Spring 2023

The Search for Metabolic Variants in Response to Climate Change in the American Pika

Tyler Stewart Trader
San Jose State University

The Search for Metabolic Variants in Response to Climate Change in the American Pika

A Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Tyler Trader

May 2023

The Designated Project Committee Approves the Project Titled

The Search for Metabolic Variants in Response to Climate Change in the American Pika

by

Tyler Trader

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

May 2023

Dr. Wendy Lee	Department of Computer Science
Dr. Jessica Castillo Vardaro	Department of Biological Sciences
Dr. William Andreopoulos	Department of Computer Science

ABSTRACT

The Search for Metabolic Variants in Response to Climate Change in the American Pika

by Tyler Trader

Climate change and rising temperatures pose a serious threat to the long term survival of American pika (*Ochotana princeps*), emphasizing the interest in the adaptive capability of the pika. This project queried single nucleotide polymorphisms in a population of American pika in Yosemite National Park using Whole Genome Sequencing data, with a specific interest in metabolic variants. The sample data included temporally separated cohorts, comparing modern population data to historical data taken before rapid anthropogenic climate change. Statistically significant variants were identified under Approximate Bayesian Computation using a population decline model. Although population statistics indicated little change between the temporal cohorts, five intergenic SNPs were identified located about 20,000 base pairs upstream from *DECR1*, a gene that plays a key role in the metabolism of polyunsaturated fatty acids. Further work is needed to investigate any link between these SNPs and *DECR1*.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Wendy Lee. This would not have been even remotely possible without your suggestions, support, and understanding.

Thank you to the members of my committee, Dr. Castillo Vardaro and Dr. Andreopoulous, for taking the time to read my report and hear my defense.

Many thanks to my manager, Dr. Tycho Speaker, and coworker Manny Flores, for the unending support and friendship.

Last but not least, thank you to my grandpa for a lifetime of support, encouragement, and teaching me how to believe in myself.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Climate change poses arguably the biggest contemporary threat to ecological systems. Although there have been natural fluxes in temperature throughout global history, it is well known that the Industrial Revolution of the 1800s produced a shift to man-made climate change due to the entrapment of heat-trapping gases in the atmosphere. This has led to a rate of change never before seen throughout history, straining organism's ability to adapt to their new surroundings and leading to the increased likelihood of extinction for over 10,000 plant and animal species [1]. Scientists and governments now are tasked to create legislation to not only mitigate the problem at its source, but also to generate evidence-based conservation and preservation efforts. The value of scientific studies that can provide insight into a species evolutionary fitness level is critical to the success of these legislative protocols. California, one of the most biodiverse areas in the world, has about thirty percent of its species in danger of extinction, and as such protects almost fifty percent of its land [2]. One of these protected areas is Yosemite National Park, a highly biodiverse preserve home to over 400 vertebrate species. The average temperature of the park is projected to rise up to ten degrees in the next century, highly impacting the park's complex ecosystems and highlighting the need for informed conservation decisions [3].

The American pika, or *Ochotana princeps*, is one such species located in Yosemite that risks extinction due to climate change. The small mammal is closely related to the rabbit and is found throughout North America, inhabiting rocky mountain areas in alpine environments mostly above the tree line. Individuals have a very high metabolic rate, estimated to be 7.5 times the basal metabolic rate of a human adult male, and an average body temperature of 104 degrees Fahrenheit [4]. Because metabolic rate increases as external temperature increases, pikas are extremely sensitive

to high temperatures and can die from heat exposure at just under eighty degrees Fahrenheit [5]. This characteristic is thought to put the species at risk of extinction due to climate change as populations are pushed to higher elevations to escape the rising average temperatures [5]. In other words, pika either must adapt to these rising temperatures or shift their range to more hospitable locations. Unfortunately, pikas exhibit low dispersal rates within a population and are therefore unlikely to shift their range, meaning that as a population they must adapt to avoid extinction [6]. Although not listed on the Endangered Species Act, the overall pika species population is decreasing as temperatures rise and their long-term survival is of great concern. Further, pikas are thought to be a model species for understanding climate change in relation to population genetics due to well documented range contraction that can be linked to climate change [7].

This research therefore aims to understand the mechanisms for adaptive change that pikas have exhibited over the last century by the identification of genetic variants present in contemporary populations as compared to historic populations. Variations that affect the animal's metabolism are of specific interest due to the connectivity between metabolic genes and a species ability to tolerate heat. Further, variations that can predict the likelihood of increased fitness within the population in the future are also of interest.

CHAPTER 2

Background

This project is influenced by a study similar in both structure and goal, conducted by researchers Bi et al in and around the Yosemite Valley. That study, published in 2019, employs both a temporal and a spatial approach to identify variants in proteins affecting metabolism in a species of chipmunk experiencing severe range contraction as temperatures rise, known as *Neotamias alpinus*, versus a chipmunk species that is stable in environmental range and size of population, known as *Neotamias speciosus*. Bi et al examined exome sequencing data and found that Alox15, a lipoxygenase, tripled in expression between the two timepoints, which is theorized to be a physiologic response to environmental warming. Their work highlighted both the ability to generate high quality data from archived specimens as well as the importance of temporal data in conservation work, particularly useful in informing potential evolutionary trajectories for species and their individual populations through historical demographic modeling [8].

Whole genome sequencing, made easily accessible by next-generation sequencing (NGS), is a technique that allows for the analysis of whole genomes, rather than just individual areas of interest [9]. This supports more thorough identification of genetic variants, generating a better understanding of the complicated multidimensional facets of living organisms as one can delve deeper into the interconnectedness of a species inner biological workings. Specifically, we are interested in the identification of single nucleotide polymorphisms (SNPs) in the germline. Whereas somatic mutations accumulate throughout an organism's lifetime in single cells or groups of cells, germline mutations are variants that are heritable and thought to be present in all cells in an organism's body from birth [10]. These are therefore variants that are likely to be present in the majority of a group of closely related individuals and can help to

identify genomic changes of a population throughout time.

One crucial analysis in population genetics is demographic modeling. Because achieving a comprehensive temporal sample set is almost impossible, population genetics relies heavily on statistical methods to predict change in a species over time, specifically through use of the Bayesian likelihood function. Real world population data is large and complex with an abundance of nuisance parameters; therefore, the probability of observing a specific variant given the true genotype is indeterminable, highlighting the need for statistical inference. Approximate Bayesian Computation (ABC) is a popular choice among bioinformaticians to bypass the need for the likelihood function by using sample parameters to produce artificial datasets that can then be analyzed to infer the evolutionary changes of a population over time [11]. For this report, the called variants were used to generate a site frequency spectrum (SFS), which describes the distribution of allele frequencies in a given genomic sample set [12]. This was then used to simulate data under several different demographic models, which was then analyzed via ABC to determine the most likely evolutionary scenario for the population. The called variants were then compared to the simulated statistical models under the most likely demographic scenario to determine the most statistically significant variants present in the sample data.

CHAPTER 3

Methods

Whole genome sequencing data was collected from American pika in and around Yosemite National Park. Ten samples were collected in 1915 from two locations, Lyell Canyon in Tuolumne County (five samples) and Vogelsang Lake in Mariposa County (five samples). These were stored as skeletal remains. Ten samples were collected in the early 2000s (2003 to 2006) from three locations, Lyell Canyon (five samples), Townsley Lake in Mariposa County (four samples), and Gardisky Trail in Mono County (one sample). These will be referred to as the historic and modern samples, respectively (see Figure 1 and Appendix B.1 Table). All twenty samples were then sequenced without replicates using the Illumina NGS platform and produced raw reads 151 base pairs long.

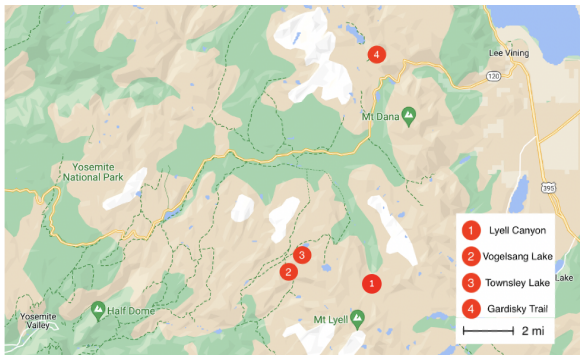


Figure 1: Sample Locations in Yosemite National Park

There were four locations where samples were collected: Lyell Canyon (5 historic, 5 modern), Vogelsang Lake (5 historic, no modern), Townsley Lake (0 historic, 4 modern), and Gardisky Trail (0 historic, 1 modern). Vogelsang Lake and Townsley Lake are the closest in proximity, located about 1 mile apart. Lyell Canyon is about five miles from all the other locations, and Gardisky Trail is about 12 miles from Vogelsang Lake.

