bases in the mitochondrial genome.
filtering at contig level
(a) Remove contigs having extremely low or high coverage relative to the empirical coverage distribution across all contigs.
(b) Remove contigs for which at least one SNP has genotype frequencies highly deviating from HardyâĂŞWeinberg equilibrium expectations ($p < 0.0001$).
(c) Retain only the contigs that pass all filters for both historic and contemporary samples.
filtering at site level
(a) Remove sites with excessively low (<1st percentile) or high (>99th percentile) coverage based on the empirical coverage distribution across all sites.
(b) Remove sites with biases associated with reference and alternate allele base quality, mapping quality and distance of alleles from the ends of reads. Also remove sites that show a bias for reads derived from the forward or reverse strand.
(c) Remove sites for which there are not at least 80% of individuals covered by at least three reads each.
(d) Remove sites with a Phred-scaled root mean square mapping quality for SNPs below 10.
(e) Due to elevated base misincorporation rates present in the historic samples, remove sites from all individuals for which C to T and G to A mutations are identified.
(f) Retain only the sites that pass all filters for both historic and contemporary samples.

Table 3.2. Population genetic summary statistics for historic and contemporary *Tamias* chipmunk populations in Yosemite National Park (YNP) and the Southern Sierras (SS).

			number				
			private			**Tajima's**	**global**
Populations	**n**	**S**	**SNPs**	$\boldsymbol{\theta_W}$	$\boldsymbol{\pi}$	**D**	$\boldsymbol{F_{ST}}$
Historic YNP *T. speciosus*	52	18,309	2,977	0.000778	0.000565	−0.3424	0.0274
Modern YNP *T. speciosus*	48	17,959	2,627	0.000601	0.000514	−0.3616	0.0296
Historic YNP *T. alpinus*	56	9,495	1,397	0.000419	0.000377	0.4330	0.0324
Modern YNP *T. alpinus*	48	8,998	900	0.000318	0.000351	0.5185	0.0575
Historic SS *T. alpinus*	52	10,211	2,232	0.000435	0.000430	0.3355	0.0344
Modern SS *T. alpinus*	38	8,722	743	0.000335	0.000380	0.5159	0.0438

Table 3.6. Statistics for ABC demographic parameter posterior and maximum likelihood estimates under the best fitting models for each *Tamias* chipmunk population. *Tamias speciosus* is appreviated as '*spec*'. The parameters $\mathrm{m_{hist}}$ and $\mathrm{m_{mod}}$ are the pairwise historic and modern migration rates between demes (when applicable, specific demes are indicated with numbers), $\mathrm{t_{mig\_change}}$ is the number of generations in the past at which migration rates change, $\mathrm{r_{grow}}$ is the intrinsic growth rate for population expansion, $\mathrm{t_{grow\_stop}}$ is the number of generations in the past that expansion stops, $\mathrm{r_{shrink}}$ is the intrinsic rate of population size decrease, $\mathrm{t_{shrink}}$ is the number of generations in the past that a bottleneck starts, $\mathrm{Ne_{mod}}$ is the modern, haploid effective size of each deme. Gray backgrounds indicate parameter values that were fixed among simulations.

parameter	**pop**	**model**	**median**	**mean**	**95% CI**	**min**	**max**	**ML**
$\mathrm{Ne_{mod}}$	YNP *alpinus*	B	2668	2659	(2382,2904)	2168	2946	2857
$\mathrm{t_{mig\_change}}$	YNP *alpinus*	B	90	90	(90,90)	90	90	90

parameter	pop	model	median	mean	95% CI	min	max	ML
m_{hist}	YNP *alpinus*	B	8.76e-5	9.28e-5	(3.63e-5, 1.74e-4)	2.64e-5	1.91e-4	1.20e-4
m_{mod}	YNP *alpinus*	B	0	0	(0,0)	0	0	0
r_{shrink}	YNP *alpinus*	B	-6.42e-7	-7.42e-6	(-5.41e-5, -1.45e-8)	-1.08e-8	-5.72e-5	-8.64e-8
t_{shrink}	YNP *alpinus*	B	90	90	(90,90)	90	90	90
t_{grow_stop}	YNP *alpinus*	B	0	0	(0,0)	0	0	0
r_{grow}	YNP *alpinus*	B	0	0	(0,0)	0	0	0
Ne_{mod}	YNP *alpinus*	F	2694	2679	(2440,2904)	2405	2963	2688
t_{mig_change}	YNP *alpinus*	F	90	90	(90,90)	90	90	90
m_{hist}	YNP *alpinus*	F	8.49e-5	9.68e-5	(4.12e-5, 1.86e-4)	3.82e-5	1.98e-4	5.14e-5
m_{mod}	YNP *alpinus*	F	0	0	(0,0)	0	0	0
r_{shrink}	YNP *alpinus*	F	0	0	(0,0)	0	0	0
t_{shrink}	YNP *alpinus*	F	0	0	(0,0)	0	0	0
t_{grow_stop}	YNP *alpinus*	F	0	0	(0,0)	0	0	0
r_{grow}	YNP *alpinus*	F	0	0	(0,0)	0	0	0
Ne_{mod}	YNP *alpinus*	H	2636	2623	(1867,3034)	1227	3177	2627
t_{mig_change}	YNP *alpinus*	H	17.5	75	(1,406)	1	494	20

parameter	pop	model	median	mean	95% CI	min	max	ML
m_{hist}	YNP *alpinus*	H	8.01e-5	2.20e-4	(2.45e-5, 1.10e-3)	1.28e-5	4.38e-3	6.78e-5
m_{mod}	YNP *alpinus*	H	1.22e-5	3.38e-4	(1.39e-7, 3.45e-3)	1.03e-7	4.78e-3	8.82e-7
r_{shrink}	YNP *alpinus*	H	-6.29e-6	-5.80e-4	(-6.78e-3, -1.34e-8)	-1.10e-8	-7.71e-3	-1.40e-5
t_{shrink}	YNP *alpinus*	H	22.5	79.28	(1,423)	1	489	19
t_{grow_stop}	YNP *alpinus*	H	1256	2287	(123,8910)	91	9990	94
r_{grow}	YNP *alpinus*	H	4.77e-7	9.02e-6	(1.14e-8,2.71e-5)	1.02e-8	1.04e-3	1.10e-8
Ne_{mod}	YNP *alpinus*	N	2664	2647	(2395,2879)	2266	2985	2868
t_{mig_change}	YNP *alpinus*	N	0	0	(0,0)	0	0	0
m_{hist}	YNP *alpinus*	N	8.18e-5	8.91e-5	(3.85e-5, 1.85e-4)	3.35e-5	2.00e-4	5.69e-5
m_{mod}	YNP *alpinus*	N	8.18e-5	8.91e-5	(3.85e-5, 1.85e-4)	3.35e-5	2.00e-4	5.69e-5
r_{shrink}	YNP *alpinus*	N	0	0	(0,0)	0	0	0
t_{shrink}	YNP *alpinus*	N	0	0	(0,0)	0	0	0
t_{grow_stop}	YNP *alpinus*	N	0	0	(0,0)	0	0	0
r_{grow}	YNP *alpinus*	N	0	0	(0,0)	0	0	0
Ne_{mod}	SS *alpinus*	A	9116	8672	(3407,10557)	1041	10860	3106
t_{mig_change}	SS *alpinus*	A	90	90	(90,90)	90	90	90

parameter	pop	model	median	mean	95% CI	min	max	ML
$m_{1,2,hist}$	SS *alpinus*	A	3.26e-4	8.82e-4	(3.31e-5, 3.96e-3)	1.95e-5	4.63e-3	1.75e-4
$m_{1,3,hist}$	SS *alpinus*	A	5.01e-7	1.44e-5	(1.58e-9;8.93e-5)	1.10e-9	9.94e-4	4.32e-7
$m_{2,3,hist}$	SS *alpinus*	A	8.57e-5	4.49e-4	(1.96e-5, 3.59e-3)	1.85e-5	4.49e-3	3.18e-5
$m_{1,2,mod}$	SS *alpinus*	A	1.23e-5	3.41e-4	(5.45e-9, 2.78e-3)	3.00e-9	4.94e-3	2.63e-9
$m_{1,3,mod}$	SS *alpinus*	A	2.15e-8	1.30e-6	(1.06e-9, 1.27e-5)	1.00e-9	4.04e-5	1.05e-9
$m_{2,3,mod}$	SS *alpinus*	A	7.94e-6	2.38e-4	(5.01e-9, 2.71e-3)	2.00e-9	4.95e-3	4.96e-9
r_{shrink}	SS *alpinus*	A	-9.82e-6	-1.20e-3	(-1.20e-2, -1.63e-8)	-1.00e-8	-2.57e-2	-1.20e-2
t_{shrink}	SS *alpinus*	A	90	90	(90,90)	90	90	90
t_{grow_stop}	SS *alpinus*	A	0	0	(0,0)	0	0	0
r_{grow}	SS *alpinus*	A	0	0	(0,0)	0	0	0
Ne_{mod}	SS *alpinus*	G	9232	8970	(5211,10495)	1884	10780	9101
t_{mig_change}	SS *alpinus*	G	29	93.3	(1,405)	1	464	5
$m_{1,2,hist}$	SS *alpinus*	G	3.53e-4	8.61e-4	(3.52e-5, 3.86e-3)	2.22e-5	4.67e-3	1.23e-4
$m_{1,3,hist}$	SS *alpinus*	G	1.77e-7	1.20e-5	(1.51e-9, 4.93e-5)	1.20e-9	5.49e-4	1.86e-9
$m_{2,3,hist}$	SS *alpinus*	G	7.42e-5	3.81e-4	(1.53e-5, 3.37e-3)	1.03e-5	4.89e-3	4.19e-5
$m_{1,2,mod}$	SS *alpinus*	G	2.53e-5	5.44e-4	(6.65e-9, 4.16e-3)	2.00e-9	4.86e-3	1.14e-4

parameter	pop	model	median	mean	95% CI	min	max	ML
$m_{1,3,mod}$	SS *alpinus*	G	2.23e-8	4.26e-6	(1.20e-9, 3.30e-5)	1.04e-9	2.75e-4	2.21e-5
$m_{2,3,mod}$	SS *alpinus*	G	1.51e-5	3.78e-4	(5.27e-9, 3.73e-3)	2.00e-9	4.84e-3	8.35e-4
r_{shrink}	SS *alpinus*	G	-1.40e-5	-2.35e-3	(-2.48e-2, -1.49e-8)	-1.00e-8	-3.79e-2	-1.57e-7
t_{shrink}	SS *alpinus*	G	17	73.28	(1,433)	1	483	2
t_{grow_stop}	SS *alpinus*	G	0	0	(0;0)	0	0	0
r_{grow}	SS *alpinus*	G	0	0	(0,0)	0	0	0
Ne_{mod}	SS *alpinus*	N	9438	9380	(7932,10674)	7134	11120	9065
t_{mig_change}	SS *alpinus*	N	0	0	(0,0)	0	0	0
$m_{1,2,hist}$	SS *alpinus*	N	3.44e-4	7.74e-4	(3.36e-5, 3.71e-3)	1.96e-5	4.39e-3	4.15e-4
$m_{1,3,hist}$	SS *alpinus*	N	3.62e-7	1.33e-5	(1.27e-9, 8.61e-5)	1.10e-9	6.46e-4	2.34e-9
$m_{2,3,hist}$	SS *alpinus*	N	6.22e-5	2.48e-4	(1.60e-5, 1.64e-3)	1.13e-5	4.91e-3	4.56e-5
$m_{1,2,mod}$	SS *alpinus*	N	3.44e-4	7.74e-4	(3.36e-5, 3.71e-3)	1.96e-5	4.39e-3	4.15e-4
$m_{1,3,mod}$	SS *alpinus*	N	3.62e-7	1.33e-5	(1.27e-9, 8.61e-5)	1.10e-9	6.46e-4	2.34e-9
$m_{2,3,mod}$	SS *alpinus*	N	6.22e-5	2.48e-4	(1.60e-5, 1.64e-3)	1.13e-5	4.91e-3	4.56e-5
r_{shrink}	SS *alpinus*	N	0	0	(0,0)	0	0	0
t_{shrink}	SS *alpinus*	N	0	0	(0,0)	0	0	0

parameter	**pop**	**model**	**median**	**mean**	**95% CI**	**min**	**max**	**ML**
t_{grow_stop}	SS *alpinus*	N	0	0	(0,0)	0	0	0
r_{grow}	SS *alpinus*	N	0	0	(0,0)	0	0	0
Ne_{mod}	YNP *spec*	D	9290	9386	(7879,11210)	7660	11770	9492
t_{mig_change}	YNP *spec*	D	33	192.4	(1,1168)	1	1918	2
m_{hist}	YNP *spec*	D	2.18e-4	2.90e-4	(2.74e-5, 9.10e-4)	1.34e-5	9.79e-4	7.64e-4
m_{mod}	YNP *spec*	D	0	0	(0,0)	0	0	0
r_{shrink}	YNP *spec*	D	0	0	(0,0)	0	0	0
t_{shrink}	YNP *spec*	D	0	0	(0,0)	0	0	0
t_{grow_stop}	YNP *spec*	D	44.5	327	(1,2105)	1	2954	4
r_{grow}	YNP *spec*	D	9.01e-6	1.07e-5	(2.79e-8, 3.01e-5)	1.00e-8	4.16e-5	1.01e-5
Ne_{mod}	YNP *spec*	J	9001	9105	(7898,10998)	7413	11840	9827
t_{mig_change}	YNP *spec*	J	19	60.86	(1,347)	1	463	341
m_{hist}	YNP *spec*	J	3.67e-4	8.73e-4	(1.83e-5, 4.09e-3)	1.28e-5	4.69e-3	3.42e-3
m_{mod}	YNP *spec*	J	7.95e-6	3.81e-4	(1.58e-8, 2.86e-3)	1.10e-8	4.91e-3	1.17e-8
r_{shrink}	YNP *spec*	J	0	0	(0;0)	0	0	0
t_{shrink}	YNP *spec*	J	0	0	(0;0)	0	0	0

parameter	**pop**	**model**	**median**	**mean**	**95% CI**	**min**	**max**	**ML**
t_{grow_stop}	YNP *spec*	J	951.5	1983	(4,8802)	3	9721	9330
r_{grow}	YNP *spec*	J	7.28e-6	1.05e-5	(3.27e-8, 4.50e-5)	1.82e-8	7.34e-5	1.47e-5
Ne_{mod}	YNP *spec*	H	9068	8979	(7171,10692)	2407	11110	9062
t_{mig_change}	YNP *spec*	H	23	72.36	(1,421)	1	478	1
m_{hist}	YNP *spec*	H	3.48e-4	8.37e-4	(2.21e-5, 3.98e-3)	1.37e-5	4.82e-3	7.82e-4
m_{mod}	YNP *spec*	H	1.47e-5	4.05e-4	(1.54e-8, 3.64e-3)	1.10e-8	4.73e-3	3.69e-4
r_{shrink}	YNP *spec*	H	-5.40e-6	-4.66e-4	(-5.41e-3, -1.53e-8)	-1.00e-8	-8.90e-3	-2.72e-7
t_{shrink}	YNP *spec*	H	18.5	65.05	(1,383)	1	471	3
t_{grow_stop}	YNP *spec*	H	1234	2446	(148,9222)	94	9949	327
r_{grow}	YNP *spec*	H	8.27e-6	1.12e-5	(3.51e-8, 3.79e-5)	1.08e-8	7.30e-5	6.77e-6

Table 3.3. ABC demographic model posterior probabilities approximated using rejection sampling whereby the closest fitting 8% of simulations were retained.

population	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**J**	**N**
YNP *speciosus*	0.0425	0.0231	-	0.2106	0.0619	-	0.0862	0.2394	0.2394	0.0969
YNP *alpinus*	0.0311	0.2011	0.0861	0.0667	0.1222	0.1794	0.0561	0.0478	-	0.2094
SS *alpinus*	0.2040	-	-	-	-	-	0.2120	0.1780	0.1700	0.2360

Table 3.4. Pairwise comparisons between $D_{ML,obs}$ empirical CDFs for the different demographic models fitted using ABC to YNP *Tamias alpinus* populations in terms of the 2-Sample Kolmogorov-Smirnov (KS) test. KS test statistic values and their corresponding p-values are displayed in the blue and red rows, respectively.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**N**
A	0.000	0.223	0.112	0.028	0.127	0.148	0.090	0.115	0.410
B		0.000	0.139	0.203	0.105	0.103	0.265	0.305	0.246
C			0.000	0.100	0.063	0.048	0.168	0.197	0.365
D				0.000	0.125	0.132	0.087	0.114	0.429
E					0.000	0.062	0.197	0.212	0.309
F						0.000	0.186	0.231	0.325
G							0.000	0.105	0.497
H								0.000	0.491
N									0.000
A	1	<2.2e-16	7.12e-6	0.828	1.98e-7	6.14e-10	6.07e-4	3.61e-6	<2.2e-16
B		1	8.13e-9	<2.2e-16	3.26e-5	4.94e-5	<2.2e-16	<2.2e-16	<2.2e-16
C			1	9.08e-5	0.03778	0.1995	1.11e-12	<2.2e-16	<2.2e-16
D				1	3.28e-7	5.42e-8	0.001032	4.54e-6	<2.2e-16
E					1	0.04282	<2.2e-16	<2.2e-16	<2.2e-16
F						1	1.89e-15	<2.2e-16	<2.2e-16
G							1	3.26e-5	<2.2e-16
H								1	<2.2e-16
N									1

Table 3.5. Confusion matrix for the demographic models fit for YNP *Tamias alpinus*. The matrix was constructed using leave-one-out cross validation with rejection sampling at a tolerance of 0.008. The matrix shows how well the different models can be differentiated based on 2D-SFS bins. The number of cross validation rounds under the different models are represented by the rows and the number of times each model was assigned as having the highest posterior probability of producing the simulated 2D-SFS bins are represented by the columns.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**N**
A	519	54	64	5	182	68	62	0	46
B	1	335	24	3	208	274	13	0	142
C	55	143	276	52	209	138	62	12	53
D	75	123	252	113	173	144	60	12	48
E	66	161	85	7	399	178	15	0	89
F	8	315	32	3	179	308	2	0	153
G	94	120	155	3	240	132	179	0	77
H	103	106	218	92	156	105	113	46	61
N	3	297	26	2	236	251	12	0	173

Table 3.7. Pairwise comparisons between $D_{ML,obs}$ empirical CDFs for the different demographic models fitted using ABC to SS *Tamias alpinus* populations in terms of the 2-Sample Kolmogorov-Smirnov (KS) test. KS test statistic values and their corresponding p-values are displayed in the blue and red rows, respectively.

	A	**G**	**H**	**J**	**N**
A	0.000	0.078	0.314	0.473	0.173
G		0.000	0.265	0.430	0.129
H			0.000	0.275	0.213
J				0.000	0.424
N					0.000
A	1	0.004558	<2.2e-16	<2.2e-16	2.01e-13
G		1	<2.2e-16	<2.2e-16	1.19e-7
H			1	<2.2e-16	<2.2e-16
J				1	<2.2e-16
N					1

Table 3.8. Confusion matrix for the demographic models fit for SS *Tamias alpinus*. The matrix was constructed using leave-one-out cross validation with rejection sampling at a tolerance of 0.008. The matrix shows how well the different models can be differentiated based on 2D-SFS bins. The number of cross validation rounds under the different models are represented by the rows and the number of times each model was assigned as having the highest posterior probability of producing the simulated 2D-SFS bins are represented by the columns.

	A	**G**	**H**	**J**	**N**
A	243	71	66	148	472
G	181	86	60	170	503
H	97	51	145	328	379
J	66	49	128	356	401
N	63	71	49	181	636

Table 3.9. Pairwise comparisons between $D_{ML,obs}$ empirical CDFs for the different demographic models fitted using ABC to YNP *Tamias speciosus* populations in terms of the 2-Sample Kolmogorov-Smirnov (KS) test. KS test statistic values and their corresponding p-values are displayed in the blue and red rows, respectively.

	A	**B**	**D**	**E**	**G**	**H**	**J**	**N**
A	0.000	0.053	0.819	0.033	0.051	0.843	0.838	0.109
B		0.000	0.821	0.070	0.032	0.844	0.842	0.084
D			0.000	0.801	0.848	0.047	0.064	0.873
E				0.000	0.061	0.824	0.824	0.114
G					0.000	0.869	0.866	0.073
H						0.000	0.049	0.886
J							0.000	0.880
N								0.000
A	1	0.1205	<2.2e-16	0.6476	0.1483	<2.2e-16	<2.2e-16	1.38e-5
B		1	<2.2e-16	0.01489	0.6852	<2.2e-16	<2.2e-16	0.001724
D			1	<2.2e-16	<2.2e-16	0.2193	0.03328	<2.2e-16
E				1	0.04842	<2.2e-16	<2.2e-16	4.54e-6
G					1	<2.2e-16	<2.2e-16	0.009698
H						1	0.1811	<2.2e-16
J							1	<2.2e-16
N								1

Table 3.10. Confusion matrix for the demographic models fit for YNP *Tamias speciosus.* The matrix was constructed using leave-one-out cross validation with rejection sampling at a tolerance of 0.008. The matrix shows how well the different models can be differentiated based on 2D-SFS bins. The number of cross validation rounds under the different models are represented by the rows and the number of times each model was assigned as having the highest posterior probability of producing the simulated 2D-SFS bins are represented by the columns.

	A	**B**	**D**	**E**	**G**	**H**	**J**	**N**
A	263	304	195	34	105	6	13	80
B	262	314	185	32	106	5	21	75
D	165	224	260	26	90	60	126	49
E	204	270	207	28	103	46	48	94
G	181	220	153	17	165	56	78	130
H	138	138	180	18	99	145	214	68
J	113	147	184	21	98	139	225	73
N	182	210	168	20	121	63	97	139

3.7 Figures

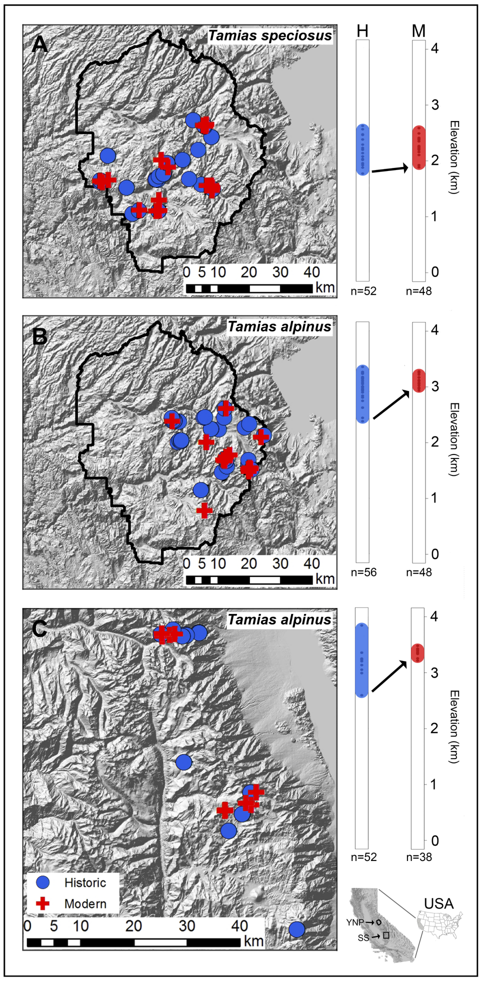


Figure 3.1. Historic and modern sampling localities. *Tamias speciosus* and *T. alpinus* specimens were collected from Yosemite National Park (A, B) and the Southern Sierras (C). Historical sampling localities (1911-1916) are shown in filled blue circles and modern (2003-2012) in filled red crosses. The vertical bar plots in the right panel show the lower elevation (kilometers) range changes of the species over the past century with sequenced samples indicated by dots within the bars. Upper elevation limits were determined for *T. speciosus* but the trapping resurvey was not designed to ascertain upper limits for *T. alpinus* from either transect area. Sample sizes (n) are indicated under the elevation bars. While *T. speciosus* maintained a stable range, *T. alpinus* populations have severely contracted upwards in elevation throughout their distribution. The inset (bottom right) USA map shows the state of California with Yosemite National Park (northern) and the Southern Sierras (southern) study areas outlined.

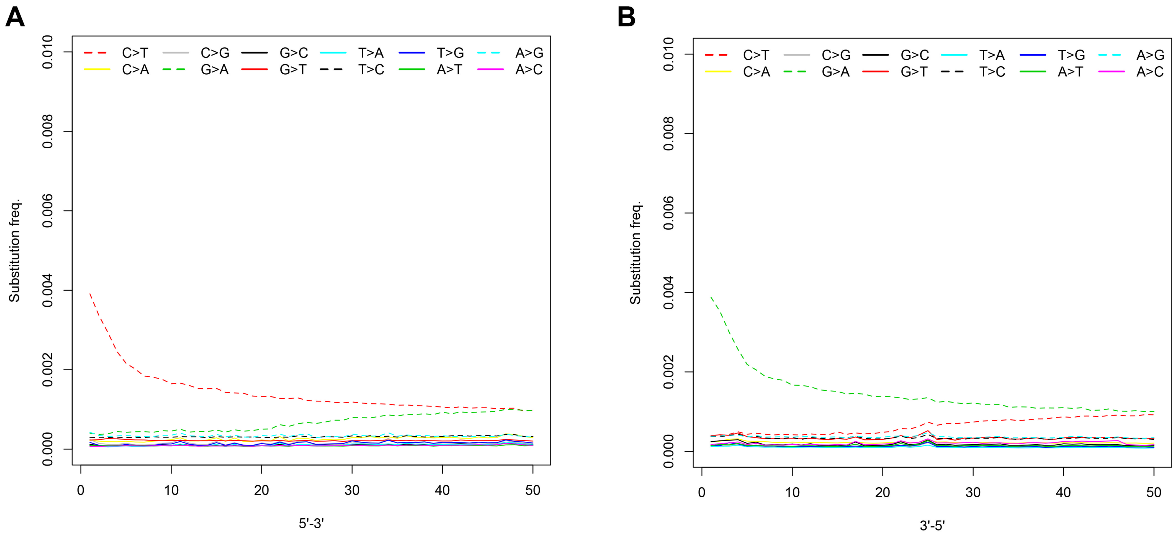
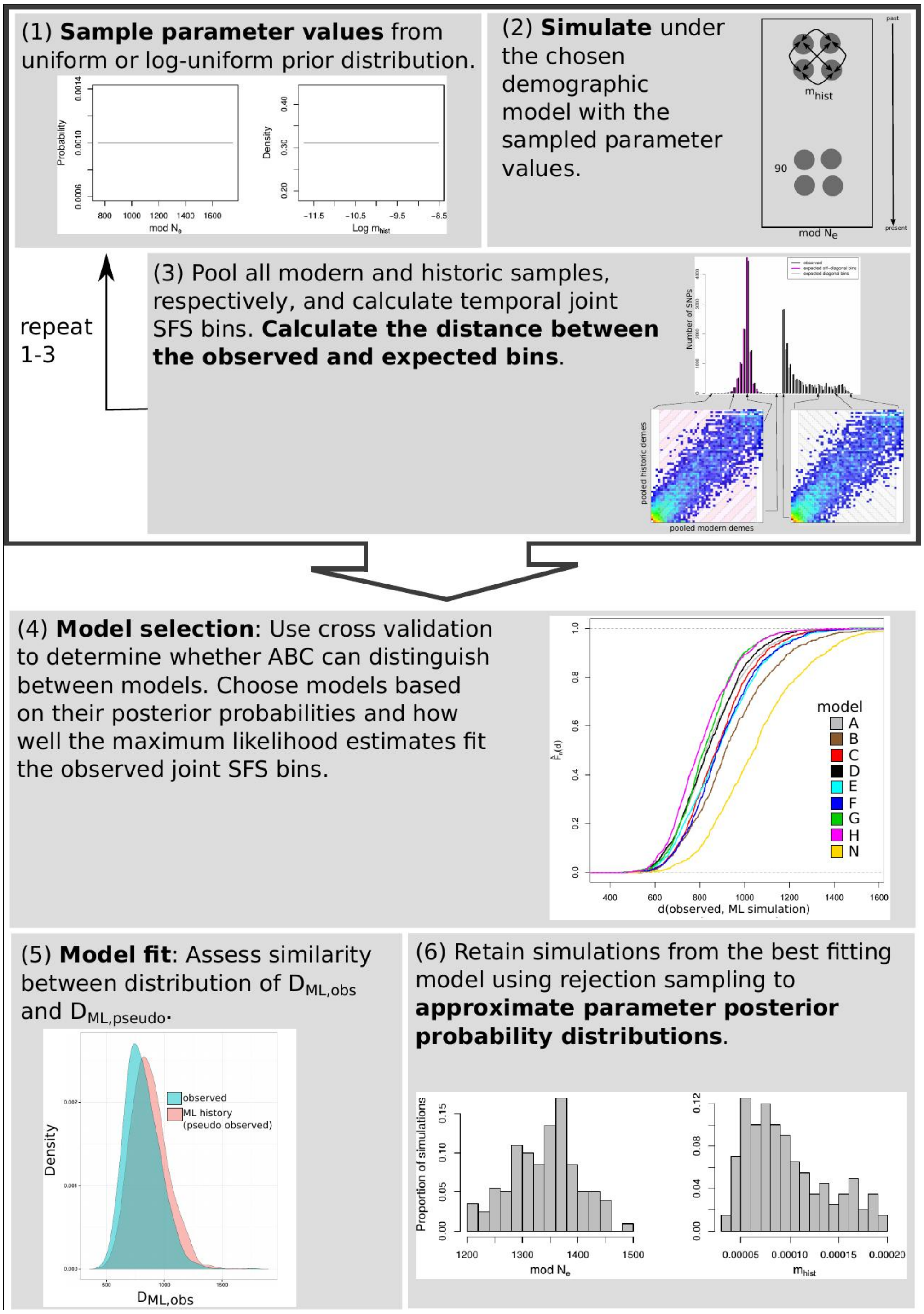


Figure 3.2. Patterns of base misincorporations in historic samples. The frequencies of the 12 types of substitutions (y-axis) are plotted as a function of distance from the 5' and 3'-ends of the DNA molecules (x-axis). The first 50 bp of the reads are shown. The substitution frequency of each particular type is calculated as the proportion of a particular alternative (non-reference) base type at a given site along the read, and is coded in different colors and line patterns explained at the bottom of the plots: 'X -> Y' indicates a change from reference base type X to alternative base type Y.



