5-2016

Molecular evolutionary analysis of the American pika (Ochotona princeps)

Andrew Michael Rankin
Northern Michigan University, arankin@nmu.edu

MOLECULAR EVOLUTIONARY ANALYSIS OF THE AMERICAN PIKA (*OCHOTONA PRINCEPS*)

By

Andrew M. Rankin

THESIS

Submitted to
Northern Michigan University
In partial fulfillment of the requirements
For the degree of

MASTER OF SCIENCE

Office of Graduate Education and Research

May 2016

SIGNATURE APPROVAL FORM

Title of Thesis: Molecular Evolutionary Analysis of the American Pika (*Ochotona princeps*)

This thesis by Andrew M. Rankin is recommended for approval by the student's Thesis Committee and Department Head in the Department of Biology and by the Assistant Provost of Graduate Education and Research.

Committee Chair: Dr. Katherine Teeter Date

First Reader: Dr. Kurt Galbreath Date

Second Reader (if required): Dr. Alec Lindsay Date

Department Head: Dr. John Rebers Date

Dr. Robert J. Winn Date
Interim Assistant Provost of Graduate Education and Research

ABSTRACT

MOLECULAR EVOLUTIONARY ANALYSIS OF THE AMERICAN PIKA (*OCHOTONA PRINCEPS*)

By

Andrew M. Rankin

The American pika (*Ochotona princeps*) is a small winter-active mammal inhabiting alpine environments. For alpine mammals, metabolic heat production and cellular mechanisms to cope with hypoxia are critical for survival and reproduction. Thus, because of the pikas' long association with alpine environments, they may have genetically adapted to alpine conditions. Here, I present evidence for positive selection acting at three loci within the *Ochotona princeps* genome. These loci were identified by first assembling a list of 54 candidate genes and extracting sequence data from two *O. princeps* draft genomes available from public databases. I then tested the possibility that positive selection acted on the *Ochotona* lineage, relative to other species in the super order Euarchontoglires, using the CODEML program in the PAML package. These analyses indicated positive selection at four loci. I then investigated rates of synonymous and nonsynonymous substitutions within the conserved domain of each of the four genes, which indicated an increased rate of nonsynonymous substitutions within pikas when compared to other mammals. To test this assumption further, I collected DNA sequence data from the four genes of interest in multiple individuals from a broad geographic range. I used the Ewens-Watterson homozygosity test, and Tajima's *D* and Fu's *Fs* statistics to test the resulting sequence data for patterns of non-neutral evolution. These tests have provided additional support for positive selection acting on three of the four genes sequences

ACKNOWLEDGMENTS

Before I begin, I owe a debt of gratitude to my advisors and mentors whom supported me throughout my academic career. Specifically, I would like to mention my undergraduate mentor Dr. Chris Floyd and graduate mentor Dr. Katherine Teeter, both of whom have influenced me greatly. I am also compelled to acknowledge Dr. Kurt Galbreath for his overwhelming encouragement, and Dr. Alec Lindsay for the reassurance he offers to NMU biology graduate students. I would also like to thank my funding sources, the American Society of Mammalogists, the Sigma Xi Research Society, and the Excellence in Education Grant, Spooner Grant, and the NMU Biology Department Developmental Fund.

This thesis follows the format of The Journal of Mammalogy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Any particular evolutionary change within a population occurs with only certain probability (Nei 1987). A mutant allele caused by nucleotide substitution may spread through the population by genetic drift and/or natural selection (Hartl and Clark 1997). It is thus important to consider the relative contribution of stochastic versus deterministic processes when reconstructing the evolutionary histories of organisms and when investigating the mechanisms of molecular evolution (Kimura 1984). Determining whether a character evolved by a particular process of natural selection or by genetic drift is of biological importance, because selection is unique among the evolutionary processes in that it always leads to adaptation (Conner and Hartl, 2004).

When comparing lineages, nucleotide and amino acid substitutions are observed more frequently than changes to protein structure and function (King and Jukes 1969; Kimura 1984). This is because most nucleotide and amino acid substitutions in evolution occur by mutation and random genetic drift, and the majority of molecular variation within a population is neutral (Nei 1987). The vast majority of mutations are deleterious and quickly eliminated by purifying selection, leaving only a small proportion of neutral mutations to account for observed polymorphisms. This would indicate that relatively few mutations are actually advantageous in nature.

Although rare, a small proportion of mutations are advantageous because they confer a fitness advantage in a changing environment or improvement of the competitive ability of individuals or phenotypes within a population (Kimura 1984; Nei 1987). We often consider that this process occurs after the initial appearance of a new allele that has

a selective advantage; however, material for adaptive evolution may also come from neutral alleles existing in a population at low frequency and which have turned out to be advantageous because of an environmental change (Dykhuizen and Hartl 1980). Either way, adaptive substitutions occur from time to time due to changes in both internal and external environment, but the ways in which mutations become advantageous are opportunistic and, therefore, difficult to predict (Kimura 1986).

For adaptive evolution to occur, a particular allele is driven to fixation due to its selective advantage instead of its frequency drifting about randomly, with the probability of its fixation equal to twice the value of its selective advantage (Kimura 1962). This allele can either represent a new mutation or an old variant existing at a low frequency. Under either circumstance, an environmental change (e.g. unpredictable change in climate or evolution of competitors) increases the selective advantage of a particular allele, thereby increasing its probability of fixation and subsequently shifting allele frequencies in favor of adaptive evolution. This is not to say that adaptation is an inevitable consequence of environmental change. Although Fisher's (1930) fundamental theorem of natural selection states that the mean fitness of a population always increases when there are advantageous mutations segregating in the population, this is only true if the environmental conditions remain constant. Natural selection is primarily a process that eliminate unfit alleles and save beneficial alleles (Nei 2013), but populations do not necessarily have the genetic variability for new adaptation. Regardless, a general pattern of adaptive evolution at the molecular level seems to be that only a small proportion of mutations are advantageous, and that is sufficient for adaptive evolution (Nei 1987), and

advantageous mutations spread through the population by selection rather than drift (Kimura and Ohta 1971).

In this study, I identify putatively adaptive loci within the American pika (*Ochotona princeps*) genome by testing for patterns of non-neutral evolution. Although adaptive evolution is proposed to occur among different individuals or phenotypes rather than among different alleles or different amino acid sites (Kimura 1984; Nei et al. 2010; but see Dawkins 2006), it is nonetheless informative to study adaptation for only a single locus as opposed to a genome-wide analysis. In a genome-wide analysis every locus is treated as equal in the statistical analysis, and therefore we may overlook the small proportion of nucleotide substitutions that may have caused innovative changes (Nei et al. 2010).

I will now turn to the organism of interest, the American pika, starting by explaining what is known of their natural history followed by a review of evolutionary studies of *O. princeps*. I end my discussion by considering some recent studies that claim to have obtained molecular evidence of an adaptive nature within the *O. princeps* lineage. I then introduce my thesis work, which is rooted in the view that present-day pika populations are isolated remnants of larger, more contiguous populations in the Pleistocene. American pikas are often considered alpine specialists, as their distribution seems to be restricted due to poor heat tolerance (Smith 1974; Smith and Weston 1990), and their persistence is facilitated by adaptation to hypoxia (Lemay et al. 2013). I test this latter notion that pikas have adapted to hypoxic environment, by performing evolutionary analyses of DNA sequences using a targeted gene approach.

CHAPTER ONE: LITERATURE REVIEW

Natural history of American pika

The American pika is a small lagomorph that is a member of the genus *Ochotona*. The genus *Ochotona* includes 30 species, has a Eurasian origin, and colonized North America via the Bering Land Bridge (Lanier and Olson 2009). In North America, there are only two extant species: the American pika, *O. princeps*, and the collared pika, *O. collaris*. Both the American pika and the collared pika have a fragmented distribution on cool, moist, montane sites, with *O. princeps* located in western North America while *O. collaris* occurs in Alaska and northwestern Canada. In these regions, pikas inhabit mountainous environments surrounded by woodland, grassland or desert, and are typically closely associated with talus slopes or rocky debris habitats (Smith and Weston 1990).