3.1 Variant Calling

Each data processing tool was run iteratively using the Python-based workflow management tool SnakeMake. All work was performed on the College of Science High

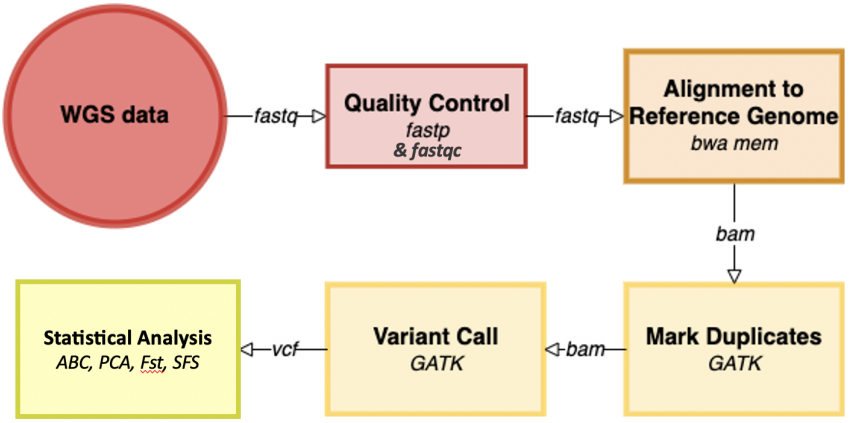


Figure 2: Overall Workflow

Performance Computer (HPC), which uses the CentOS 7 package and has Slurm as the resource manager and job scheduler. The raw reads from the Illumina sequencer, in compressed fastq format, were quality checked using the Java based analytical tool FastQC and then processed using the C++ based tool fastp to trim for the universal Illumina adapter and correct for mean quality scores using the default settings [13, 14]. Results were then aggregated using the report generating tool multiQC [15]. The resulting processed compressed fastq files were then aligned to a reference genome using the software package Burrows-Wheeler Aligner, particularly the BWA-MEM algorithm [16]. The reference genome, OchPri4.0, was taken from the liver of a single male American pika organism from the Beaverhead-Deerlodge National Forest in Montana. It was sequenced using Illumina HiSeq and assembled using Dovetail HiRise version 2017 for a total genome size of 2.23 billion base pairs made up of 33 chromosomes, with a genome coverage level of 23.68x. The reference genome was submitted to the National Center for Biotechnology Information (NCBI) by the University of British Columbia Okanagan in 2020 (GenBank assembly accession GCA014633375.1). There are 29,701 protein coding genes and 9,351 scaffolds, with a

GC concentration of 44.21 percent [17].

After mapping, duplicate reads were marked using MarkDuplicates from the Genome Analysis Toolkit (GATK) and resulting bam files were index and coordinate sorted in preparation for variant calling. Variant calling pipeline followed the GATK best practices pipeline. Variants were called separately for each cohort using the GATK-based programs HaplotypeCaller and GenotypeGVCFs. HaplotypeCaller was ran in emit reference confidence (ERC) mode using the compressed variant call format (GVCF) option. Each sample was run iteratively to generate separate GVCF files. These GVCFs were passed by cohort into GenomicsDBImport to generate a datastore, which contained variant information for each chromosome for each sample in that cohort. This datastore was then used as input for GenotypeGVCFs, which joint-called the variants in the cohort. Joint calling generates a variant call format (VCF) file with every SNP site where any individual in the cohort has evidence of variation. The two resulting VCF files (one per cohort) were then filtered using the GATK tool SelectVariants to select for the most statistically significant variants, with thresholds outlined in Table 1 and explained in Appendix A.3 [10]. Filters were chosen via graphical interpretation and applied using Java Expression Language (JEXL) (see Appendix C.12 Figure). This produced a final vcf file for each cohort which was then used in downstream analyses. For more in-depth explanations of each tool, see Appendix A.1 Text.

3.2 Population Genetics Statistical Analyses

Cohorts were first analyzed for global and per-site nucleotide diversity and Tajima's D values (using a 10,000 base pair sliding window) using the Popgen Pipeline Platform (PPP), a set of python-based scripts written to aid in the incorporation of commonly used population genetic software packages into pipelines [18]. Intra-

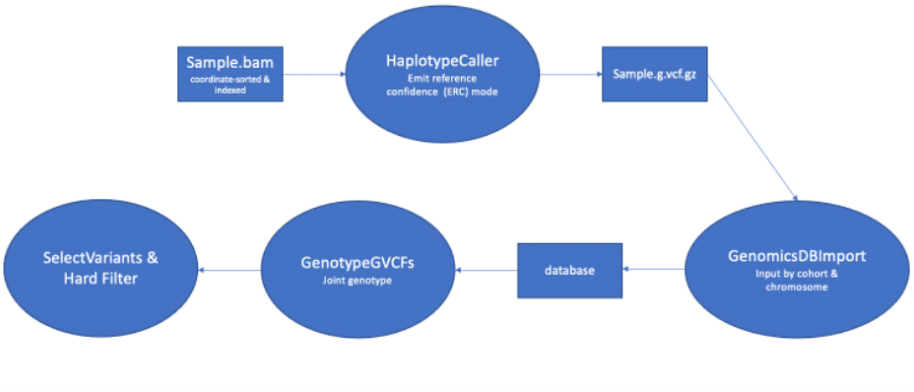


Figure 3: Variant Call Workflow

population genetic structure was analyzed via PCA using SNPRelate, a module within the R-based Bioconductor package, and inter-population global, per site, and per chromosome Fst values were calculated for the two cohorts using the PPP [19]. The chromosome with the highest weighted Fst value and the two 10,000 bp sections on that chromosome with the highest Tajima's D values were chosen for further analysis. Joint-SFS matrices (2D-SFS) were calculated using the PPP and used to model observed data in ABC analysis [20].

To determine the best demographic model for use in ABC analysis, five runs of 100,000 simulations of SNP data were generated using fastsimcoal2, a coalescent demographic modeling software, to model neutral molecular diversity under three different demographic historical scenarios (see Appendix A.4 Text) [21]. The algorithm then performed 40 maximum likelihood iterations per run to estimate parameters for each scenario, and the demographic model with the highest estimated log likelihood was chosen. Next, 5,000 2D-SFS matrices were simulated under the most likely demographic scenario. In an ABC rejection framework, rejection sampling based on Euclidean distance was performed on the simulated matrices using custom scripts with an acceptance rate of 0.1. As described in Bi et al, custom bins in both the diagonal and off-diagonal axes of the 2D-SFS were then computed and used as the summary

statistic for a given matrix. These statistics were used to check the goodness of fit of the chosen model onto the observed data. The five simulations with the best fit to the observed data were chosen, with the acceptance rate of 0.01 similar to that used in Bi et al (0.008). The simulations then were analyzed via analysis of molecular variance (AMOVA) using Arlsumstat to generate the neutral Fst distribution in order to build an approximation of the posterior distribution [22]. This distribution was then plotted and compared to the observed Fst values to identify outliers.

CHAPTER 4

Results

4.1 Variant Calling

Initial quality control of the raw reads showed high per sequence quality scores. The per sequence quality report allows for the identification of sequences or subsets of sequences with low-quality values. Quality scores by position indicated 99.9 percent accurate calls at all read positions after base correction. The statistics warrant the assumption of high-quality imaging and read calls such that no further processing was necessary and reads could be used in variant calling pipeline (see Appendix C.10 Figure) [13]. Reads were then mapped to the reference genome in pairs using BWA-MEM due to its usefulness in mapping high-quality sequences above 70 base pairs in length. Most reads were properly paired to and mapped with their mate (see Appendix C.11 Figure). On average, samples had 91 percent of reads properly paired with a median of 96 percent. Further, 70 percent of samples had less than 2 percent of reads with mates mapped to a different chromosome. Two samples, both in the modern cohort, had about 10 percent of reads with mates mapped to a different chromosome. Supplementary alignments, occurring when different sections of a read maps to completely distinct and separate sections of the genome, made up less than 4 percent of the total sample for almost 80 percent of samples. Finally, singletons, or reads whose mate was unmapped, make up very little of the population, representing less than 1 percent of total reads for any given sample. Using the Lander-Waterman equation, total sequencing coverage (ie read depth) was 20x per individual on average ($s = 4$), meaning that each base in the genome was sequenced about 20 times each [23].

Next, duplicate reads were marked using query-sorted mapped data which allowed the MarkDuplicates tool to include both unmapped mates and supplementary reads in analysis, which helped to account for those two samples in the modern cohort with a

large portion of reads with mates mapped to a different chromosome. MarkDuplicates found an average of 950,355 optical duplicates per sample with a range of 45,796 to 6,556,847 optical duplicates. Overall, there was on average 38 percent duplication with a standard deviation of 7 percent and ranging from 15 to 50 percent duplication. The mapped bam files were then coordinate-sorted and indexed, and SNPs were called using HaplotypeCaller. HaplotypeCaller produced 63 GB of data for the modern cohort (with file sizes ranging from 2 GB to 9 GB) and 53 GB of data for the historic cohort (with file sizes ranging from 5 GB to 6 GB). These files were used to create a 1 GB datastore for both the modern and historic cohorts. Each datastore was input into GenotypeGVCFs to joint-call variants, which produced 4.2 GB of data and 7.1 GB of data, respectively. The modern cohort contained 35,066,905 SNPs while the historic cohort contained 64,628,272 SNPs.