(1) **Sample parameter values** from uniform or log-uniform prior distribution.

(2) **Simulate** under the chosen demographic model with the sampled parameter values.

(3) Pool all modern and historic samples, respectively, and calculate temporal joint SFS bins. **Calculate the distance between the observed and expected bins**.

repeat 1-3

(4) **Model selection**: Use cross validation to determine whether ABC can distinguish between models. Choose models based on their posterior probabilities and how well the maximum likelihood estimates fit the observed joint SFS bins.

(5) **Model fit**: Assess similarity between distribution of $D_{ML,obs}$ and $D_{ML,pseudo}$.

(6) Retain simulations from the best fitting model using rejection sampling to **approximate parameter posterior probability distributions**.

Figure 3.3. Diagram of the demographic inference method based on fitting bins of the 2D-SFS with Approximate Bayesian Computation (ABC) that was used to infer the *Tamias* population histories. For assessing model fit in step 5, $D_{ML,obs}$ is the Euclidean distance between the observed 2D-SFS bins and bins from the maximum likelihood (ML) history under the chosen model, while $D_{ML,pseudo}$ is the same except that the bins from a single simulation under the ML history serve as the observed bins.

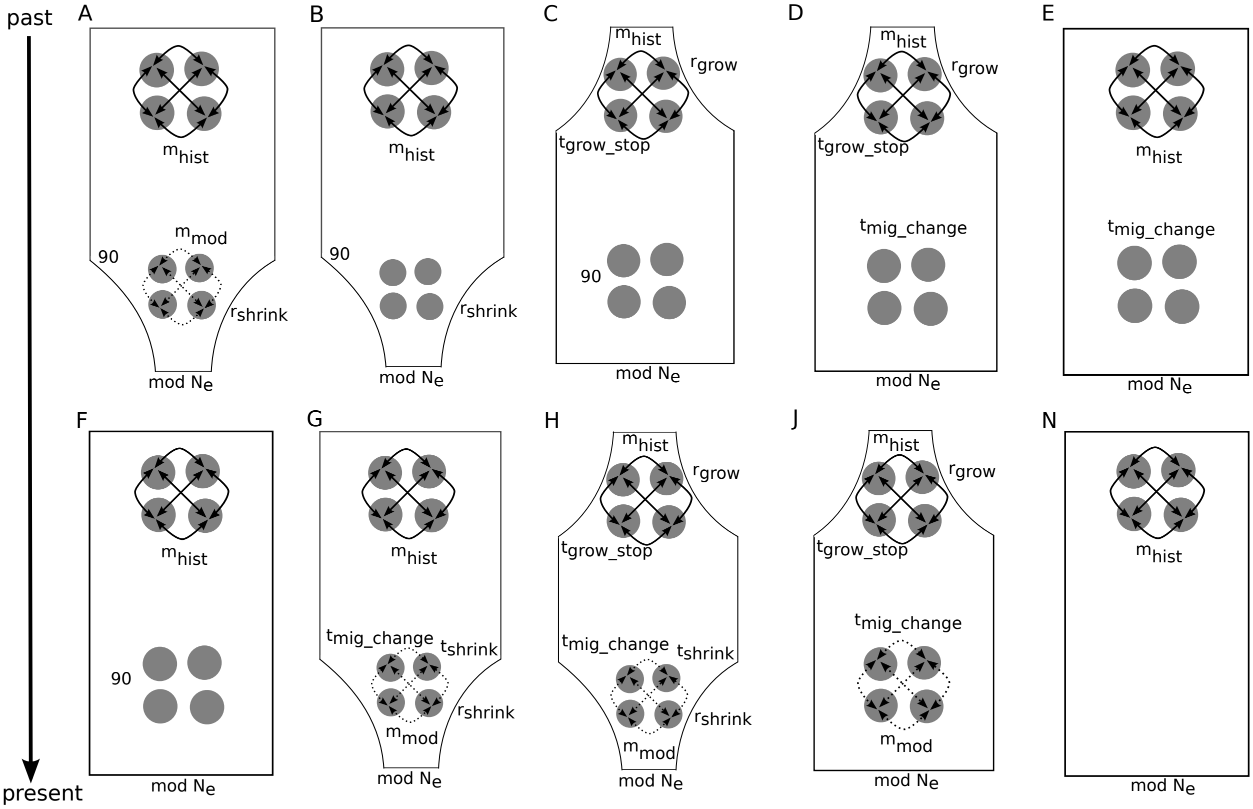


Figure 3.4. The general demographic models that were fit with ABC included histories with demes bottlenecking into the present (A, B, G), historic population expansion (C, D, J), historic population expansion followed by a bottleneck (H), and constant deme size (E, F, N). The dark circles within the history outlines represent different demes. The actual number of simulated demes equaled the number of sampling localities, which were 10 for YNP *T. alpinus*, 3 for SS *T. alpinus*, and 8 for YNP *T. speciosus*. The demes for YNP *T. alpinus* and *T. speciosus* were simulated under an island model (symmetric migration and identical deme sizes), while SS *T. alpinus* demes were simulated under a stepping stone model (isolation-by-distance and identical deme sizes). Solid versus dotted arrows between demes represent different migration rates, while no arrows represent no migration between demes. The parameter m_{hist} is the historic migration rate between two demes, m_{mod} is the modern migration rate between a pair of demes, t_{mig_change} is the number of generations in the past at which the migration rate changes, r_{grow} is the intrinsic growth rate for population expansion, t_{grow_stop} is the number of generations in the past that expansion stops, r_{shrink} is the intrinsic growth rate of population decline, t_{shrink} is the number of generations in the past that a bottleneck starts, and 'mod Ne' is the effective size of each deme at the present.

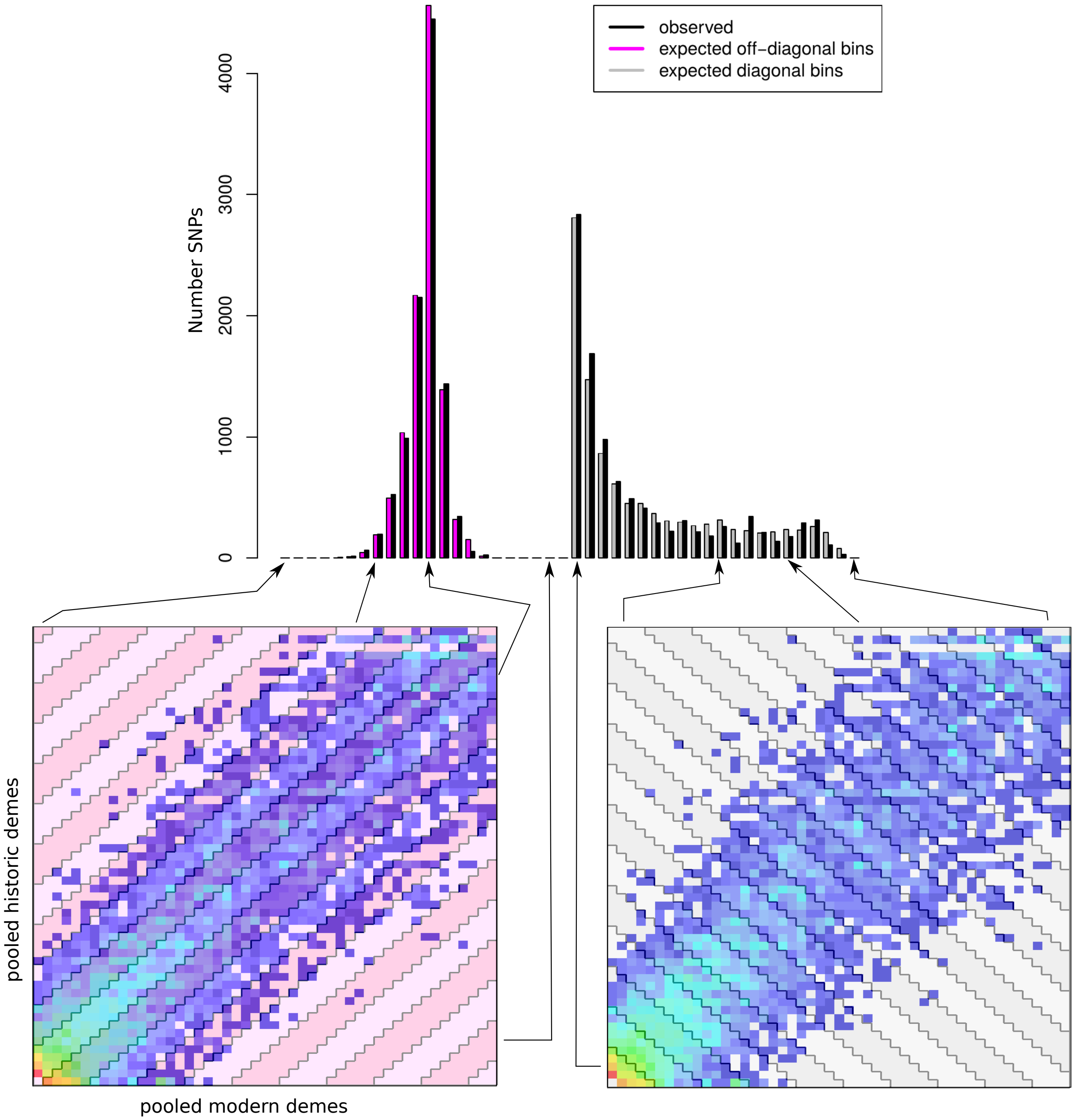


Figure 3.5. Diagram showing how bins of the 2D-SFS were fit in the ABC inferrence procedure. The 2D-SFS was constructed using the historic (pooled historic demes) and modern (pooled modern demes) metapopulations. The summary statistic used in the ABC procedure was the set of off-diagonal (left) and diagonal (right) bins of the 2D-SFS, where each bin's value is the sum of the counts contained within the bin. The width of each bin refers to the number of 2D-SFS categories on either side of the diagonal/off-diagonal and determines the amount of resolution (finer bins for higher resolution) and noise dampening (wider bins) when fitting the 2D-SFS. For each ABC simulation, the set of simulated 2D-SFS bins was compared to the observed bins (top) by calculating the Euclidean distance between them.

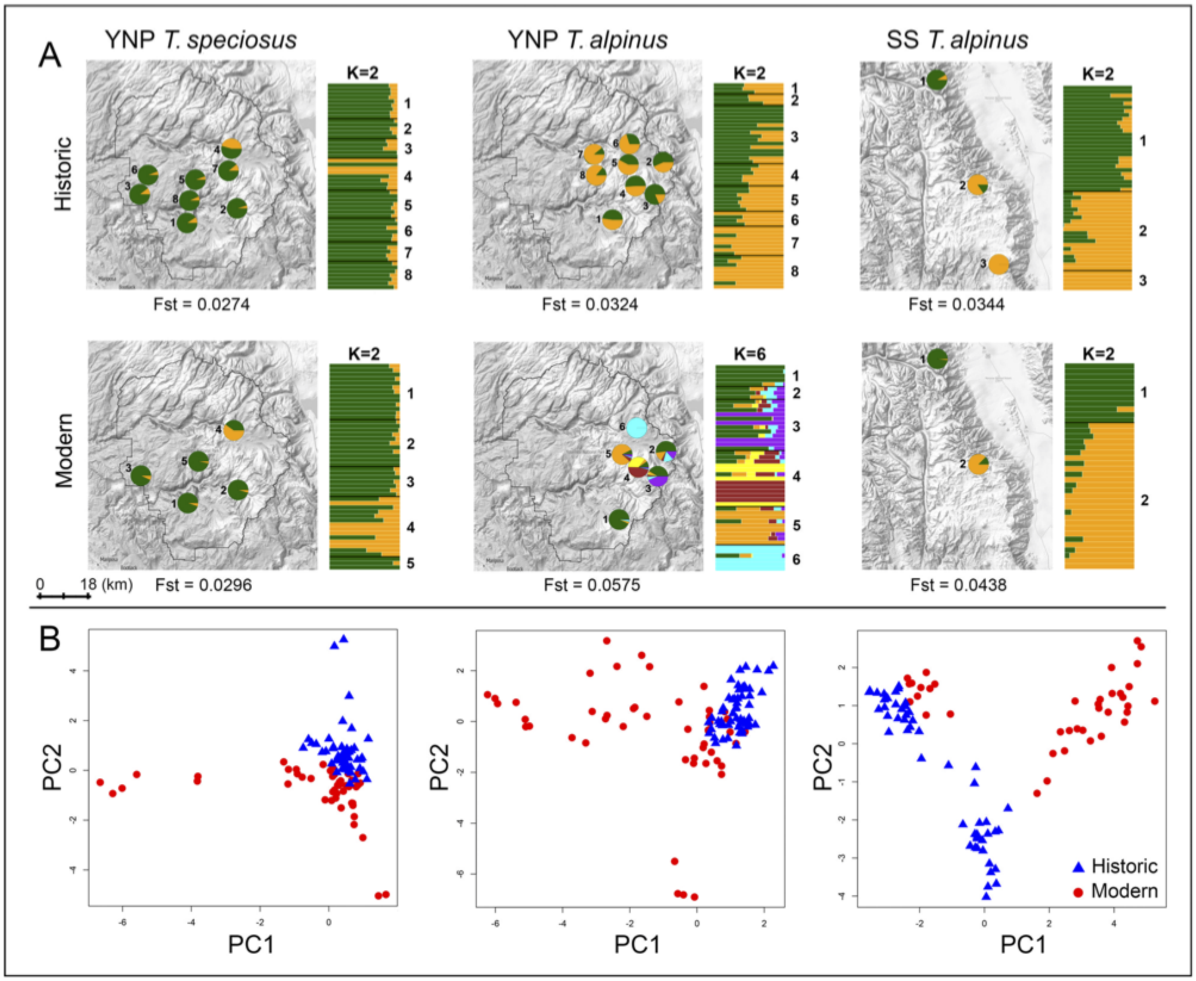


Figure 3.6. Temporal and spatial population genetic structure among *Tamias* chipmunk populations. A) Genetic clustering by general sampling locality (left, pie charts) and individual (bar graph to the right of each map) based on NGSadmix analyses. Each individual is partitioned into colored segments that indicate cluster membership. Pie charts represent the sum of all individualsâĂŹ membership in each cluster at each general locality on the map. Proximate individual sample localities were pooled for clarity following Rubidge and colleagues[14]. The inferred best number of clusters (K) is shown on the top of each bar graph. The global F_{ST} estimate for each population is indicated at the bottom of each map. B) Each data point in the PCA plot represents an individual specimen. The first two principal components (PCs) are shown.

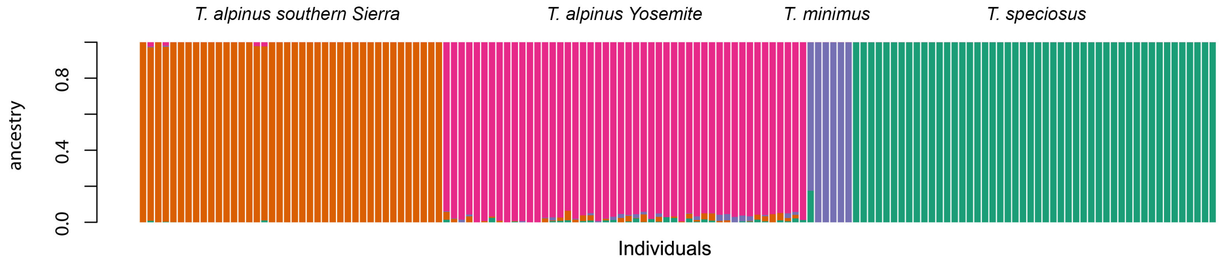


Figure 3.7. Inferred genetic ancestry of three *Tamias* chipmunk species. Individual ancestries were inferred based on randomly sampling one SNP per contiguous sequence to minimize the effect of linkage disequilibrium. Each individual specimen is represented by a horizontal bar partitioned into colored segments indicating their proportion of ancestry from each species.

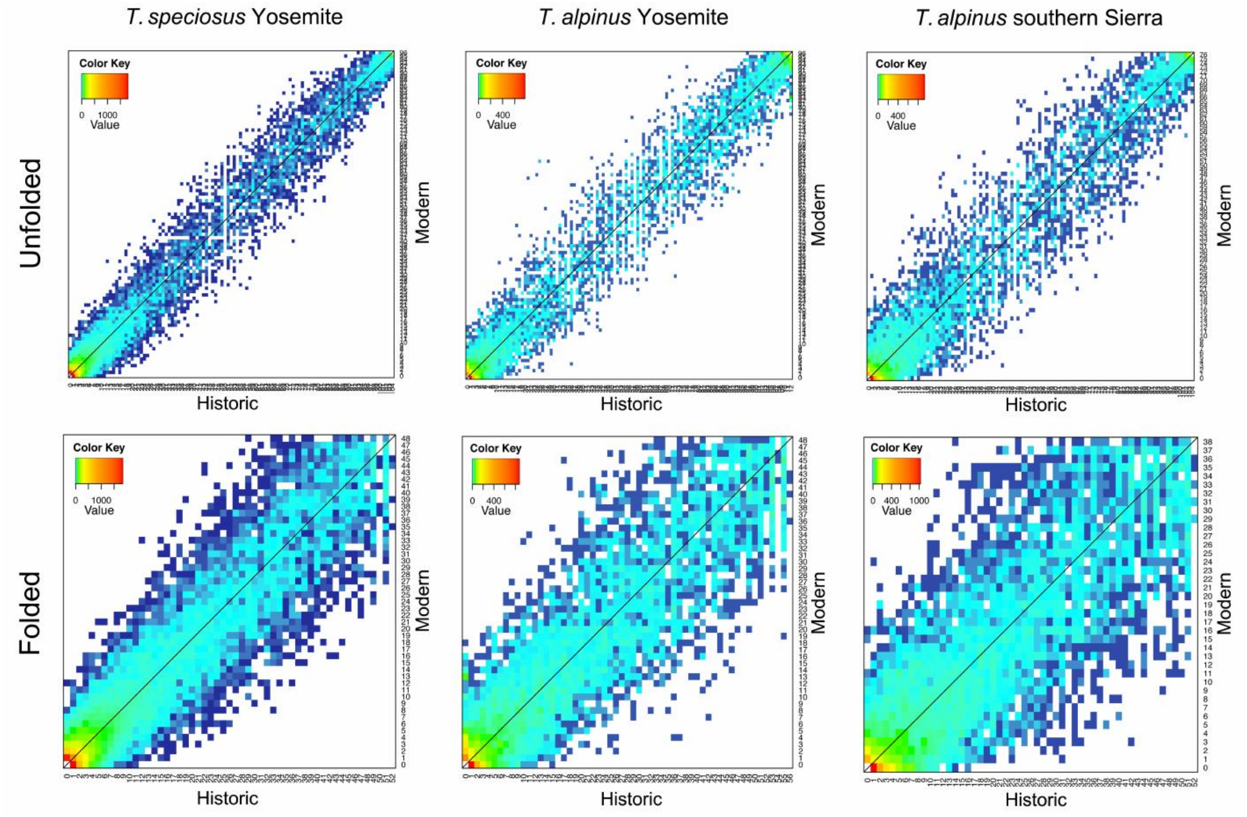


Figure 3.8. Unfolded and folded two-dimensional site frequency spectrum (2D-SFS) for SNPs between historic (x-axis) and modern (yÂŋÂŋÂŋÂŋ-axis) *Tamias* chipmunk populations. The color of each data point represents the number of SNPs belonging to that particular category in the 2D-SFS, which is specified by the color key inset. Folded spectra were used in demographic analyses and the unfolded spectra were used to identify outlier SNPs.

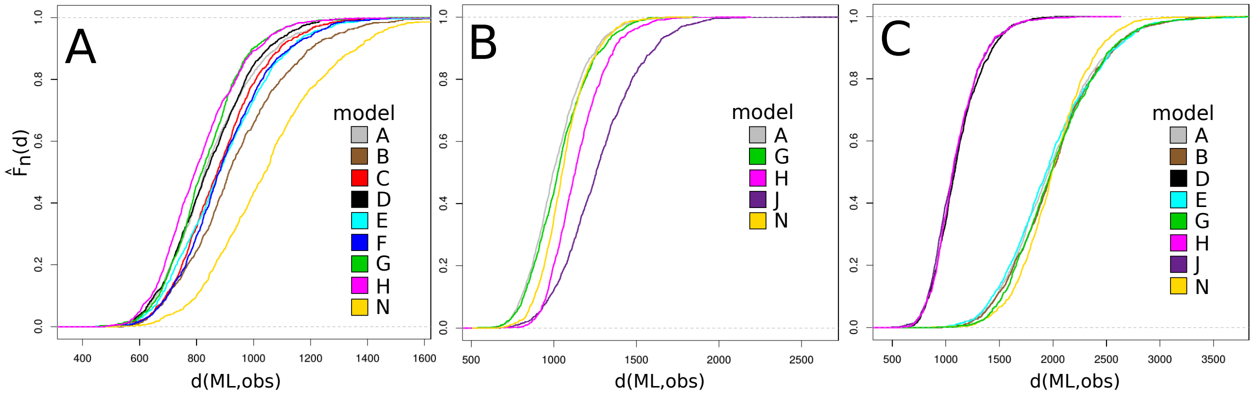


Figure 3.9. Empirical cumulative distribution functions (ECDF) for the Euclidean distance between the observed and expected joint SFS bins, d(ML,obs), for the different demographic models tested for (A) YNP *T. alpinus*, (B) SS *T. alpinus*, and (C) YNP *T. speciosus*. For each model 1,000 simulations under the ML history were performed to generate the expected joint frequency spectra from which the distribution of d(ML,obs) was calculated. Models with the leftmost ECDF curves are relatively most likely to resemble the true demographic history.

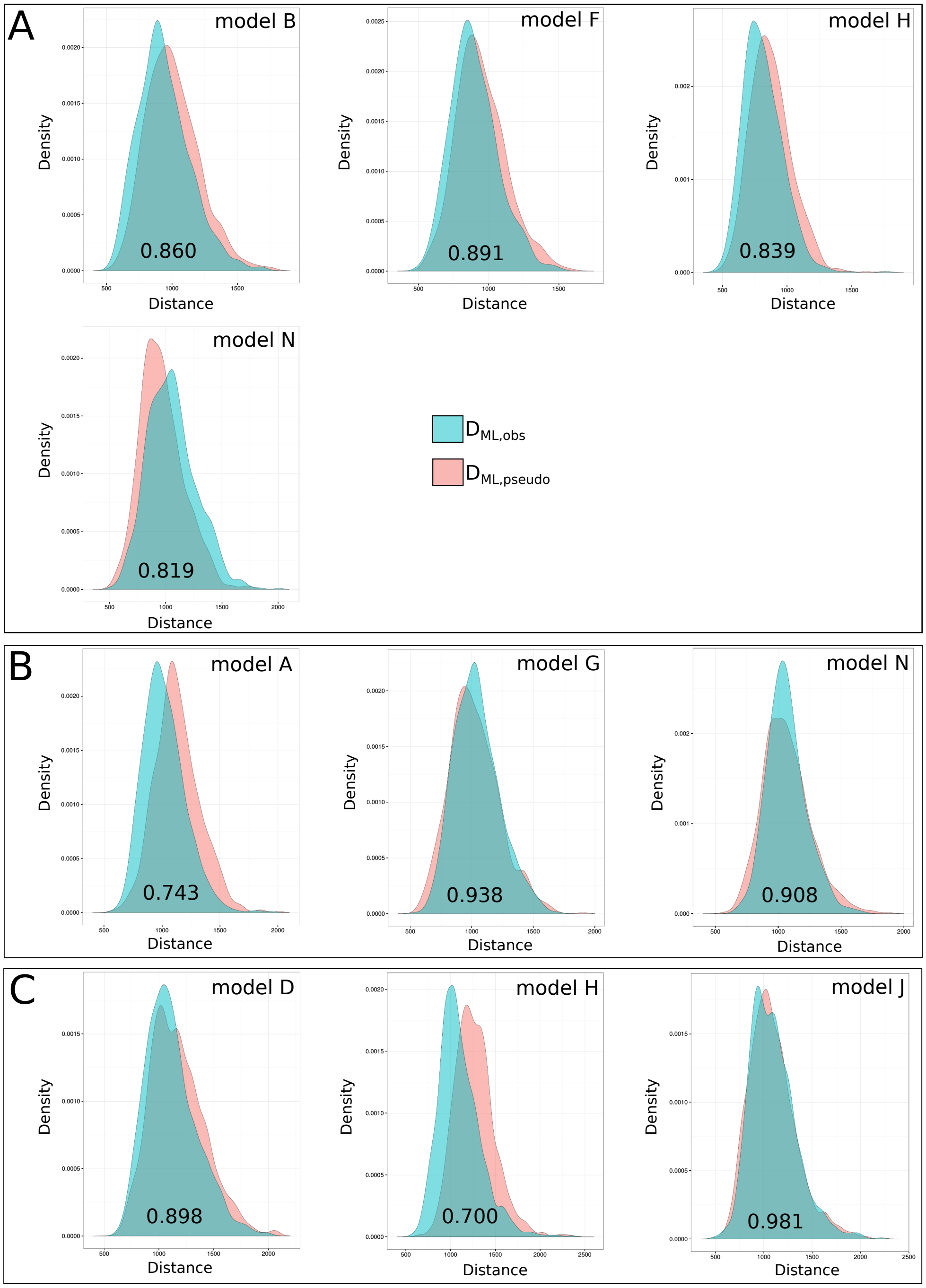


Figure 3.10. $D_{ML,obs}$ and $D_{ML,pseudo}$ kernel density estimates (KDE) for the best fitting demographic models for (A) YNP *T. alpinus*, (B) SS *T. alpinus*, and (C) *T. speciosus*. $D_{ML,obs}$ is the Euclidean distance between the observed and expected 2D-SFS bins under the ML history for each model, while $D_{ML,pseudo}$ is the distance between a single set of 2D-SFS bins under the ML history (pseudo observed) and the expected joint SFS bins. For each ML history, 1,000 simulations were performed to generate the expected joint spectra. The values within the distributions are Weitzman's coefficient of overlapping (OVL), which ranges from 0 to 1 and quantifies the area of overlap between the two KDEs. More overlap between the two distributions indicates greater similarity between the ML and true demography.