Despite contemporary distributions, *O. princeps* populations have shifted toward lower elevations during past glacial maxima (Hafner and Sullivan 1995; Galbreath et al. 2009). This is supported by fossil evidence which documents pikas at lower elevations during the most recent (Wisconsinan) glaciation, however, the warming trend that followed during the Holocene extirpated these low elevation populations starting about 11,000 years ago (Hafner and Sullivan 1995; Grayson 2005). This event fragmented a more continuous distribution, resulting in the current isolation on montane islands. This view that assumes the current fragmented distribution is a relict of a more continuous distribution during the last glacial period was well established during the mid-20th century (Brown 1971), but studies focused on evolutionary relationships among isolated

populations and the historical events that drove these processes have received little attention until recently.

Several researchers have utilized pikas as an indicator species for deducing the genetic consequences of historical alpine biogeography (Hafner and Sullivan 1995; Galbreath et al. 2009). An initial study based on allozymes identified four major pika lineages distributed across mountain systems (Hafner and Sullivan 1995). A subsequent mtDNA and nDNA phylogenetic analysis supported the previous four major lineages and also discovered a fifth clade (Galbreath et al. 2009, 2010). These five lineages match particular geographic regions and are named as follows: the Cascade Range lineage, Sierra-Nevada lineage, Northern Rocky Mountain lineage, Southern Rocky Mountain lineage, and the Central Utah lineage. Estimation of divergence times using molecular data suggested these lineages were established before the Last Glacial Maximum (LGM) (Galbreath et al. 2010), and that genetic cohesion within mountain systems has been maintained by gene flow during periods of history with cooler climates (Galbreath et al. 2009).

These studies also confirmed that the American pika is monophyletic and sister to the collared pika. The deep differentiation between *O. princeps* and *O. collaris* suggests that these two species diverged first, which was then followed by the intraspecific divergence of *O. princeps* (Galbreath et al. 2010). Unfortunately, this deep divergence between *O. princeps* and *O. collaris* precludes any inference as to the colonization history of Nearctic pikas. It was originally suggested that *O. collaris* descended from the ancient Beringian species, while *O. princeps* originated from individuals that immigrated southward (Guthrie 1973). An alternative hypothesis is that the ancestral northern

population went extinct and left no descendants, and that *O. collaris* is derived from southern immigrants (Mead and Grady 1996; Galbreath and Hoberg 2012). This question has been somewhat answered by a host-parasite comparative phylogeographic study, which suggested that the ancestral Beringian population went extinct and that *O. collaris* was derived from southern individuals that immigrated northward (Galbreath and Hoberg 2012).

Genetic diversity in natural populations of American pika

Preliminary attempts to evaluate genetic relationships among disjunct populations of *O. princeps* revealed extremely low levels of variability. One of the first published studies aimed at determining geographic variation in *O. princeps* attempted to investigate population differentiation among Colorado pikas by electrophoretic mobility of serum proteins (Brunson et al. 1977). The results showed significant, although minor, variation between local geographic areas, which the authors attributed to differential migration rates between localities (Brunson et al. 1977). This study thus provided the first genetic evidence suggesting lack of gene flow between isolated populations of pika.

In another study, researchers investigated protein variation encoded by 26 loci using horizontal starch-gel electrophoresis from four Colorado populations and one Montana population (Glover et al. 1977). This study revealed a mean heterozygosity for the five populations examined to be 1.1%, which is below values reported for other mammals (Wooten and Smith 1985). From these results it appeared that populations had experienced some level of genetic differentiation but that differentiation was most likely due to isolation and drift. In a follow-up effort, Tolliver et al. (1985) reported a similar

low level of mean heterozygosity (1.0%) in addition to low percent loci polymorphic (6.5%). Overall, the data suggested that stochastic processes may play an important role in population divergence of *O. princeps*

The first range-wide study of the American pika examined allozymes at 26 loci among 59 populations representing the entirety of the species' range (Hafner and Sullivan 1995). Expanding the geographic sampling revealed similar patterns that were reported on local scales. Populations exhibited low levels of genetic variation with average values for heterozygosity (1%) and polymorphism (4.6%) below average mammalian levels of heterozygosity (5%) and polymorphism (15%) (Selander and Johnson 1973). Populations also exhibited positive inbreeding coefficients (0.160) and high population differentiation (*Fst* = 0.856) (Hafner and Sullivan 1995).

The above suggests that stochastic processes such as founder effects, genetic drift and bottlenecks have acted to reduce genetic variation in pikas. This is consistent with pika ecology: habitat specialists distributed across small, patchy sites (Smith and Weston 1990). These aspects of pika ecology likely contribute to stochastic processes in reducing genetic variation. Indeed, several studies have investigated the impact of fragmented habitat on dispersal patterns in pika (Peacock and Smith 1997; Peacock et al. 2002; Henry et al. 2012; Castillo et al. 2014; Robson et al. 2015). In these studies, shared genetic variation using neutral loci was assessed to infer dispersal across local landscapes. Results from these studies indicated that small spatial scales do not appear to prevent dispersal in pikas but distances of 10 km or greater seem to be a formidable barrier to gene flow (Peacock and Smith 1997; Peacock et al. 2002). Restricted dispersal between mountain systems separated by low elevation valleys also supports the

hypothesis that geographical isolation is the environmental variable that correlates most strongly with genetic differentiation in pika (Henry et al. 2012; Castillo et al. 2014; Robson et al. 2015).

It is reasonable to assume that both physical and physiological (Smith 1974) constraints represent barriers to gene flow, which would support the general observation that, under current conditions, pikas seem to be isolated within individual sky-islands. In other words, due to both physical and physiological constraints, pikas are currently unable to establish populations between mountain systems that could act as intermediates to facilitate gene flow between source populations within sky-islands. This idea assumes that contact between ranges may be more likely to occur via the island-hopping method, which I hypothesize was the mechanism that connected regional populations during the LGM. As landscapes and climates change through time, sky-island populations may once again coalesce as many did during prior glacial periods.

Pika adaptation

There are many descriptive studies on the physiological (MacArthur and Wang 1973, 1974) and behavioral (Conner 1983, 1985; Ivins and Smith 1983; Holmes 1991; Dearing 1996; Varner and Dearing 2014) adaptations of *O. princeps*. For example, the unusually high body temperature (40.1^{o} C) of pikas is suggestive of an adaptive response to cool, mesic conditions (MacArthur and Wang 1973, 1974) while the amount of time and energy, and level of selectivity, that pikas exhibit when caching suggests that there has been strong selection for caching behavior (Conner 1983). Phenotypic plasticity also seems to play a role in *O. princeps'* ability to persist across a range of environmental

gradients (Varner and Dearing 2014). Indeed, Sheafor (2003) suggested differing selective regimes across elevational gradients to be the primary selective force in driving pika adaptation. Sheafor (2003) examined differences in metabolic enzyme oxidative capacity and lactate dehydrogenase (LDH) isozyme activity among different pika species inhabiting a range of altitudes: *O. princeps*, *O. collaris*, and *O. hyperborea*. Because *O. princeps* inhabits relatively higher elevations than its congeners *O. collaris* and *O. hyperborea*, it was assumed that there may be differences in the activities of metabolic enzymes or in the proportions of metabolic isozymes between these species due to differences in hypoxic conditions (Sheafor 2003). Interestingly, *O. princeps* did exhibit greater metabolic enzyme activities in heart and diaphragm tissue. In addition, *O. princeps* was also found to have greater total LDH activities when compared to *O. collaris*.

Although many assumptions about pika adaptation seem intuitive and correlations are often found between pika phenotypes and the environment, it is difficult to pinpoint at the molecular level which loci have been affected. Thus, in order to reconcile the disparities between phenotypic and molecular evidence of adaptation, several researchers have taken advantage of the recent advances in biochemical studies and the current availability of DNA sequence data, both of which have been used to test hypotheses of adaptive evolution at the molecular level in *O. princeps*.

One of the first bioinformatic studies of American pika evolution compared the Tibetan antelope (*Pantholops hodgsonii*) genome to that of *O. princeps*, both of which are native to high altitudes. Here, Ge et al., (2013) used the CODEML algorithm in the PAML package, which is based on the ratio of non-synonymous to synonymous

substitutions rate, as an approach to identify potential positive selection at the level of individual genes. The results showed that gene ontology categories of putatively adaptive pika genes included ATPase activity, regulation of angiogenesis, oxidoreductase activity, and DNA repair. However, we might consider that the results showed that the pika lineage exhibited >5,000 genes under positive selection. If we take into account the annotation of the pika genome, which reports 25,691 predicted genes, then this would indicate that approximately 20% of pika genes have experienced positive selection and this value seems to exceed expectations. Regardless, seven potential hypoxia genes showed significant signals of positive selection for both the Tibetan antelope and American pika: *ADORA2A, CCL2, ENG, PIK3C2A, PKLR, ATP12A,* and *NOS3* (Ge et al. 2013).