Variants were then filtered based on their various statistical measurements including strand bias and quality score using SelectVariants and JEXL (see Appendix A.3 Text). Variants were first filtered using the GATK recommended filters, which are by design very lenient [24]. Remaining variants were then mapped using the Python data visualization library seaborn to generate density curves and final filters were chosen based on the proportion of variants above that statistical threshold (see Table 1 and Appendix C.12 Figure). After filtering, the modern cohort contained 16,061,661 SNPs and the historic cohort contained 36,695,297 variants, a 54.2 percent decrease and a 43.2 percent decrease, respectively.

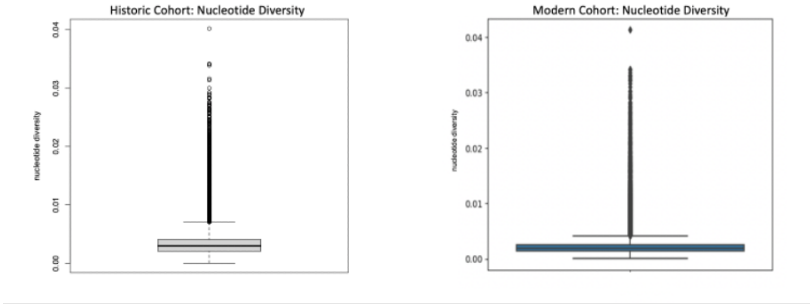
Threshold(Pass)						
	QD	FS	SOR	MQ	MQRankSum	ReadPosRankSum
Custom Filter	>2	<10	<3	>50	>-5	>-3

Table 1: Final Statistical Thresholds for Variant Calling

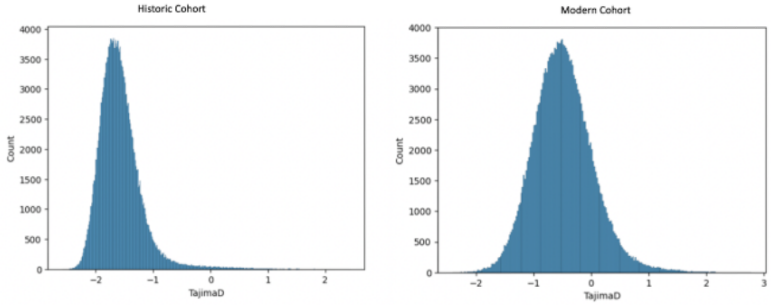
from [10]: 'QD' [QualByDepth]: Normalized variant quality score; 'FS' [FisherStrand]: Phred-scaled strand bias probability; 'SOR' [StrandOddsRatio]: Strand bias probability score by ratio of reads on both alleles; 'MQ' [RMSMappingQuality]: Square root of average of squares of mapping quality at given site; 'MQRankSum' [MappingQualityRankSumTest]: u-based z-approximation from Rank Sum Test for mapping qualities; 'ReadPosRankSum' [ReadPosRankSumTest]: u-based z-approximation from Rank Sum Test per base

4.2 Population Genetics Statistical Analyses

Filtered cohorts were analyzed for global and per site nucleotide diversity and Tajima's D values. Both are most accurately calculated in a sliding window to average over several variants at a time; therefore, a sliding window of 10kb was chosen [25]. Nucleotide diversity, symbolized as π, measures genetic diversity. Global nucleotide diversity for the modern cohort was 0.002 (standard deviation 0.0013) and for the historic cohort was 0.003 (standard deviation 0.0017), showing a thirty-three percent decrease in population diversity over time. Next, Tajima's D is a summary statistic of the SFS and is used to aid in the detection of areas of the genome that are under selective pressures and are therefore not evolving neutrally [26]. Global Tajima's D for the modern cohort was -0.4844 (standard deviation 0.512) and for the historic cohort was -1.5732 (standard deviation 0.377). In the modern cohort, there were 133 sections of genome 10kb long that had evidence of a non-neutral Tajima's D. This represented 9228 SNPs across 29 chromosomes, or 0.057 percent of the filtered variants. Of these 9228 SNPs, 3335 were indicative of balancing selection (TD>2). In the historic cohort, there were 9736 sections of genome 10kb long that had evidence of a non-neutral Tajima's D, representing 1,865,733 SNPs or 5.08 percent of filtered variants. All of these markers in the historic cohort showed evidence of positive selection (TD<-2).



(a) Nucleotide Diversity by Cohort



(b) Tajima's D by Cohort

Figure 4: Windowed Nucleotide Diversity and Tajima's D

(a) Boxplots of the nucleotide diversity for the historic and modern cohort. Both populations have low nucleotide diversity with similar tail lengths. (b) Histograms for Tajima's D in the historic and modern cohorts. Historic cohort shifts more toward positive selection, whereas the modern cohort is centered closer to zero.

Transition and transversion frequency was analyzed to explore the possibility of DNA degradation as seen in the Bi et al study (see Figure 5). The historic cohort shows a 61 percent relative increase in both G to A as well as C to T conversions compared to the modern cohort, which may suggest hydrolytic deamination, a sign of degradation in ancient DNA samples [27]. SNPs that indicated a G to A conversion in the historic cohort were then removed and principal component analysis was performed to analyze intra-population genetic structure (see Figure 6). The modern cohort clustered tightly while the historic cohort was slightly less genetically similar. However, there was seemingly no distinguishable differences between cohorts.

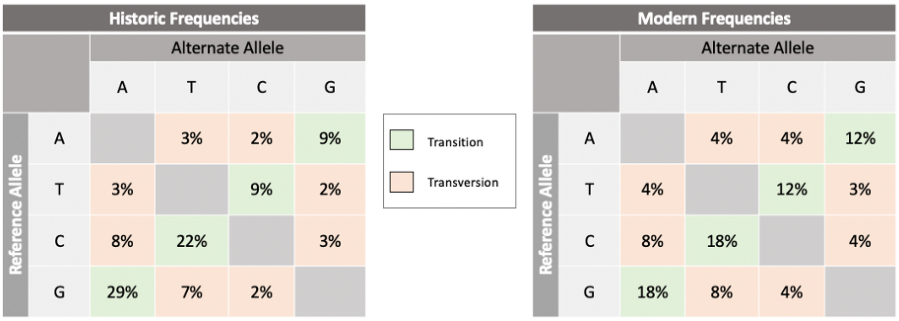


Figure 5: Allele Frequency per Cohort

G to A SNPs are seen 1.6 times more frequently and C to T SNPs are seen 1.2 times more frequently in the historic cohort when compared to the conversion rate in the modern cohort.

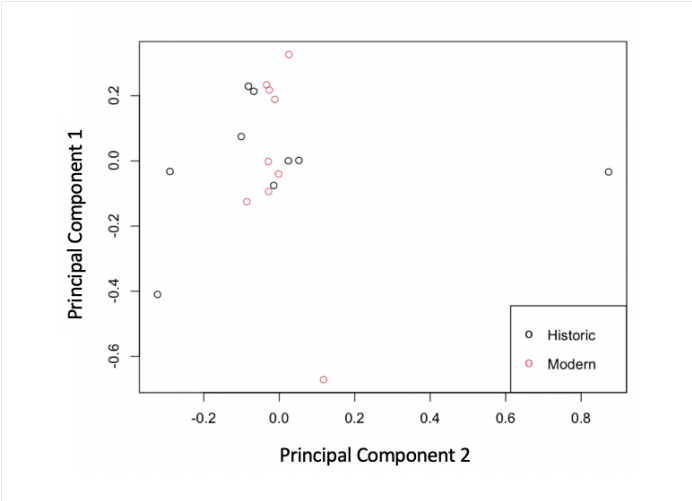
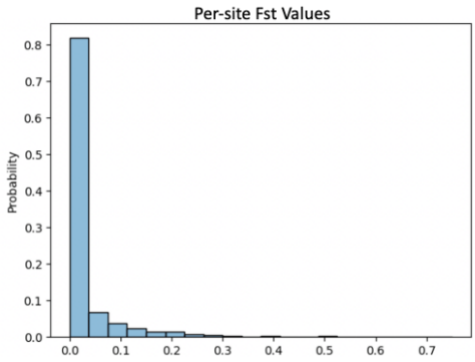


Figure 6: Principal Component Analysis

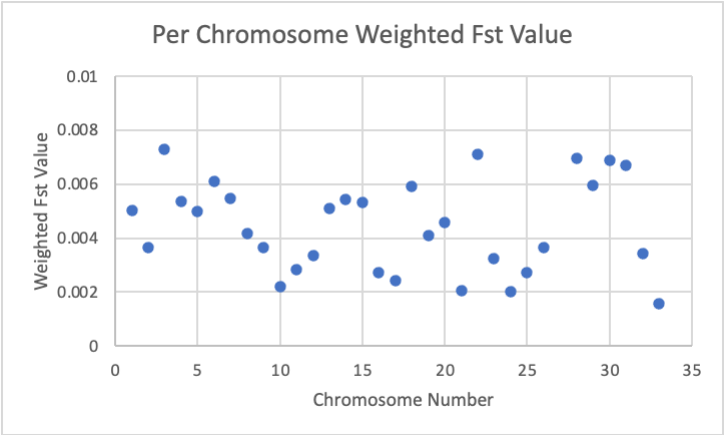
Each data point represents an individual. The first two principal components comprise 17.08 percent of total genetic variance (9.05 percent and 8.03 percent, respectively).