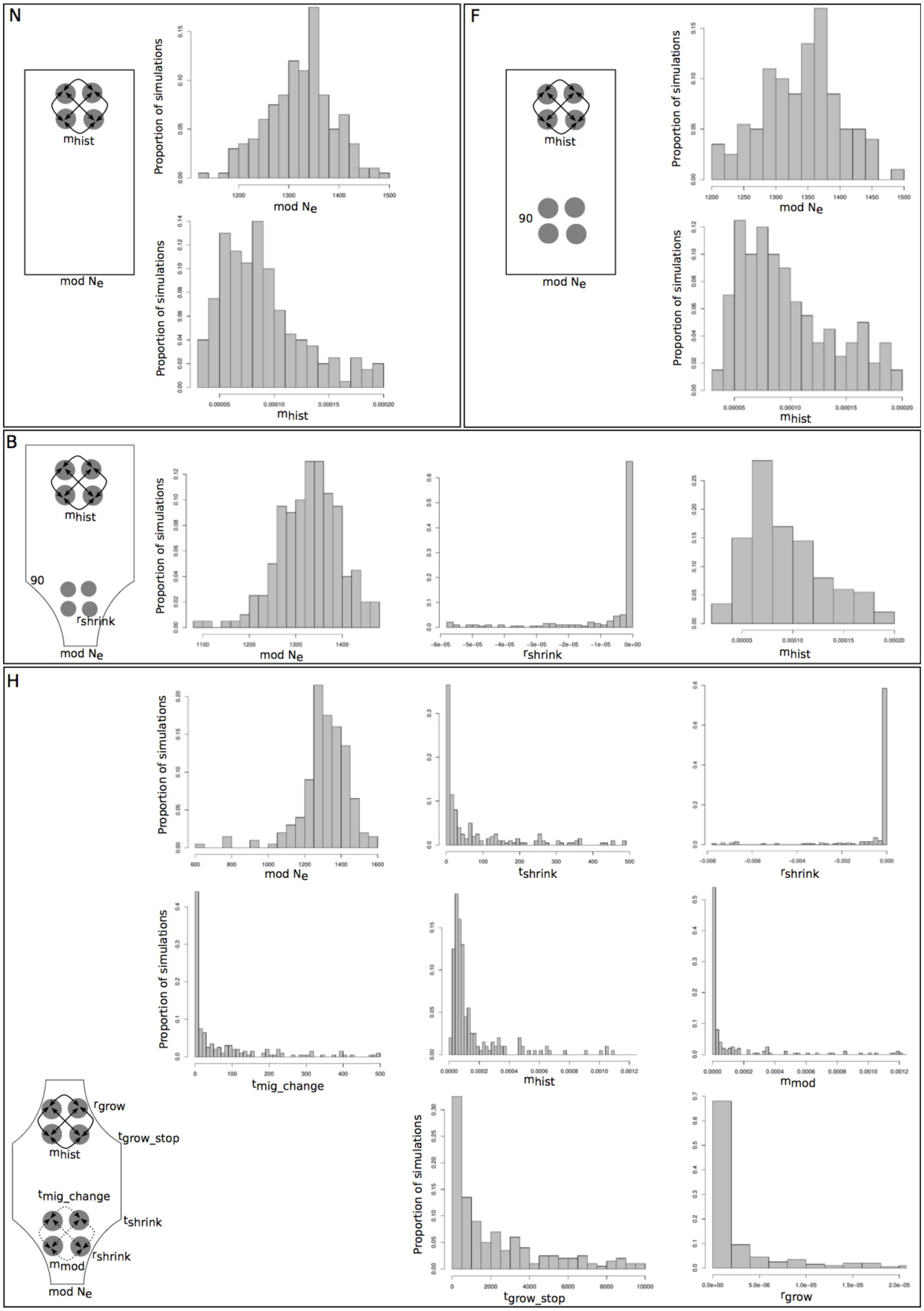


Figure 3.11. Posterior probability distributions for the demographic parameters of the best fitting models, B, F, H, and N, for YNP *Tamias alpinus* obtained using ABC. The posterior distributions reflect the set of parameter values after retaining 8% of the total 25,000 ABC simulations for each model that produced 2D-SFS bins most closely matching the observed 2D-SFS bins.

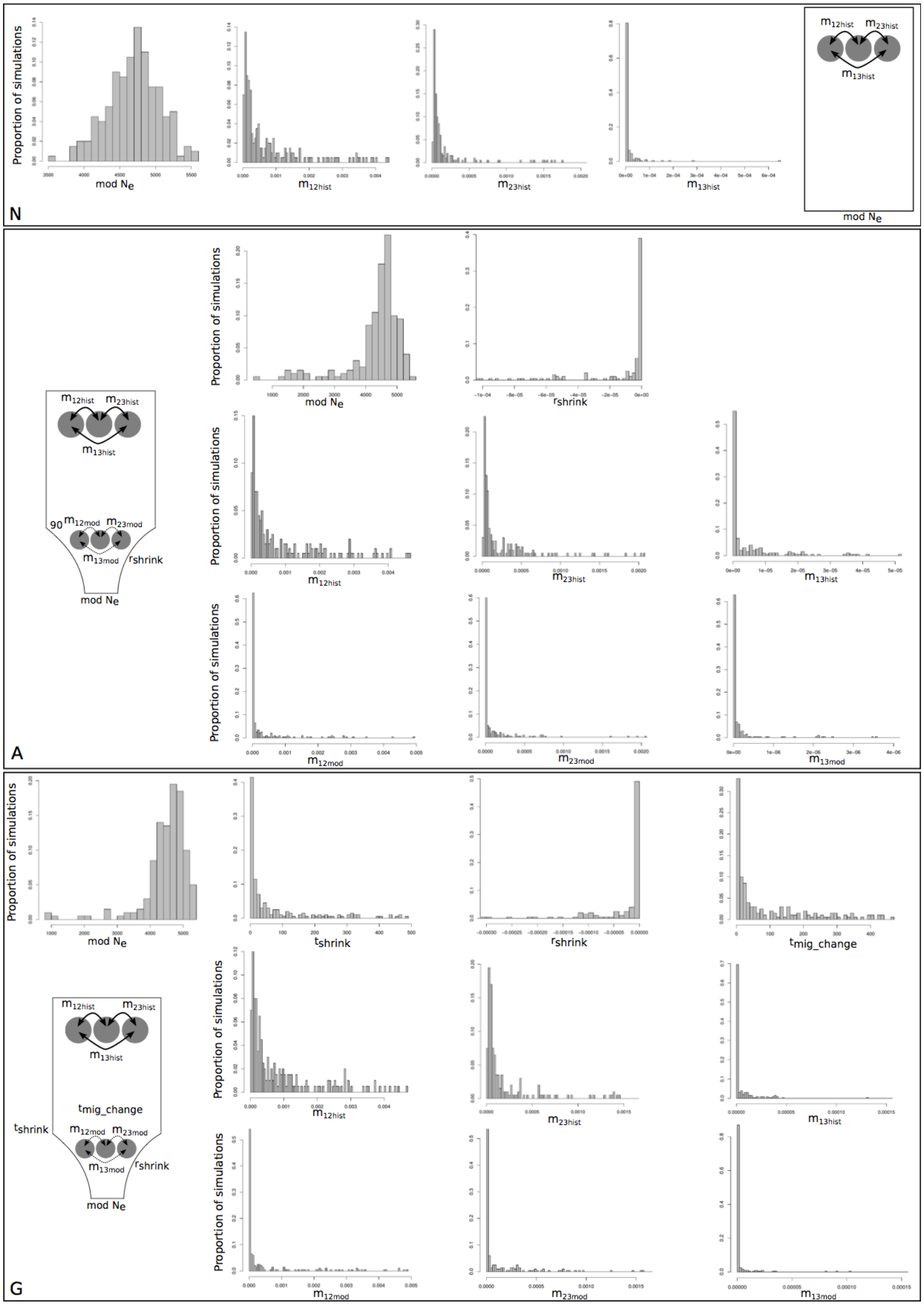


Figure 3.12. Posterior probability distributions for the demographic parameters of the best fitting models, A, G, and N, for SS *Tamias alpinus* obtained using ABC. The posterior distributions reflect the set of parameter values after retaining 8% of the total 25,000 ABC simulations under each model that produced 2D-SFS bins most closely matching the observed 2D-SFS bins.

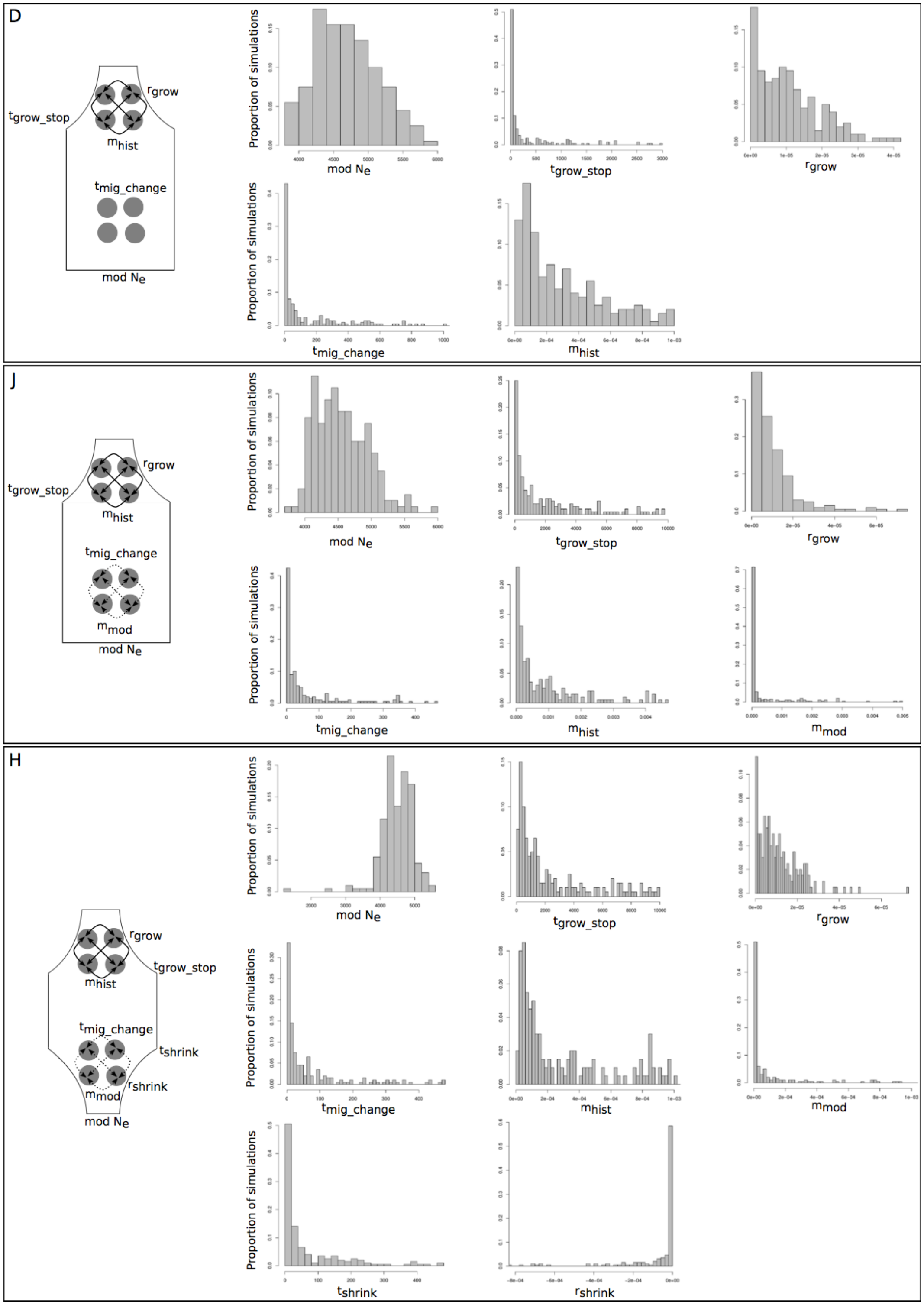


Figure 3.13. Posterior probability distributions for the demographic parameters of the best fitting models, D, J, and H, for YNP *Tamias speciosus* obtained using ABC. The posterior distributions reflect the set of parameter values after retaining 8% of the total 25,000 ABC simulations under each model that produced 2D-SFS bins most closely matching the observed 2D-SFS bins.

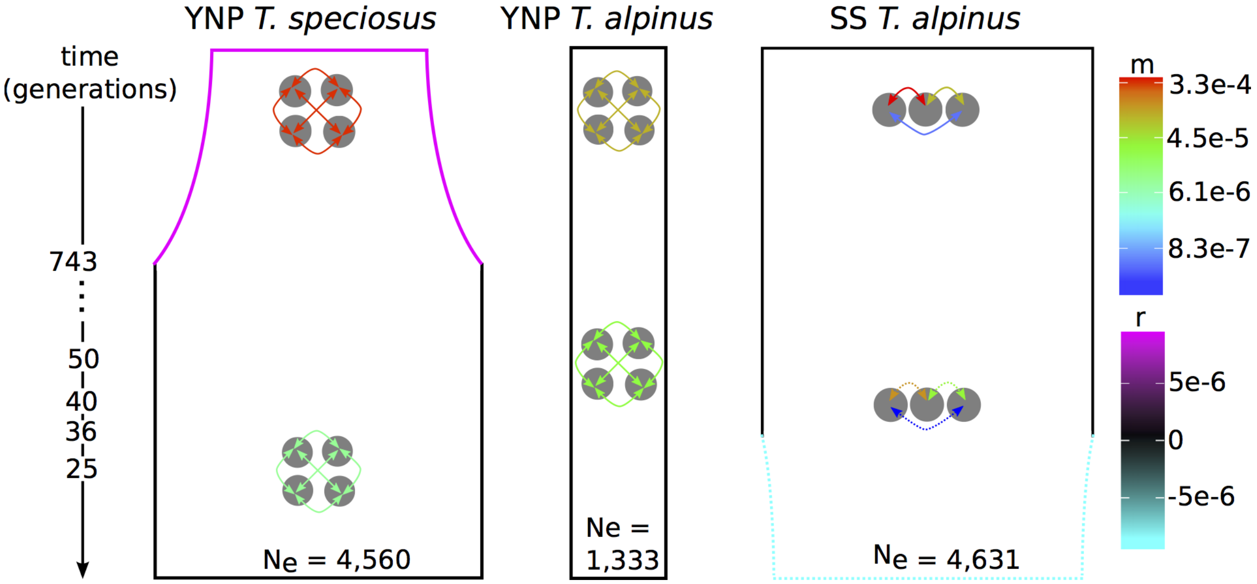


Figure 3.14. Population histories fitted with ABC showing the general topology and posterior median demographic parameter values averaged across the best fitting models for YNP *T. speciosus*, YNP *T. alpinus*, and SS *T. alpinus*, respectively. The fitted parameters are the modern effective number of individuals per deme (Ne), migration rates (m), intrinsic growth rates (r), and the timing of demographic events (1 generation = 1 year). History widths are proportional to the deme effective sizes though time. The four-deme depiction represents histories fitted with an island model, where the actual number of islands equaled the number of sampled demes. The three-deme depiction indicates that histories were fitted with a stepping stone model involving three demes. The events depicted with dashed lines for SS *T. alpinus* are relatively uncertain (see 'SS *T. alpinus* demography'), but if any demographic changes did occur they are as shown (note that a bottleneck would have been weak and the size is not to scale).

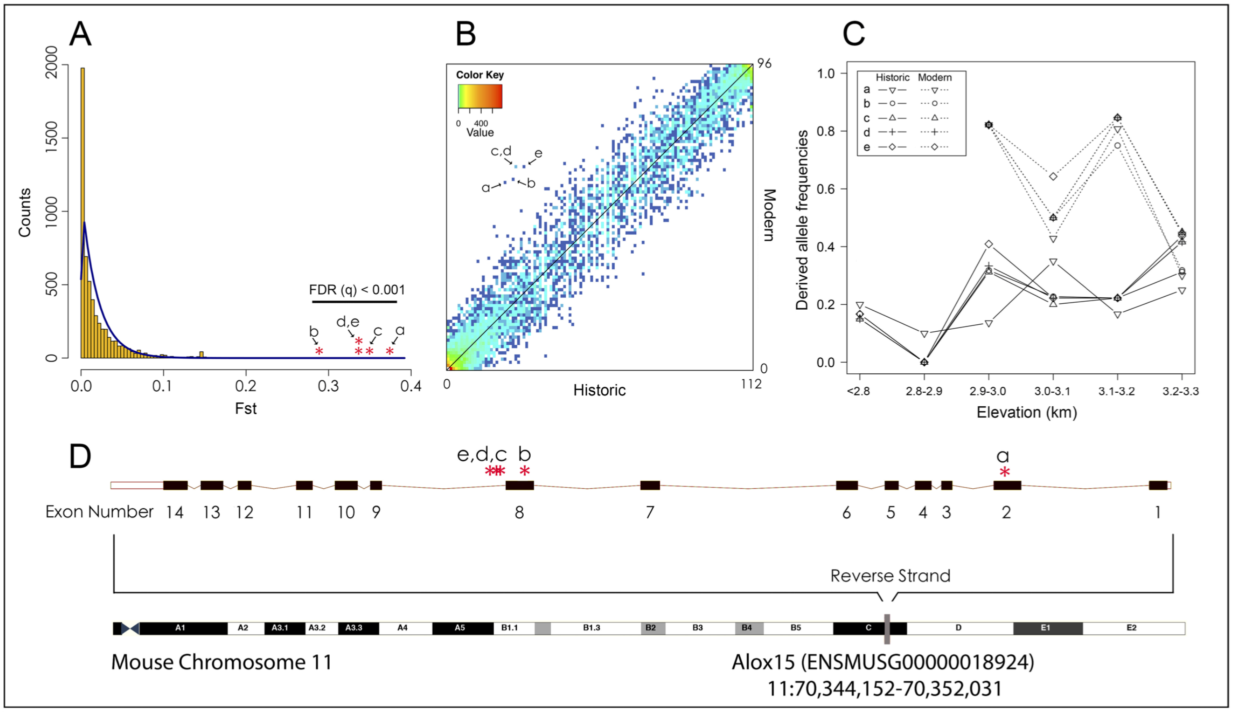


Figure 3.15. Derived alleles showing significant frequency shifts between historic and modern populations of *Tamias alpinus* in YNP. (A) Five outlier SNPs (a-e, FDR q < 0.001) are labeled on a plot of the neutral per site temporal F_{ST} distribution (modern versus historic) estimated by OutFLANK. The histogram of observed F_{ST} (yellow bins) is shown with the inferred neutral distribution (blue line). (B) Unfolded two-dimensional site frequency spectrum (2D-SFS) for SNPs between historic (x-axis) and modern (y-axis) YNP *Tamias alpinus* specimens. The color of each data point represents the number of SNPs (specified by the color key) belonging to that particular 2D-SFS category. Arrows point to the five outliers (a-e) showing the only significant allele frequency shifts over time. (C) Derived allele frequencies of the five outliers SNPs plotted against sample elevation. Individual sample localities were pooled into 100-meter elevational bands to enable allele frequency estimation. (D) The position of the five outliers mapped onto the mouse Alox15 gene.

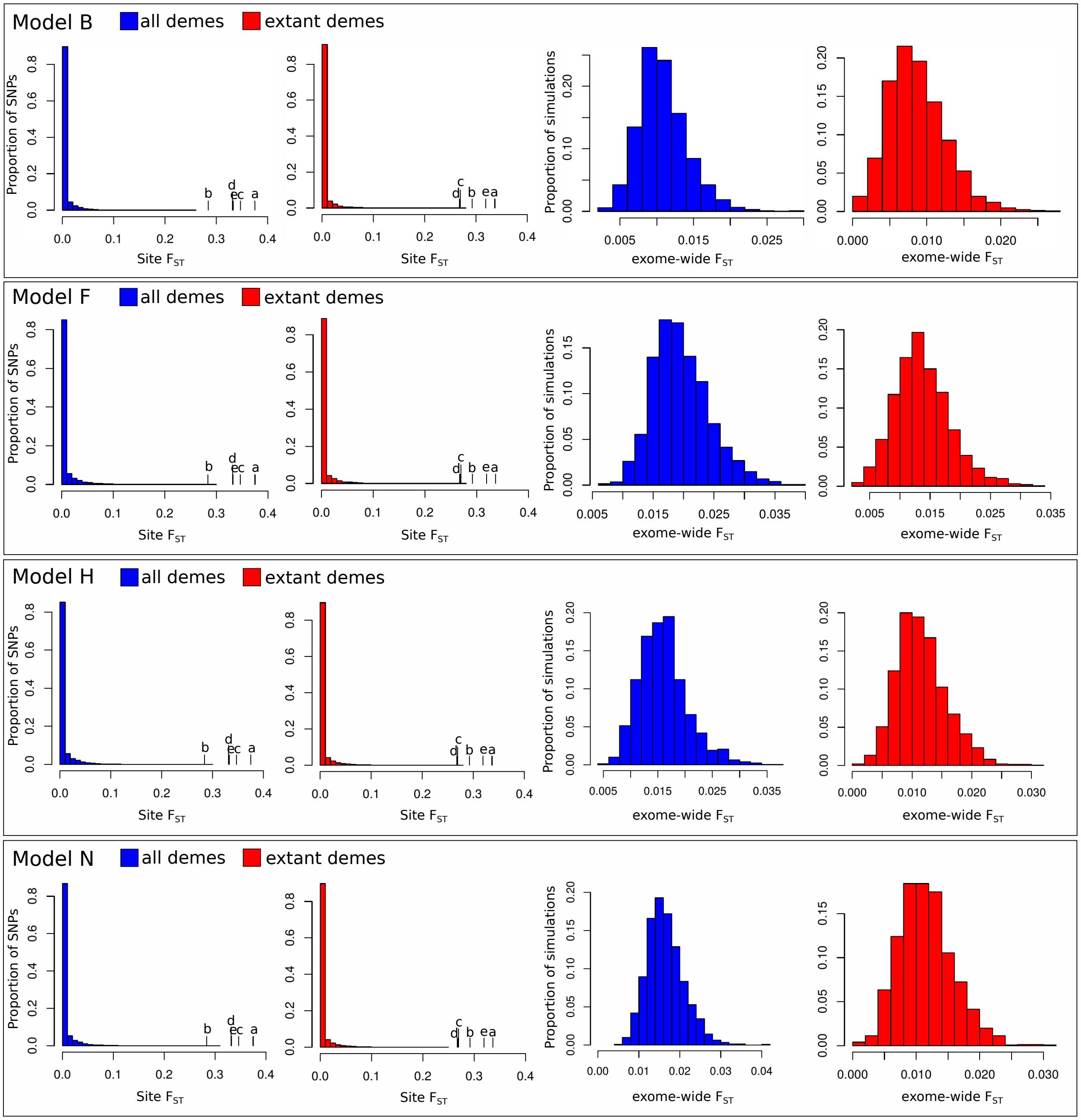


Figure 3.16. The expected distribution of neutral F_{ST} per site and exome-wide under the best fitting demographic histories for YNP *T. alpinus* inferred with ABC. Each F_{ST} distribution was generated from 1,500 simulations under the maximum likelihood histories for the best fitting demographic models for YNP *T. alpinus*, B, F, H, and N. The blue distributions were determined from all sampled demes. Calculation of the red distributions was limited to demes that provided samples in both the historic and modern periods (the extant demes). The values of the five observed F_{ST} outlier SNPs estimated from all and extant demes are plotted onto their respective expected per site F_{ST} distributions.

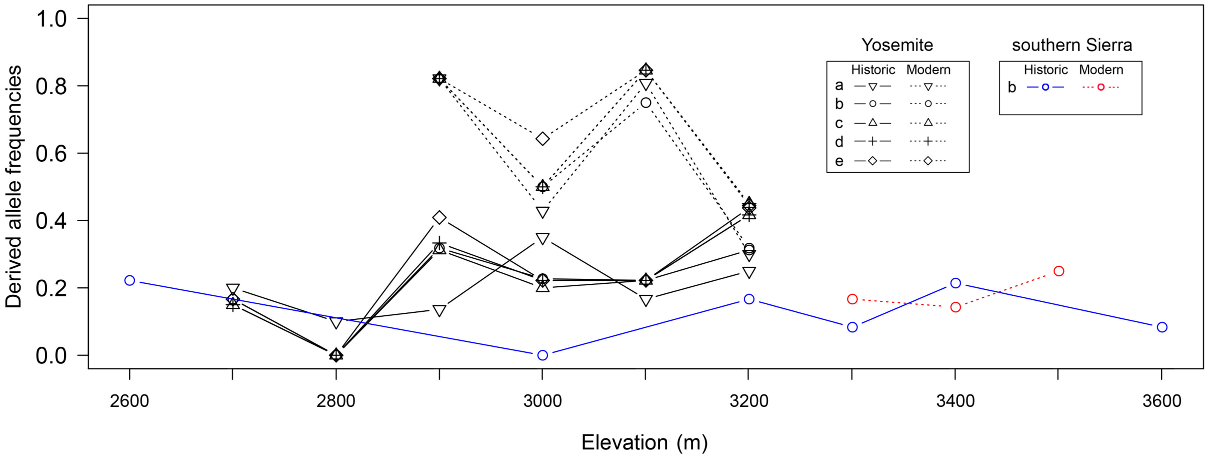


Figure 3.17. Site b derived allele frequencies in YNP and southern Sierra *T. alpinus* plotted against sample elevation. Among sites a-e that are variable in YNP *T. alpinus*, only b was segregating in SS *T. alpinus* for which it had similar frequency across temporal contrasts.

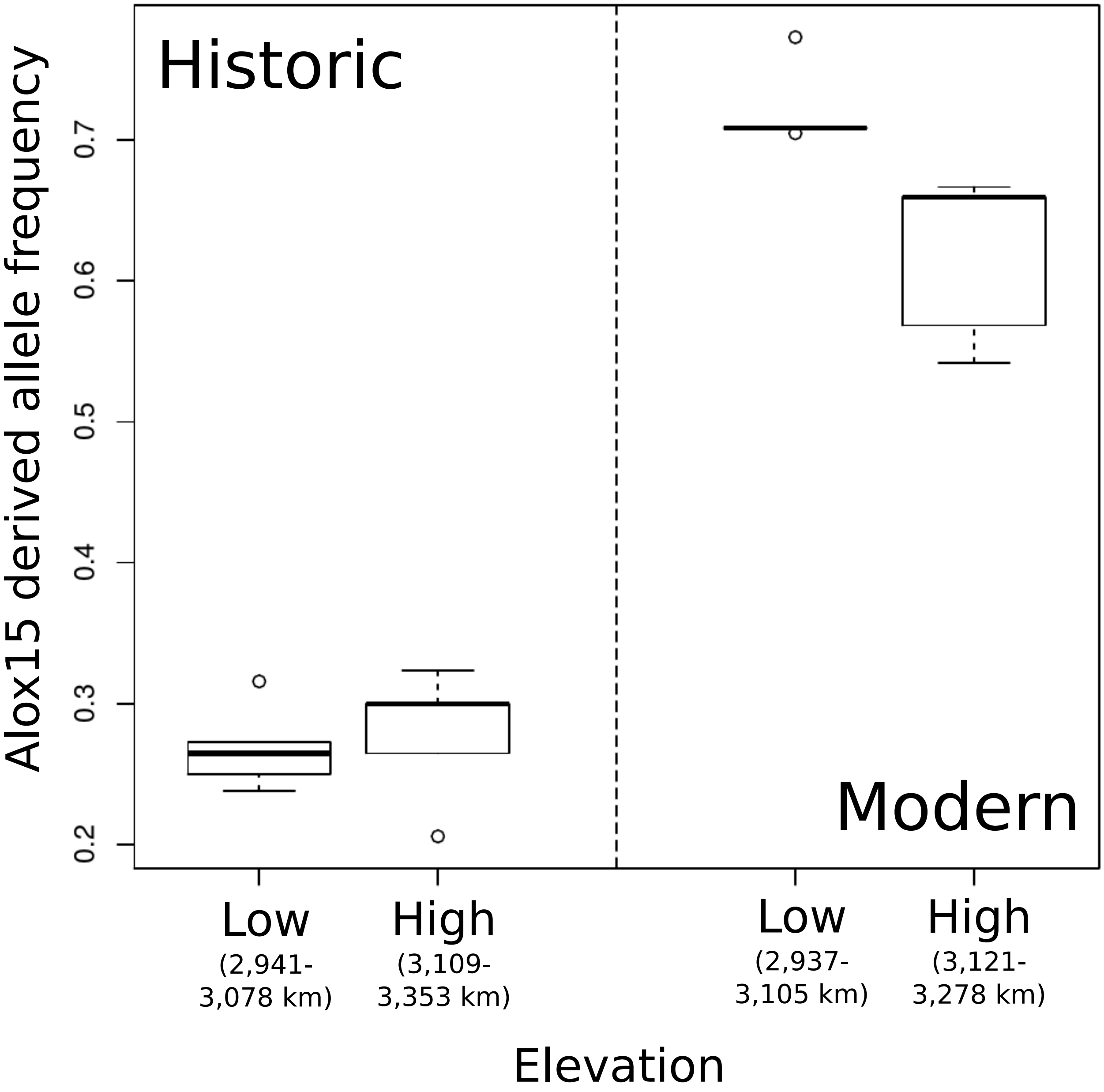


Figure 3.18. Modern and historic derived allele frequencies at the five Alox15 outlier SNPs among YNP *T. alpinus* sampled at high and low elevations. Calculation of the allele frequencies was restricted to individuals from demes that provided samples in both the modern and historic time periods. Low elevation individuals are those from the lower half of the modern-sample elevational range, while high elevation individuals are from the upper half. Sample sizes at the five SNPs were n = 17-22 for historic low, n = 15-17 for historic high, n = 22-24 for modern low, and n = 21-24 for modern high.