Another bioinformatic study of pika evolution described selection of a single gene occurring along the *Ochotona* lineage. When comparing the T-complex testis expressed protein 1 domain containing 4 (*TCTEX1D4*) binding motif in >20 mammalian species, the motif was observed to be absent in *O. princeps* (Korrodi-Gregório et al. 2013), which was confirmed by the sequencing of additional individuals. A subsequent comparison of d_N/d_S ratios supported the hypothesis that this region has been evolving under positive selection in pika. That is, when comparing ratios between pikas and each of the mammalian sequences, d_N/d_S was always >1, suggesting that this gene-region has evolved under positive selection in pika (Korrodi-Gregório et al., 2013). In addition, a biochemical assay showed that pika *TCTEX1D4* maintained its original function, which provides some evidence for positive selection in that the observed changes in pika *TCTEX1D4* may have allowed the protein to acquire an additional function.

Studies examining local adaptation are useful because they can provide information of how species and populations have responded to differing environmental conditions across their geographic range. Henry and Russello (2013) investigated adaptive population divergence in pika populations found in the Coast Mountains of British Columbia, Canada. The authors used a F_{ST} outlier detection method to identify loci potentially subject to selection in this system, and followed up with an investigation of which environmental variables might be associated with adaptive divergence along independent elevational and longitudinal gradients. To do so they used AFLP-based genomic scans, multiple F_{ST} outlier detection methods, and linear regression models to correlate AFLP band frequencies with environmental variables. The authors graphically depicted two loci whose allele frequencies show the highest correlation with an environmental gradient. To summarize, linear regression of the frequency of one allele against mean annual precipitation depicts a significant negative relationship across a longitudinal gradient from coast to interior. In addition, linear regression of the frequency of a different allele against summer mean maximum temperature depicts a significant negative relationship across an elevational gradient. The authors then conclude to have identified loci responsible for local adaptation and that two main environmental variables—mean annual precipitation and summer mean maximum temperatures—were identified as being associated with adaptive divergence (Henry and Russello 2013).

It is important to note that the F_{ST} outlier detection method assumes that no mutation or selection affects differentiation (Hartl 2000; Conner and Hartl 2004). If the assumption of no mutation is met, then an F_{ST} for one locus that differs from the other loci suggests that selection is acting at that locus. However, because AFLPs are based on

random genomic sequences, we might assume that the likelihood of a locus representing a particular region in the genome would be equal to its frequency within the genome, which, for protein-coding loci, is relatively low. This means that the probability of an AFLP band representing protein-coding loci, or linked regions, is also low. On the other hand, non-coding DNA sequences, which are noted for being highly variable, comprise a large majority of mammalian genomes. Therefore, due to the nature of AFLP data and the expected composition of mammalian genomes, it is not necessarily a safe assumption to exclude mutation affecting differentiation of outlier loci as opposed to selection.

Looking closely at the data presented by Henry and Russello (2013) we note that the frequency of one allele shows a correlation with mean annual precipitation across the longitudinal gradient from coast to interior. It is widely known that the Pacific Northwest receives moist winds from the west which, as it flows upward upon encountering coastal mountains, cools at higher elevations leading to high levels of precipitation on coastal land. Further inland, precipitation decreases and conditions become drier on the leeward side. Thus, it seems that a simpler explanation is that allele frequencies are explained by isolated populations across longitudinal transects simply due to mutation and drift in the absence of migration. However, because precipitation is also associated with longitude in this case, we see a correlation between precipitation and allele frequencies as a causal relationship. Indeed, we find that if we perform a linear regression of longitude against mean annual precipitation using the author's data, we see a very significant relationship across the longitudinal gradient from coast to interior.

In addition to this, we can analyze the other reported locus in a similar way. That is, the second reported locus shows the frequency of one allele to be correlated with

summer mean maximum temperature across an elevation gradient. Again, using their reported data and performing a linear regression of elevation against summer mean maximum temperature, we see a very significant relationship across the elevation gradient. Once again, it is well established that temperature decreases as elevation increases. This would indicate that allele frequencies are likely to be correlated with altitude and only casually related to summer mean maximum temperature. In other words, populations in high elevations and those in low elevations are not exchanging migrants and stochastic processes are causing allele frequencies to drift along independent trajectories.

In a follow up to the previous study, Lemay et al. (2013) sampled pikas from both high and low elevation sites along the same elevation gradient in the Coast Mountains of British Columbia and produced transcriptome-wide sequence data for pooled cDNA libraries. These data have been uploaded to GenBank and they acted as a valuable resource for my own work. They report the discovery of two distinct haplotypes within the mitochondrial encoded NADH dehydrogenase subunit 5: one associated with high-elevation and the other with low-elevation, and a high elevation-specific hemoglobin alpha chain allele. These findings may indicate elevation-specific patterns in the distribution of genetic variation within pika populations, but from previous studies we know that there is a high likelihood of geographic patterns of genetic variation due to mutation, isolation and drift. In addition, a recent study found a modest amount of variation of pika hemoglobin alpha chain and no evidence for positive selection (Tufts et al. 2014).

Tufts et al. (2014) used protein synthesis methods to estimate the structural and functional effects of mutations in *O. princeps* hemoglobin beta chain. The authors reported that the hemoglobins of *O. princeps* and *O. collaris* were separated by five amino acid substitutions in the beta chain. They then measured oxygen affinities (P50) of purified hemoglobins of both species and claim that *O. princeps* hemoglobin exhibited a higher oxygen affinity than that of *O. collaris,* although they do not report a *P*-value despite reporting a mean and replicate number for each species. Regardless, this experiment suggested important differences driving divergence between *O. princeps* and *O. collaris*. The authors reported a mean P50 value for *O. princeps* to be 5.20 torr (n = 4) and a mean P50 value for *O. collaris* to be 6.23 torr (n = 4). Indeed, it appears that *O. princeps* has a higher affinity than that of *O. collaris*, but both of these values are far below means reported for other mammals, including *Cammelus dromedaries* (21.5 torr), *Lama guanicoe* (22.2 torr), *L. glama* (20.3 torr), *L. pacos* (20.3 torr), *Vicugna vicugna* (17.5 torr), and *Homo sapiens* (26.0 torr) (Storz 2007). Thus, I suggest that a more likely explanation is that the hemoglobin beta chain may have experienced positive selection in the ancestor of *Ochotona*, which would explain the exceptional low P50 of pikas when compared to other mammals, and that the five amino acid substitutions separating *O. princeps* and *O. collaris* are effectively neutral.

Future Direction

A convincing study includes both empirical data and a deductive interpretation. However, because we often lack the means necessary to compile a comprehensive data-set, we are often left to piece together a story with only limited insight. Although

physiological, behavioral, and distributional evidence would suggest pikas are well adapted to hypoxic environments, up to this point, there has been limited genetic evidence to support this assumption. Resolving this gap between apparent phenotypic evolution and molecular signals of adaptation is the focus of this present study.

CHAPTER TWO: MINING THE AMERICAN PIKA (*OCHOTONA PRINCEPS*) GENOME FOR PUTATIVELY ADAPTIVE LOCI

Introduction

Present-day American pika populations are isolated remnants of larger, more contiguous populations in the Pleistocene. This view is supported by both molecular and paleontological studies that indicate distributions shifted upslope and became progressively fragmented as climate warmed following the Last Glacial Maximum (Grayson 1987, 2005; Hafner and Sullivan 1995; Galbreath et al. 2009, 2010). American pika are now considered alpine specialists: restricted to their high-altitude distribution due to poor heat tolerance (Smith 1974), while their persistence is facilitated by adaptation to hypoxia (Lemay et al. 2013). This model suggests that, as populations shifted their distributions upslope, pikas would have experienced increased selective pressure to develop adaptations to hypoxia. However, there has been limited genetic evidence to support the latter assumption (but see Luo et al. 2008; Yang et al. 2008; Ge et al. 2013).