Global, per-site, and per-chromosome Fst values were then calculated. The two cohorts displayed a global weighted Fst of 0.0138 (unweighted 0.0092), per-site Fst values were all less than 0.1, and per-chromosome weighted Fst values ranged from 0 to 0.0073 with a standard deviation of 0.00184 (see Figure 7). Chromosome 3 showed the highest Fst value and was therefore chosen for further analysis. This chromosome

displayed 1,671,501 statistically filtered SNPs between the two cohorts and is 83 Mb in length. This chromosome was then scanned for sections of DNA (10 kb sliding window) above a Tajima's D equal to 2 threshold value (as described above), which resulted in two sections of genome (starting at base 3,230,000 and base 35,080,000). These sections were identified as the DNA of interest in subsequent analysis. Finally, joint 2D-SFS matrices were generated for chromosome 3 and for the sections of interest for use in posterior probability generation via ABC analysis (see Appendix C.13 Figure).



(a) Per-site Weighted Fst Values



(b) Per-Chromosome Weighted Fst Values

Figure 7: Weighted Fst Values

(a) Most per-site weighted Fst values show a Fst of 0, with 90.1 percent of values below 0.1. (b)Highest chromosome Fst value was on chromosome 3 (Fst = 0.00729). Standard deviation was 0.00184 and the average was 0.00429.

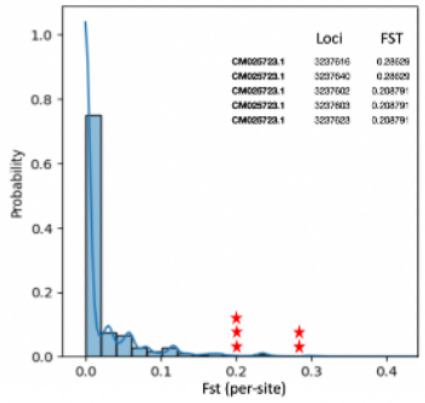
As described previously, ABC analysis is centered toward estimating the posterior probability function without use of the likelihood function. Rather, the ABC framework consists of generating large amounts of data and culling by some rejection method based on some summary statistic or statistics in order to use a subset of the simulations closest to the observed data to estimate the posterior distribution. In other words, accepted simulations under some tolerance can be thought of as a sample of the approximate posterior distribution [28]. This project uses an ABC framework to generate a neutral Fst distribution, or the expected distribution if no selecting factors were occurring within the genome. Observed data can then be analyzed against this distribution to identify outliers [29]. The ABC workflow employed in this project was influenced heavily by the workflow outlined in Bi et al.

Under ABC analysis, three demographic models were chosen to investigate: a constant population size, a declining population, and an expanding population. For parameter estimation and model likelihood inquiry, five runs of 100,000 simulated 2D-SFS matrices were generated per model using fastsimcoal2, with 40 iterations per run using likelihood maximization to estimate parameters. Parameters are drawn from set priors and simulates data to build a likelihood function, which it then uses to estimate parameters using the conditional maximization (ECM) algorithm [21] (see Appendix A.4 Table). The fifteen runs were then compared, and the population decline model had the highest log likelihood across all five runs (see Appendix C.14 Figure).This model suggests a current population size of 2,820 individuals on average (range: 1045 – 4087) and a growth rate of 0.0816 (the growth rate is positive because the algorithm moves backward in time, such that the population is growing as the algorithm moves into the past and therefore shrinking forward in time).

Random DNA data (10 kb per chromosomal area of interest) was then simulated under the population decline model and output as SNP data using the -s option of

fastsimcoal. Five simulations were run for 10,000 randomly drawn sets of parameters from the defined priors, generating 1,000,000 total bases. Rejection in an ABC rejection framework was then performed, with simulated 2D-SFS matrices accepted based on Euclidean distance to the observed matrix. An acceptance threshold of 0.1 was used for this step. As was performed in the Bi et al study, diagonal and off-diagonal bins were then calculated as the summary statistics on the accepted simulations with a bin width of 2 (for 12 total bins per 2D-SFS) (see Appendix C.15 Figure). Goodness-of-fit calculations were then performed on each bin, and the five simulations with the highest overall goodness-of-fit scores were chosen to build the neutral Fst distribution. The genetic site information for these simulations was ran through AMOVA in the Arlequin population genetics software package to generate the expected Fst distribution. AMOVA is modeled after ANOVA analysis (analysis of variance) and allows for the use of ANOVA statistical analysis on population wide molecular data [22]. The observed Fst outliers were finally identified based on their probability under the neutral distribution (see Figure 8).

a



b

	Expected Fst	Observed Fst
Mean [95% C.I.]	0.0210 [0.0206 – 0.0214]	0.0266 [0.0197 – 0.0334]
Standard Deviation	0.047	0.054
Range	0 – 0.6088	0 – 0.2863
97.5% Quantile	0.165	0.194

Figure 8: Neutral Fst Distribution

(a) The neutral distribution displayed as a probability histogram with a kernel density estimate overlay (alpha of 0.95). The p-value for the neutral Fst estimates was an average of .645 with a confidence interval of .643 - .648, meaning the null hypothesis is accepted and these Fst values occur by random chance (thus, a neutral distribution). Observed loci with significant Fst values (values where the probability distribution is equal to 0) are labeled as red stars on the graph (height of star stack is arbitrary and therefore not correlated to the y-axis). (b) Various Fst statistics on both the expected Fst (the neutral Fst) and the observed Fst as computed in JMP (a statistical analysis software). It is important to note that differences in mean and standard deviation may likely be an artifact of dataset size (the observed Fst dataset was much smaller than the expected Fst dataset).

Five observed loci on chromosome 3 were found to have significant Fst values: positions 3237602, 3237603, 3237616, 3737623, and 3237640. These are sites where the probability of observing these values under the neutral distribution was very low (see Figure 8). These sites were then located on the reference genome using the NCBI Genome Data Viewer application. All SNPs were found to be intergenic and about 20,000 base pairs upstream from the closest gene, known as DECR1 (see Figure 9).

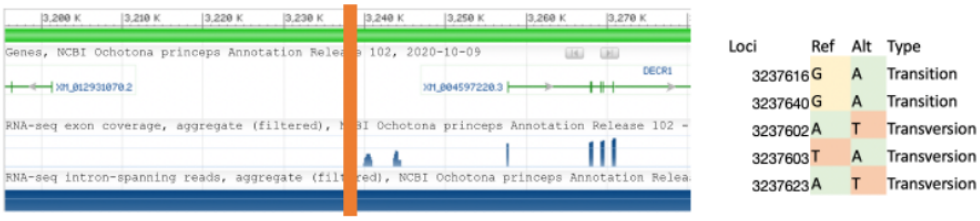


Figure 9: Identified Outliers on the Reference Genome

The location of the loci is denoted on the reference genome by an orange rectangle (loci all within 38 bases). It is important to note that the orange rectangle is a general reference point and therefore is not to scale. The gene upstream of the loci, XM012931070.2, is thought to be a zinc finger protein, but no specific function was indicated by NCBI.

CHAPTER 5

Discussion

5.1 Variant Calling

Initial quality metrics, including high per sequence quality scores and quality scores by position, warrant the assumption of high-quality imaging and read calls such that no further preprocessing was necessary and reads could be used in variant calling pipeline (see Appendix C.10 Figure) [13]. Further, mapping quality was high and read depth averaged 20x per sample. The recommended coverage for variant calling in a non-clinical setting is about 20x; thus, no individuals needed to be filtered out on the basis of extremely high or low coverage levels (see Appendix C.11 Figure) [30]. Overall, mapping and coverage was sufficient and mapped reads were able to be used in further processing.

After duplicate reads were marked using MarkDuplicates, SNPs were called using HaplotypeCaller. HaplotypeCaller generates haplotypes, which are possible genotypes for a given section of DNA, from the existing mapped read information and then employs de-novo reassembly of the haplotypes in each region that contains evidence of variation. This means that the program generates possible sequences for a given section of DNA from the mapped reads, and then ignores the existing mapped read information to map the generated sequences to the reference genome. This reassembly approach allows HaplotypeCaller to be more accurate than other callers. HaplotypeCaller is very computationally expensive, however, and is a bottleneck in most variant calling processes [10]. Due to the lack of large datasets of known variants within the pika population, which exist for both human and model organism data, HaplotypeCaller was chosen for this step for its accuracy despite the process bottleneck.

Following GATK best practices after variant call, variants should be filtered by

Variant Quality Score Recalibration (ApplyVQSR), which is a machine-learning based approach to retain the most statistically significant variants. Unfortunately, this requires a large known variant set ("truth set") for the population, which is unavailable for most non-model organisms. Instead, hard filters consisting of flat thresholds were used to cull the datasets. One caveat of this approach is that each statistical dimension is analyzed individually, such that the analysis of variant clustering among multidimensional statistical variables is impossible. Because of this, false positives will exist within the dataset to keep true positives along certain dimensions. Likewise, a certain percentage of true positives will be culled out of the population because of one statistically poor variable [10]. Further downstream statistical analysis will ideally cull these false positives out from analysis, although any true positives culled at this step would be lost.