4. Mapping the genetic basis of coloration in the mimic poison frog

Tyler Linderoth, Evan Twomey, Adam Stuckert, Ke Bi, Amy Ko, Joana Rocha, Jason Chang, Matthre MacManes, Kyle Summers, Rasmus Nielsen

4.1 Introduction

A crux of modern evolutionary biology is discovering adaptive genomic regions, the pursuit of which is constantly being propelled forward by the latest molecular methods of the time starting with allozymes [91, 92], Sanger and length polymorphism sequencing of a few loci [93, 94, 95], and now next generation sequencing (NGS) on a genomic scale. New sequencing approaches and technologies based on NGS have also rapidly expanded the taxonomic breadth for mapping, at least putative, adaptive loci [96, 97, 28, 98]. Since animal color (including its distribution on the body in terms of pattern) has a major role on survival and reproduction [99, 100], it has been the focus of many mapping studies in non-model organisms [101, 102, 44]. Among the different ways in which color can influence fitness [103, 104], mimicry is perhaps one of the most blatant examples of just how intimately related color and selection are in nature [105, 106, 107]. For this reason, wing pattern mimicry has been extensively studied in *Heliconius* butterflies [108, 109, 110], and genomic approaches have revealed genes underlying wing color [111, 112, 113].

In regions of Peru proximal to where *Heliconius melpomene* and *H. erato* have undergone a mimetic radiation, the poison frog, *Ranitomeya imitator*, has also undergone a similar radiation, as its common name, the mimic poison frog, suggests. *Ranitomeya imitator* is a Müllerian mimic of three congeners (*R. summersi*, *R. variabilis*, and *R. fantastica*), representing four distinct color morphs. Phylogenetic analyses [114, 115] indicating that the congeners diverged prior to the *R. imitator* morphs support *R. imitator* as being the mimic, versus the model, species. Between the two geographic regions where *R. imitator* is sympatric with, and mimics, banded *R. summersi* and the region where it mimics striped *R. variabilis* there exists an introgression zone with admixed individuals between the banded and striped *R. imitator* morphs that exhibit intermediate color and pattern phenotypes [116]. The genes responsible for the different *R. imitator* color morphs are entirely unknown, which is mostly true for dendrobatids in general despite showing arguably some of the most striking color diversity among terrestrial vertebrates (one exception is Posso-Terranova & Andres[117] who recently showed MC1R to be responsible for darker phenotypes in *Oophaga histrionica*). The *R. imitator* complex, with its divergent and admixed morphs provides an excellent stage for mapping color genes and using them to understand what kind of evolutionary mechanisms drive mimetic radiations and maintain introgression zones.

Mapping genes in *R. imitator*, however, with a genome size (12 GB) that is 30-40x larger than *Heliconius* and 4x larger than humans, and completely lacking any prior genomic resources, is not a trivial task. We sought to overcome these challenges by designing a custom exon capture system for *R imitator*, which we used to survey over 13K genes in 124 individuals representing the banded, striped, and admixed morphs. We used a combination of divergence and admixture mapping to reveal candidate genes, showing enrichment for melanogenensis pathways, that we believe are likely influencing pattern, dorsal color, and/or leg color in *R. imitator*.

4.2 Methods

Ranitomeya imitator exome capture system design

Our own sequencing results have revealed that *Ranitomeya imitator* has a ∼12 GB genome that is large due to extensive paralogy. To minimize the obstacle of mapping genes in a very large and highly duplicated genome, we decided to focus on the exome, which could feasibly be assembled and was likely directly, via the encoded proteins, or indirectly, as targets of expression regulation, relevant to the phenotypes that we sought to identify the genetic basis for. There are no pre-existing reference genomes to design suitable capture probes from however, so we designed our capture system from *R. imitator* transcriptomes in a manner inspired by Bi *et al.*[58]. Specifically, we sequenced barcoded cDNA libraries prepared separately using Illumina TruSeq stranded total RNA kits for mRNA isolated from various tissues sampled from two, young adult, Tarapoto morph *R. imitator*. The different libraries for these samples were generated from the brain and eyes, and skin patches sampled from the dorsal trunk, nape and dorsal head, ventral jaw, and rear leg areas. These four skin areas differed amongst each other in terms of color and pattern. We sequenced additional barcoded cDNA libraries prepared from the skin for individuals sampled at four developmental stages: Week 1 (n = 1), 2 (n = 3), 4 (n = 3), and 5 (n = 3) tadpoles. Various developmental stages were sequenced since we had no *a priori* knowledge for when the relevant color and patterns genes may be expressed, and color becomes visible at around week 5 (Figure 4.1). The different skin patches plus brain and eye libraries were multiplelxed and sequenced across two lanes of the Illumina HiSeq 2000 using 100 bp paired-end reads. All of the tadpole skin libraries were multiplexed and sequenced in the exact same way.

The sequencing reads from each cDNA library were jointly assembled using Trinity, yielding 273,039 transcripts. We used the transcriptome annotation procedure explained in Singhal [63] to annotate the R. imitator transcripts based on BLASTX comparisons to *Xenopus tropicalis* proteins downloaded from Ensemble (transcriptome annotation methods are implemented in the script available at `https://github.com/CGRL-QB3-UCBerkeley/MarkerDevelopmentPylogenomics/blob/master/5-Annotation`. By requiring a maximum blast e-value of 1e-10 and at least 50% protein similarity to *X. tropicalis* we were able to assign Ensemble gene identifiers to 12,305 unique transcripts, which were used as the template sequence to design exon capture probes from. In order to ensure that we would not miss designing probes on any potentially relevant transcripts that we failed to annotate due to the considerable divergence between *R. imitator* and *X. tropicalis*, we looked for differential expression between the different skin patches and brain/eyes from the juvenile and between developmental stages as well as transcripts that were constitutively highly expressed throughout development. Despite having small sample sizes, we used EdgeR [118] to identify transcripts with the most extreme differences in the number of mapped read counts between the different body parts and developmental stages as well as transcripts that had TPM values calculated with Kallisto [119] that were above the 95th percentile across all stages of development. Any transcripts that we were unable to annotate but that showed interesting expression patterns were added to the set of transcripts to design capture probes from. To be confident that these unannotated candidate transcripts, or at least a portion of each assembly, represented genuine genes we ran transDecoder [120] on them to identify likely coding regions. We retained only transcripts having open reading frames of at least 100 bp and trimmed away any non-gene-like portions.

We applied a suite of quality control filters to the set of probe design transcripts to arrive at a final set that would result in optimal capture performance. This filtering entailed first performing a reciprocal blast of the transcripts and retaining one isoform per gene. For the retained transcripts we trimmed any untranslated regions (UTRs) greater than 500 bp down to this threshold in order to avoid designing probes on exceptionally long UTRs that could potentially represent assembly artifacts. We then filtered away any low-complexity and repetitive regions occurring in either *Xenopus* or the database for all vertebrates using repeatMasker [40] since these types of regions would promote off-target capture. After repeat masking, we only kept transcript regions of at least 80 contiguous base pairs of unmasked sequence. We then removed any sequences with GC content outside of the range of 40-70% since capture becomes relatively ineffective at these more extreme levels [58]. Lastly, since mitochondrial genes were unlikely to be of relevance to the phenotypes we wanted to map, we removed all mitochondrial transcripts except for cytochrome B, which we kept in order to ensure that we had some haploid representation to check things like sequencing error rates. After all quality controls, our total capture target was 28,281,490 bp, representing 13,265 genes. Among these genes, 10,904 were annotated (or at least had Ensemble IDs) and had a total length of 24,530,337 bp, while the remaining 2,361 unannotated genes with intriguing expression patterns had a total length of 3,751,153 bp. In-solution, sequence capture probes were designed from this target sequence and synthesized as a NimbleGen SeqCap EZ Developer Library kit.

Data collection

Ranitomeya imitator sample collection

We opportunistically sampled *R. imitator* from sites throughout the San Martin province of Peru (Figure 4.2) throughout the morning and early afternoon in 2010, 2012, 2013, and 2014. We hand-caught frogs as we encountered them out foraging during the day or resting in plant axils. Frogs were placed individually into film canisters or 15 mL Falcon tubes and transported back to a field-based laboratory where we collected phenotypic data and toe clips, which were stored in 96% ethanol. On the following day, frogs were released to the location from which they were caught. We collected samples from each location in one to three consecutive days.

Genomic library preparation, exome capture, and sequencing

We extracted DNA from a single toe clip for each of the 124 *R. imitator* samples using either Qiagen DNeasy Blood and Tissue kits following the manufacturer's protocols or using standard salt precipitation. For the salt precipitation method we first incubated the toes in 500 μL cell lysis buffer (1 mM Tris pH 8, 100 mM NaCl, 10 mM EDTA pH 8, 0.5% SDS) and 10 μL 20 mg/mL proteinase K at 55 °C for ~24 hours. After the tissues were completely digested we added 3 μL of 10 mg/mL RNase A to the samples and incubated them for an additional 30 minutes at 37°C. Following this RNA digestion the samples were placed in a freezer at -20 °C for 5 minutes. We then added 200 *u*L of 5 M NaCl to each sample accompanied by vortexing for 20 seconds, and then placed them back in the freezer for another 5 minutes. Samples were then centrifuged at 11,000 RPM for eight minutes to pellet the proteins. The supernatant containing the DNA was then transferred to new tubes, leaving behind the protein pellet, and 600 μL of 100% isopropanol was added to the tubes containing DNA, followed by 50 gentle inversions of the tubes. The samples were then placed in the freezer for ~1 hour and then centrifuged at 10,000 RPM for 10 minutes to pellet the DNA. After pouring off the supernatant, being careful not lose the DNA pellet, we washed the DNA with 600 μL of fresh 70% ethanol. The tubes were inverted to dislodge the pellet in the 70% ethanol, and then centrifuged for eight minutes at 10,000 RPM. After centrifugation, the supernatant was poured off, leaving the DNA pellet behind, which was allowed to dry for ~12 hours or until completely dry. Lastly, the samples were resuspended in 50 μL TE buffer.

For samples that had been extracted using the Qiagen kits, we performed an additional post-extraction, RNase A digestion to ensure complete removal of RNA. Specifically, 3 μL of 10 mg/mL RNase A was added to DNA suspended in Qiagen AE buffer and incubated at 37 °C for 30 minutes. The DNA was then recovered by adding 0.1x volume of 3 M sodium acetate (pH 5.2) and 2 volumes of 100% isopropanol to the samples, followed by gently inverting the samples 50 times. The samples were then placed into a -20 °C freezer for 15 minutes. After removing the samples from the freezer, we centrifuged them at 10,000 RPM for 12 minutes at room temperature to pellet the DNA. We then discarded the supernatant and washed the DNA pellet by adding 600 μL of fresh 70% ethanol to each sample followed by inverting the tubes to dislodge the pellet. The samples were then centrifuged at 10,000 RPM for 12 minutes. Lastly, we poured off the supernatant being careful not to lose the DNA pellet, which we allowed to dry for ~12 hours. After the samples were completely dry, we eluted them in 50 μL TE buffer.

We prepared genomic libraries for 124 *R. imitator* following the protocol of Meyer & Kircher [57]. We started the library preparation with 1-1.2 μg DNA per sample for individuals from the striped and banded populations and 0.92 - 1.4 μg per sample for the admixed population as measured by a Nanodrop fluorospectrometer. The one exception was for admixed sample CH-14-5, for which we used 0.4 μg of initial DNA because less overall DNA was available for this sample. The DNA for each sample was first sheared using a Diagenode Bioruptor to an average fragment size of ~250 bp for striped and banded samples and ~300 bp for the admixed samples. For striped and banded population samples this was achieved by shearing samples at the 'high' Bioruptor setting for a total of 7 minutes using cycles defined by 30 seconds of shearing followed by 30 seconds of pause (each sample was in the Bioruptor for a total of 13 minutes). The admixed samples were sheared at the 'medium' Bioruptor setting for a total of 4 minutes using the same 30 seconds on / 30 seconds pause cycle scheme (each sample was in the Bioruptor for a total of 8 minutes). Following each enzymatic reaction up until the indexing PCR step, all samples were cleaned using 1.6x volume of Sera-Mag bead solution. A sample-specific 7-nt barcode was incorporated into the library sequences for each sample via indexing PCR using 4 μL of template library DNA and 10 PCR cycles under the reaction conditions specified in Meyer & Kircher [57]. We performed three independent indexing PCR reactions for the striped and banded individuals followed by bead purification using 1.3x volume Sera-Mag beads. For the admixed individuals we performed five independent PCR reactions, each of which were purified using 0.8x volume Sera-Mag beads. All of the PCR products from each of the independent PCRs were pooled for each individual respectively.

Equal amounts of DNA from barcoded libraries for striped and banded samples were pooled to generate three separate banded/striped pools for capture with each pool comprised of 22 individuals. Likewise, equal amounts of DNA from all 58 admixed individual barcoded libraries were pooled to form an admixed capture pool. We isolated the targeted exons from these capture pools using our custom *R. imitator* capture kit following NimbleGen's protocols with slight modifications. In light of *R. imitator*'s large genome size we increased the amount of input DNA for capture compared to the 1 μg called for by the NimbleGen protocol in order to preserve the complexity of the captured libraries. For each of the three banded/striped capture pools we slightly more than doubled the called for amount of DNA (2.522, 2.388, 2.663 μg of DNA), and for the admixed capture pool we used 3 μg of DNA for hybridization. Accordingly, we also doubled and tripled the amount of barcode blockers and COT-1 DNA used in the banded/striped and admixed hybridization reactions, respectively, relative to the amounts specified in the NimbleGen protocols. For COT-1 blocking DNA we used a cocktail of chicken Hybloc, human COT-1, and mouse COT-1 combined in equal amounts. Both the banded/striped and admixed libraries hybridized for ~75 hours. The captured libraries were LM-PCR amplified in three separate reactions using 20 μL of the template solution and the number of PCR cycles was catered to yield concentrations between 20-25 ng/μL as assessed using the Nanodrop. Each of the banded/striped captured libraries were initially amplified using 14 PCR cyles, which was then reduced to 12 cycles for the second set of PCRs, and then to 11 cycles for the final set of PCR reactions for two of the captured libraries (the 3rd continued to use 12 cycles). The captured admixed libraries were amplified using one 12-cycle PCR and two 13-cycle PCRs. The products from the three separate PCR reactions for each capture were pooled by equal amounts based on Qubit fluorometer measurements. The enrichment efficiency for the four different capture experiments was assessed using qPCR with negative and positive targeted control loci. The three banded/striped captures were pooled equally based on qPCR measurements and the resulting library pool was sequenced on two lanes of the Illumina HiSeq 4000 using 100 bp paired-end reads. The captured admixed libraries were also sequenced using two Illumina HiSeq 4000 lanes but with 150 bp paired-end reads.

Sequencing data quality control

Raw reads

The raw sequencing reads were processed for quality prior to assembly or using them for any analyses with the program readCleaner `https://github.com/tplinderoth/ngsQC/tree/master/readCleaner`. This entailed applying, in the following order, specific quality control measures to the raw fastq files: 1) remove any reads spawned from PCR duplicates from the dataset, 2) trim adapter sequences from the ends of reads, 3) trim bases from the ends of reads with average Phred Quality below 20 in a 4 bp sliding window as well with the BWA algorithm implemented in cutadapt [121] using the same quality threshold of 20, 4) remove low complexity reads having a DUST score [122] above 4, 5) merge any read pairs that overlap by at least 6 base pairs and that have an observed expected alignment score [123] p-value less than 0.01, 6) remove reads that may be derived from potential contaminants identified as those that map to the human GRCh38 or *Escherichia coli* (NCBI GenBank accession U00096 AE000111-AE000510) reference genomes, 7) remove reads that are shorter than 36 bp long and/or have at least 50% of their sequence comprised of ‘N’s.

Mapped data

Prior to performing any population genetic or association analyses we filtered the mapped data for quality, initially processing data from the banded and striped population capture experiment and the admixed population capture separately. We pooled the data for the banded and striped populations to which we applied snpCleaner (`https://github.com/tplinderoth/ngsQC/tree/master/snpCleaner`) in order to retain sites that met the following criteria: 1) At least 70% of the individuals had to have data, meaning that they were covered by at least 1 sequencing read, 2) p-value for a test of strand bias above 0.0001, 3) Phred root mean square (RMS) mapping quality of at least 15, 4) base quality bias test p-value above 1e-100, 5) mapping quality bias test p-value above 1e-100, 6) distance from end-of-read bias test p-value above 0.0001. Criteria 3-6 applied only to potentially variable sites identified with samtools and bcftools since, other than for RMS mapping quality, these filters are geared to looking for systematic bias between reads with the reference and alternative alleles. We refer to the subset of retained sites from this filtering as the SBall set of sites. The divergence mapping used SNPs called from the SBall set. We filtered sites for the admixed population dataset using the same criteria as for SBall except for relaxing the coverage requirement by requiring at least ∼25% of individuals to have data (at least one sequencing read) to retain the site. This generated a quality controlled subset of admixed population sites referred to as ADMIXall. The intersection of ADMIXall and a second subset of sites for the banded and striped dataset generated using the same quality filtering parameters as SBall except for requiring only ∼25% of individuals to have data in order to keep sites, yielded the subset of quality controlled sites called SBAall. SBAall, representing sites that passed quality control requirements in both the admixed and striped/banded datasets, were used for admixture mapping in the introgression zone.

We applied more stringent quality filtering for sites used for population genetic characterization analyses to ensure highy accurate results. We chose to use the more liberally filtered SBall and SBAall sets of sites for gene mapping so as to not risk removing any potentially associated SNPs from the analysis; we ensured that these particular SNPs were of reliable quality after identifying potential candidates. Additional filtering of the SBall and ADMIXall sites for population genetic analyses involved removing any sites from SBall for which either the banded or striped population did not have at least 90% of individuals covered by at least 3 reads and removing any sites from ADMIXall for which these same minimum coverage requirements were not met among the admixed individuals. From these trimmed versions of SBall and ADMIXall we additionally removed any sites showing evidence of being paralogous. We ran ngsParalog (`https://github.com/tplinderoth/ngsParalog`) on the banded, striped, and admixed populations separately in order to calculate population-specific duplication likelihood ratios at sites contained in the trimmed SBall and ADMIXall sets. We purged any sites from the coverage-trimmed SBall subset that had a significant likelihood ratio of being duplicated at a 0.008 significance level after a Bonferroni correction for multiple testing in either the banded or striped population. Using a Bonferroni adjusted 0.008 significance level, we removed sites from the coverage-trimmed ADMIXall subset that had significant p-values of duplication for the admixed population. We then took the intersection of the remaining SBall and ADMIXall sites to obtain a subset of paralog-free, high quality sites refered to as SBAall$_{\text{strict}}$. This represented sites that are high quality in the striped, banded, and admixed populations and that are appropriate for population genetic inferrence.

Reference exome assembly and mapping

We assembled the subset of exome capture reads that passed all quality controls for six individuals separately using Spades with all default parameters. Three of the six individuals used for assembly were from the banded population (ET13012, ET13017, ET13018) while the other three were from the striped population (ET13029, ET13032, ET13033). We blasted the assembled contigs to the *R. imitator* transcriptome sequences used for probe design to identify in-target assemblies based on if they matched the targeted sequences by at least 80%. The in-target assemblies among the six individuals were then clustered using CD-HIT [65] and, if possible, assembled with CAP3 [66]. Any unassembled contigs belonging to the same gene were joined together using a span of 39 Ns between them to prevent reads from mapping across the gaps. Scripts used or assembly are available at `https://github.com/CGRL-QB3-UCBerkeley/seqCapture`. We then mapped the quality controlled reads for each of the 124 individuals to this *de novo* exome reference using NovoAlign (`http://www.novocraft.com/products/novoalign`) requiring at least 30 good quality (option l) bases per read and a minimum Phre-scaled mapping quality of 20 to retain reads, and setting the highest acceptable alignment score for the best alignment at 150. The average insert size and standard deviation mapping parameters were set according to values determined from bioanalyzer traces for the libraries. Mapping for the striped and banded population dataset used an insert size of 220 bp with a standard deviation of 58, while for the admixed population dataset the insert size was set to 271 bp with a standard deviation of 63. We then used samtools to merge the resulting unpaired and paired alignments for each individual into a single, sorted file, and converted the SAM files into BAM format.

Population genetic analyses

We used the program ANGSD [37] to estimate the respective site frequency spectrum (SFS) for the banded, striped, and admixed populations for the SBAall$_{\text{strict}}$ set of quality controlled sites. The shape of the SFS were used to gain insight into the demographic histories of the populations and also served as a proxy for the adequacy of our data quality control measures. The three SFS suggested that there were no apparent data quality problems with the SBAall$_{\text{strict}}$ sites and so all population genetic analyses were based on this subset. ANGSD performs population genetic analyses in a probabilistic framework, based on genotype and allele frequency likelihoods, thereby avoiding compounding inaccuracy due to genotyping errors in downstream analyses, which is a particularly pronounced problem for low-medium coverage datasets like ours. It also takes an empirical Bayes approach to many analyses that use allele frequencies, for which it requires an estimate of the site frequency spectrum as a prior. Accordingly, each population's respective SFS served as a prior to obtain estimates of genetic diversity in terms of Watterson's estimator, θ_W, and nucleotide diversity, π, for the different *R. imitator* populations. We also used ANGSD to calculate p-values for whether sites in the SBAall$_{\text{strict}}$ set were variable, and called SNPs based on a p-value cutoff of 1e-6 for the banded, striped, and admixed populations separately.

We quantified the degree of genetic differentation between populations in terms of F_{ST} with ANGSD as well. We estimated the joint SFS between each pair of populations, which served a prior for estimating F_{ST} for each of the corresponding pairwise comparisons. In addition, we examined the genetic relationship among individuals by using ANGSD to estimate the genetic covariance among individuals based on genotype posterior probabilities assuming a Hardy-Weinberg equilibrium prior. Eigen decomposition was performed on the resulting covariance matrix using the 'eigen' function in R [79] in order to perform PCA analysis. We also estimated ancestry proportions in the banded, striped, and admixed populations using NGSadmix [74] run with two ancestral populations, considering only sites with a minimum minor allele frequency (MAF) of 5%. The PCA and admixture analyses were performed using SNPs called from the SBAall$_{\text{strict}}$ set for the three populations pooled.

Color and pattern phenotyping

In order to determine which genes may be underlying specific colors and pattern in *R. imitator*, we quantified pattern in terms of its degree of stripedness or bandedness, the proportion of the dorsum that was black, and five aspects of color in the Lab color space; 1) dorsal color along the red-green color axis, DC_a, 2) dorsal color along the lightness axis, DC_L, 3) hind leg color along the red-green color axis, LC_a, 4) hind leg color along the lightness axis, LC_L. These features were quantified from digital photographs taken with a Nikon D7000 camera using a Nikkor 85 mm macro lens. Images were converted from RAW to the highest quality JPEG format using Nikon ViewNX2 software. Prior to quantifying the phenotypes the JPEG images were converted to PNG format using the software GIMP (`www.gimp.org`). *Ranitomeya imitator* from the striped population are characterized by segments of black running parallel to the anteroposterior axis along the dorsum versus individuals from the banded population which have segments of black running perpendicular to this axis. *Ranitomeya imitator* from the introgression zone exhibit a gradient of patterns intermediate between the pure striped and banded morphs. To quantify pattern from digital images, we used a continuous measure of the degree of stripedness or bandedness referred to as the pattern rotation measure, pattern$_{rot}$. When considering a raster image the number of pattern gaps, γ, is the count of the number of transitions between pattern color and non-pattern color pixels when traversing the image in a straight line. We denote γ calculated along the horizontal axis as γ_r and along the vertical axis as γ_c. Note that throughout we use the subscript r to refer to the horizontal axis of a raster image (the rows) and c to refer to the vertical axis (the columns). Pattern track length, τ, is defined as the count of directly adjacent pattern color pixels along a straight line. We use τ_r^{max} to denote the maximum τ in a set of tracks for a given raster image row, and similarly, τ_c^{max} for the maximum τ among the set of tracks for a given column. The effective segment length, ϕ, is the count of all pixels in a given row or column of a raster image to be analyzed and so is bounded from above by the image's width and height, respectively. For a raster image of width, $\mathbf{w}$, and height, $\mathbf{h}$,

$$\eta = \sum_{j=1}^{\mathbf{w}} \gamma_{cj} + \sum_{i=1}^{\mathbf{h}} \gamma_{ri}$$

$$\xi = \frac{1}{\mathbf{h}} \sum_{i=1}^{\mathbf{h}} \tau_{ri}^{max}/\phi_i - \frac{1}{\mathbf{w}} \sum_{j=1}^{\mathbf{w}} \tau_{cj}^{max}/\phi_j$$

$$\text{pattern}_{rot} = (\sum_{j=1}^{\mathbf{w}} \gamma_{cj} - \sum_{i=1}^{\mathbf{h}} \gamma_{ri})/\eta + \xi$$

Increasing pattern$_{rot}$ in the positive direction corresponds to more stripedness, while increasing magnitude in the negative direction indicates more bandedness. A pattern$_{rot}$ value of 0 corresponds to a frog that is perfectly intermediate between striped and banded.

To quantify color from PNG images we determined the average red, green, and blue values in the RGB color space over all non-tranparent pixels. Specifically, for a raster image with d non-tranparent pixels having color values $(\psi_{v1}, ..., \psi_{vd})$ for color channels v, $v \in \{\text{red}, \text{green}, \text{blue}\}$, the average for each RGB channel, Ψ_v, was calculated as

$$\Psi_v = (\frac{1}{d} \sum_{k=1}^{d} \psi_{vk}^2)^{\frac{1}{2}}$$

These average RGB values were then converted to Lab color space using CIE Standard Illuminant D65 as the reference white for both color spaces.

To calculate the proportion of the dorsum that was black we simply took the fraction of the number of black pixels out of all non-transparent pixels. Specifically, for a PNG image with d non-transparent pixels having fractional RGB values as outlined above, the proportion of black is given by

$$\frac{\sum_{k=1}^{d} \mathbb{1}_{\{\psi_{red,k}=0,\, \psi_{green,k}=0,\, \psi_{blue,k}=0\}}}{d}$$

Prior to quantifying the phenotypes we preprocessed the JPEG images by first standardizing the dimensions of the frogs in each image. The anteroposterior length of each *R. imitator* sample as well as its orientation to the horizontal axis was measured using imageJ software. These measurements were then used in R to transform each image such that the anteroposterior axis of each frog was oriented horizontally and the length of the frogs along this axis from the tip of the rostrum to pelvis was the same across all photographs. These standardized photos were imported into GIMP, which we used to manually fill in black portions of the frogs. This was done to reduce noise from glare, debris, and the occasional tissue damage. We also used GIMP to extract regions of the photos that we wanted to restrict our phenotyping to, which were exported in PNG format for input into our phenotyping scripts. All scripts used for image processing and phenotyping are available at `https://github.com/tplinderoth/image_phenotype`.

Association mapping

Divergence mapping between striped and banded morphs

We used the likelihood ratio test for allelic association described in Kim *et al.* [35] to look for allelic divergence between the 33 banded and 33 striped *R. imitator*. This test compares the null hypothesis that the frequency, p_A, of the total population minor allele, denoted A, is the same among treatments to the alternative hypothesis that it is different, where in our case the treatments are banded and striped morph. That is, we test the null situation in which $p_A = p_A^{band} = p_A^{stripe}$ against the alternative case in which $p_A^{band} \neq p_A^{stripe}$, where p_A^{band} and p_A^{stripe} are the frequency of the A allele in the banded and striped population, respectively. We implemented this likelihood ratio test as a C++ program, ngsAssociation (`https://github.com/tplinderoth/ngsAssociation`), which accounts for individual base qualities, $\mathcal{Q}_r^m$, that translate into the probability that an observed read, $X_r^{(m)}$, in pool m, is an error, $\mathcal{E}_r^{(m)}$. Assuming uniform error among the three possible incorrect alleles gives

$$P\Big(X_r^{(m)} = A|G^{(m)} = k, \mathcal{Q}_r^{(m)}\Big) \\ = \left(\frac{S_{\text{pool}} - k}{S_{\text{pool}}}\right)\left(1 - \mathcal{E}_r^{(m)}\right) + \left(\frac{k}{S_{\text{pool}}}\right)\left(\frac{\mathcal{E}_r^{(m)}}{3}\right)$$

and

$$P\Big(X_r^{(m)} \neq A|G^{(m)} = k, \mathcal{Q}_r^{(m)}\Big) \\ = \left(\frac{S_{\text{pool}} - k}{S_{\text{pool}}}\right)\mathcal{E}_r^{(m)} + \frac{k}{S_{\text{pool}}}\left(\left(1 - \mathcal{E}_r^{(m)}\right) + \frac{2}{3}\mathcal{E}_r^{(m)}\right)$$

where S_{pool} is the haploid sample size of each pool and G_m is the true number of type A alleles that went into constructing pool m. Note that a pool of sequencing reads can represent a single individual or multiple, pooled, individuals, and we have chosen to use the term pool here only because it is most general.