Resolving this gap between apparent phenotypic evolution and molecular signals of adaptation thus represents a challenge to the field of ecological genetics. This challenge may be approached by first identifying the selective force and subsequently investigating the molecular targets of selection (Conner and Hartl, 2004). When considering pika populations, it is reasonable to assume that they experienced strong selective pressure to develop adaptations to hypoxia as distributions shifted upslope and became progressively fragmented as climate warmed following the Last Glacial

Maximum. We would thus expect to find molecular signals of this adaptive event in the genome. We then must identify a small number of genes with signals of positive selection among the many thousands of genes in the genome. This issue can be somewhat alleviated, however, if we have prior reason to believe that a particular gene has been a target of selection. That is, if a gene's product is an important component of metabolic regulation, then perhaps a new high-fitness allele arose and was strongly selected for in recent evolutionary time.

Following this assumption, a reasonable approach would be to select genes that have previously shown evidence of contributing to pika-specific adaptation and then investigating those loci for adaptive evolution. Indeed, there is some evidence for genes involved in energy metabolism and hypoxia-response contributing to pika adaptation to a hypoxic environment (Li et al. 2001; Zhao et al. 2004). For example, in a comparative genomic study between Tibetan antelope (*Pantholops hodgsonii*) and *O. princeps*, several hypoxia genes were identified as strong candidates for convergent evolution (Ge et al. 2013), while another study cited several metabolic enzymes whose cellular activity in *O. princeps* correlated to altitude (Sheafor, 2003). In addition, extensive research on the plateau pika (*Ochotona curzoniae*) has demonstrated adaptive responses to both cold and hypoxia via: (1) an increase in non-shivering thermogenesis (Wang et al. 1999, 2006; Yang et al. 2008, 2011; Bai et al. 2015) and (2) transcriptional regulation of genes important for maintaining oxygen homeostasis (Chen et al. 2007; Li et al. 2009, 2013; Pichon et al. 2009, 2013; Xie et al. 2014). To date, a number of candidates have been identified for *Ochotona* as genes involved in adaptive thermogenesis and adaptation to hypoxia. These particular loci form the basis of this research with the prediction that, if

positive selection acted to increase the frequency of an allele conferring high fitness, then we should observe elevated rates of nonsynonymous substitutions in pika coding sequence relative to other mammals.

Materials and Methods

Through a literature search for putatively selected loci for *Ochotona*, 54 genes were selected for subsequent analysis (Table 1). In order to investigate if these genes have indeed experienced selection in *Ochotona*, I performed evolutionary sequence analyses using publically available data from two *O. princeps* draft genomes (GCA_000292845 OchPri3.0, and GCA_000164825 OchPri2.0), together representing two lineages (Hafner and Smith 2010), as well as sequence data representing various mammalian taxa located in the GenBank database.

In order to isolate *O. princeps* sequences from each draft genome, I used tools available through the UCSC genome browser database (http://genome.ucsc.edu/). The BLAT tool was used to identify pika homologs that had aligned with European rabbit (*Oryctolagus cuniculus*) mRNA sequences obtained from GenBank. After both OchPri3 and OchPri2 sequences were obtained, for each gene I generated multiple alignments between *Or. cuniculus* and *O. princeps* using the MUSCLE (Edgar 2004) application in the software program MEGA6 (Tamura et al. 2013). For each alignment, *Or. cuniculus* mRNA sequence was used as a reference to identify open reading frames, remove introns, and to clip 3' and 5' untranslated regions.

Phylogenetic analysis

To analyze whether genes included in this data set evolved at a different rate in *Ochotona* relative to other mammalian taxa, I first performed reconstruction of each gene phylogeny. I added to each *O. princeps*/*Or. cuniculus* alignment additional sequence data from mammalian taxa belonging to the Euarchontoglires clade (Rodentia, Lagomorpha, and Primates). I used these data to construct genealogies by Maximum Likelihood using GARLI 2.01 (Zwickl 2006), and then tested the possibility that positive selection acted on the *Ochotona* lineage relative to the Euarchontoglires clade using the CODEML program in the PAML package. When constructing phylogenies I considered that, because instances of selection can be clade-specific or occur along particular branches within the phylogeny, including a wide range of species in the analysis may preclude detecting any signal of selection (Grayson and Civetta 2012). Thus, I expanded the phylogenetic analysis to only include species within the superorder Euarchontoglires and chose species belonging to the superorder Laurasiatheria as outgroups (Springer et al. 2007).

Phylogenetic reconstruction was accomplished by first querying NCBI database to obtain mammalian sequences and then performing multiple alignments using the MUSCLE (Edgar 2004) application of the software program MEGA6 (Tamura et al. 2013). Additionally, partial sequences for *O. princeps* (SRX277346 and SRX277347) and European hare (*Lepus europaeus*) (SRX910418, SRX910419, and SRX894642) were obtained through Sequence Read Archive (SRA)-BLAST searches. Multiple alignments were evaluated by eye and adjusted manually when necessary. Each alignment was then used to identify the appropriate nucleotide substitution model for each gene separately

using the Akaike Information Criterion (AIC) (Akaike 1981) available in jModelTest (Darriba et al. 2012). Phylogenetic trees were constructed according to the Maximum Likelihood algorithm using GARLI 2.01 (Swickl, 2006), applying 10 generations. Maximum Likelihood trees were displayed using Figtree (Rambaut 2007).

Evolutionary analysis

To analyze the possibility that positive selection acted on the *Ochotona* lineage, I used the Maximum Likelihood codon model from the CODEML program in the PAML package (Yang 1997, 2007). For each gene tree and codon model, the CODEML program computes the likelihood score, which can then be compared between models using a likelihood ratio test (LRT) (Yang 1998). The LRT statistic was calculated as twice the difference in likelihood values between models. Significance was then determined using a chi-squared distribution with one degree of freedom.

I evaluated two codon substitution models in maximum likelihood analysis: the branch-specific and branch-site models. Both of these analyses measure variation in d_N/d_S using Maximum Likelihood (Yang 1998). The branch-specific models allow for d_N/d_S ratios to vary among branches but not sites (Yang et al. 2008). This allows one to investigate accelerated branch evolution in specific lineages (Grayson and Civetta 2013). The null (one ratio) model assumes the same d_N/d_S for all lineages while the alternative (two ratio) assigns one d_N/d_S each to the foreground branch and background branches. The branch site models (model A and model A/null) allow the d_N/d_S to vary both among sites and among lineages and can detect positive selection localized on a few sites along a specific lineage (Yang et al. 2008). Tests of selection treated the ancestral branch leading

to the *Ochotona* lineage as the foreground, thereby assessing the rate of evolution in *Ochotona* relative to other mammals.

Synonymous and nonsynonymous nucleotide substitutions

Four genes that looked most significant from the PAML analyses were investigated further by estimating synonymous and non-synonymous substitution rates along the active/functional region of each gene. In the study of adaptive evolution, it is necessary to compare values of d_N and d_S across different nucleotide regions of a gene because some regions of a gene are subject to natural selection more often than other regions (Hughes and Nei 1988; 1989), and the average evolutionary rate across the entire sequence is not always very informative. Therefore, I identified protein sequence motifs using the NCBI Conserved Domain Search tool (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and calculated d_N/d_S for these isolated regions.

There are several different methods of calculating the rates of synonymous and nonsyononymous substitution, but the most commonly used is that suggested by Nei and Gojobori (1986). This method considers all possible evolutionary pathways in which a codon position can mutate but excludes all pathways that go through stop codons. However, one issue with this method is that the number of synonymous substitutions can be underestimated and nonsynonymous substitutions overestimated (Nei and Kumar 2000). To account for this, I included only phylogenetically close lineages, that is, four lineages belonging to the clade Glires (*Ochotona, Oryctolagus*, *Mus*, and *Rattus*) and a single primate (*Homo*).

I also determine d_N/d_S along the branches leading to *Ochotona*, *Oryctolagus, Mus, and Rattus* by estimating the ancestral Lagomorph and ancestral Rodentia nucleotide sequence using a Maximum Likelihood method to infer the ancestral state for

each nucleotide site (Hall 2004). To do this, I first constructed a gene tree by Maximum Likelihood in GARLI 2.01 (Zwickl 2006) and then imported that tree to the Ancestor application in MEGA6 to estimate the sequence represented by the internal/ancestral Lagomorph and Rodentia node. A single sequence from *Homo sapiens* was used to root the tree.