5.2 Population Genetics Statistical Analyses

Nucleotide diversity in the modern cohort ($\pi = 0.002$) is similar to the nucleotide diversity values estimated in other studies on the Yosemite Valley American pika population. For example, Klinger et al placed the genetic diversity of this population around 0.0019 [31]. Overall, nucleotide diversity in both cohorts was low, with a slight decline in diversity over time. This may indicate a poor ability to adapt as well as a low population fitness level. The summary statistic Tajima's D was used to determine any areas within the genome under selective pressures. A high positive Tajima's D value indicates balancing selection, which in turn indicates genetic variation and possible adaptations in the population. In contrast, a low negative value indicates positive selection, which reduces genetic variation [32]. A threshold of +/- 2 was chosen as the cutoff for the Tajima's D statistic due to a consensus that areas above or below these values indicate genomic regions not evolving neutrally [33]. Global

Tajima's D relayed a 70 percent positive relative change between the two populations, with the modern cohort shifting toward zero. This could indicate a shift toward genomic equilibrium in the modern population, which would further corroborate the idea of declining nucleotide diversity (see Figure 4). However, Tajima's D is affected by population size and tends to increase as population size decreases, such that this change could also be due to a declining population [34].

Initial principal component analysis indicated high similarity between individuals in the modern cohort (see Appendix C.16 Figure). The historic cluster, however, showed considerably less similarity to itself, which could be indicative of DNA degradation in the historic samples as was seen in the Bi et al study [8]. This prompted the investigation into allele frequency per cohort, which further indicated DNA degradation. Historic G to A SNPs were then removed from analysis, which produced the final PCA plot seen in Figure 6 above. This plot indicates high intra-population genetic similarity for the modern cohort. The historic cohort clustered closer without the G to A SNPs, but still had a few individuals that exhibited some variation. Overall, there was very little distinction between the modern and historic cohort on the plot, indicating genetic similarity between the two cohorts and a lack of significant genetic change over the last century.

As described previously, Fst, or global fixation index, is a summary statistic of the site frequency spectra (SFS) and helps to describe inter-population separation via allele frequency variance [25, 35]. An Fst value of 0 indicates genetically identical populations, whereas an Fst of 1 indicates genetically distinct populations. Global weighted Fst between the two cohorts was low (Fst = 0.0318), indicating low genetic change over the last century. Likewise, per-site weighted Fst average values were low, with most Fst values less than 0.1. Chromosome 3 showed the highest Fst value and is therefore thought to be the most genetically distinct chromosome between the two

cohorts. This chromosome was then scanned for areas showing evidence of selective pressure, defined as having a Tajima's D value above 2 or below -2, resulting in two areas of interest for further analysis.

ABC analysis is centered toward estimating the posterior probability function without use of the likelihood function. Rather, the ABC framework consists of generating large amounts of data and culling by some rejection method based on some summary statistic or statistics in order to use a subset of the simulations closest to the observed data to estimate the posterior distribution. In other words, accepted simulations under some tolerance can be thought of as a sample of the approximate posterior distribution [28]. This project uses an ABC framework to generate a neutral Fst distribution, or the expected distribution if no selecting factors were occurring within the genome. Observed data can then be analyzed against this distribution to identify outliers [29]. The ABC workflow employed in this project was influenced heavily by the workflow outlined in Bi et al. To perform ABC analysis, a demographic historical model must be chosen under which to generate large amount of simulated data. A population decline model was chosen due to the log likelihood of the model across five runs. This model choice is unsurprising, as the increase in Tajima's D between the historic and modern population hinted at a declining population. A neutral Fst distribution was then determined under this model via ABC, and outliers were identified as observations where the neutral probability distribution was zero.

All five outlier loci identified are located about 20,000 base pairs upstream from *DECR1*. According to NCBI, the protein encoded from this gene, also known as *DECR1*, is an ortholog of the human DECR1 protein and plays a role in fatty acid beta-oxidation. In fact, DECR1 is the rate-limiting enzyme in the metabolism of polyunsaturated fatty acids. Fatty acid beta-oxidation plays a key role in energy production for the heart and other organs in humans. Further, there is evidence to

suggest that alterations in fatty acid beta-oxidation in cancer cells enables tumors the ability to survive through periods of increased metabolic stress [36].

CHAPTER 6

Conclusion and Future Directions

As with any bioinformatics study, further investigation into any finding is required, particularly with the use of wet lab experiments. It is most likely of interest to study any potential link between the identified intergenic region and the *DECR1* gene. Further, this study was limited in scope to chromosome 3 because of the indication of increased selective factors, but a more expansive study to cover a broader range of the genome in the downstream population genetic analyses is warranted. The discovery of potential historical sample degradation, as evidenced in 5, also poses an interesting hurdle. Although G to A SNPs were removed from the historic cohort to account for hydrolytic deamination in PCA analysis, ABC analysis was ran with an uncorrected data set. This change may affect the neutral Fst distribution as the five closest simulated data sets would vary with a change in observed 2D-SFS. Another future inquiry includes analyzing the modern genome directly to the historic genome in downstream analysis rather than to the reference genome. Because the reference genome was built from a single individual, there is associated uncertainty about the accuracy of the genome. Further, since the reference genome was built from a contemporary individual (that is, an individual from a population that has also experienced rapid anthropogenic climate change), the modern cohort may be genetically more similar to the reference genome than the historic cohort is to the reference genome, which would also pose a source of error. Finally, a broader study on demographic model choice for this population is recommended. As previously mentioned, this study explored three basic models: historical population decline, historical population expansion, and a constant population size. There is interest in the effect of an increase in model complexity on both the maximum estimated likelihood as well as the fit of the model to the observed data. Although further inquiry

into this data is necessary, this project allowed for the development of a Snakemake variant calling pipeline and a downstream statistical workflow, both geared toward the analysis of SNP data. More importantly, it delved into some of the possibilities of temporal data in population genetics, including the ability to build models that can allow for insight into future evolutionary trajectories of a given population [8].

Overall, the contemporary pika population exhibited low nucleotide diversity, slightly lower than the historic cohort with a similar standard deviation. Tajima's D for the modern cohort shifted toward zero, which could be further indication of either genomic equilibrium, as evidenced by low nucleotide diversity, or indicative of a declining population, as indicated by statistical modeling. Likewise, inter-cohort allele frequency variance described by the global Fst value was low, indicating genetic similarity between the two temporal cohorts. Unfortunately, these global statistics, taken together, indicate little genomic change or adaptation over the last century. However, homing in on the areas with the highest Tajima's D on the chromosome with the highest Fst value led to the discovery of five variant loci when compared to the expected neutral Fst distribution for a population with similar size and demographic history. These loci were all intergenic, although they mapped near a metabolic gene, *DECR1*. Hopefully, the discovery of these variants can be indicative of other, more impactful variants within the population.

LIST OF REFERENCES

[1] ''Species and climate change,'' https://www.iucn.org/, International Union for Conservation of Nature and Natural Resources, 2021, (Accessed on 03/02/2023).

[2] ''Biodiversity – the variety of life on earth,'' https://wildlife.ca.gov/, California Department of Fish and Wildlife, (Accessed on 04/10/2023).

[3] ''Environmental issues,'' https://www.nps.gov/yose/learn/nature/, National Park Service, 2021, (Accessed on 03/02/2023).

[4] ''Anage entry for *Ochotana princeps*,'' https://genomics.senescence.info/species/, Human Ageing Genomic Resources, 2017, (Accessed on 04/10/2023).

[5] A. Smith, ''Conservation status of american pikas (*(Ochtoana princeps)*,'' *Journal of Mammalogy*, vol. 101, no. 6, pp. 1466--1488, 2020. [Online]. Available: doi:https://doi.org/10.1093/jmammal/gyaa110.

[6] J. C. Vardaro, ''Project description: Brc-bio: Adaptive variation through space and time in american pikas (*Ochtoana princeps*).''

[7] E. Beever and P. Brussard, ''Patterns of apparent extirpation among isolated populations of pika (*Ochtoana princeps*) in the great basin,'' *Journal of Mammalogy*, vol. 84, no. 1, pp. 37--54, 2003. [Online]. Available: doi:https://doi.org/10.1644/1545-1542(2003)084<0037:POAEAI>2.0.CO;2.

[8] K. B. et al, ''Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change,'' *PLOS Genetics*, vol. 15, no. 5, 2019. [Online]. Available: doi:https://doi.org/10.1371/journal.pgen.1008119.

[9] ''Whole-genome sequencing (wgs),'' https://www.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing.html, Illumina Inc, (Accessed on 05/6/2022).

[10] G. V. der Auwera and B. O'Connor, *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition).* Sebastopol, CA: O'Reilly Media, 2020.

[11] M. S. et al, ''Approximate bayesian computation,'' *PLOS Computation Biology*, vol. 9, no. 1, 2012. [Online]. Available: doi:https://doi.org/10.1371/journal.pcbi.1002803.

[12] E. Jewett, ''Fast and accurate approximation of the joint site frequency spectrum of multiple populations,'' (Accessed on 04/01/2023). [Online]. Available: DOI:https://doi.org/10.1101/2020.05.01.073213.