Assuming that sites are diallelic and no population structure, the probability of the data for an observed pool, $O^{(m)}$, with V^m sequencing reads, given p_A, is then

$$P(O^{(m)}|p_A) = \sum_{k=0}^{S_{\text{pool}}}\Big(\text{Binom}(k; S_{\text{pool}}, p_A)\prod_{r=1}^{V^{(m)}} P\Big(X_r^{(m)}|G^{(m)} = k, \mathcal{Q}_r^{(m)}\Big)\Big)$$

The likelihood of p_A for an entire site with N pools (or individuals) is

$$L(p_A|O) = \prod_{m=1}^{N} P(O^{(m)}|p_A)$$

In our particular case, this is compared to the joint likelihood

$$L(p_A^{band}, p_A^{stripe}|O) = \prod_{m=1}^{N_{band}} P(O^{(m)}|p_A^{band}) \prod_{m=N_{band}+1}^{N} P(O^{(m)}|p_A^{stripe})$$

Maximum likelihood estimates for the allele frequencies $\hat{p}_A$, $\hat{p}_A^{band}$, and $\hat{p}_A^{stripe}$ are obtained using the bounded, limited-memory, Broyden-Fletcher-Goldfarb-Shanno (L-BFGS-B) algorithm. The likelihood ratio statistic is then

$$LR_{allele} = -2 \log \left(\frac{L(\hat{p}_A|O)}{L(\hat{p}_A^{band}, \hat{p}_A^{stripe}|O)} \right)$$

While we have described the implementation of a slightly modified version of the likelihood ratio test from Kim *et al.* [35] in terms of allele frequencies corresponding to striped and banded frogs, the implementation is general and ngsAssociation can be used for any similar divergence mapping experiment. We used ngsAssociation to identify variable sites among the SBall set after filtering out any sites with total coverage greater than 6000X (corresponding to the 99.7th percentile of the empirical coverage distribution) because these sites are likely within low mappability regions for reasons such as paralogy or low complexity. The test for whether a site is variable implemented in ngsAssociation is a likelihood ratio test that compares the likelihood under the null that $p_A = 0$ to the $\hat{p}_A$ likelihood using the same likelihood function for p_A described above. We calculated LR_{allele}, comparing the banded to the striped population, at all called SNPs. The LR_{allele} value at each SNP was divided by the mean LR_{allele} value across all SNPs in order to control for inflation of the likelihood ratios due to population structure between the striped and banded populations [124]. We compared these genomic-controlled LR_{allele} values to a χ_1^2 distribution in order to obtain corresponding p-values.

Admixture mapping in the introgression zone

We used a general linear model (GLM) framework to look for associations of genotype with pattern, leg color, and dorsal color among the 58 *R. imitator* sampled from the introgression zone between the striped and banded morphs. Specifically, we used ANGSD to perform the score statistic test for association described in Skotte *et al.* [125], thereby calculating likelihood ratios, LR_{geno}, that compare maximum likelihood estimates for the effect of genotype on phenotype, β, in the linear predictor to the null case where $\beta = 0$. This was done for sites in the SBAall subset that had an associated p-value for being variable of at most 1e-6 for the admixed population as determined using ANGSD. By using a GLM framework we were able to account for genetic ancestry and other phenotypic features that could confound the specific genotype-phenotype relationships of interest by including the confounders as covariates. By assuming that the phenotypes are normally distributed our models for the distribution of each of the phenotypes conditional on genotype, $\mathbf{g}$, genetic admixture proportion, $\mathbf{m}$, and other phenotypic features, for individual i were

$$\mathbf{pattern_{rot}}_i \sim N(\alpha_0 + \alpha_1 \mathbf{m}_i + \alpha_2 \mathbf{black}_i + \beta \mathbf{g}_i^\top, \sigma^2)$$

$$\mathbf{LC_a}_i \sim N(\alpha_0 + \alpha_1 \mathbf{m}_i + \alpha_2 \mathbf{LC_L}_i + \beta \mathbf{g}_i^\top, \sigma^2)$$

$$\mathbf{DC_a}_i \sim N(\alpha_0 + \alpha_1 \mathbf{m}_i + \alpha_2 \mathbf{DC_L}_i + \beta \mathbf{g}_i^\top, \sigma^2)$$

where $\mathbf{g}_i \in \{0, 1, 2\}$, that is, the genotype can have 0, 1, or 2 copies of the minor allele. The admixture proportion, $\mathbf{m}_i$, is the proportion of striped morph ancestry for individual i estimated with NGSadmix. For modeling the distribution of color quantified along the red-green color axis, we included lightness as a covariate because no color checker card was used and the lighting conditions among photos was not standardized. By including lightness as a covariate we reduced the effect of noise caused by variable lighting conditions. Using these models, we calculated LR_{geno} for each of the SNPs having at least 10 individuals in two of the three genotypic categories with genotype posterior probabilities greater than 0.9. We then calculated p-values for the LR_{geno} values based on a χ^2_1 distribution.

We combined the results from the divergence and admixture mapping using Fisher's method; $S_F = -2\big(log(p_{allele}) + log(p_{geno})\big)$, where p_{allele} and p_{geno} are the p-values correponding to LR_{allele} and LR_{geno}, respectively. We obtained p-values corresponding to S_F through comparison to a χ^2_4 distribution, which is the theoretical asymptotic distribution of S_F under the null. However, distortion of the null p_{allele} distribution caused by population structure even after applying genomic control resulted in inflation of S_F. To obtain a more appropriate distribution for S_F under the null, we fit a maximum likelihood estimate for the number of degrees of freedom, δ, of a χ^2_δ to the empirical distribution of S_F, thereby assuming that there are no genetic associations with phenotype. This is a reasonable assumption because we expect a negligible number of allelic associations due to selection across the exome. We then compared S_F to the χ^2_δ distribution to obtain a slightly improved distribution of p-values for S_F.

4.3 Results

Exome assembly and capture efficiency

We used a custom capture sytem to target a little over 28.28 megabases of the *R. imitator* exome in 124 individuals, the resulting libraries from which were sequenced on the Illumina Hiseq 4000. After performing quality control on the resulting sequencing reads we used Spades to generate separate assemblies for three banded population individuals and three striped population individuals, representing on average 61,471,541 reads per individual (range 55,318,560 - 67,710,836). These six assemblies were then merged, resulting in a final in-target exome assembly of 76,414,788 base pairs, 87.7% of which belonged to annotated genes, with the remaining 12.3% assigned to unannotated genes that showed either constantly high or differential expression in our set of transcripts used for probe design (Table 4.1). We were able to generate assemblies for 98.7%, that is 13,086, out of the 13,265 targeted genes. The assembly was comprised of 101,607 contigs having an N50 of 786 bp, which represent exons and a portion of their flanking sequences. The average gene length among all genes represented in the assembly was 5,825 bp.

We mapped the quality-controlled reads for each individual to our exome assembly using NovoAlign. For the banded and striped population capture experiment, on average 66.9% (variance = 3.3) of reads mapped uniquely to the exome assembly, with the range of unique mapping among individuals spanning 63.9 - 68.8%. For the admixed population capture, an average of 69.3% (variance = 0.2) of reads mapped uniquely to the exome assembly with unique mapping ranging from 68.1% - 70.4% among individuals. On average, 4.8% and 3.2% of reads mapped to multiple locations in the exome reference from the striped/banded and admixed population captures, respectively. This resulted in an average sequencing coverage of 14.5X and 12.2X per individual for the banded and striped populations, respectively, among the SBall set of quality-controlled sites. The average sequencing coverage per individual in the admixed population was 16.1X for the ADMIXall set of sites.

Phenotyping

We quantified the degree of pattern rotation (stripedness versus bandedness), dorsal and leg color along both the red-green and lightness color axes of the Lab color space, and the proportion of black on the dorsum for 85 frogs using digital images. Although only the 58 admixed individual's phenotypes were used for genetic mapping, we included 15 banded and 12 striped population individuals to ensure the accuracy of our phenotyping methods since these morphs should show clear separation in terms of their phenotypic values. Our automated phenotyping worked well as demonstrated by logical ordering of the *R. imitator* images based on the phenotypic values for traits that we sought to map genes for (Figures 4.3 - 4.5) and those which we used as covariates in the association analysis (Figures 4.6 - 4.8). There is a negative relationship between leg color and pattern$_{rot}$, meaning that increasingly banded frogs tend to have redder legs, while all other traits of interest appeared independent (Figure 4.9). As expected for uniform lighting conditions across the entire body of a single individual, there was a strong relationship between leg and dorsal color measured along the lightness axis, which was quantified to control for lighting variability between individuals. Interestingly, while lighting conditions appear to have a strong influence on dorsal color, that is brighter lights make frogs appear less red, pushing them towards the green end of the color spectrum, there appears to be no relationship between lighting conditions and leg color.

Population genetic characterization

We sequenced the exomes from 33 striped morph, 33 banded morph, and 58 striped/banded admixed *R. imitator* individuals. Our quality controlled consensus dataset purged of paralogs consisted of 26,910,103 sites. The site frequency spectra generated from these sites do not show any abnormalities, and so were deemed reliable to make population genetic characterizations from (Figure 4.10). Among all 124 individuals we identified 292,159 sites with a p-value of $\leq 10^{-6}$ of being variable. Genetic diversity is highest in the striped population for which we identified 2.7x as many segregating sites as in the banded population (193,249 striped SNPs and 70,580 banded SNPs), while the admixed population with 188,011 SNPs is more similar to the striped population in terms of the number of segregating sites. Estimates of Watterson's theta, θ_W, and nucleotide diversity, π, were highest for the striped population, followed by the the admixed population, and lowest in the banded population (Table 4.2).

We also found evidence for exome-wide genetic structure between the striped and banded populations with an F_{ST} of 0.137. The admixed population, while genetically intermediate between the striped and banded populations along the primary axis of genetic variation in a principle components analysis, is substructured into two dinstinct groups (Figure 4.11). Measures of F_{ST} (Table 4.3) indicate that the admixed population is slightly more genetically similar to the striped than banded population, which is corroborated by a greater proportion of the admixed sample showing more striped ancestry than banded (Figure 4.11).

Candidate gene identification

To reveal potential color and pattern genes in *R. imitator* we identified SNPs having exceptionally strong alleleic associations between the striped and banded population using a likelihood ratio test. Even after applying genomic control based on the mean exome-wide LR_{allele}, genetic structure between the two populations (Figure 4.11) caused inflation of the LR_{allele} values, thus biasing formal statistical testing (Figures 4.12 & 4.13). Consequently, we limited our analysis to exploring the relative ranking of allelic association. For the divergence mapping, we examined 563,187 variable sites among the striped and banded morphs, representing 13,086 genes (83% of which we were able to annotate using *Xenopus*). Table 4.4 lists the relative ranking for the top 35 most highly associated genes from the divergence analysis. Among the 10 genes with SNPs having the strongest associations (Table 4.5), the first and fourth ranked genes, MC1R and ASIP, are well known melanogenesis genes. The third and eighth ranked genes, ARG2 and PTCH2, we also noted as being potentially strong candidates due to their relevant biological functions.

Given the challenge of formal statistical testing using LR_{allele}, we also looked for genotype-phenotype associations among 58 admixed individuals at ~143.9K SNPs while controlling for genetic ancestry, to provide evidence that the divergence at the most strongly associated SNPs was not confounded by exome-wide divergence. We focused on 3 traits; pattern, leg color, and dorsal color, that exhibit phenotypic divergence between the striped and banded morphs. Inclusion of the admixture proportion in the admixture mapping model effectively controlled for the influence of ancestry on the association between SNP-specific genotypes and phenotypes (4.12). We combined the results from the divergence and admixture mapping using Fisher's combined p-value statistic, S_F, which for the highest values indicates

SNPs showing both high allelic divergence between the striped and banded morphs and strong genotype-phenotype associations among admixed individuals while controlling for background genetic structure. Any candidates from the divergence analysis that were confounded by population structure should not intersect with the combined test candidates since this test controls for ancestry. We examined S_F at ∼141.1K SNPs representing 12,544 genes. Among the most highly associated 35 genes for each phenotype in terms of S_F (Tables 4.6 - 4.8), we identified seven color and pattern candidates (Table 4.9) based on either highly relevant biological function and/or being ranked among the top 10 genes in both the divergence and the combined test (Figure 4.14). Among these seven genes, five were present among the top 10 from the divergence analysis, while RETSAT was the most highly associated gene with dorsal color and KRT8.2 was the second most associated gene with leg color (after MC1R) according to S_F. Each of the seven candidate genes' SNPs with the greatest S_F all had both LR_{allele} values above the 99th percentile across all associated phenotypes, except for KRT8.2 (93rd LR_{allele} percentile) and LR_{geno} values at or above the 99th percentile for at least one phenotype, except for ASIP which fell into the 86th and 81st LR_{geno} percentiles for pattern and leg color, respectively (Table 4.10). Among the top 35 most highly associated genes from the divergence analysis (Table 4.4), none of the non-candidate genes are in the intersection between the top 10 divergence and combined test genes. Out of the top 10 combined test genes we found that the only ones that showed evidence for effecting multiple phenotypes (Figure 4.15) were those which were present in the intersection of the genes found by the divergence and combined divergence/admixture analyses respectively.

4.4 Discussion

We used a custom exon capture system to survey over 13K genes in 124 samples of the mimic poison frog, *R. imitator*, representing banded, striped, and admixed morphs. We found twice as much genetic diversity in the striped population as compared to the banded, and intermediate levels of diversity in the admixed population. We found genetic substructure among the admixed individuals resembling two strata of striped and banded ancestry. We also found clear exome-wide differentation between the striped versus banded populations, though with an F_{ST} of 0.137, we ought to have power to map genes underlying the color and pattern differences between the two populations without excessive genomic background noise according to a simulation study by Crawford & Nielsen [126]. When we compared 33 striped to 33 banded *R. imitator* we found that in practice the population structure distorted the distribution of our association statistic, LR_{allele}, even after applying genomic control, making formal statistical outlier detection difficult. Consequently, to avoid making any invalid statistical claims we decided to focus on the relative rank ordering of LR_{allele}, which indicates the relatively most divergent genes in the *R. imitator* exome. In order to be sure that we were not confounding genes associated with the different phenotypes due to background structure, we incorporated genotype-phenotype correlation information from the admixture zone in a way that successfully controlled for genetic ancestry. Using S_F

as a measure of association that reflects both the degree of divergence between the striped and banded morphs and association of genotype with specific phenotypes while controlling for genetic structure we identified seven potential color and pattern candidate loci: MC1R, ASIP, ARG2, BSN, PTCH2, RETSAT, and KRT8.2. Each of these candidates has independent measures of allelic divergence between the striped and banded population, LR_{allele}, and genotype-phenotype association among admixed individuals, LR_{geno}, for at least one phenotype in the most extreme upper tails of their empirical distributions, indicating that it was not just exome-wide divergence driving their high S_F association ranking.

MC1R, ASIP, ARG2, BSN, and PTCH2 are among the top 35 genes showing the strongest allelic exome-wide divergence, with MC1R ranked number one. RETSAT, a retinol metabolism gene, is the strongest associated gene with dorsal color quantified along the red-green color axis, and is exclusively associated with this phenotype. Retinol directly influences epithelial cell function and certain classes of retinoids, specfically β-carotene, are involved in the transition from yellow to more red phenotypes in birds [127, 128]. Therefore, the association of RETSAT with the transition of yellowish, striped to orange, banded *R. imitator* seems plausible. KRT8.2, a keratin encoding gene responsible for producing keratin filaments in epithelial cells, was the second most strongly associated gene for leg color, after MC1R, and was exclusively associated with this phenotype. The association of KRT8.2 with leg color would make biological sense, particularly in light of MC1R's involvement, because the interplay of melanin with keratin is known to produce the green structural colors in birds [129, 130, 131]. There is most certainly a structural color component to leg color in *R. imitator*, which changes from green in the striped morph to orange in the banded morph, and so it is possible that a similar interaction of MC1R with keratin could in part account for the difference in leg color.

PTCH2 and ASIP, while among the top 10 most divergent genes between the striped and banded morphs (particularly ASIP with six out of the 17 most highly associated SNPs among the 10 genes), show the weakest signal of association among all of the candidates when involving direct comparison with the admixed phenotypes. If these genes are in fact related to color and pattern, this weaker signal may just be because we did not quantify the phenotypes in a manner as directly related to the specific effect of these genes as compared to the other candidates. Nevertheless, PTCH2 and ASIP, which are involved in skin development and melanogenesis, respectively, may be effecting pattern formation in *R. imitator*. ASIP, an antagonist of MC1R, additionally shows evidence of influencing leg color, for which MC1R is a very strong candidate. Though PTCH2 and ASIP rank less strongly than the other candidate genes in their association with the particular phenotypes that they may be influencing, among all of the top 35 most divergent genes between the striped and banded morph, they are still more strongly associated with their respective phenotypes than any other annotatable, non-candidate, gene.

The genes for which we have the most evidence for influencing color and pattern in *R. imitator* are MC1R, ARG2, and BSN. In our analysis of over 13K genes, these were the only ones among the top 10 most highly divergent genes between the striped and banded morphs to show both this excessively high divergence, and to also be among the the top 10 most strongly associated genes in our combined divergence/admixture analysis using S_F. This means that the evidence is very strong that the extreme allelic association of these genes with their respective phenotypes is not just a consequence of genetic background structure. Furthermore, MC1R, ARG2, and BSN are the only genes among the 10 most divergent and associated genes in our analysis that combined the divergence with admixture mapping to show an association with multiple phenotypes, with MC1R being associated with all three phenotypes that we examined. Given the biological function of MC1R and ARG2 this seems quite feasible, and should support their role as candidates. MC1R is directly involved in melanogenesis, and has a well documented role in influencing color across many taxa [132, 133] including dendrobatids [117]. According to Posso-Terranova & Andrés [117], the density of melanosomes influences the dorsal background color in the harlequin poison frog, *Oopaga histrionica*. Apart from MC1R playing a role in melanosome aggregation, ARG2 may also play a relevant role. Kim *et al.* [134] recently experimentally showed that upregulation of ARG2 in humans results in skin pigmentation by reducing the degradation of melansomes, therefore, it seems plausible that ARG2 could be playing a role in regulating the melanosome density known to influence color in poison frogs. Heterogeneous melanin production and melanosome degradation across the frog body could also certainly produce pattern differences. BSN encodes a scaffolding protein involved in organizing the presynpatic cytoskeleton of axons, making its mechanistic role in how it could be influencing color and pattern less clear despite its very convincing association with these phenotypes.

While evidence for all of these candidate color and pattern genes is currently circumstantial, we believe that the support for particularly MC1R and ARG2 is very compelling. We are currently underway with experiments to validate these genes. Importantly, our exome capture approach and analytical framework demonstrates an effective methodology for identifying candidate loci for a non-model species with a daunting genome and lacking any pre-existing genomic resources.

4.5 Acknowledgements

This research was supported by the NSF grant RUI: Collaborative Research: "Genomic and phenotypic analyses of color pattern divergence in a mimetic radiation of poison frogs". KS would like to express gratitude for a grant from the National Geographic Society (1375-13): "Population genomics of geographic divergence in a mimetic radiation in Peruvian poison frogs" and research funding from a Thomas Harriot College Advancement Council Distinguished Professorship in the Natural Sciences and Mathematics at East Carolina University. TL was supported by an NIH Biomedical Big Data training grant 1T32LM1241-7.

4.6 Tables

Table 4.1. Statistics for the *R. imitator* exome assembly discounting the Ns between unmerged contigs belonging to the same gene. The gene statistics use the summed length across all unmerged contigs belonging to a gene

assembly subset	**number genes**	**assembled sequence (bp)**	**gene N50 (bp)**	**gene mean length (bp)**	**gene median length (bp)**	**number contigs**	**contig N50 (bp)**	**contig mean length (bp)**	**contig median length (bp)**
all genes	13,086	76,414,788	7,336	5,825	4,681	101,607	786	752	605
annotated genes	10,842	67,050,479	7,569	6,184	5,044	90,053	774	745	603
unannotated genes	2,244	9,364,309	5,695	4,113	3,015	11,554	885	810	624

Table 4.2. Estimates of θ_W and π for the banded, striped, and admixed *R. imitator* populations.

population	$\boldsymbol{\theta_W}$	$\boldsymbol{\pi}$
striped	0.001932	0.001036
banded	0.000719	0.000434
admixed	0.001413	0.000686

Table 4.3. Pairwise F_{ST} between the banded, striped, and admixed *R. imitator* populations.

	banded	**striped**	**admixed**
banded		0.1372	0.0668
striped	0.1372		0.0564
admixed	0.0668	0.0564	

Table 4.4. Ranked association strengths for SNPs belonging to the top 35 genes showing the greatest divergence between the striped and banded *R. imitator* morphs. Genes considered to be candidates for influencing color and pattern are denoted in bold.

			banded	**striped**
	SNP		**allele**	**allele**
gene	**position**	$\mathbf{LR_{allele}}$	**frequency**	**frequency**
mc1r	1662	27.364	0.00	0.90
mc1r	1813	26.300	0.00	0.88
arg2	2193	25.935	0.00	0.89
asip	219	25.739	0.02	0.93
asip	1303	25.732	0.00	0.88
asip	1231	24.907	0.00	0.85
asip	1312	24.774	0.00	0.86
bsn	1764	24.682	0.05	0.93
itpkc	573	24.151	0.02	0.88
asip	1500	23.862	0.00	0.87
asip	1327	23.808	0.00	0.84
spire2	1532	23.714	0.00	0.83
ptch2	8521	23.258	0.03	0.91
rnf182	198	23.155	0.02	0.88
zbtb40	1173	23.086	0.02	0.86
mphosph10	5850	23.078	0.03	0.90
soat2	405	23.063	0.00	0.82
asip	1031	23.038	0.00	0.88
tm4sf4	3032	22.846	0.02	0.85
gpr61	824	22.763	0.02	0.85
colgalt2	7481	22.744	0.00	0.81
asip	1542	22.703	0.00	0.88
contig6769	405	22.688	0.92	0.06
fbxo41	6663	22.500	0.06	0.91
ahcy	5337	22.440	0.00	0.81
nicn1	843	22.437	0.90	0.03
sfxn5	5338	22.431	0.02	0.85
fam179a	926	22.233	0.00	0.90
asip	306	22.149	0.02	0.85
eps8	5615	21.983	0.00	0.81
bsn	6952	21.868	0.88	0.04
slc22a31	1794	21.852	0.00	0.86
app	311	21.843	0.02	0.85
tenc1	5730	21.839	0.05	0.89
asip	1582	21.749	0.00	0.90

contig3030	5178	21.662	0.02	0.83
asip	1553	21.602	0.00	0.87
tll1	461	21.519	0.89	0.06
mb21d2	1080	21.476	0.03	0.86
contig12957	3141	21.390	0.02	0.83
cdh15	5206	21.366	1.00	0.23
cdh5	6797	21.346	0.93	0.09
zcchc14	8028	21.281	0.02	0.82
nrcam	17755	21.147	0.00	0.81
vegfc	313	21.145	0.02	0.83
phlpp2	7275	21.123	0.00	0.80
rc3h1	1101	21.03	0.00	0.77

Table 4.5. Ranked order of SNPs showing the most divergence between the striped and banded R. imitator morphs based on LR_{allele}, listed up until the 10th-ranked gene. Candidate genes are in bold.

gene	SNP position	$\mathbf{LR_{allele}}$	banded allele frequency	striped allele frequency
mc1r	1662	27.364	0.00	0.90
mc1r	1813	26.300	0.00	0.88
arg2	2193	25.935	0.00	0.89
asip	219	25.739	0.02	0.93
asip	1303	25.732	0.00	0.88
asip	1231	24.907	0.00	0.85
asip	1312	24.774	0.00	0.86
bsn	1764	24.682	0.05	0.93
itpkc	573	24.151	0.02	0.88
asip	1500	23.862	0.00	0.87
asip	1327	23.808	0.00	0.84
spire2	1532	23.714	0.00	0.83
ptch2	8521	23.258	0.03	0.91
rnf182	198	23.155	0.02	0.88
zbtb40	1173	23.086	0.02	0.86
mphosph10	5850	23.078	0.03	0.90

Table 4.6. Ranked association strengths for SNPs belonging to the top 35 genes showing both the greatest divergence between the striped and banded *R. imitator* morphs and genotype association with pattern in the admixed population. Genes considered to be pattern candidates are denoted in bold. $*$ = gene is ranked among the top 35 most divergent genes between the striped and banded morphs.

			banded	**striped**
	SNP		**allele**	**allele**
gene	**position**	$\mathbf{S_F}$	**frequency**	**frequency**
mc1r*	1662	32.767	0.00	0.90
pitpnm2	382	32.387	0.89	0.08
mc1r*	1813	31.961	0.00	0.88
arg2*	2193	31.081	0.00	0.89
bsn*	1764	30.779	0.05	0.93
htr3a	5212	28.299	0.03	0.79
clip1	9830	28.035	0.05	0.84
fam178a	677	27.254	0.05	0.79
synj1	13999	26.960	0.09	0.47
arg2*	2158	26.872	0.01	0.87
contig6769*	405	26.623	0.92	0.06
contig12985	4310	26.474	0.02	0.81
decr2	930	26.383	0.03	0.85
lcorl	4389	26.298	0.00	0.65
contig6742	866	26.291	0.09	0.55
asip*	1303	26.013	0.00	0.88
sorcs3	11795	25.699	0.22	0.63
mb21d2*	1080	25.180	0.03	0.86
atp13a1	5008	25.127	0.03	0.56
fkbp8	361	25.036	0.00	0.69
tmem63b	7653	25.028	0.02	0.79
barx1	447	24.880	0.09	0.59
mphosph9	355	24.856	0.88	0.07
mb21d2*	2501	24.852	0.03	0.82
contig7184	3859	24.794	0.06	0.67
fancd2	14551	24.755	0.05	0.87
tceb2	1033	24.687	0.00	0.62
rnf182*	198	24.570	0.02	0.88
ppip5k2	2162	24.566	0.89	0.85
syngr2	1562	24.509	0.05	0.78
tbc1d4	294	24.405	0.13	0.81
ckmt1b	943	24.351	0.00	0.83
ptch2*	6811	24.252	0.03	0.69

asip*	1500	24.197	0.00	0.87
cilp	4471	24.131	0.08	0.83
aifm1	405	24.120	0.07	0.83
kcnk15	411	24.085	0.80	0.10
fkbp8	281	24.029	0.00	0.72
nicn1*	843	23.960	0.90	0.03
fam126a	463	23.848	0.97	0.17

Table 4.7. Ranked association strengths for SNPs belonging to the top 35 genes showing both the greatest allelic divergence between the striped and banded *R. imitator* morphs and genotype association with leg color quantified along the red-green axis in the admixed population. Genes considered to be leg color candidates are denoted in bold. $*$ = gene is ranked among the top 35 most divergent genes between the striped and banded morphs.

			banded	**striped**
	SNP		**allele**	**allele**
gene	**position**	$\mathbf{S_F}$	**frequency**	**frequency**
mc1r*	1662	36.372	0.00	0.90
mc1r*	1813	35.567	0.00	0.88
krt8.2	2306	33.967	0.09	0.67
EpCAM	3181	32.798	0.00	0.68
jak1	1426	32.125	0.14	0.82
arg2*	2193	29.502	0.00	0.89
kda2a	3553	29.294	0.06	0.80
dbnl	3593	27.803	0.00	0.44
unc119b	263	27.682	0.80	0.08
tpi1	1172	27.626	0.20	0.79
contig3238	2438	27.560	0.00	0.58
spon1	8539	27.393	0.00	0.66
cep83	8901	27.135	0.04	0.56
akap11	5190	27.099	0.00	0.77
heatr5a	4484	26.719	0.00	0.64
ttc7b	4575	26.662	0.02	0.72
heatr5a	4483	26.572	0.00	0.63
chsy3	3270	26.465	0.02	0.79
spire2*	1532	26.325	0.00	0.83
dbnl	3602	26.312	0.00	0.57
arg2*	2158	26.288	0.01	0.87
tpi1	1163	26.278	0.19	0.78
usp5	7298	26.189	0.17	0.66
pgs1	3715	26.035	0.86	0.09
trappc6a	759	26.004	0.73	0.07
hepacam	4209	25.943	0.00	0.41
contig8410	2730	25.899	0.13	0.85
lhx8	4094	25.867	0.49	0.05
lmnb3	6286	25.843	0.03	0.61
slc6a16	1317	25.697	0.32	0.08
ttc7b	7941	25.687	0.00	0.83
strap	2761	25.675	0.02	0.75
lasp1	3430	25.617	0.00	0.69

cdc42bpb	4578	25.561	0.02	0.73
nin	16754	25.475	0.73	0.24
contig12475	8235	25.219	0.04	0.85
dip2a	10212	25.180	0.03	0.73
mok	5289	24.980	0.06	0.82
krt8.2	3260	24.930	0.12	0.54
asip*	1231	24.900	0.00	0.85
clstn2	8384	24.886	0.00	0.54
cep83	8915	24.735	0.05	0.58
tmem63b	7653	24.454	0.02	0.79