The previous analysis is helpful in determining the level of purifying selection since divergence from an ancestor, however, it is also useful to know the absolute number of amino acid substitutions which have accumulated in separate lineages since divergence from a common ancestor. For this analysis, I followed the methodology proposed by Kimura (1987). I first calculated the observed number of amino acid differences between species using the Distance application in MEGA6 and then used that information to obtain an estimation of the number of amino acid substitutions (*Xi*) which have accumulated in each lineage since divergence from the common ancestor. In order to account for multiple substitutions and reversions, the following formula is used,

$$D`_{ij} = -n_{aa}\ln(1 - D_{ij}/n_{aa})$$

where n_{aa} is the total number of amino acid sites compared between lineages. Then, we determine X_i's by the following formula:

$$X_i = 1/(L - 2)\ (D_i - D../2(L - 1))$$

Where L denotes the number of species used in the analysis and

$$D_i = \text{sum of all } D`_{ij}$$

and

$$D.. = \text{sum of all } D_i$$

Results

Initial comparisons between OchPri3 and OchPri2 assemblies at candidate genes revealed 26 loci exhibiting only synonymous differences (hereby referred to as monomorphic) and 28 loci that exhibited at least one nonsynonymous difference (hereby referred to as polymorphic) (Table 1). For the 28 polymorphic sequences, I calculated two metrics of whole-gene evolutionary rate comparing the OchPri3 and OchPri2 sequences: (1) number of nonsynonymous substitutions/number of nonsynonymous sites and (2) rate of nonsynonymous substitutions/rate of synonymous substitutions (d_N/d_S) (Table 2). The first comparison allows us to assess the overall divergence at potentially functional sites whereas the second allows us to examine possible differences in selection pressures between homologs (positive, neutral, or negative). The average nonsynonymous divergence was 0.003232, with a range of 0.0003 - 0.0093, and the average d_N/d_S was 0.2444 with a range of 0.0246 - 0.7568.

Maximum Likelihood phylogenetic reconstruction was performed for all 54 genes separately, and topologies were used in CODEML analyses. Of 54 genes, two showed a significantly elevated rate of evolution for both the branch-specific and branch-site model—insulin-like growth factor 1 (*IGF1*) and enoyl-CoA hydratase/3-hydroxyacyl CoA dehydrogenase (*EHHADH*)—while an additional two showed significantly elevated rate of evolution for only the branch-site model—transforming growth factor beta-1 (*TGFB1*) and inducible nitric oxide synthase 2 (*NOS2*).

IGF1 Locus

A multiple alignment was constructed among 22 sequences of *IGF1* representing 15 taxa, which were collected from NCBI GenBank BLAST and SRA-BLAST searches.

Maximum Likelihood phylogenetic reconstruction was performed using the HKY+G+I nucleotide substitution model. The topology obtained from the reconstruction of the ML gene tree differed slightly with the mammalian taxonomy currently accepted (Grayson and Civetta 2012); that is, Lagomorpha is shown to be sister to Primates (Figure 2.1).

The topology of the ML tree was used in all CODEML analyses. In the branch-specific likelihood analysis, the LRT statistic for the comparison of the one ratio (null) model vs the two ratio model was 6.1716 with one degree of freedom ($P = 0.01$). Therefore, the d_N/d_S ratio for the *Ochotona* ancestral branch (0.3237) was significantly different from that for all other branches (0.0691). The branch-site likelihood analysis indicated that a small proportion of sites in the foreground branch have experienced positive selection. The LRT test statistic of comparison of Model A/null vs Model A was 6.688962 with 1 degree of freedom ($P = 0.01$). Model A suggested a d_N/d_S ratio of 26.67324 for the branch leading to the *Ochotona* lineage and identified six codon sites under positive selection (29, 49, 77, 87, 114, and 145) (Table 2.3).

A conserved domain search indicated nucleotides 145-348 matching to insulin-like growth factors; these are specific to vertebrates. The d_N/d_S ratio within this gene region, calculated from pairwise comparisons between *O. princeps*, *Or. cuniculus*, *R. norvegicus*, *M. musculus*, and *H. sapiens*, averaged about 0.0998, while the average for only comparisons including *O. princeps* averaged about 0.1817 (Table 2.4). I used the HKY+G nucleotide substitution model to reconstruct the Maximum Likelihood phylogeny of the *IGF1* conserved domain. As in the previous *IGF1* analysis, the topology obtained differed slightly from expectation, with *Or. cuniculus* more closely related to rodents (Figure 2.2). The resulting topology was used to estimate the ancestral

Lagomorph and Rodentia sequence by Maximum Likelihood using the same substitution model. The d_N/d_S ratio between *O. princeps* and its ancestral node was 0.02913/0.05051 (0.568) while that between *Or. cuniculus* and its ancestor node was 0.00/0.06 (0.00). Likewise, the rate between *M. musculus* and its ancestral node was 0.00/0.000 (0.00) while that between *R. norvegicus* and its ancestral node was 0.00/0.1556 (0.00). Lastly, the estimated numbers of amino acid substitutions for these five lineages indicated more substitutions in *O. princeps*, with pikas accruing 4.732 amino acid substitutions compared to the average of 1.765 and median of 1.568 (Table 2.6).

EHHADH Locus

A multiple alignment was constructed among 15 sequences of *EHHADH* representing 12 taxa, which were collected from NCBI GenBank BLAST and SRA-BLAST searches (Figure 2.3). Maximum Likelihood phylogenetic reconstruction was performed using the GTR+G nucleotide substitution model. The topology obtained conformed to the mammalian taxonomy currently accepted, with Lagomorpha sister to Rodentia forming the Glires clade.

The topology of the ML tree above was used in all CODEML analyses. In the branch-specific likelihood analysis, the LRT statistic for the comparison of the one ratio model (null) vs the two ratio model was 10.5348972 with 1 degree of freedom (P = 0.0001). Therefore, the d_N/d_S ratio for the *Ochotona* ancestral branch (0.4489) was significantly different from that for all other branches (0.2634). The branch-site likelihood analysis indicated that a small proportion of sites in the foreground branch have experienced positive selection. The LRT test statistic of comparison of Model A/null vs the Model A was 4.75761 with 1 degree of freedom (P = 0.03). Model A

suggested a d_N/d_S ratio of 2.99701 for the branch leading to the *Ochotona* lineage and identified 12 codon sites under positive selection (50, 176, 187, 350, 352, 356, 363, 586, 610, 641, 657, and 716) (Table 2.7)

A conserved domain search indicated nucleotides 13-564 matching to the Crotonase/Enoyl-Coenzyme A (CoA) hydratase superfamily. The d_N/d_S ratio within this gene region, calculated from pairwise comparisons between *O. princeps*, *Or. cuniculus*, *R. norvegicus*, *M. musculus*, and *H. sapiens*, averaged about 0.3377, while the average for only comparisons including *O. princeps* averaged about 0.3760 (Table 2.8). I used the GTR+G nucleotide substitution model to reconstruct the Maximum Likelihood phylogeny of the *EHHADH* conserved domain (Figure 2.4). The resulting topology was used to estimate the ancestral Lagomorph and Rodentia sequence by Maximum Likelihood using the same substitution model. The d_N/d_S ratio between *O. princeps* and the Lagomorph ancestor was 0.1146/0.1949 (0.5881) while that between *Or. cuniculus* and the Lagomorph ancestor was 0.0223/0.06738 (0.331). Likewise, the rate between *M. musculus* and the Rodentia ancestor was 0.0984/0.2869 (0.292) while that between *R. norvegicus* and the Rodentia ancestor was 0.0968/0.2188 (0.401). Lastly, the estimated numbers of amino acid substitutions for these five lineages indicated more substitutions in *O. princeps*, with pikas accruing 38.93 amino acid substitutions compared to the average of 26.502 and median of 24.06 (Figure 2.10).

NOS2 Locus

A multiple alignment was constructed among 11 sequences of *NOS2* representing 10 taxa, which were collected from NCBI GenBank BLAST searches (Figure 2.5). Maximum Likelihood phylogenetic reconstruction was performed using the GTR+G

nucleotide substitution model. The topology obtained conformed to the mammalian taxonomy currently accepted, with Lagomorpha sister to Rodentia forming the Glires clade.

The topology of the ML tree above was used in all CODEML analyses. In the branch-specific likelihood analysis, the LRT statistic for the comparison of the one ratio model (null) vs the two ratio model was 1.214452 with 1 degree of freedom (P = 0.27). Therefore, the d_N/d_S ratio for the *Ochotona* ancestral branch (0.1289) was not significantly different from that for all other branches (0.1539). The branch-site likelihood analysis indicated that a small proportion of sites in the foreground branch have experienced positive selection. The LRT test statistic of comparison of Model A/null vs Model A was 5.203264 with 1 degree of freedom (P = 0.02). Model A suggested a d_N/d_S ratio of 999 for the branch leading to the *Ochotona* lineage and identified five codon sites under positive selection (58, 505, 641, 863, and 1031) (Table 2.11). In the CODEML package, when there are only few synonymous differences, the d_N/d_S ratio will be estimated to be infinity (999). The LRT statistic obtained from such a result is still valid and can be used to test for significance (Yang 2007).