[13] S. Andrews, ‘‘Fastqc,’’ https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, (Accessed on 03/10/2023).

[14] S. Chen and Y. Z. et al, ‘‘fastp: an ultra-fast all-in-one fastq preprocessor,’’ *Bioinformatics*, vol. 34, no. 17, pp. i884--i890, 2018. [Online]. Available: doi:https://doi.org/10.1093/bioinformatics/bty560.

[15] P. E. et al, ‘‘Mutliqc: summarize analysis for multiple tools and samples in a single report,’’ *Bioinformatics*, vol. 32, no. 19, pp. 3047--3048, 2016. [Online]. Available: doi:https://doi.org/10.1093/bioinformatics/btw354.

[16] H. Li and R. Durbin, ‘‘Fast and accurate long-read alignment with burrows-wheeler transform,’’ *Bioinformatics*, vol. 26, no. 5, pp. 3589--595, 2010. [Online]. Available: doi:https://doi.org/10.1093/bioinformatics/btp698.

[17] B. S. et al, ‘‘Chromosome-level reference genome assembly for the american pika (*Ochtoana princeps*),’’ *Journal of Heredity*, vol. 112, no. 6, pp. 549--557, 2021. [Online]. Available: doi:https://doi.org/10.1093/jhered/esab031.

[18] A. W. et al, ‘‘The pop-gen pipeline platform: A software platform for population genomic analyses,’’ *Molecular Biology and Evolution*, vol. 38, no. 8, pp. 3478--3485, 2021. [Online]. Available: doi:https://doi.org/10.1093/molbev/msab113.

[19] X. Z. et al, ‘‘A high-performance computing toolset for relatedness and principal component analysis of snp data,’’ *Bioinformatics*, vol. 18, no. 24, pp. 3326--3328, 2012. [Online]. Available: doi:10.1093/bioinformatics/bts606.

[20] P. D. et al, ‘‘The variant call format and vcftools,’’ *Bioinformatics*, vol. 27, no. 15, pp. 2156--2158, 2011. [Online]. Available: doi:https://doi.org/10.1093/bioinformatics/btr330.

[21] L. E. et al, ‘‘*fastsimcoal2*: demographic inference under complex evolutionary scenarios,’’ *Bioinformatics*, vol. 37, pp. 4882--4885, 2021. [Online]. Available: doi:https://doi.org/10.1093/bioinformatics/btab468.

[22] L. Excoffier and H. Lisher, ‘‘Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under linux and windows,’’ *Molecular Ecology Resources*, vol. 10, pp. 564--567, 2010. [Online]. Available: doi:10.1111/j.1755-0998.2010.02847.x.

[23] ‘‘Estimating sequence coverage,’’ https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf, Illumina Inc, (Accessed on 03/22/2023).

[24] GATK Broad Institute. ‘‘Hard-filtering germline short variants.’’ https://gatk.broadinstitute.org/. (Accessed on 03/20/2023).

[25] K. Dasmahapatra. University of York. ‘‘Workshop 4: Population genomics.’’ https://www.york.ac.uk/res/dasmahapatra/teaching/MBiol_sequence_analysis/workshop4_2019.html. (Accessed on 03/18/2023).

[26] T. K. et al, ‘‘Calculation of tajima’s d and other neutrality test statistics from low depth next-generation sequencing data,’’ *BMC Bioinformatics*, vol. 14, 2013. [Online]. Available: doi:https://doi.org/10.1186/1471-2105-14-289.

[27] J. Dabney and M. Meyer, ‘‘Ancient dna damage,’’ *Cold Spring Harb Perspect Biology*, vol. 5, no. 7, 2013. [Online]. Available: doi:10.1101/cshperspect.a012567.

[28] K. C. et al, ‘‘abc: an r package for approximate bayesian computation (abc),’’ *Methods in Ecology and Evolution*, vol. 3, pp. 475--479, 2012. [Online]. Available: doi:10.1111/j.2041-210X.2011.00179.x.

[29] E. W. et al, ‘‘Abctoolbox: a versatile toolkit for approximate bayesian computations,’’ *BMC Bioinformatics*, vol. 11, 2010. [Online]. Available: doi:https://doi.org/10.1186/1471-2015-11-116.

[30] ‘‘Estimating sequence coverage,’’ https://uofuhealth.utah.edu/huntsman/shared-resources/gba/bioinformatics/analysis/germline-analysis, University of Utah Health, (Accessed on 03/22/2023).

[31] K. K. et al, ‘‘Genomic variation in the american pika: signature of geographic isolation and implications for conservation,’’ *BMC Ecology and Evolution*, vol. 21, 2021. [Online]. Available: doi:https://doi.org/10.1186/s12862-020-01739-9.

[32] W. Q. et al, ‘‘Long-term balancing selection contributes to adaptation in arabidopsis and its relatives,’’ *Genome Biology*, vol. 18, 2017. [Online]. Available: doi:https://doi.org/10.1186/s13059-017-1342-8.

[33] N. Eckshtain-Levi and B. Vinatzer, ‘‘The population genetic test tajima’s d identifies genes encoding pathogen-associated molecular patterns and other virulence-realted genes in *Ralstonia solanacearum*,’’ *Molecular Plant Pathology*, vol. 19, no. 9, pp. 2187--2192, 2018. [Online]. Available: doi:10.1111/mpp.12688.

[34] S. Subramanian, ‘‘The effects of sample size on population genomic analyses – implications for the tests of neturality,’’ *BMC Genomics*, vol. 123, no. 17, 2016. [Online]. Available: doi:10.1186/s12864-016-2441-8.

[35] K. Holsinger, ‘‘Genetics in geographically structured populations: defining, estimating, and interpreting fst,’’ *Nature Reviews Genetics*, vol. 10, pp. 639--650, 2009. [Online]. Available: doi:https://doi.org/10.1038/nrg2611.

[36] J. S. et al, ‘‘Chapter 2: Altered metabolism of leukemic cells: New therapeutic opportunity,’’ *International Review of Cell and Molecular Biology*, vol. 336, pp. 93--147, 2018. [Online]. Available: doi:10.1016/bs.ircmb.2017.012.

[37] M. Burrows and D. Wheeler, ‘‘A block-sorting lossless data compression algorithm,’’ https://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-124.pdf, Tech. Rep., 1994.

[38] M. Beaumont and W. Zhang, ‘‘Approximate bayesian computation in population genetics,’’ *Genetics Society of America*, vol. 162, pp. 2025--2035, 2002. [Online]. Available: doi:10.1093/genetics/162.4.2025.

[39] Illumina Inc. ‘‘Qual, qd, and gq formation.’’ https://support.illumina.com/content/dam/illumina-support/help/Illumina_DRAGEN_Bio_IT_Platform_v3_7_1000000141465/Content/SW/Informatics/Dragen/QUAL_QD_GQ_Formulation_fDG.htm. (Accessed on 04/07/2023).

[40] K. Galbreath and K. R. Zamudio, ‘‘When cold is better: climate-driven elevation shifts yield complex patterns of diversification and demography in an alpine specialist (american pika, *Ochotana princeps*),’’ *Evolution*, vol. 63, no. 11, 2009. [Online]. Available: doi:https://doi.org/10.1111/j.1558-5646.2009.00803.x.

APPENDIX A

Supporting Information: Text

A.1 Technical Background - Variant Calling

After read cleanup and annotation, reads are mapped to a reference genome and checked for repetition before genomic variants are called and analyzed. First, BWA, or Burrows Wheeler Alignment tool, allows for the direct comparison of sequence reads to a reference genome, performing a local alignment. It is based on the Burrows Wheeler transform, which is a block-sorting data compression algorithm that allows for the reordering of input text [37]. More specifically, BWA-MEM looks for "maximal exact matches" (MEM), meaning the algorithm searches for the most exact matches between a read and the reference genome and then uses the Smith-Waterman algorithm to fill in gaps, thus "mapping" the read to the genome. BWA-MEM, as opposed to the other iterations of the tool, is used for longer high-quality sequences [16]. Next, MarkDuplicates is a java-based tool in the Picard toolkit, accessible through GATK. It is used to tag duplicate reads stemming from a single DNA fragment, present either as a PCR duplication artifact or an optical duplicate which occurs when the optical sensor identifies a single amplification cluster as multiple clusters. The tool compares the sequence of reads and read-pairs located at the 5' position and identifies the duplicate read based on sum of the base-quality score. A SAM flag is then created for each duplicate, marked as the hexadecimal value 0x0400. HaplotypeCaller, like MarkDuplicates, is also a java-based tool available in GATK. The tool aims to identify variants, including single nucleotide polymorphisms (SNPs) and insertion-deletion variants (indels), using haplotypes and the local reassembly of the haplotypes onto the reference genome. A haplotype is a possible sequence that represents a given section of DNA. HaplotypeCaller uses the input reads to generate these haplotypes, which are then analyzed to find the genotype that is most likely. There are four main

steps in this process: defining the active regions, local assembly of the active regions to generate the haplotypes, analyzing the possible haplotypes, and then assigning genotypes. First, the tool looks for "active regions", which are regions of DNA that have a high likelihood of variation from the reference genotype. After examining each genomic position, the tool creates an "activity profile" using an "activity score" assigned to each position. The shape and local maxima of this activity profile are then used to determine the location and size of the active regions. Then, these active regions are evaluated alongside the input reads that mapped to that specific region to generate haplotypes. It is common to see several haplotypes generated per active region due to read diversity as well as sequencing and mapping errors. Next, the individual reads are then aligned to the haplotypes using PairHMM, a hidden Markov model that determines the amount of evidence that exists for each allele by generating a score that expresses "the likelihood of observing that read given that haplotype" [10]. Finally, the algorithm uses Baye's theorem to select the most likely genotype for that section of DNA, which then is placed in an output gVCF (genomic variant call format) file.