Table 4.8. Ranked association strengths for SNPs belonging to the top 35 genes showing both the greatest allelic divergence between the striped and banded *R. imitator* morphs and genotype association with dorsal color quantified along the red-green color axis in the admixed population. Genes considered to be dorsal color candidates are denoted in bold. * = gene is ranked among the top 35 most divergent genes between the striped and banded morphs.

			banded	**striped**
	SNP		**allele**	**allele**
gene	**position**	$\mathbf{S_F}$	**frequency**	**frequency**
retsat	6779	32.518	0.00	0.75
arhgef25	7447	30.354	0.03	0.77
tmem63b	7653	30.309	0.02	0.79
rps13	1592	29.253	0.03	0.76
bsn*	6952	28.656	0.88	0.04
clip1	9830	28.272	0.05	0.84
tenc1*	5730	27.972	0.05	0.89
mc1r*	1662	27.830	0.00	0.90
hnrnpa1	1777	27.700	0.11	0.89
strap	2761	27.422	0.02	0.75
arhgef28	6305	27.221	0.06	0.84
contig3228	8822	27.181	0.05	0.82
mc1r*	1813	27.025	0.00	0.88
plekha7	11555	26.853	0.06	0.74
contig6288	1749	26.659	0.02	0.89
contig12922	927	26.235	0.01	0.55
pik3c2a	5961	26.133	0.08	0.73
contig3228	5091	25.981	0.30	0.09
hap1	6525	25.785	0.74	0.19
plekha7	9452	25.705	0.04	0.83
herc1	31561	25.531	0.09	0.60
herc1	31560	25.525	0.09	0.60
vps37b	348	25.479	0.03	0.84
dlg1	9015	25.361	0.76	0.08
heatr5a	4484	25.247	0.00	0.64
sfxn5*	5338	25.214	0.02	0.85
herc1	31559	25.196	0.09	0.58
heatr5a	4483	25.099	0.00	0.63
pnpla6	7978	25.080	0.05	0.70
med20	2290	25.035	0.02	0.56
arhgap44	10566	25.021	0.83	0.08
contig12922	1022	24.857	0.02	0.58

plekha7	10124	24.832	0.06	0.77
contig5910	1832	24.787	0.77	0.09
clns1a	3147	24.742	0.05	0.75
pik3c2a	7852	24.724	0.08	0.76
snx18	1856	24.684	0.07	0.81
st6galnac5	789	24.681	0.10	0.78
syde2	770	24.674	0.08	0.69
adal	5264	24.575	0.06	0.78
contig8410	2730	24.549	0.13	0.85
unc119b	263	24.543	0.80	0.08
contig12477	1377	24.444	0.03	0.81
rc3h1*	1101	24.379	0.00	0.77

Table 4.9. Candidate color and pattern genes identified among the top 35 genes with highest S_F out of ∼14.1K SNPs comprising 12,544 genes.

gene	biological function	associated phenotypes	S_F rank
mc1r	melanogenesis	{pattern, leg color, dorsal color}	{1, 1, 8}
asip	melanogenesis	{pattern, leg color}	{14, 33}
arg2	keratinocyte autophagy	{pattern, leg color}	{3, 5}
ptch2	skin development and epidermal cell fate	pattern	30
bsn	organization of axon presynaptic cytoskeleton	{pattern, dorsal color}	{4, 5}
krt8.2	keratin filament formation in epithelial cells, cellular structural integrity, signal transduction, and differentiation	leg color	2
retsat	retinol metabolism	dorsal color	1

Table 4.10. Association strengths of the top SNP belonging to each potential candidate color/pattern gene in terms of allelic divergence between the striped and banded morphs (LR_{allele}), genotype-phenotype correlation among the admixed individuals (LR_{geno}), and these two measures combined (S_F). The max LR columns refer to the values for whichever SNP belonging to the particular gene has the maximum value with respect to the given LR measure of association.

pattern

gene	SNP position	$\mathbf{LR_{allele}}$	$\mathbf{LR_{allele}}$ percentile	$\mathbf{LR_{geno}}$	$\mathbf{LR_{geno}}$ percentile	$\mathbf{S_F}$	max $\mathbf{LR_{allele}}$	max $\mathbf{LR_{allele}}$ percentile	max $\mathbf{LR_{geno}}$	max $\mathbf{LR_{geno}}$ percentile
mc1r	1662	19.773	100	6.864	99.100	32.767	19.773	100	7.063	99.225
arg2	2193	18.740	99.998	6.327	98.820	31.081	18.740	99.999	6.327	98.820
bsn	1764	17.834	99.994	6.904	99.142	30.779	17.834	99.994	6.904	99.142
asip	1303	18.593	99.998	2.191	85.899	26.013	18.593	99.996	2.316	87.017
ptch2	6811	10.433	99.140	8.121	99.580	24.252	16.806	99.991	8.121	99.580

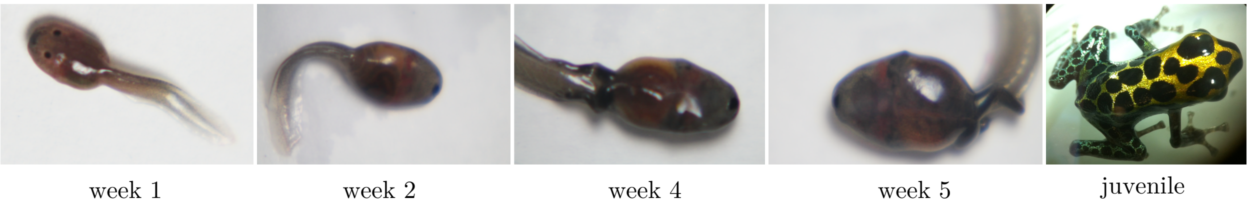
leg color

gene	SNP position	$\mathbf{LR_{allele}}$	$\mathbf{LR_{allele}}$ percentile	$\mathbf{LR_{geno}}$	$\mathbf{LR_{geno}}$ percentile	$\mathbf{S_F}$	max $\mathbf{LR_{allele}}$	max $\mathbf{LR_{allele}}$ percentile	max $\mathbf{LR_{geno}}$	max $\mathbf{LR_{geno}}$ percentile
mc1r	1662	19.773	100	10.142	99.841	36.372	19.773	100	10.142	99.841
krt8.2	2306	5.536	92.541	22.361	100	33.967	5.536	92.541	22.361	100
arg2	2193	18.740	99.998	4.943	97.041	29.502	18.740	99.998	4.943	97.041
asip	1231	17.997	99.997	1.822	81.123	24.900	18.593	99.998	2.222	85.312

dorsal color

gene	SNP position	$\mathbf{LR_{allele}}$	$\mathbf{LR_{allele}}$ percentile	$\mathbf{LR_{geno}}$	$\mathbf{LR_{geno}}$ percentile	$\mathbf{S_F}$	max $\mathbf{LR_{allele}}$	max $\mathbf{LR_{allele}}$ percentile	max $\mathbf{LR_{geno}}$	max $\mathbf{LR_{geno}}$ percentile
retsat	6779	13.092	99.825	13.115	99.970	32.518	13.092	99.826	13.115	99.969
bsn	6952	15.801	99.972	6.922	98.860	28.656	17.834	99.995	6.922	98.860
mc1r	1662	19.773	100	2.646	87.559	27.830	19.773	100	5.314	97.272

4.7 Figures



week 1 week 2 week 4 week 5 juvenile

Figure 4.1. Photos of the developmental stage individuals for which transcriptomes were generated. Transcriptomes were sequenced for the entire body for individuals from weeks 1 through 5. Six separate cDNA libraries were prepared from mRNA for the juvenile stage, representing various tissues with a focus on areas of the skin representing different color and pattern; 1) dorsal trunk, 2) nape and dorsal head, 3) ventral jaw, 4) hind leg skin areas. A library representing the brain and eyes was also prepared for the juvenile stage.

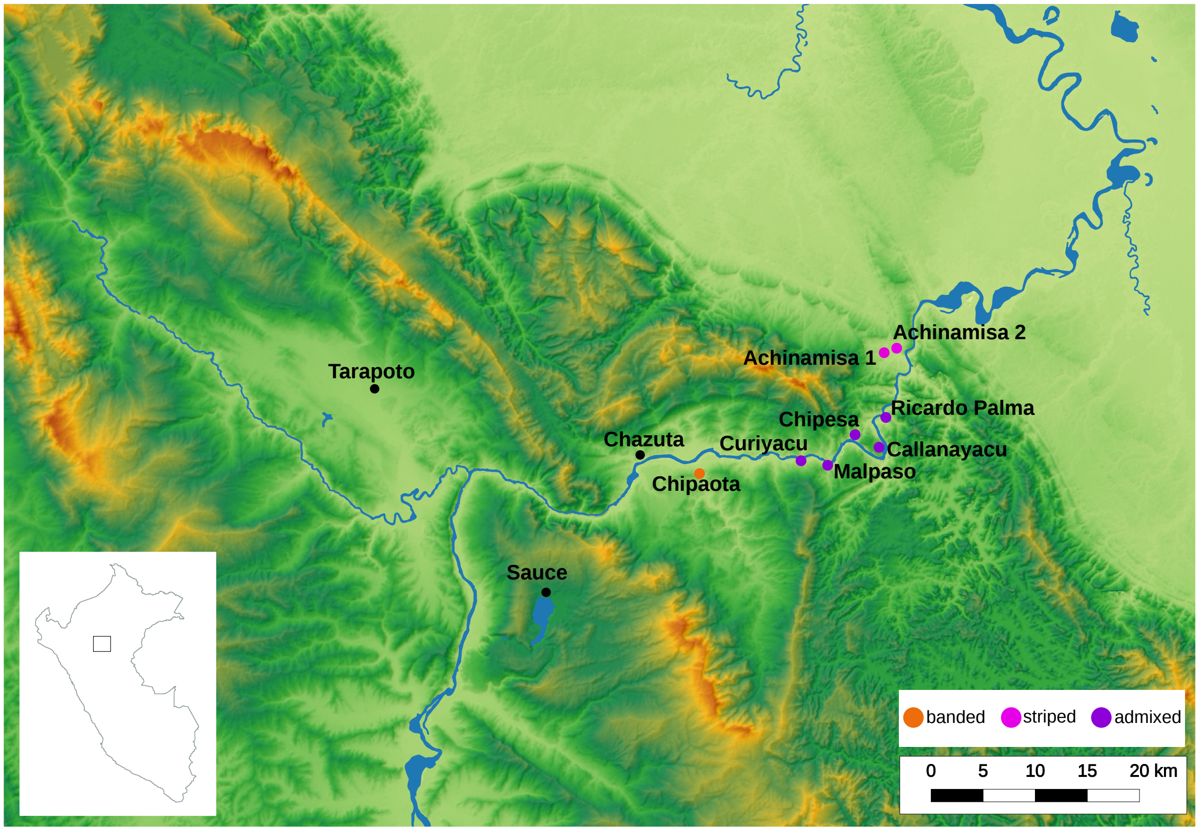


Figure 4.2. Map of the San Martin province in Peru with sampling localities for the banded (orange circles), striped (magenta circles), and admixed (purple circles) morph *Ranitomeya imitator* samples.

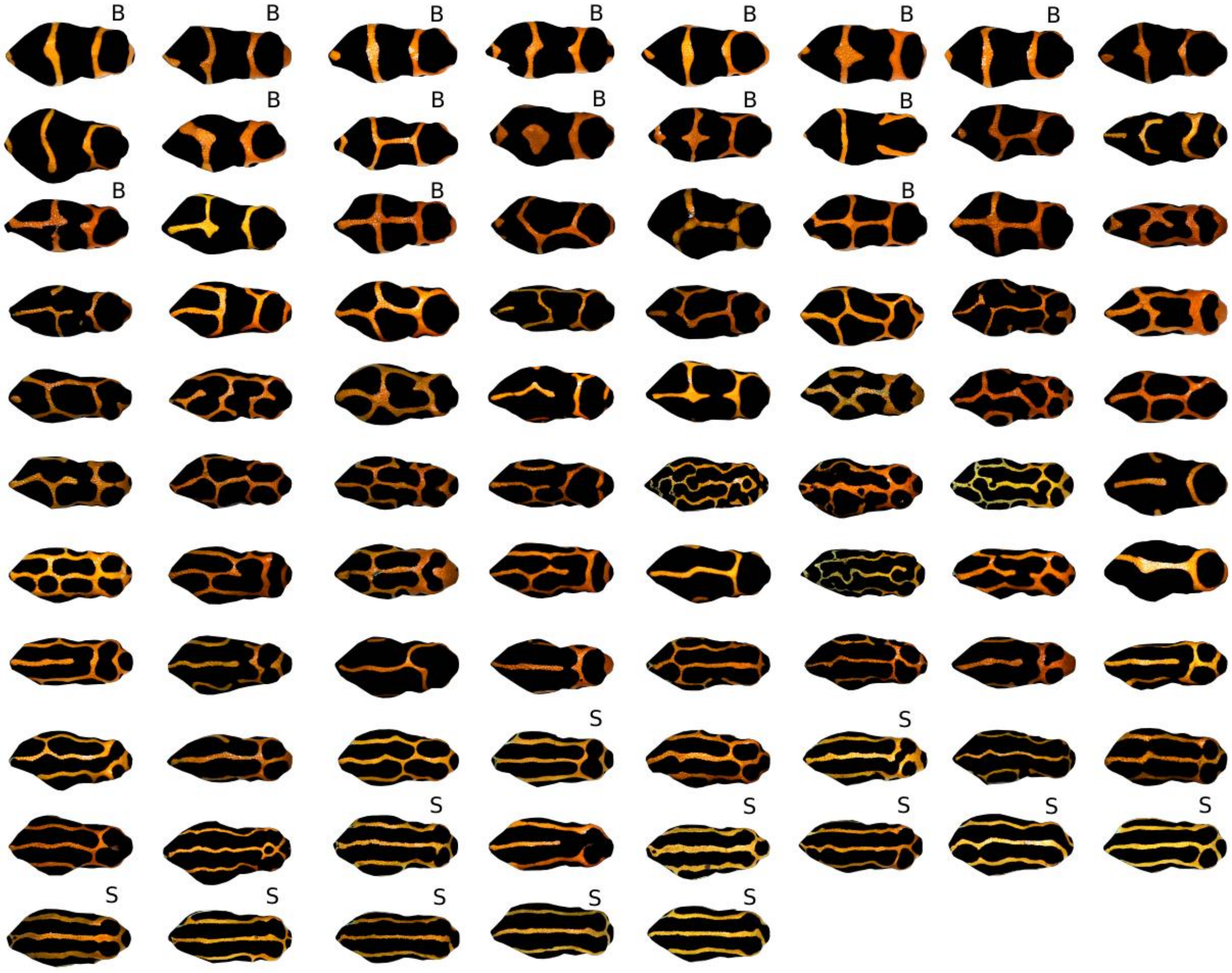


Figure 4.3. Dorsal photos for 58 admixed, 12 pure striped, and 15 pure banded morph *R. imitator* sorted according to pattern$_{rot}$. The images represent values for pattern$_{rot}$ increasing from lowest (most banded) in the upper left-hand corner to highest (most striped) in the lower right. Banded and striped population individuals are indicated with 'B' and 'S', respectively.

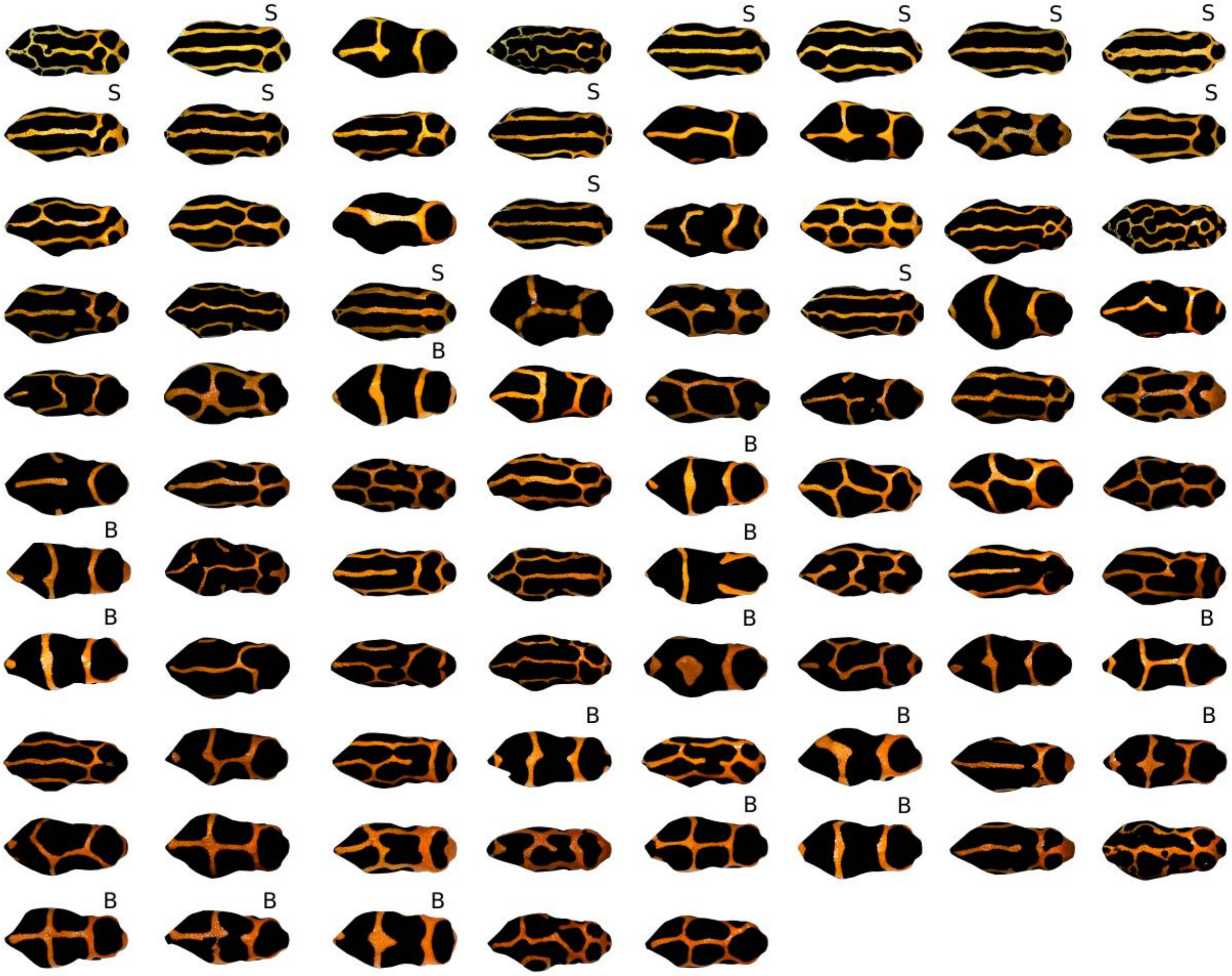


Figure 4.4. Dorsal photos for 58 admixed, 12 pure striped, and 15 pure banded morph *R. imitator* sorted in increasing order of DC_a, which quantifies color along the red-green Lab color axis with green at negative DC_a values and red at positive DC_a values. Banded and striped population individuals are indicated with 'B' and 'S', respectively.

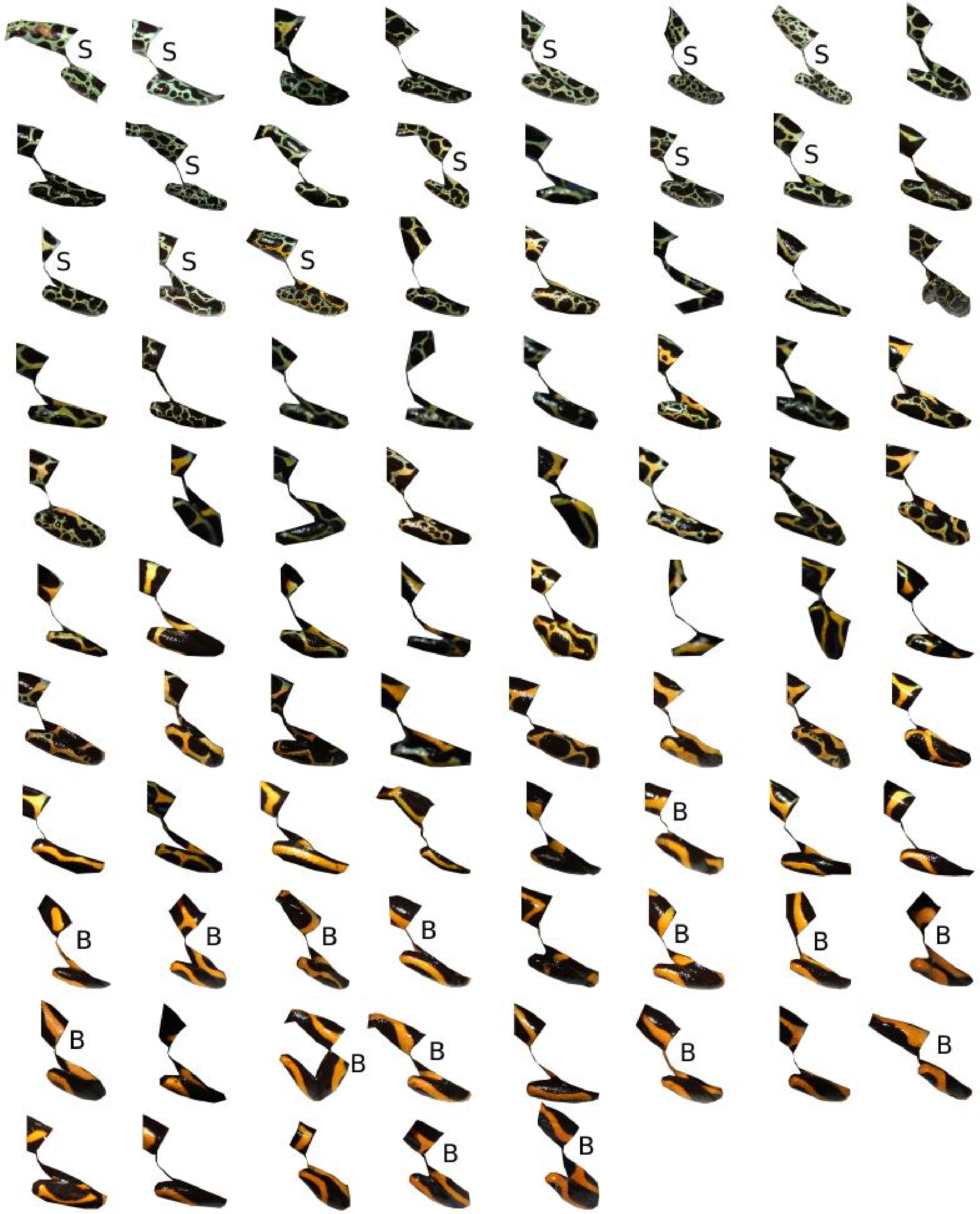


Figure 4.5. Hind leg images for 58 admixed, 12 pure striped, and 15 pure banded morph *R. imitator* sorted in increasing order of LC_a, which quantifies color along the red-green Lab color axis with green at negative LC_a values and red at positive LC_a values. Banded and striped population individuals are indicated with 'B' and 'S', respectively.

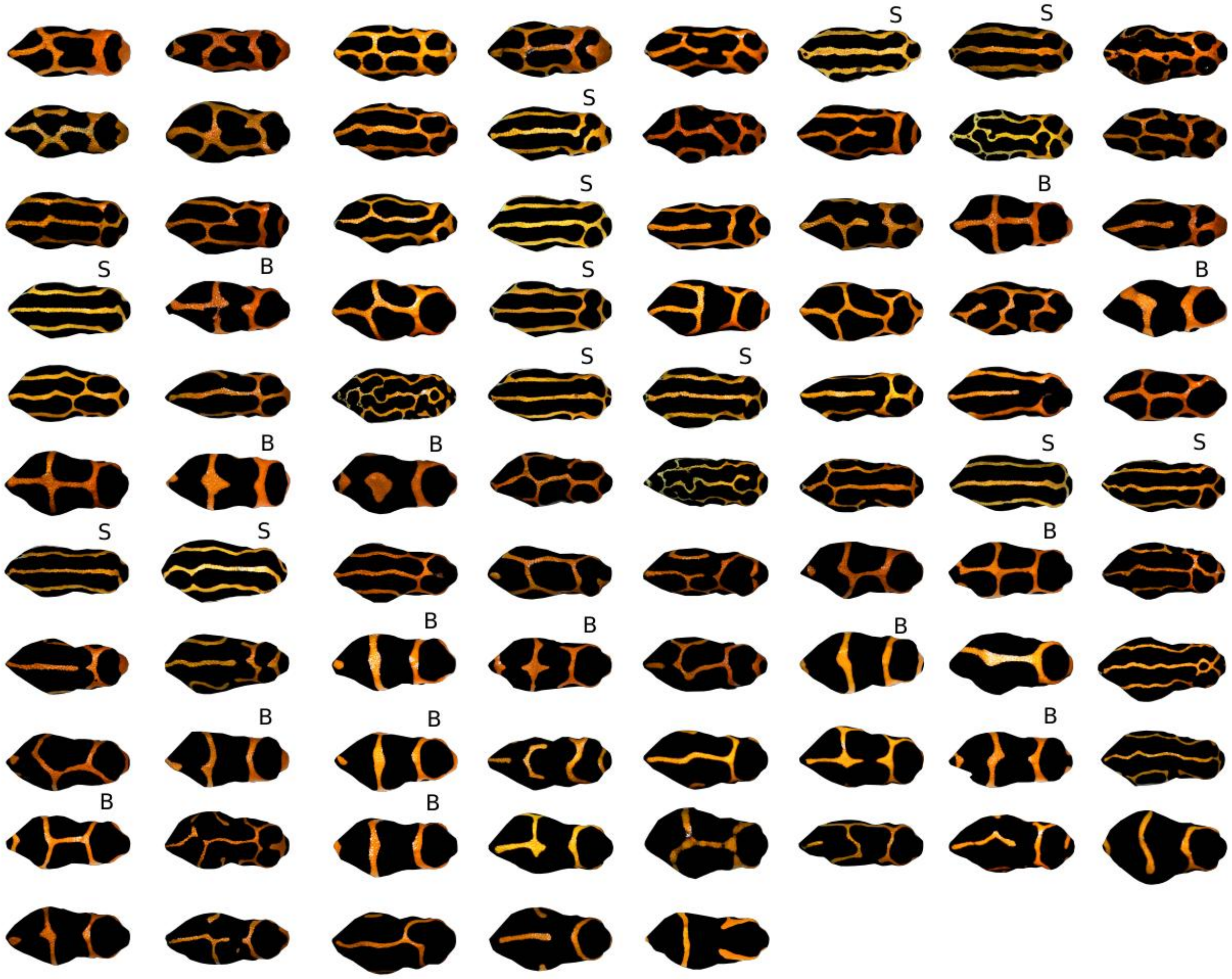


Figure 4.6. Dorsal photos for 58 admixed, 12 pure striped, and 15 pure banded morph *R. imitator* sorted in increasing order of the proportion of the dorsum that is black. Banded and striped population individuals are indicated with 'B' and 'S', respectively.