A conserved domain search indicated nucleotides 251-1489 matching to the Nitric oxide synthase (*NOS*) eukaryotic oxygenase domain. The d_N/d_S ratio within this gene region, calculated from pairwise comparisons between *O. princeps*, *Or. cuniculus*, *R. norvegicus*, *M. musculus*, and *H. sapiens*, averaged about 0.1174, while the average for only comparisons including *O. princeps* averaged about 0.1053 (Table 2.12). I used the GTR+G nucleotide substitution model to reconstruct the Maximum Likelihood phylogeny of the *NOS2* conserved domain. The resulting topology was used to estimate

the ancestral Lagomorpha and Rodentia sequence by Maximum Likelihood using the same substitution model. The d_N/d_S ratio between *O. princeps* and the Lagomorph ancestor was 0.0242/0.2636 (0.076) while that between *Or. cuniculus* and the Lagomorph ancestor was 0.0152/0.1096 (0.130). Likewise, the ratio between *M. musculus* and the Rodentia ancestor was 0.0063/0.0877 (0.068) while that between *R. norvegicus* and the Rodentia ancestor was 0.0142/0.0965 (0.138). Lastly, the estimated numbers of substitutions for these five lineages indicated more substitutions in *O. princeps*, with pikas accruing 29.711 amino acid substitutions compared to the average of 26.099 and median of 25.924 (Table 2.14).

TGFB1 Locus

A multiple alignment was constructed among 12 sequences of *TGFB1* representing 11 taxa, which were collected from NCBI GenBank BLAST and SRA-BLAST searches (Figure 2.7). Maximum Likelihood phylogenetic reconstruction was performed using the GTR+G nucleotide substitution model. The topology obtained conformed to the mammalian taxonomy currently accepted, with Lagomorpha sister to Rodentia forming the Glires clade.

The topology of the ML tree above was used in all CODEML analyses. In the branch-specific likelihood analysis, the LRT statistic for the comparison of the one ratio model (null) vs the two ratio model was 0.147114 with 1 degree of freedom ($P = 0.97$). Therefore, the d_N/d_S ratio for the *Ochotona* ancestral branch (0.0640) was not significantly different from that for all other branches (0.0570). The branch-site likelihood analysis indicated that a small proportion of sites in the foreground branch have experienced positive selection. The LRT test statistic of comparison of Model

A/null vs Model A was 11.85355398 with 1 degree of freedom (P = 0.0006). Model A suggested a d_N/d_S ratio of 22.62598 for the branch leading to the *Ochotona* lineage and identified six codon sites under positive selection (14, 125, 139, 140, 175, and 278) (Table 2.15).

A conserved domain search indicated nucleotides 84-785 matching to *TGF*-beta propeptide; latency associated peptide (*LAP*) in *TGF-beta*. The d_N/d_S ratio within this gene region, calculated from pairwise comparisons between *O. princeps*, *Or. cuniculus*, *R. norvegicus*, *M. musculus*, averaged about 0.1466, while the average for only comparisons including *O. princeps* averaged about 0.1574 (Table 2.16). I used the GTR+I nucleotide substitution model to reconstruct the Maximum Likelihood phylogeny of the *TGFB1* conserved domain (Figure 2.8). The resulting topology was used to estimate the ancestral Lagomorpha and Rodentia sequence by Maximum Likelihood using the same substitution model. The d_N/d_S ratio between *O. princeps* and the Lagomorph ancestor was 0.0361/0.1802 (0.179) while that between *Or. cuniculus* and the Lagomorph ancestor was 0.0038/0.0878 (0.041). Likewise, the ratio between *M. musculus* and Rodentia ancestor was 0.0114/0.1226 (0.086) while that between *R. norvegicus* and the Rodentia ancestor was 0.00/0.0174 (0.000). Lastly, the estimated numbers of substitutions for these five lineages indicated more substitutions in *H. sapiens*, with humans accruing 19.91 amino acid substitutions compared to the average of 16.104 and median of 14.76 (Table 2.18). However, the analysis did indicate that *O. princeps* (17.87) has accumulated more substitutions as compared to other members of Glires.

Gene	Polymorphic	Monomorphic	Reference
ADORA2A	X		(Ge et al. 2013)
ADIPq	X		(Bai et al. 2015)
ADRB1		X	(Pichon et al. 2013)
ADRB2	X		(Pichon et al. 2013)
ADRB3	X		(Yang et al. 2011)
ANGPT1		X	(Li et al. 2013)
ATP12A	X		(Ge et al. 2013)
ATP1B2		X	(Ge et al. 2013)
EHHADH	X		(Sheafor 2003)
CCL2	X		(Ge et al. 2013)
CEBPA		X	(Bai et al. 2015)
CIDEA		X	(Bai et al. 2015)
CRH	X		(Jizeng et al. 1996; Du et al. 1992)
CRHR	X		(Jizeng et al. 1996; Du et al. 1992)
CS		X	(Sheafor 2003)
DIO2	X		(Li et al. 2001; Yang et al. 2011; Bai et al. 2015)
EDG1		X	(Awad et al. 2006)
ELOVL3		X	(Yang et al. 2011)
ENG	X		(Ge et al. 2013)
EPAS1	X		(Fang et al. 2011)
EPO		X	(Li et al. 2009)
ERRA		X	(Yang et al. 2011)
FABP3		X	(Yang et al. 2011)
FGF21	X		(Lee et al. 2014; Bai et al. 2015)
GCGR	X		(Kietzmann et al. 2002)
HMOX1	X		(Du et al. 2013)
HIF		X	(Zhao et al. 2004b; Li et al. 2009)
IGF1		X	(Chen et al. 2007b)
LDHA		X	(Sheafor 2003; Sigeng 1988; Xiaomei 1985; Xz et al. 2008)
LDHB		X	(Sheafor 2003; Sigeng 1988; Xiaomei 1985 Xz et al. 2008)
LDHC	X		(Wang et al. 2013; Zhang et al. 2014; b)
LPL	X		(Yang et al. 2008c, 2011)
MCR4		X	(Yang et al. 2011)
MDH1		X	(Rj et al. 2012)
MDH2	X		(Rj et al. 2012)
MTNR1B		X	(Wang et al. 1999, 2006)
Myoglobin	X		(Avivi et al. 2010)
NOS1	X		(Pichon et al. 2009)
NOS2	X		(Xie et al. 2014)
NOS3	X		(Ge et al. 2013)
NPY		X	(Yang et al. 2011)
PGC1A	X		(Bai et al. 2015)
PIK3	X		(Ge et al. 2013)
PKLR		X	(Ge et al. 2013)
RANK	X		(Yang et al. 2011)
STAT3		X	(Yang et al. 2011)
TGFB1	X		(Ge et al. 1998)
TRH	X		(Li et al. 2001)
TRHR		X	(Li et al. 2001)
TSH		X	(Li et al. 2001)
TSHR		X	(Li et al. 2001)
UCP		X	(Li et al. 2001; Wang et al. 2006; Bai et al. 2015)
VEGF 165		X	(Li et al. 2013)
VEGF 189	X		(Li et al. 2013)

Table 2.1. Comparison of pairwise alignments between two pika sequences originating from the OchPri3 and OchPri2 draft genomes. Pairwise comparisons exhibiting only synonymous substitutions were considered monomorphic while those exhibiting one or more nonsynonymous substitutions were considered polymorphic.

Gene	Nonsynonymous Divergence (d_N)	d_N/d_S
ADORA2A	0.0022	0.077465
ADIPq	0.0017	0.076923
ADRB2	0.0041	0.394231
ADRB3	0.0062	0.484375
ATP12A	0.0004	0.028169
BHCAD	0.0024	0.250000
CCL2	0.0064	0.275862
CRH	0.0093	0.744000
CRHR	0.0011	0.062500
DIO2	0.0033	0.212903
ENG	0.0047	0.291925
EPAS1	0.0017	0.229729
FGF21	0.0023	0.193277
GCGR	0.0036	0.169811
HMOX1	0.0071	0.162844
LDHC	0.0020	0.156250
LPL	0.0031	0.125506
MDH2	0.0013	0.030232
Myoglobin	0.0037	Undefined
NOS1	0.0003	0.024590
NOS2	0.0008	0.216216
NOS3	0.0016	0.192771
PGC1A	0.0028	0.756757
PIK3	0.0030	0.555556
RANK	0.0073	0.223926
TGFB1	0.0011	0.107843
TRH	0.0048	0.311688
VEGF 189	0.0022	Undefined

Table 2.2. Using the 28 polymorphic sequences between OchPri3 and OchPri2, I calculated (1) the number of nonsynonymous substitutions/number of nonsynonymous sites (Nonsynonymous Divergence) and (2) rate of nonsynonymous substitutions/rate of synonymous substitutions (d_N/d_S).