A.2 Technical Background: Statistical Analysis

After the data is processed, it can then be analyzed. Approximate Bayesian Computation, or ABC, is a common analysis in population genetics and can be used to understand the evolutionary changes a population experiences over time by inferring the population size history. Stemming from Bayes theorem, it uses sample parameters to produce artificial datasets which mitigates the need for the likelihood function P(A|B), which can be difficult to determine in real-world problems where datasets are normally very large and complex [38]. Another common statistic in population genetics is site frequency spectra, or SFS. This provides information about genetic

variation within and between populations by summarizing the distribution of allele frequencies. Next, principal component analysis, or PCA, is a feature extraction technique and is a dimension reduction algorithm. It compares genetic covariance of individuals and provides data visualization of the relatedness of individuals based on clustering. The global fixation index, or Fst, describes the distribution of allele frequencies among populations and can identify outlier loci. In this case, outlier loci would be any loci that were subjected to different patterns of selection or a different demographic process [35]. Tajima's D and Fst are summary statistics of the SFS.

A.3 Statistical Tests for Variant Calling

See also: Table 1 and Appendix C.12 Figure. According to GATK best practices, variants are filtered according to six statistical variables to account for quality score, stand bias, and mapping quality. These are QualByDepth (QD), FisherStrand (FS), StrandOddsRatio (SOR), RMSMappingQuality (MQ), MappingQualityRankSumTest (RQRankSum), and ReadPosRankSumTest (ReadPosRankSum). The generic filtering recommendation are based on analysis tests ran by the Broad Institute in comparison with VQSR results [24]. QD is the normalized variant quality score and expresses the variant quality (QUAL) as normalized by the read depth. The QUAL metric describes the Phred-scaled probability that a given site does not have a variant [39]. GATK best practices recommends use of QD for filtering rather than QUAL directly to account for differences in coverage level throughout the genome which affect the QUAL metric and suggests the filtering out of variants with a QD below 2. FS is a measure of strand bias probability computed by the Fisher Exact test. No strand bias relates to an FS of 0 and the generic filtering recommendation suggests a threshold of 60. SOR is also a measure of strand bias, but it is instead estimated by the Symmetric Odds Ratio test and can better handle variants at the end of exons than

the Fisher Exact test. The generic filtering recommendation is to remove variants with an SOR value above 3. MQ is the root mean square mapping quality calculated over all reads at a given site, which allows for the inclusion of the standard deviation in the metric. An MQ of 60 indicates high quality mapping at a given site and the generic filtering recommendation indicates a threshold of 40. MQRankSum is the u-based z-approximation from the Rank Sum Test and describes the quality of reads supporting the reference or alternate allele. A negative value indicates higher mapping qualities for reads supporting the reference allele, a positive value supports the alternate allele, and a value of 0 indicates similar mapping quality at a given site. The GATK recommendation is to filter out MQRankSum values below -12.5. Finally, ReadPosRankSum is the u-based z-approximation from the Rank Sum Test by position. This describes the location within the read that the reference and alternate alleles occur. A negative value means that the end of the read sees the alternative allele more frequently than the reference allele, which can indicate a sequencing error. GATK recommends keeping variants above a ReadPosRankSum value of -8 [24].

A.4 Demographic History Modeling

For each model, a template and estimation file were written in fastsimcoal2 format. The population size parameter and growth rate parameter were passed as ranges from which different sets of parameters could be chosen for each simulation (except for constant population size, where the growth rate was set at 0). Population size was set from 100 to 5000, although the top parameter acts more as a suggestion and the algorithm can exceed that as needed. Growth rate for population decline was set from 0.001 to 0.1, with a positive value being necessary because the algorithm simulates population histories backward in time, meaning that population size would grow as the algorithm works toward the past. Similarly, growth rate for population

expansion was -0.001 to -0.1. DNA mutation rate was set at 1.58e-8 mutations per generation as per findings from Galbreath et al, and no transition bias was set [40].

APPENDIX B

Supporting Information: Tables

B.1 Table B.1: Sample Information

Sequence ID	Cohort	County	Location	Date
JCV1	Historic	Tuolumne	Lyell Canyon (YNP)	14-Jul-15
JCV2	Historic	Tuolumne	Lyell Canyon (YNP)	16-Jul-15
JCV3	Historic	Tuolumne	Lyell Canyon (YNP)	16-Jul-15
JCV4	Historic	Tuolumne	Lyell Canyon (YNP)	17-Jul-15
JCV5	Historic	Tuolumne	Lyell Canyon (YNP)	22-Jul-15
JCV6	Historic	Mariposa	Vogelsang Lake	31-Aug-15
JCV7	Historic	Mariposa	Vogelsang Lake	31-Aug-15
JCV8	Historic	Mariposa	Vogelsang Lake	1-Sep-15
JCV9	Historic	Mariposa	Vogelsang Lake	1-Sep-15
JCV10	Historic	Mariposa	Vogelsang Lake	2-Sep-15
JCV11	Modern	Tuolumne	Lyell Canyon (YNP)	29-Jul-03
JCV12	Modern	Tuolumne	Lyell Canyon (YNP)	31-Jul-03
JCV13	Modern	Tuolumne	Lyell Canyon (YNP)	30-Jul-03
JCV14	Modern	Mariposa	Townsley Lake (YNP)	15-Jul-04
JCV15	Modern	Mariposa	Townsley Lake (YNP)	15-Jul-04
JCV16	Modern	Mariposa	Townsley Lake (YNP)	16-Jul-04
JCV17	Modern	Mariposa	Townsley Lake (YNP)	16-Jul-04
JCV18	Modern	Mono	Gardisky Lake	12-Jul-05
JCV19	Modern	Tuolumne	Lyell Canyon (YNP)	14-Aug-05
JCV20	Modern	Tuolumne	Lyell Canyon (YNP)	12-Aug-06

Table B.2: Sample Information

B.2 Table B.2: Read Depth

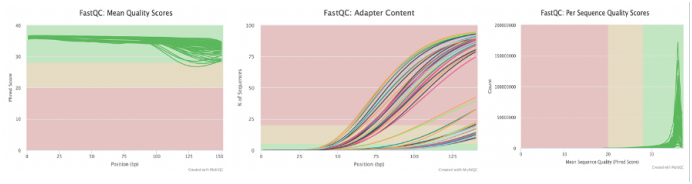
Read Depth = (read count * read length)/total genome size			
	Sample #	Coverage	Percent of Average
Historic	1	20x	110%
	2	18x	90%
	3	16x	80%
	4	20x	100%
	5	19x	100%
	6	22x	110%
	7	21x	100%
	8	17x	90%
	9	17x	90%
	10	21x	110%
Modern	11	19x	100%
	12	19x	100%
	13	22x	110%
	14	19x	90%
	15	14x	70%
	16	22x	110%
	17	19x	100%
	18	18x	90%
	19	33x	160%
	20	22x	110%

Table B.3: Read Depth

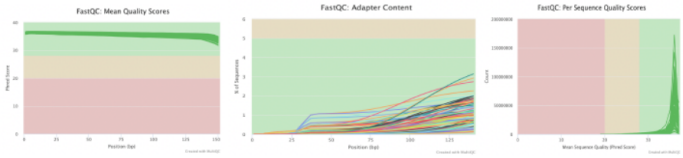
Sample average ($\bar{x}$) = 20, sample standard deviation (s) = 4, range = 14-33x, median = 19x.
Read length = 151 bp, Genome size (reference) = 2,231,492,728 bp

APPENDIX C

Supporting Information: Figures



(a) Raw data quality



(b) Corrected data quality

Figure C.10: Quality Control using FastQC and Fastp

This figure is highlighting some filters of read quality control and their effect on the data. The raw data for both cohorts is pictured first in (a), whereas the filtered data for both cohorts is captured on bottom in (b). (a) The first graph (far left) shows the raw mean quality scores of the reads by position. Reads are 151 base pairs long, represented in the X-axis. The y-axis shows quality scores in Phred-scale. The Phred-score is the logarithmic probability that a base was called incorrectly by the sequencer. A Phred-score of 20 is deemed acceptable and means that a base was called with 99 percent accuracy [10]. The green section of the graph indicates a Phred score above 30 (99.9 percent accurate call) and the red section indicates a fail, as there is a higher probability that call was incorrect. Most reads pass for this filter initially, although some sections of some reads are in the orange warning zone. The second graph (middle) shows adapter content and indicates that the Illumina adapter is present on each read. The third graph (far right) shows the per sequence quality score report. This allows us to see if we have a subset of sequences with universally low-quality values. The y-axis shows the number of sequences with a given Phred-score average. The raw reads had high per sequence quality scores indicating negligible DNA degradation, which could be a potential worry for the historic samples. (b) Fastp performs base correction for the reads in the warning range of the first graph (far left) such that all reads pass for mean quality score after processing. The second graph (middle) indicates all Illumina adapter content was filtered out. The third graph (far right) does not change, as initial per sequence quality scores were high and correction was therefore unnecessary.