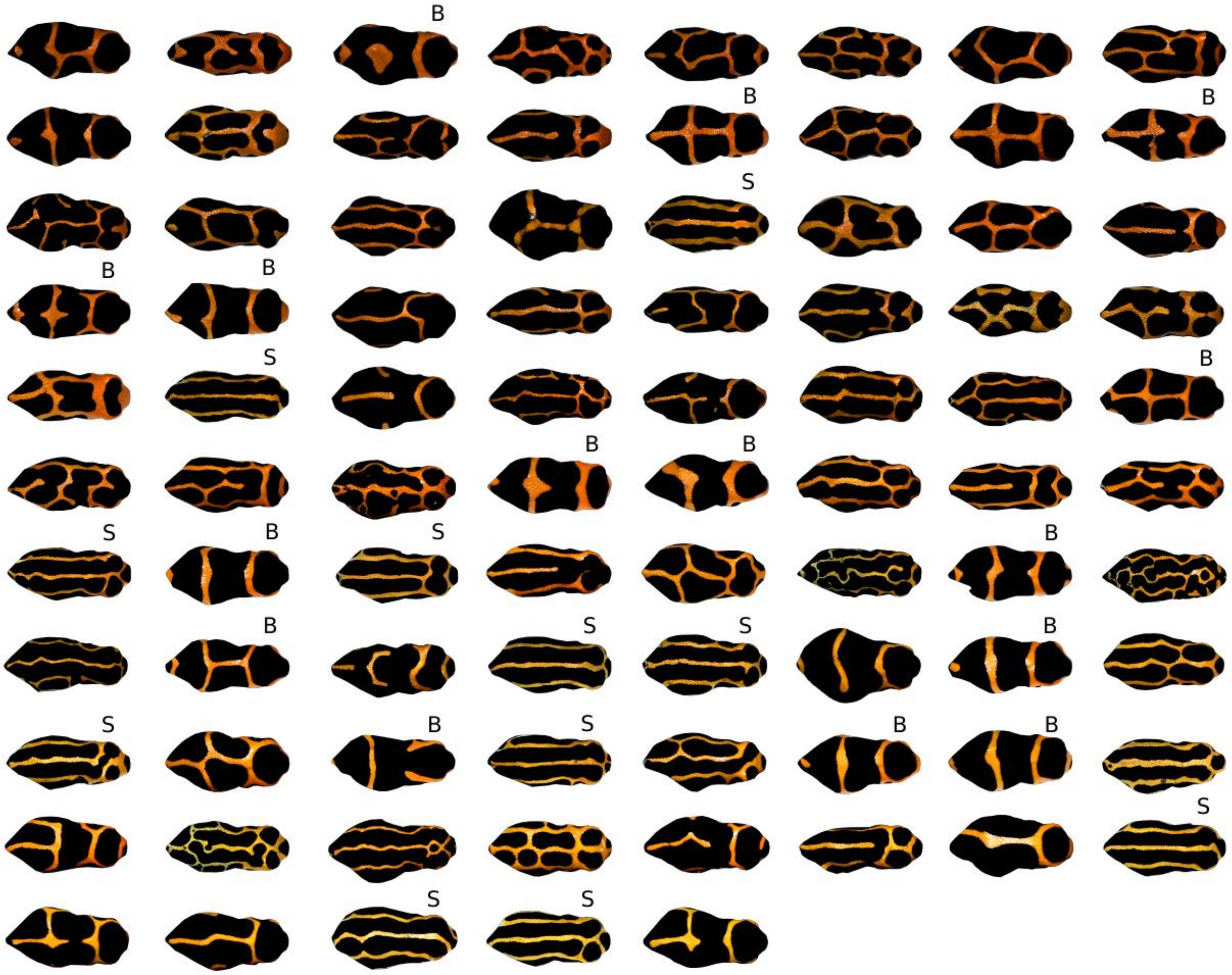


Figure 4.7. Dorsal photos for 58 admixed, 12 pure striped, and 15 pure banded morph *R. imitator* sorted in increasing order of DC_L, which quantifies color along the lightness Lab color axis with black at negative $DC_L = 0$ and white at $DC_L = 100$. Banded and striped population individuals are indicated with 'B' and 'S', respectively.

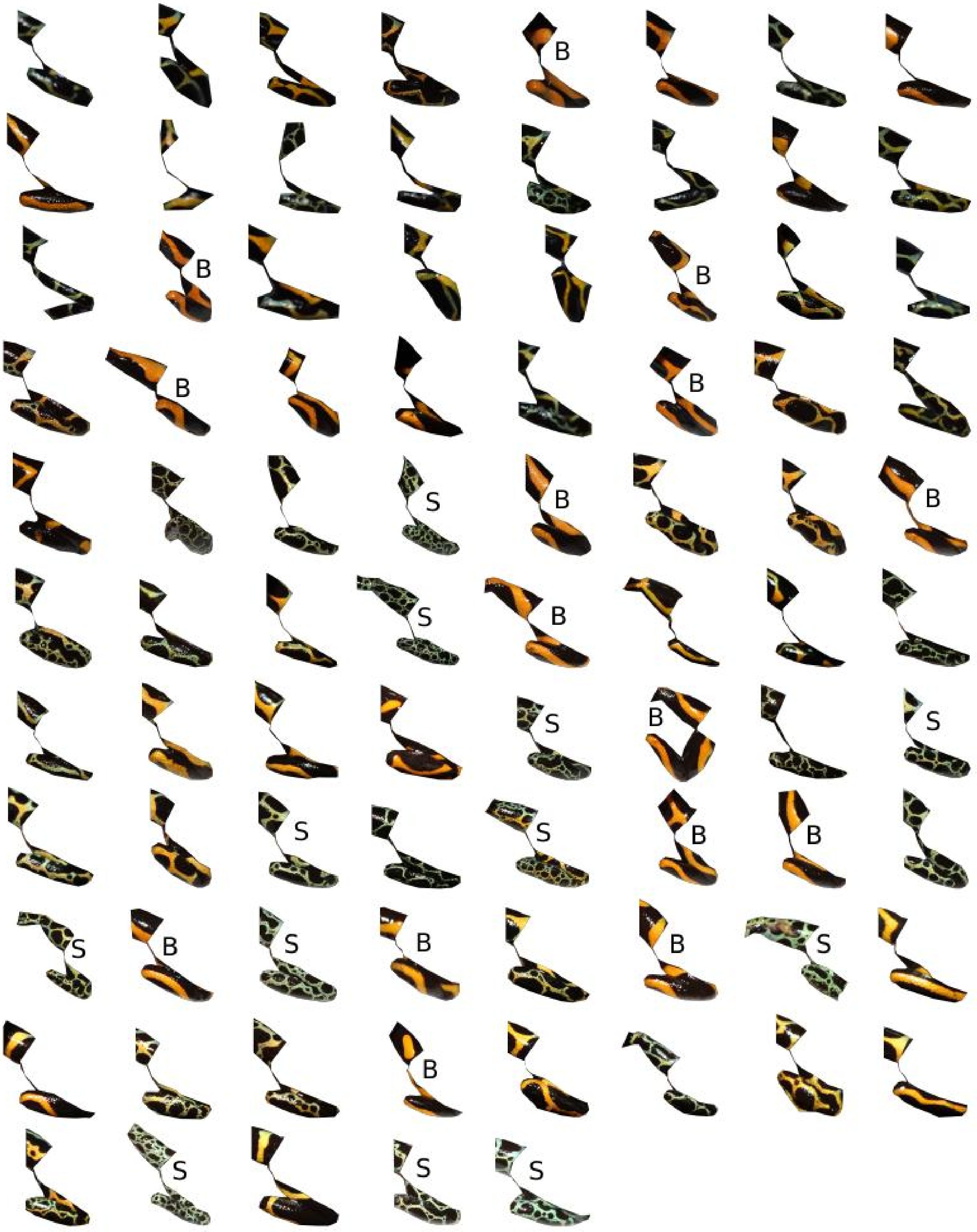


Figure 4.8. Hind leg images for 58 admixed, 12 pure striped, and 15 pure banded morph *R. imitator* sorted in increasing order of LC_L, which quantifies color along the lightness Lab color axis with black at negative $LC_L = 0$ and white at $LC_L = 100$. Banded and striped population individuals are indicated with 'B' and 'S', respectively.

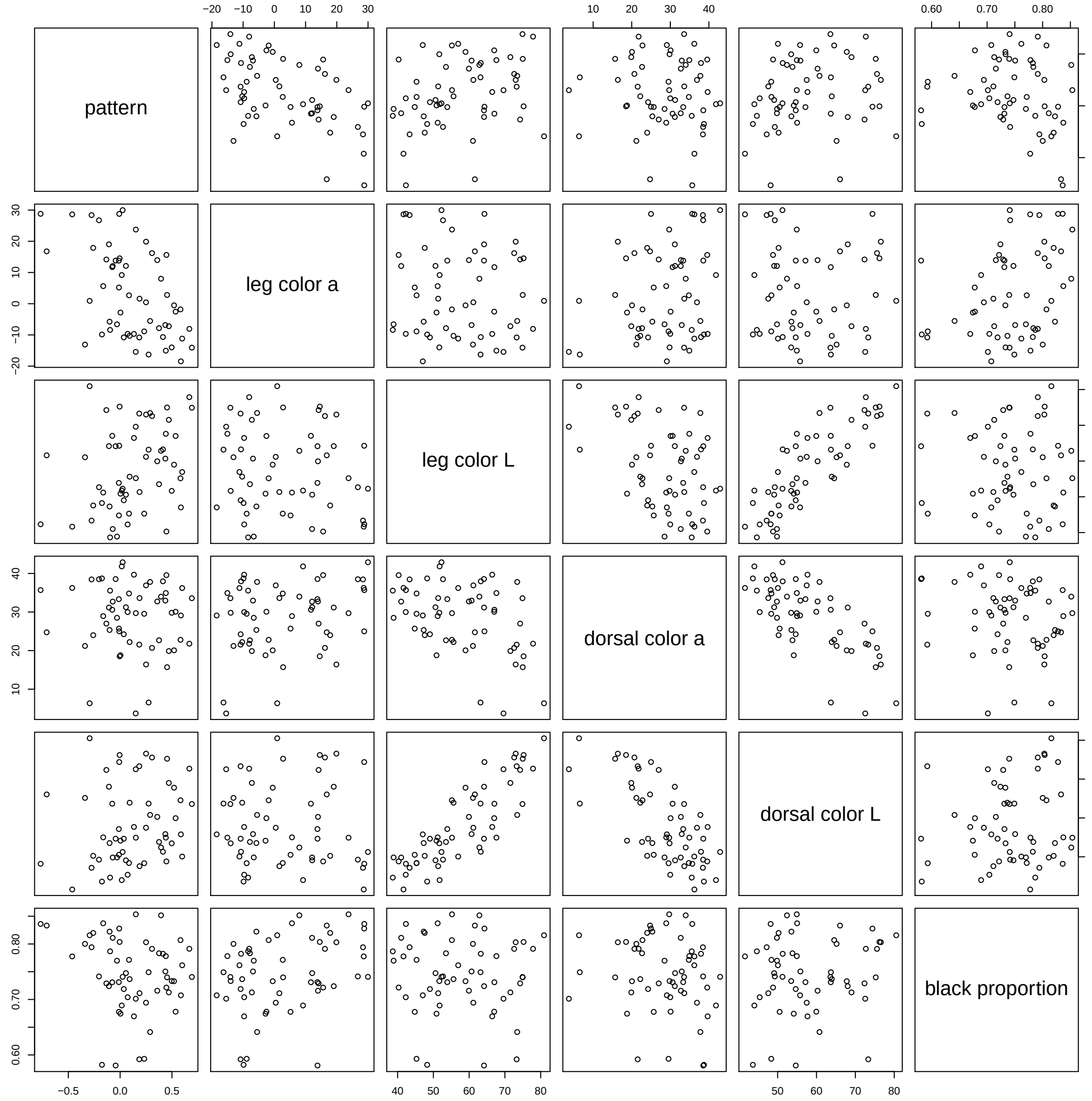


Figure 4.9. Correlations between six phenotypic features that were quantified from the *R. imitator* photos. Each point represents the measurement for an individual. Dorsal/leg color *a* and dorsal/leg color *L* correspond to the color of the frogs quantified along the red-green and lightness axes of the Lab color space. The greater pattern is in the positive direction the more striped are the frogs, while the opposite direction indicates more banded frogs.

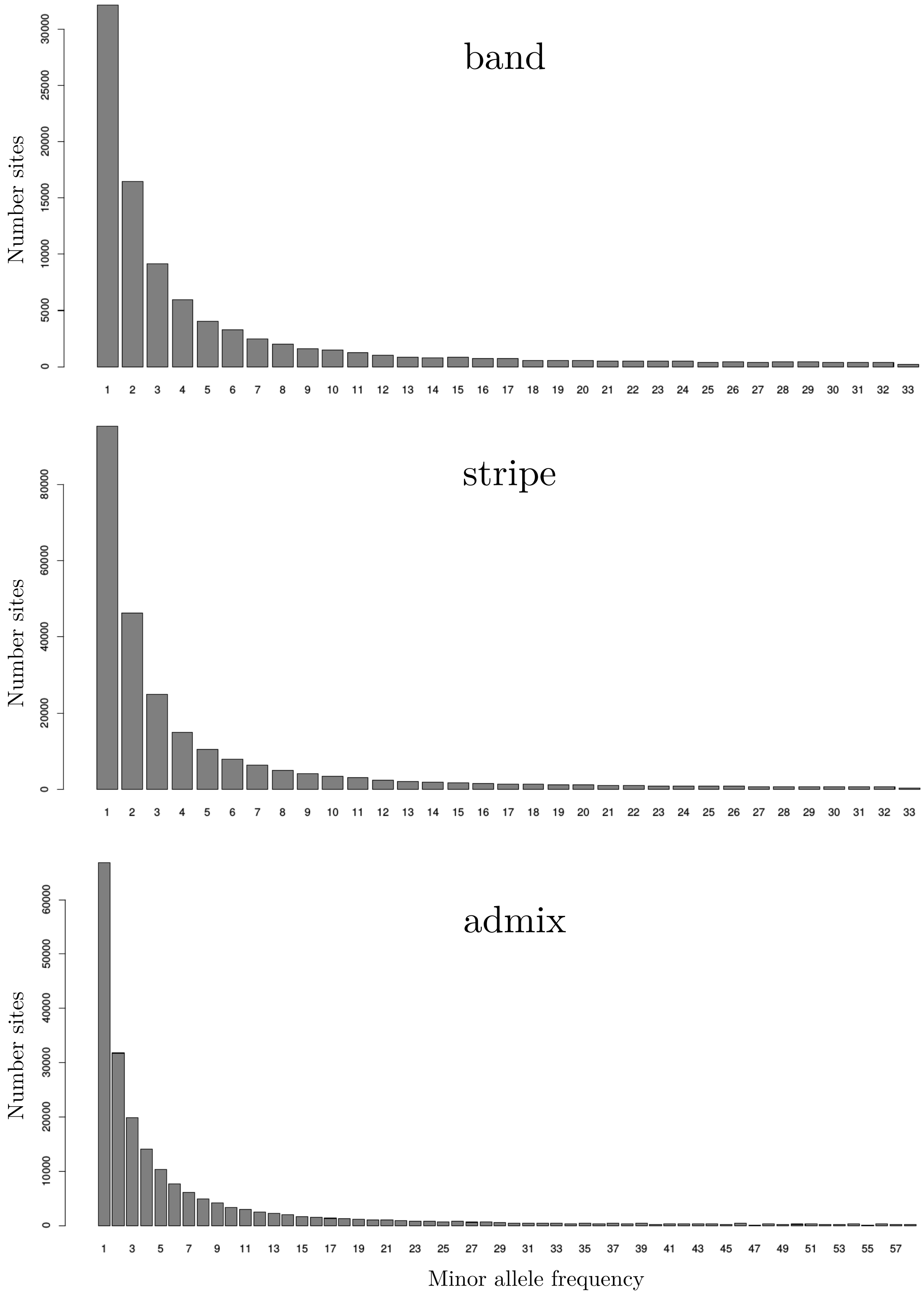


Figure 4.10. Site frequency spectrum for the striped, banded, and admixed *R. imitator* populations generated using a set of 26,910,103 high quality sites among all three populations.

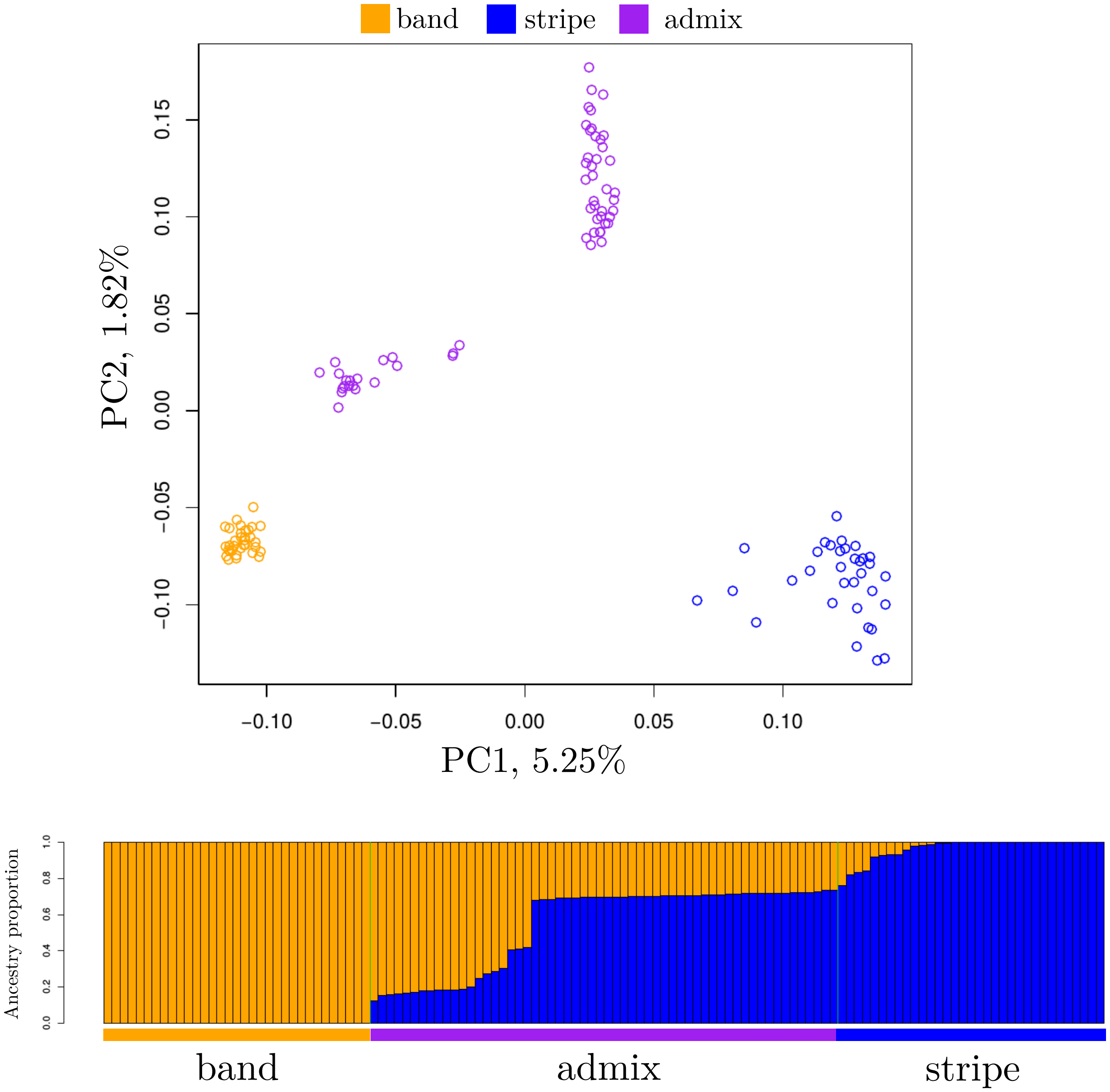


Figure 4.11. Principle components analysis (top) showing the genetic relationship among banded, striped, and admixed *R. imitator* individuals. Each point in the PCA represents an individual, while the axes label percentages denote the amount of variance explained by the respective principle component. In the admixture plot (bottom) each vertical bar represents an individual with their proportion of genetic ancestry indicated by different colors. From left to right in the admixture plot, the first 33 individuals are banded, the following 58 are admixed, and the last 33 are striped.

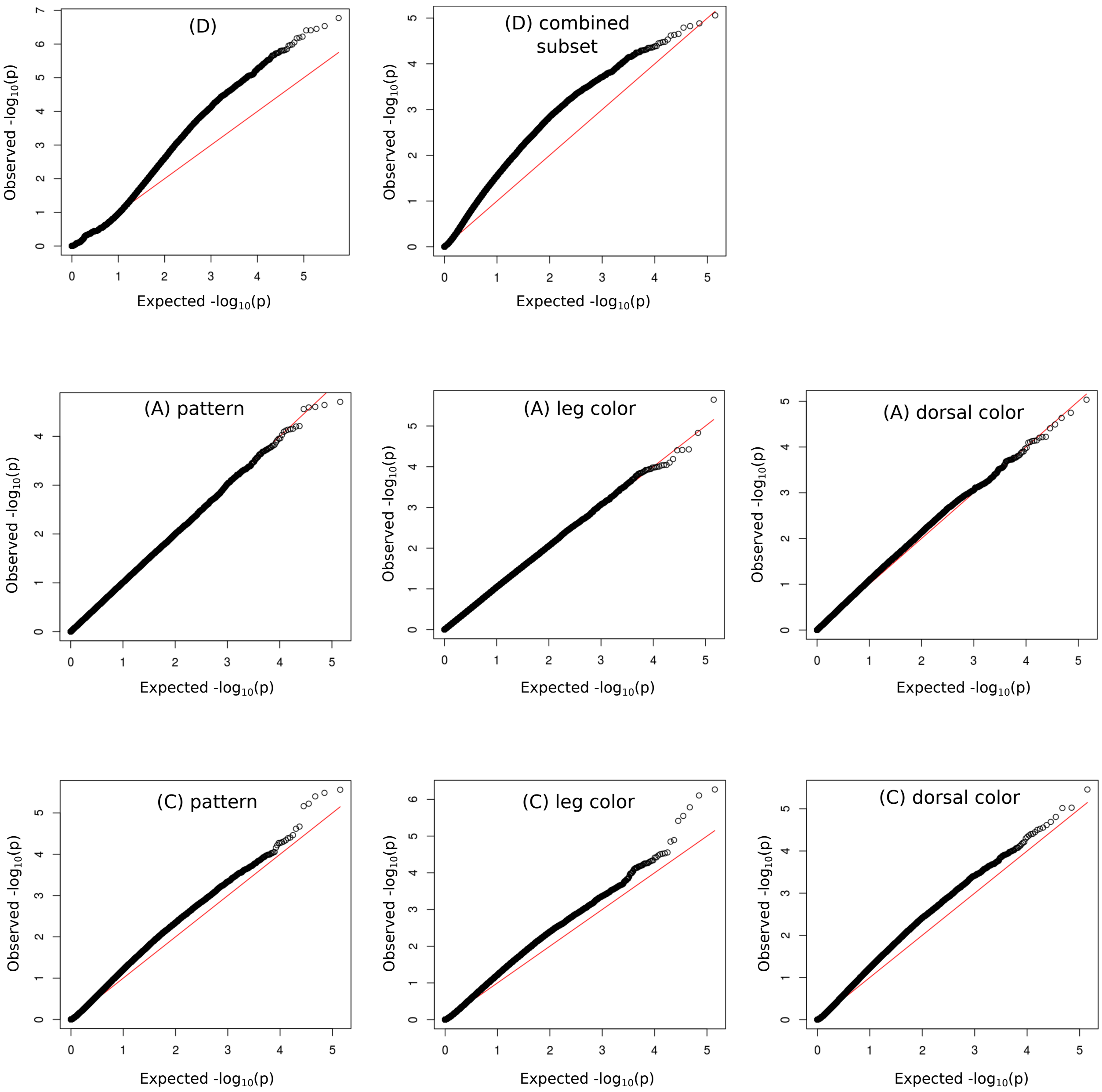


Figure 4.12. Quantile-quantile plots of the observed versus expected p-values corresponding to the likelihood ratios from the divergence test between the banded and striped population after applying genomic control (D) and the score statistic test performed on the admixed population (A), and S_F (C), which is combines the p-values from the divergence and score statistic tests using Fisher's method. The divergence 'combined subset' plot is for LR_{allele} calculated from the subset of consensus quality sites between the banded, striped, and admixed population used to construct S_F. Combined test (C) p-values were calculated by comparing S_F to a χ^2_δ distribution, where a maximum likelihood estimate for the number of degrees of freedom, δ, was obtained by fitting δ to the empirical distribution of S_F.

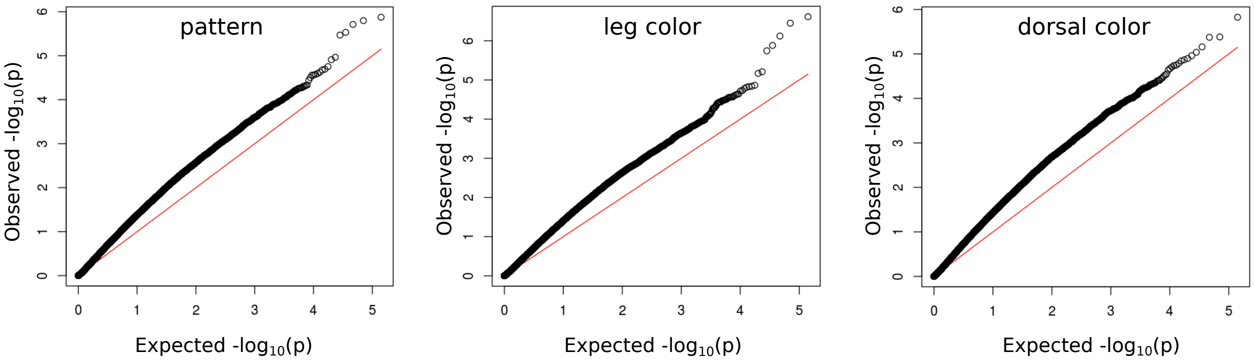


Figure 4.13. Quantile-quantile plots of the observed versus expected p-values corresponding Fisher's combined p-value statistic, S_F, constructed from the p-values for the divergence and admixture mapping analyses. Observed p-values were calculated by comparing S_F to a χ_4^2 distribution.

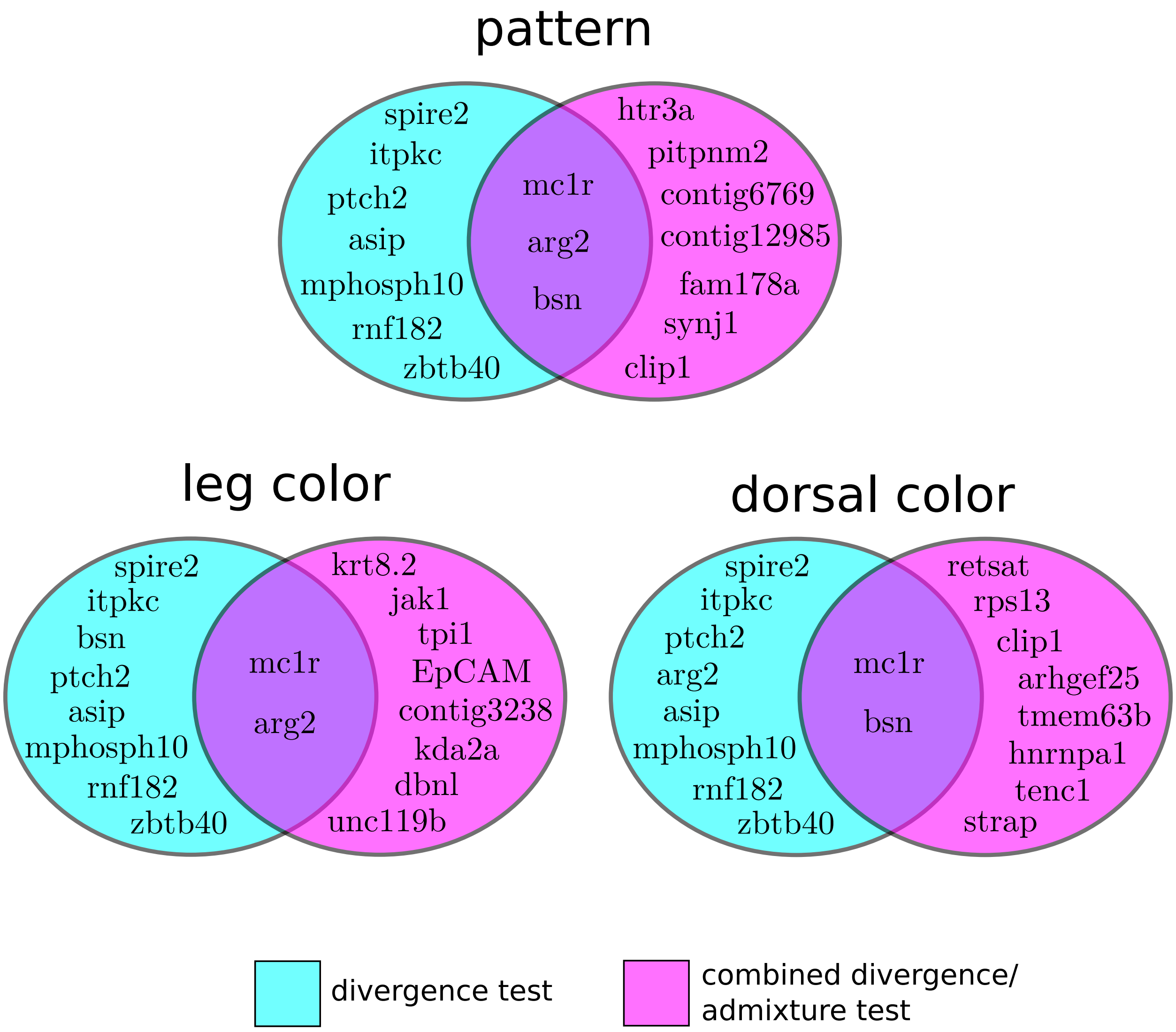


Figure 4.14. Venn diagrams showing overlap between the top 10 most highly associated genes from the respective divergence and combined divergence/admixture tests. Any divergence mapping gene confounded by population structure should not overlap with genes from the combined analysis because a requisite for high S_F is a strong genotype-phenotype association while controlling for genetic ancestry.