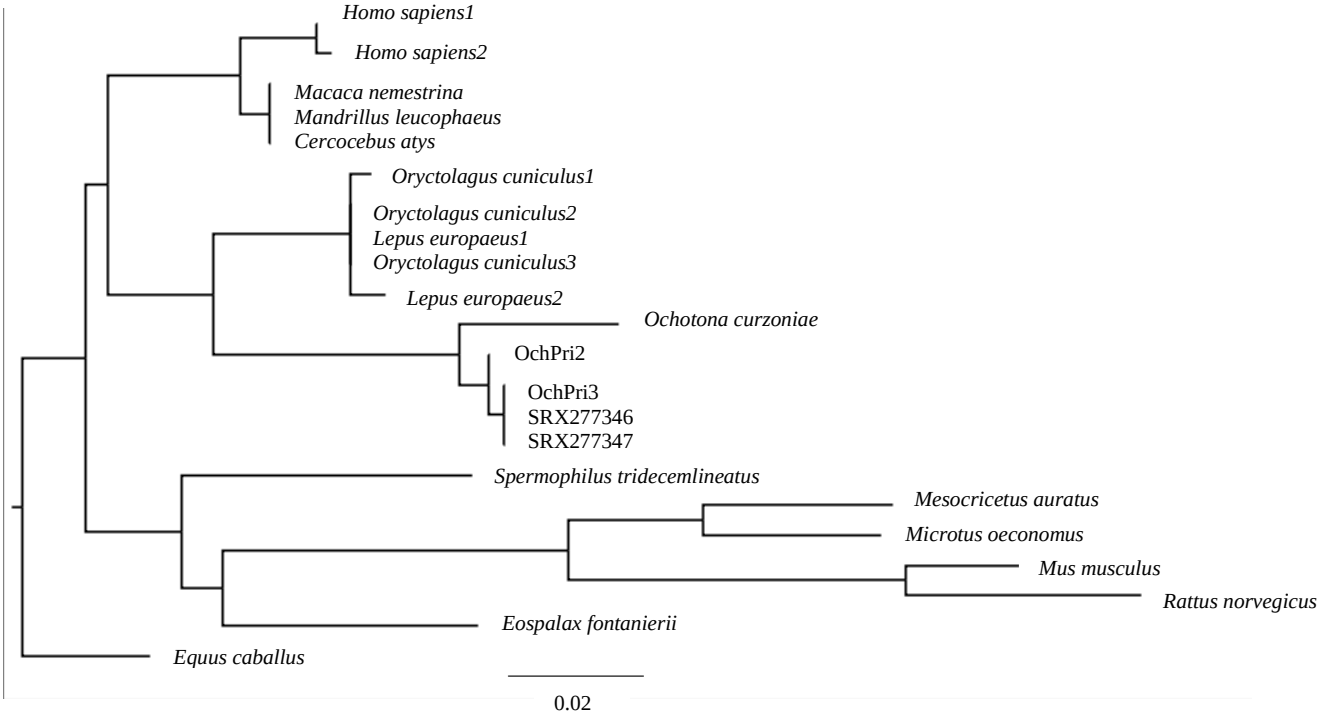

Figure 2.1. Maximum likelihood phylogeny of the *IGF1* gene. Species in this phylogeny include human (*H. sapiens*), macaque (*M. nemastrina*), drill (*M. leucophaeus*), mangabey (*C. atys*), rabbit (*Or. cuniculus*), hare (*L. europaeus*), plateau pika (*O. curzoniae*), ground-squirrel (*S. tridecemlineatus*), hamster (*M. auratus*), vole (*M. oeconomus*), mouse (*M. musculus*), rat (*R. norvegicus*), zokor (*E. fontanierii*), and horse (*E. caballus*). American pika sequences are represented by the OchPri3 and OchPri2 genome assemblies and the SRX277346 and SRX277347 transcriptome assemblies.

Model	lnL	No. parameters	Parameter estimates				
Branch-specific							
1 ratio	-1497.037592	38	Background w	0.0802			
2 ratio/*Ochotona*	-1493.951792	39	Background w	0.0691			
			Foreground w	0.3237			
Branch-site							
Model A null/*Ochotona*	-1490.214985	40	site class	0	1	2a	2b
			proportion	0.70795	00.03388	0.24637	0.01179
			background w	0.04485	1	0.04485	1
			foreground w	0.04485	1	1	1
Model A/*Ochotona*	-1486.870477	41	site class	0	1	2a	2b
			proportion	0.91222	0.04376	0.04201	0.00202
			background w	0.04462	1	0.04462	1
			foreground w	0.04462	1	26.67324	26.67324

Table 2.3. Results of the CODEML branch-specific and branch-site analyses implemented for the *IGF1* alignment.

Seq 1	Seq 2	SynDif	SynPos	d_S	NSynDif	NSynPos	d_N	d_N/d_S
O. princeps	*Or. cuniculus*	8.5	49.50	0.1950	4.5	154.50	0.0297	0.1523
O. princeps	*M. musculus*	17.0	50.25	0.4498	9.0	153.75	0.0609	0.1354
O. princeps	*R. norvegicus*	16.5	50.17	0.4329	8.5	153.83	0.0574	0.1326
O. princeps	*H. sapiens*	4.5	49.50	0.0969	4.5	154.50	0.0297	0.3065
Or. cuniculus	*M. musculus*	15.5	50.75	0.3922	4.5	153.25	0.0300	0.0765
Or. cuniculus	*R. norvegicus*	16.5	50.67	0.4271	4.5	153.33	0.0299	0.0700
Or. cuniculus	*H. sapiens*	9.0	50.00	0.2058	0.0	154.00	0.0000	0.0000
M. musculus	*R. norvegicus*	8.0	51.42	0.1744	0.0	152.58	0.0000	0.0000
M. musculus	*H. sapiens*	17.5	50.75	0.4618	4.5	153.25	0.0300	0.0650
R. norvegicus	*H. sapiens*	18.5	50.67	0.5004	4.5	153.33	0.0299	0.0598

Table 2.4. Ratio of synonymous (d_S) and nonsynonymous (d_N) substitutions between pairwise comparisons of five mammals for the *IGF1* conserved domain.