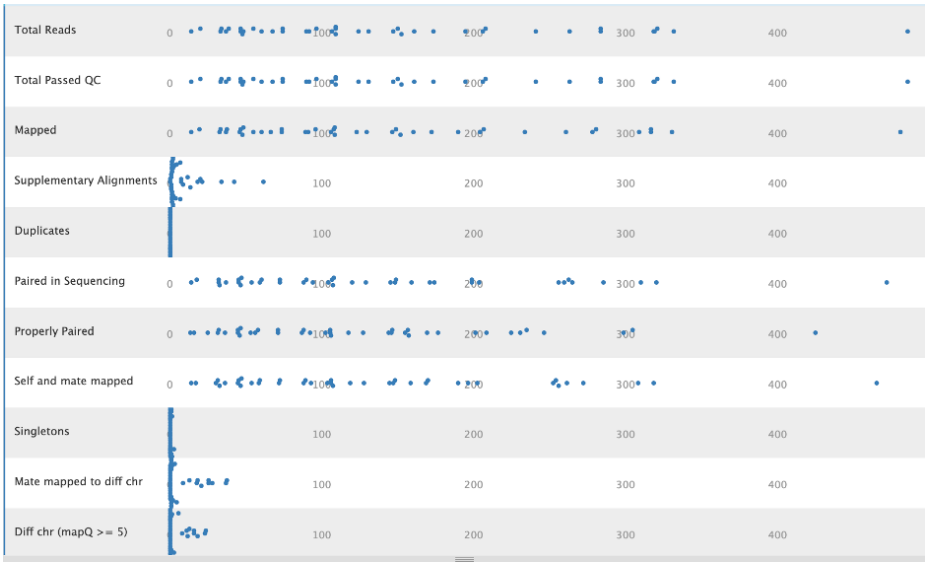


Figure C.11: Mapping Quality

Report generated using MultiQC. Each point represents a single bam file produced from the mapping step. In total, there were 41 bam files generated, one for each fastq file from sequencing. Samples 2-10 and Sample 12 all had three fastq files, each representing a different section of genome, and therefore had three bam files each. Therefore, number of total reads is not equal among all bam files. Therefore, the various mapping statistics should be analyzed based on percentage of total reads in that respective bam file, not comparatively across bam files. Almost 75 percent of samples have under 200 million total reads, with similar results for both number of mapped reads as well as the total number of reads passing quality control. Total reads per sample range from 14 million to 485 million, with the files with higher number of reads corresponding to samples that only have one file associated with the sample, whereas files with smaller number of reads correspond to the samples with three files of read data. For example, sample 20 has 330 million reads present in one file, whereas sample 6 has 320 million reads spread out over three files, ranging in size from 50 million to 160 million reads per file. The modern samples tend to be bigger in file size whereas the historic samples are spread out over multiple runs. Therefore, it is important to look at the data relative to total file size rather than just the raw number of reads.

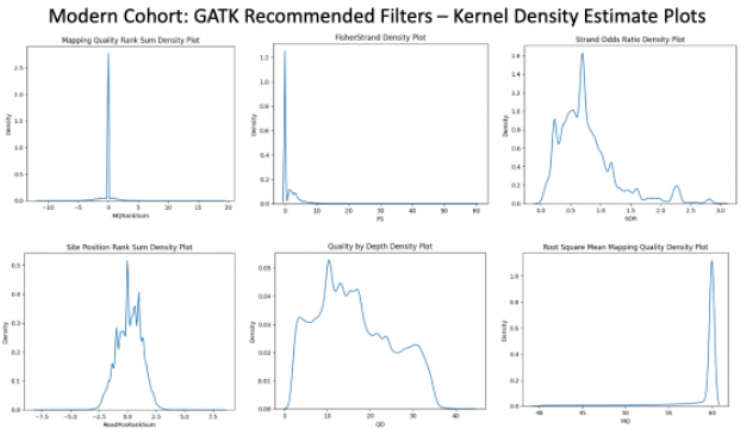
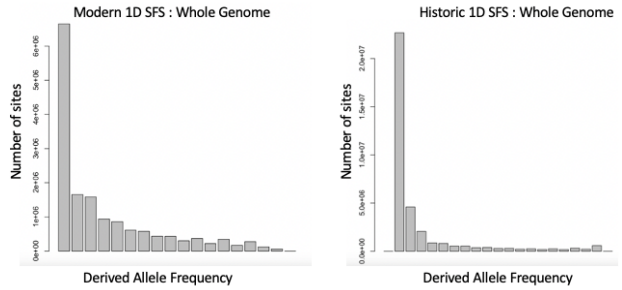
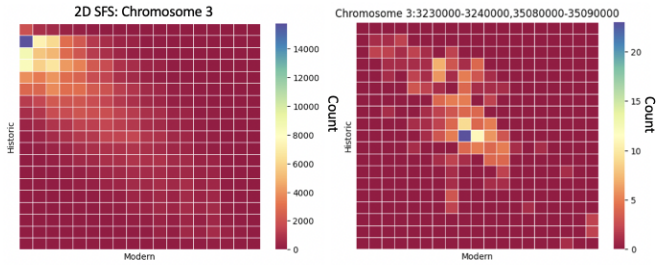


Figure C.12: Graphical Interpretation for Filter Selection

Plots show density of SNPs in the modern cohort with various statistical measurements. Each plot represents the modern cohort after being filtered with the GATK recommended filters. Plots were then visually inspected and custom, more stringent filters were chosen based on the point at which the graph flattens (per GATK recommendation [24]). The same filters were then used on the historic cohort(see Table C.12 for final filters).



(a) One-dimensional Site Frequency Spectra



(b) Two-dimensional Site Frequency Spectra

Figure C.13: Various Site Frequency Spectra

(a) The one-dimensional site frequency spectra for the modern and historic populations across the entire genome. Note the differences in y-axis scale between the two graphs. 'Count' refers to the number of sites seen within the genome with a given derived allele frequency, regardless of where it is in the genome. (b) Joint matrix for allele counts in the entire chromosome 3 and just the areas of interest. Color of each bin is relative to the number of SNPs within that. Bins relate to the number of SNPs that are seen with a frequency of x in the modern samples and y in the historic samples. For example, the blue bin in the 2D-SFS for chromosome 3 (column 0, row 1) means that there are about 15,000 SNPs that are seen no times in the modern cohort and once in the historic cohort.

		Pop size	Grow rate	MaxEstLhood
Constant Population	Run 1	47287	0	-7855
	Run 2	47195	0	-7855
	Run 3	47261	0	-7859
	Run 4	47345	0	-7855
	Run 5	47369	0	-7859
Population Expansion	Run 1	51028	-0.00124	-7884
	Run 2	50849	-0.00157	-7886
	Run 3	51147	-0.00111	-7886
	Run 4	51366	-0.00102	-7877
	Run 5	50521	-0.00103	-7895
Population Decline	Run 1	3343	0.078	-4464
	Run 2	1045	0.091	-4443
	Run 3	4087	0.077	-4467
	Run 4	3878	0.077	-4465
	Run 5	1750	0.085	-4452

Figure C.14: Maximum Estimated Demographic Likelihoods

Top five ECM algorithm derived parameter values per model. Maximum estimated likelihood is in logscale, with values closest to 0 denoting the highest likelihood.

Figure C.15: Bin Formation

Six diagonal and six off-diagonal bins were calculated. Bin width refers to the number of boxes on either side of the box on the diagonal or off-diagonal. A bin width of 2 was used in this analysis. The gray area of the graph mostly represents a box of 0, although about thirty percent of the gray area contained numbers in the range of 1 to 4. Overall, about eighty percent of the observed data was contained within the bins and therefore was used for goodness of fit estimation.

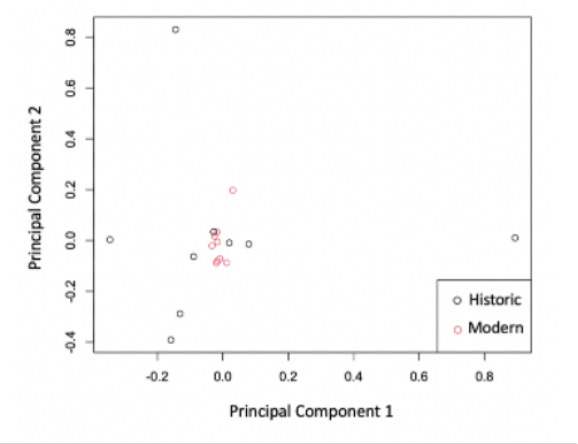


Figure C.16: Initial Principal Component Analysis

Initial principal component analysis indicated DNA degradation in the historic cohort. Historic G to A conversions were removed from final PCA results (shown in Figure 6).