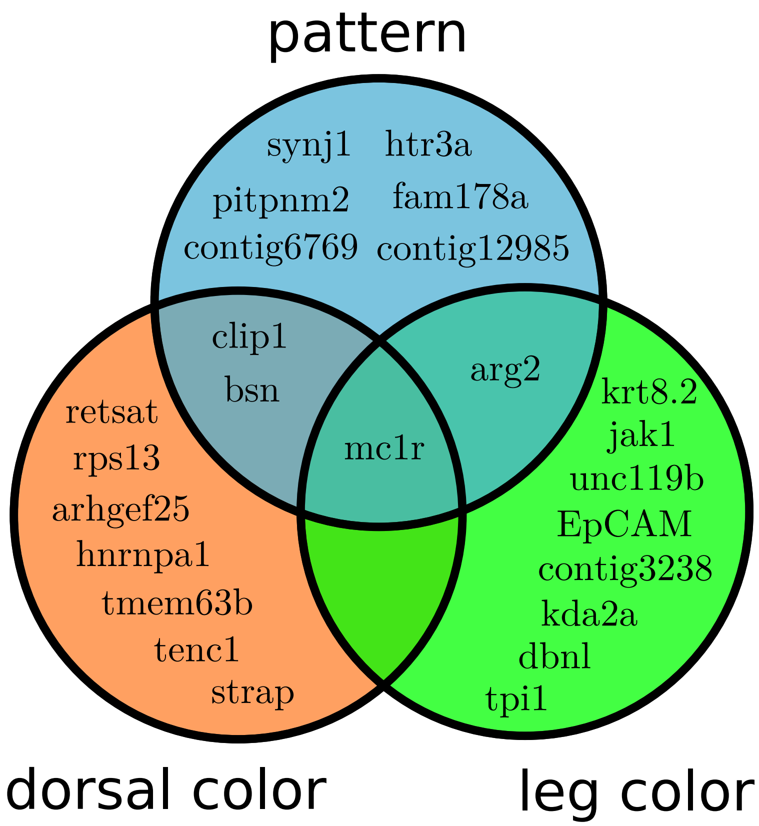


Figure 4.15. Venn diagrams showing which phenotypes the top 10 most highly associated genes identified from the combined divergencce/admixture analysis are associated with.

5. Conclusion

The computational methods and general approaches for analyzing next-generation sequencing (NGS) data discussed in this thesis are intended to extend the utility of NGS for identifying adaptive genes into two insightful yet challenging areas: Natural history museum collections and taxa with genomes that have previously obstructed genomic-scale population genetics. The first step towards achieving this goal was developing the computational tool ngsParalog that identifies paralogy from NGS data. If not addressed, paralogy can obscure population genetic and selection inference by falsely inflating levels of heterozygosity. This can then lead to inaccurate estimates of things like mutation rates which bias demographic inference, including the type that we demonstrated can be done using museum collections. Apart from the quality control aspect, finding genomic regions that have been duplicated is by itself interesting from an evolutionary perspective since paralogs can effect fitness through sub and neo functionalization and gene dosage effects [32, 33]. While other computational approaches for detecting paralogy do exist [6, 7, 5, 4, 8, 9], ngsParalog is the only one among them to jointly leverage information from both sequencing coverage and read proportions within and across individuals, which we showed provides superior true positive versus false positive rates than either of these signals independently. At the same time, the applicability and effectiveness of ngsParalog is not limited by sequencing scheme, since it works for both single or paired-end read sequencing data at low to high coverage. The high power of ngsParalog enabled us to find previously unrecognized regions in the human genome that show evidence of low mappability and that were not included in the 1000 Genomes masks.

Other new computational tools spawned from this thesis work are a C++ program for probabilistic association mapping from pooled or unpooled NGS data, ngsAssociation, and a suite of programs, ABCutils, that implement an Approximate Bayesian Computation (ABC) approach for inferring population demography through fitting the joint site frequency spectrum. I used ABCutils to infer population histories from museum time series exome comparisons of *Tamias* chipmunk species, which revealed that metapopulation structure in *T. alpinus* had increased due to decreased migration among demes over the past century. Additionally, I found evidence that migration may also be slowing among other chipmunk populations with much larger inferred effective sizes, and for which genetic changes are not yet as apparent. This is significant as it demonstrates that high resolution demographic inference using museum genomics may be useful for early detection of demographic responses to environmental change that are of interest for wildlife conservation, thus avoiding relatively ineffective late-stage intervention into biodiversity loss. A second very important aspect of our ability to infer recent population histories from museum specimens is that it allows us to identify which genes have undergone large allele frequency shifts as a consequence of selection, without being confounded by demography, which points to genes possibly involved in rapid adaptation to climate change. Accordingly, we were able to pinpoint one such gene in *T. alpinus*, Alox15, which may helping alpine chipmunk populations persist at high elevations.

The research focused on mapping color genes in the mimic poison frog, *Ranitomeya imitator*, reflects a culmination of the methods and approaches developed during work surrounding the first two chapters. I used the same exon capture approach that colleagues and I developed for *Tamias* chipmunks to reduce the ∼12 gb *R. imitator* genome down to a little over 13K genes, which could be assembled and tested for having associations with color and pattern. Targeting the exome also allowed us to partly circumvent the problems posed to inference by the extensive duplication which led to the large genome size. Then I used ngsParalog to fully identify remaining paralogs. Having accounted for paralogy, I was able to use ngsAssociation in conjunction with admixture mapping to identify seven candidate genes that show strong evidence for influencing color and pattern in a species in which gene mapping seemed to be an extremely intimidating endeavor.

In summary, while NGS now makes genomic-scale sequencing theoretically possible for nearly any organism, including extinct taxa preserved naturally or held in museums, a lack of high-quality preexisting genomic resources can impede effectively using the data for population genetic, selection, and demographic inference. This is compounded for organisms with large,highly duplicated genomes and by pre molecular era sampling schemes. The research presented in this thesis highlights a particular combination of experimental and new computational tools that can be used to overcome all of these challenges and discover adaptive loci.

Bibliography

[1] Marcel Margulies et al. "Genome sequencing in open microfabricated high density picoliter reactors". In: *Nature* 437.7057 (2005), pp. 376–380.

[2] Evan E. Eichler, Royden A. Clark, and Xinwei She. "An assessment of the sequence gaps: Unfinished business in a finished human genome". In: *Nature Reviews Genetics* 5.5 (2004), pp. 345–354.

[3] Mark J. P. Chaisson, Richard K. Wilson, and Evan E. Eichler. "Genetic variation and the de novo assembly of human genomes". In: *Nature Reviews Genetics* 16.11 (2015), pp. 627–640.

[4] Gustave Djedatin et al. "DuplicationDetector, a light weight tool for duplication detection using NGS data". In: *Current Plant Biology* 9-10 (2017), pp. 23–28.

[5] Philippe Gayral et al. "Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap". In: *PLOS Genetics* 9.4 (2013), e1003457.

[6] Seungtai Yoon et al. "Sensitive and accurate detection of copy number variants using read depth of coverage". In: *Genome Research* 19.9 (2009), pp. 1586–1592.

[7] Xihong Wang et al. "CNVcaller: highly efficient and widely applicable software for detecting copy number variations in large populations". In: *GigaScience* 6.12 (2017), pp. 1–12.

[8] Yavas et al. "DB2: a probabilistic approach for accurate detection of tandem duplication breakpoints using paired-end reads". In: *BMC genomics* 15 (2014), p. 175.

[9] Tobias Rausch et al. "DELLY: structural variant discovery by integrated paired-end and split-read analysis". In: *Bioinformatics* 28.18 (2012), pp. i333–i339.

[10] C. Moritz et al. "Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA". In: *Science* 322.5899 (2008), pp. 261–264.

[11] P. Taberlet et al. "Reliable genotyping of samples with very low DNA quantities using PCR". In: *Nucleic Acids Research* 24.16 (1996), pp. 3189–3194.

[12] Kevin C. Rowe et al. "Museum genomics: low-cost and high-accuracy genetic data from historical specimens". In: *Molecular Ecology Resources* 11.6 (2011), pp. 1082–1092.

[13] Russell Higuchi et al. “DNA sequences from the quagga, an extinct member of the horse family”. In: *Nature* 312.5991 (1984), pp. 282–284.

[14] W. Kelley Thomas et al. “Spatial and temporal continuity of kangaroo rat populations shown by sequencing mitochondrial DNA from museum specimens”. In: *Journal of Molecular Evolution* 31.2 (1990), pp. 101–112.

[15] A. Cooper et al. “Independent origins of New Zealand moas and kiwis.” In: *Proceedings of the National Academy of Sciences* 89.18 (1992), pp. 8741–8744.

[16] A. C. Taylor, W. B. Sherwin, and R. K. Wayne. “Genetic variation of microsatellite loci in a bottlenecked species: the northern hairy-nosed wombat Lasiorhinus krefftii”. In: *Molecular Ecology* 3.4 (1994), pp. 277–290.

[17] G. L. Harper, N. Maclean, and D. Goulson. “Analysis of museum specimens suggests extreme genetic drift in the adonis blue butterfly (Polyommatus bellargus)”. In: *Biological Journal of the Linnean Society* 88.3 (2006), pp. 447–452.

[18] Emily M. Rubidge et al. “Climate-induced range contraction drives genetic erosion in an alpine mammal”. In: *Nature Climate Change* 2.4 (2012), pp. 285–288.

[19] Ke Bi et al. “Unlocking the vault: next-generation museum population genomics”. In: *Molecular Ecology* 22.24 (2013), pp. 6018–6032.

[20] Chih-Ming Hung et al. “Drastic population fluctuations explain the rapid extinction of the passenger pigeon”. In: *Proceedings of the National Academy of Sciences* 111.29 (2014), pp. 10636–10641.

[21] John E. McCormack, Whitney L. E. Tsai, and Brant C. Faircloth. “Sequence capture of ultraconserved elements from bird museum specimens”. In: *Molecular Ecology Resources* 16.5 (2016), pp. 1189–1203.

[22] Michael Hofreiter et al. “DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA”. In: *Nucleic Acids Research* 29.23 (2001), pp. 4793–4799.

[23] Kym M. Boycott et al. “Rare-disease genetics in the era of next-generation sequencing: discovery to translation”. In: *Nature Reviews Genetics* 14.10 (2013), pp. 681–691.

[24] Felicity C. Jones et al. “The genomic basis of adaptive evolution in threespine sticklebacks”. In: *Nature* 484.7392 (2012), pp. 55–61.

[25] E. H. Leder, R. G. Danzmann, and M. M. Ferguson. “The candidate gene, Clock, localizes to a strong spawning time quantitative trait locus region in rainbow trout”. In: *Journal of Heredity* 97.1 (2006), pp. 74–80.

[26] Milan Malinsky et al. “Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake”. In: *Science* 350.6267 (2015), pp. 1493–1498.

[27] Wei Zhang et al. “Genome-wide introgression among distantly related Heliconius butterfly species”. In: *Genome Biology* 17 (2016), p. 25.

[28] Dorothea Lindtke et al. “Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect”. In: *Molecular Ecology* 26.22 (2017), pp. 6189–6205.

[29] Joachim Weischenfeldt et al. “Phenotypic impact of genomic structural variation: insights from and for human disease”. In: *Nature Reviews Genetics* 14.2 (2013), pp. 125–138.

[30] Wei-Hua Chen et al. “Human monogenic disease genes have frequently functionally redundant paralogs”. In: *PLOS Computational Biology* 9.5 (2013), e1003073.

[31] Jason S. Williams et al. “Requirement of zebrafish pcdh10a and pcdh10b in melanocyte precursor migration”. In: *Developmental Biology* (2018).

[32] Benjamin Schuster-Bockler, Donald Conrad, and Alex Bateman. “Dosage sensitivity shapes the evolution of copy-number varied regions”. In: *PLOS ONE* 5.3 (2010), e9474.

[33] Reiner A. Veitia, Samuel Bottani, and James A. Birchler. “Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation”. In: *Trends in Genetics* 29.7 (2013), pp. 385–393.

[34] D. V. Lindley. “A statistical paradox”. In: *Biometrika* 44.1/2 (1957), pp. 187–192.

[35] Su Yeon Kim et al. “Design of association studies with pooled or un-pooled next-generation sequencing data”. In: *Genetic Epidemiology* 34.5 (2010), pp. 479–491.

[36] Matteo Fumagalli et al. “Quantifying population genetic differentiation from next-generation sequencing data”. In: *Genetics* 195.3 (2013), pp. 979–992.

[37] Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. “ANGSD: analysis of next generation sequencing data”. In: *BMC Bioinformatics* 15.1 (2014).

[38] Heng Li et al. “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[39] Peter H. Sudmant et al. “An integrated map of structural variation in 2,504 human genomes”. In: *Nature* 526.7571 (2015), pp. 75–81.

[40] AFA Smit, R Hubley, and P Green. “RepeatMasker Open-4.0. 2015”. In: *Google Scholar* (2016).

[41] W. James Kent. “BLAT-The BLAST-like alignment tool”. In: *Genome Research* 12.4 (2002), pp. 656–664.

[42] Aaron A. Comeault et al. “Color phenotypes are under similar genetic control in two distantly related species of Timema stick insect”. In: *Evolution* 70.6 (2016), pp. 1283–1296.

[43] Rudiger Riesch et al. “Transitions between phases of genomic differentiation during stick-insect speciation”. In: *Nature Ecology & Evolution* 1.4 (2017), p. 0082.

[44] Patrik Nosil et al. "Natural selection and the predictability of evolution in (Timema) stick insects". In: *Science* 359.6377 (2018), pp. 765–770.

[45] B. Sinervo et al. "Erosion of lizard diversity by climate change and altered thermal niches". In: *Science* 328.5980 (2010), pp. 894–899.

[46] Jeff E Houlahan et al. "Quantitative evidence for global amphibian population declines". In: *Nature* 404 (2000), p. 4.

[47] Ary A Hoffmann and Carla M Sgro. "Climate change and evolutionary adaptation". In: *Nature* 470.7335 (2011), p. 479.

[48] P. Gienapp et al. "Climate change and evolution: disentangling environmental and genetic responses". In: *Molecular Ecology* 17.1 (2008), pp. 167–178.

[49] Juha Merila and Andrew P. Hendry. "Climate change, adaptation, and phenotypic plasticity: the problem and the evidence". In: *Evolutionary Applications* 7.1 (2014), pp. 1–14.

[50] Michael W. Holmes et al. "Natural history collections as windows on evolutionary processes". In: *Molecular Ecology* 25.4 (2016), pp. 864–881.

[51] Qiaomei Fu et al. "The genetic history of Ice Age Europe". In: *Nature* 534.7606 (2016), pp. 200–205.

[52] Iain Mathieson et al. "Genome-wide patterns of selection in 230 ancient Eurasians". In: *Nature* 528.7583 (2015), pp. 499–503.

[53] K. C. Rowe et al. "Spatially heterogeneous impact of climate change on small mammals of montane California". In: *Proceedings of the Royal Society B: Biological Sciences* 282.1799 (2014), pp. 20141857–20141857.

[54] Rachel E. Walsh et al. "Morphological and dietary responses of chipmunks to a century of climate change". In: *Global Change Biology* 22.9 (2016), pp. 3233–3252.

[55] A. P. A. Assis et al. "Directional selection effects on patterns of phenotypic (co)variation in wild populations". In: *Proc. R. Soc. B* 283.1843 (2016), p. 20161615.

[56] Emily M. Rubidge, James L. Patton, and Craig Moritz. "Diversification of the Alpine Chipmunk, Tamias alpinus, an alpine endemic of the Sierra Nevada, California". In: *BMC Evolutionary Biology* 14.1 (2014), pp. 1–28.

[57] M. Meyer and M. Kircher. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing". In: *Cold Spring Harbor Protocols* 2010.6 (2010), pdb.prot5448–pdb.prot5448.

[58] Ke Bi et al. "Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales". In: *BMC genomics* 13.1 (2012).

[59] Jeffrey M. Good et al. "Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks". In: *Evolution* 69.8 (2015), pp. 1961–1972.

[60] Jeffrey M. Good et al. “Ancient hybridization and mitochondrial capture between two species of chipmunks”. In: *Molecular Ecology* 17.5 (2008), pp. 1313–1327.

[61] Noah Reid, John R. Demboski, and Jack Sullivan. “Phylogeny estimation of the radiation of Western North American chipmunks (Tamias) in the face of introgression using reproductive protein genes”. In: *Systematic Biology* 61.1 (2012), p. 44.

[62] Vladimir A. Trifonov et al. “FISH with and without COT1 DNA”. In: *Fluorescence In Situ Hybridization (FISH)*. Springer, 2017, pp. 123–133.

[63] Sonal Singhal. “De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set”. In: *Molecular Ecology Resources* 13.3 (2013), pp. 403–416.

[64] I. Birol et al. “De novo transcriptome assembly with ABySS”. In: *Bioinformatics* 25.21 (2009), pp. 2872–2877.

[65] Weizhong Li and Adam Godzik. “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. In: *Bioinformatics* 22.13 (2006), pp. 1658–1659.

[66] X. Huang. “CAP3: A DNA sequence assembly program”. In: *Genome Research* 9.9 (1999), pp. 868–877.

[67] E. Axelsson et al. “The effect of ancient DNA damage on inferences of demographic histories”. In: *Molecular Biology and Evolution* 25.10 (2008), pp. 2181–2187.

[68] A. W. Briggs et al. “Patterns of damage in genomic DNA sequences from a Neandertal”. In: *Proceedings of the National Academy of Sciences* 104.37 (2007), pp. 14616–14621.

[69] M. Stiller et al. “Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA”. In: *Proceedings of the National Academy of Sciences* 103.37 (2006), pp. 13578–13584.

[70] Susanna Sawyer et al. “Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA”. In: *PLoS ONE* 7.3 (2012), e34131.

[71] Rasmus Nielsen et al. “SNP Calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data”. In: *PLoS ONE* 7.7 (2012), e37558.

[72] Fumio Tajima. “Statistical method for testing the neutral mutation hypothesis by DNA polymorphism”. In: *Genetics* 123.3 (1989), pp. 585–595.

[73] Matteo Fumagalli et al. “ngsTools: methods for population genetics analyses from next-generation sequencing data”. In: *Bioinformatics* 30.10 (2014), pp. 1486–1487.

[74] Line Skotte, Thorfinn Sand Korneliussen, and Anders Albrechtsen. “Estimating individual admixture proportions from next generation sequencing data”. In: *Genetics* 195.3 (2013), pp. 693–702.

[75] G. Evanno, S. Regnaut, and J. Goudet. “Detecting the number of clusters of individuals using the software structure: a simulation study”. In: *Molecular Ecology* 14.8 (2005), pp. 2611–2620.

[76] David H. Alexander, John Novembre, and Kenneth Lange. “Fast model-based estimation of ancestry in unrelated individuals”. In: *Genome Research* 19.9 (2009), pp. 1655–1664.

[77] Mark A. Beaumont. “Approximate Bayesian computation in evolution and ecology”. In: *Annual Review of Ecology, Evolution, and Systematics* 41.1 (2010), pp. 379–406.

[78] Laurent Excoffier and Matthieu Foll. “fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios”. In: *Bioinformatics* 27.9 (2011), pp. 1332–1334.

[79] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[80] Katalin Csillery, Olivier Francois, and Michael G. B. Blum. “abc: an R package for approximate Bayesian computation (ABC)”. In: *Methods in Ecology and Evolution* 3.3 (2012), pp. 475–479.

[81] Murray S. Weitzman. *Measures of overlap of income distributions of white and Negro families in the United States*. Vol. 22. US Bureau of the Census, 1970.

[82] M. C. Whitlock and K. E. Lotterhos. “Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F(ST).” In: *The American naturalist* 186 Suppl 1 (2015), S24–36.

[83] Eric Frichot et al. “Testing for associations between loci and environmental gradients using latent factor mixed models”. In: *Molecular Biology and Evolution* 30.7 (2013), pp. 1687–1699.

[84] Rasmus Nielsen. “Molecular signatures of natural selection”. In: *Annual Review of Genetics* 39.1 (2005), pp. 197–218.

[85] Katie E. Lotterhos and Michael C. Whitlock. “Evaluation of demographic history and neutral parameterization on the performance of F-ST outlier tests”. In: *Molecular Ecology* 23.9 (2014), pp. 2178–2192.

[86] Sean Hoban et al. “Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions”. In: *The American Naturalist* 188.4 (2016), pp. 379–397.

[87] Yosuke Kayama et al. “Cardiac 12/15 lipoxygenase-induced inflammation is involved in heart failure”. In: *The Journal of Experimental Medicine* 206.7 (2009), pp. 1565–1574.

[88] Annika Lundqvist et al. “The arachidonate 15-lipoxygenase enzyme product 15-HETE is present in heart tissue from patients with ischemic heart disease and enhances clot formation”. In: *PLOS ONE* 11.8 (2016), e0161629.

[89] Lan Yao et al. “Reciprocal regulation of HIF-1α and 15-LO/15-HETE promotes anti-apoptosis process in pulmonary artery smooth muscle cells during hypoxia”. In: *Prostaglandins & Other Lipid Mediators* 99.3 (2012), pp. 96–106.

[90] G. Coop et al. “Using environmental correlations to identify loci underlying local adaptation”. In: *Genetics* 185.4 (2010), pp. 1411–1423.

[91] P. S. Schmidt, M. D. Bertness, and D. M. Rand. “Environmental heterogeneity and balancing selection in the acorn barnacle Semibalanus balanoides”. In: *Proceedings of the Royal Society B: Biological Sciences* 267.1441 (2000), pp. 379–384.

[92] Paul S. Schmidt et al. “Genetic heterogeneity among intertidal habitats in the flat periwinkle, Littorina obtusata”. In: *Molecular Ecology* 16.11 (2007), pp. 2393–2404.

[93] Jay F Storz et al. “The molecular basis of high-altitude adaptation in deer mice”. In: *PLoS genetics* 3.3 (2007), e45.

[94] Andres Aguilar et al. “High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal”. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.10 (2004), pp. 3490–3494.

[95] Elizabeth P Dahlhoff and Nathan E Rank. “The role of stress proteins in responses of a montane willow leaf beetle to environmental temperature variation”. In: *Journal of biosciences* 32.3 (2007), pp. 477–488.

[96] Guofan Zhang et al. “The oyster genome reveals stress adaptation and complexity of shell formation”. In: *Nature* 490.7418 (2012), p. 49.

[97] Benjamin C. Hecht et al. “Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout”. In: *Molecular Ecology* 22.11 (2013), pp. 3061–3076.

[98] Misty R Riddle et al. “Insulin resistance in cavefish as an adaptation to a nutrient-limited environment”. In: *Nature* (2018).

[99] Tim Caro. “The adaptive significance of coloration in mammals”. In: *BioScience* 55.2 (2005), pp. 125–136.

[100] H. E. Hoekstra. “Genetics, development and evolution of adaptive pigmentation in vertebrates”. In: *Heredity* 97.3 (2006), pp. 222–234.

[101] Michael W. Nachman, Hopi E. Hoekstra, and Susan L. D’Agostino. “The genetic basis of adaptive melanism in pocket mice”. In: *Proceedings of the National Academy of Sciences* 100.9 (2003), pp. 5268–5273.

[102] Erica Bree Rosenblum et al. “Adaptive reptile color variation and the evolution of the mc1r gene”. In: *Evolution* 58.8 (2004), pp. 1794–1808.

[103] Edward Bagnall Poulton. *The colours of animals: their meaning and use, especially considered in the case of insects.* D. Appleton, 1890.

[104] Innes C Cuthill et al. "Disruptive coloration and background pattern matching". In: *Nature* 434.7029 (2005), p. 72.

[105] HENRY WALTER BATES. "Contributions to an insect fauna of the Amazon valley (Lepidoptera: Heliconidae)". In: *Biological Journal of the Linnean Society* 16.1 (1981), pp. 41–54.

[106] Chris D Jiggins et al. "Reproductive isolation caused by colour pattern mimicry". In: *Nature* 411.6835 (2001), p. 302.

[107] Joseph S. Wilson et al. "North American velvet ants form one of the world's largest known Mullerian mimicry complexes". In: *Current Biology* 25.16 (2015), R704–R706.

[108] Keith S. Brown and Woodruff W. Benson. "Adaptive polymorphism associated with multiple Mullerian mimicry in Heliconius numata (Lepid. Nymph.)" In: *Biotropica* 6.4 (1974), pp. 205–228.

[109] James Mallet and Mathieu Joron. "Evolution of diversity in warning color and mimicry: polymorphisms, shifting balance, and speciation". In: *Annual Review of Ecology and Systematics* 30.1 (1999), pp. 201–233.

[110] Marcus R. Kronforst et al. "Unraveling the thread of nature's tapestry: the genetics of diversity and convergence in animal pigmentation". In: *Pigment Cell & Melanoma Research* 25.4 (2012), pp. 411–433.

[111] Nicola J. Nadeau et al. "Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 367.1587 (2012), pp. 343–353.

[112] Arnaud Martin et al. "Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.31 (2012), pp. 12632–12637.

[113] Nicola J. Nadeau et al. "Genome-wide patterns of divergence and gene flow across a butterfly radiation". In: *Molecular Ecology* 22.3 (2013), pp. 814–826.

[114] Rebecca Symula, Rainer Schulte, and Kyle Summers. "Molecular phylogenetic evidence for a mimetic radiation in Peruvian poison frogs supports a Mullerian mimicry hypothesis". In: *Proceedings of the Royal Society of London B: Biological Sciences* 268.1484 (2001), pp. 2415–2421.

[115] R Symula, R Schulte, and K Summers. "Molecular systematics and phylogeography of Amazonian poison frogs of the genus Dendrobates". In: *Molecular Phylogenetics and Evolution* 26.3 (2003), pp. 452–475.

[116] Evan Twomey et al. "Phenotypic and genetic divergence among poison frog populations in a mimetic radiation". In: *PLOS ONE* 8.2 (2013), e55443.

[117] Andres Posso-Terranova and Jose Andres. “Diversification and convergence of aposematic phenotypes: truncated receptors and cellular arrangements mediate rapid evolution of coloration in harlequin poison frogs”. In: *Evolution; International Journal of Organic Evolution* 71.11 (2017), pp. 2677–2692.

[118] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[119] Nicolas L Bray et al. “Near-optimal probabilistic RNA-seq quantification”. In: *Nature biotechnology* 34.5 (2016), p. 525.

[120] Brian J Haas et al. “De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis”. In: *Nature protocols* 8.8 (2013), p. 1494.

[121] Marcel Martin. “cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1 (201‘), pp. 10–12.

[122] Aleksandr Morgulis et al. “A fast and symmetric DUST implementation to mask low-complexity DNA sequences”. In: *Journal of Computational Biology* 13.5 (2006), pp. 1028–1040.

[123] Jiajie Zhang et al. “PEAR: a fast and accurate Illumina Paired-End reAd mergeR”. In: *Bioinformatics (Oxford, England)* 30.5 (2014), pp. 614–620.

[124] B. Devlin, K. Roeder, and L. Wasserman. “Genomic control, a new approach to genetic-based association studies”. In: *Theoretical Population Biology* 60.3 (2001), pp. 155–166.

[125] Line Skotte, Thorfinn Sand Korneliussen, and Anders Albrechtsen. “Association testing for next-generation sequencing data using score statistics”. In: *Genetic Epidemiology* 36.5 (2012), pp. 430–437.

[126] Jacob E. Crawford and Rasmus Nielsen. “Detecting adaptive trait loci in nonmodel systems: divergence or admixture mapping?” In: *Molecular Ecology* 22.24 (2013), pp. 6131–6148.

[127] Jocelyn Hudon et al. “Plumage pigment differences underlying the yellow-red differentiation in the Northern Flicker (Colaptes auratus)”. In: *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 183 (2015), pp. 1–10.

[128] Ricardo J. Lopes et al. “Genetic basis for red coloration in birds”. In: *Current Biology* 26.11 (2016), pp. 1427–1434.

[129] Geoffrey E. Hill, Geoffrey Edward Hill, and Kevin J. McGraw. *Bird Coloration: Mechanisms and measurements*. Harvard University Press, 2006.

[130] S. Kinoshita, S. Yoshioka, and J. Miyazaki. “Physics of structural colors”. In: *Reports on Progress in Physics* 71.7 (2008), p. 076401.

[131] Rafael Maia et al. “Iridescent structural colour production in male blue-black grassquit feather barbules: the role of keratin and melanin”. In: *Journal of The Royal Society Interface* 6.Suppl 2 (2009), S203–S211.

[132] Michael W Nachman, Hopi E Hoekstra, and Susan L D’Agostino. “The genetic basis of adaptive melanism in pocket mice”. In: *Proceedings of the National Academy of Sciences* 100.9 (2003), pp. 5268–5273.

[133] Margaret G. Mills and Larissa B. Patterson. “Not just black and white: pigment pattern development and evolution in vertebrates”. In: *Seminars in Cell & Developmental Biology* 20.1 (2009), pp. 72–81.

[134] Nan-Hyung Kim et al. “Arginase-2, a miR-1299 target, enhances pigmentation in melasma by reducing melanosome degradation via senescence-induced autophagy inhibition”. In: *Pigment Cell & Melanoma Research* 30.6 (2017), pp. 521–530.