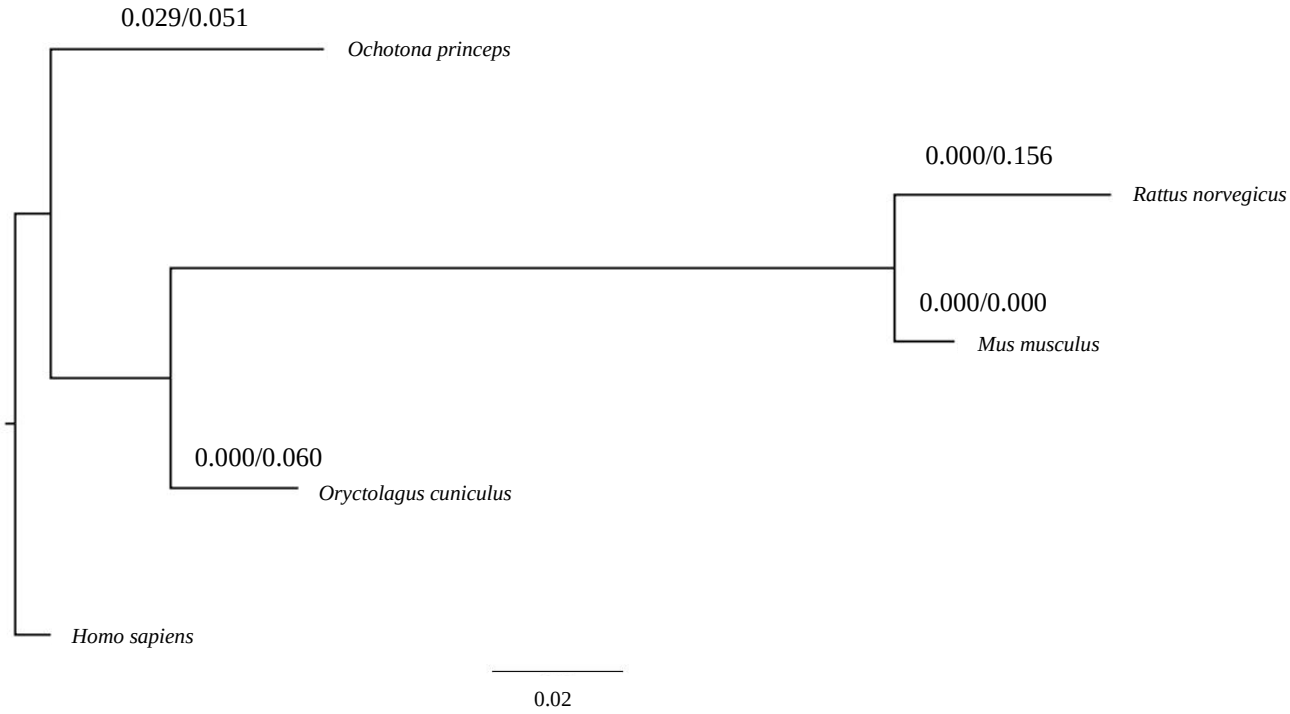

Figure 2.2. Maximum likelihood phylogeny of the *IGF1* conserved domain; nucleotides 145-348. Species in this phylogeny include American pika (*O. princeps*), rat (*R. norvegicus*), mouse (*M. musculus*), rabbit (*O. cuniculus*), and human (*H. sapiens*). The estimated d_N/d_S ratio since divergence from a common ancestor is listed above each branch.

	Or. cuniculus	*M. musculus*	*R. norvegicus*	*H. sapiens*
O. princeps	4	7	7	4
Or. cuniculus		3	3	0
M. musculus			0	3
R. norvegicus				3

Table 2.5. The observed number of amino acid differences between *IGF1* of five mammals: n_{aa} = 68, *L*= 5.

Species	***Xi***
O. princeps	4.732
Or. cuniculus	0.478
M. musculus	0.478
R. norvegicus	1.568
H. sapiens	1.568

Table 2.6. The estimated numbers of substitutions within the *IGF1* conserved domain for five mammals. The mean number of amino acid substitutions is about 1.765 while the median is 1.568.

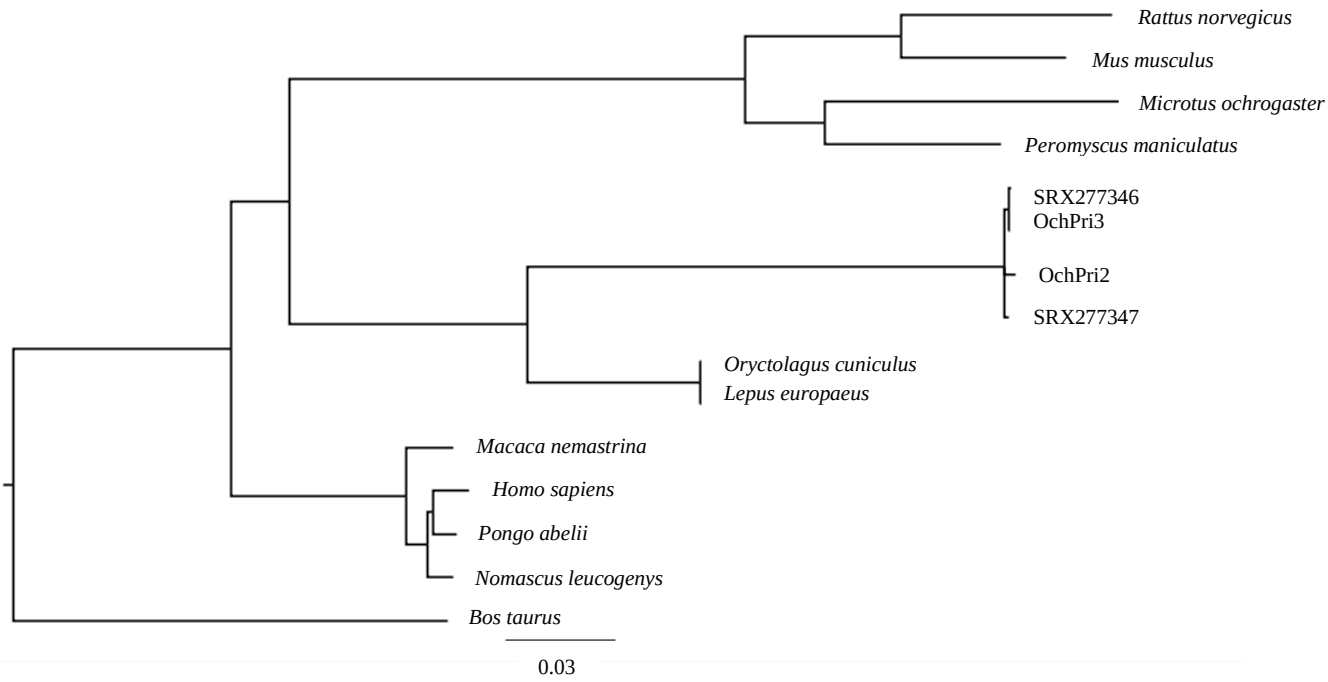

Figure 2.3. Maximum likelihood phylogeny of the *EHHADH* gene. Species in this phylogeny include rat (*R. norvegicus*), mouse (*M. musculus*), vole (*M. ochrogaster*), deer-mouse (*P. maniculatus*), rabbit (*Or. cuniculus*), hare (*L. europaeus*), macaque (*M. nemestrina*), human (*H. sapiens*), orangutan (*P. abelii*), gibbon (*N. leucogenys*), and cattle (*B. taurus*). American pika sequences are represented by the OchPri3 and OchPri2 genome assemblies and the SRX277346 and SRX 277347 transcriptome assemblies.

Model	lnL	No. parameters	Parameter estimates				
Branch-specific							
1 ratio	-10165.3225	28	Background w	0.2825			
2 ratio/*Ochotona*	-10160.0551	29	Background w	0.2634			
			Foreground w	0.4489			
Branch-site							
Model A null/*Ochotona*	-10022.1873	30	site class	0	1	2a	2b
			proportion	0.6626	0.19335	0.1115	0.0325
			background w	0.1034	1	0.1034	1
			foreground w	0.1034	1	1	1
Model A/*Ochotona*	-10019.8085	31	site class	0	1	2a	2b
			proportion	0.7213	0.20803	0.0548	0.0158
			background w	0.1051	1	0.1051	1
			foreground w	0.1051	1	2.997	2.997

Table 2.7. Results of the CODEML branch-specific and branch-site analyses implemented for the *EHHADH* alignment.

Seq 1	Seq 2	Syn Dif	Syn Pos	d_S	NSyn Dif	NSyn Pos	d_N	d_N/d_S
O. princeps	*R. norvegicus*	59.33	143.33	0.6021	76.67	402.67	0.2196	0.3647
O. princeps	*M. musculus*	68.00	142.75	0.7562	73.00	403.25	0.2072	0.2740
O. princeps	*H. sapiens*	48.00	143.67	0.4422	69.00	402.33	0.1947	0.4403
O. princeps	*O.cuniculus*	39.00	144.50	0.3346	52.00	401.50	0.1422	0.4250
Or. cuniculus	*R. norvegicus*	55.33	145.33	0.5314	58.67	400.67	0.1629	0.3065
Or. cuniculus	*M. musculus*	64.50	144.75	0.6763	57.50	401.25	0.1590	0.2351
Or. cuniculus	*H. sapiens*	39.17	145.67	0.3330	46.83	400.33	0.1272	0.3820
R. norvegicus	*M. musculus*	30.50	143.58	0.2497	24.50	402.42	0.0635	0.2543
R. norvegicus	*H. sapiens*	49.00	144.50	0.4513	58.00	401.50	0.1605	0.3556
M. musculus	*H. sapiens*	50.67	143.92	0.4753	58.33	402.08	0.1612	0.3391

Table 2.8. Ratio of synonymous (d_S) and nonsynonymous (d_N) substitutions between pairwise comparisons of five mammals for the *EHHADH* conserved domain.

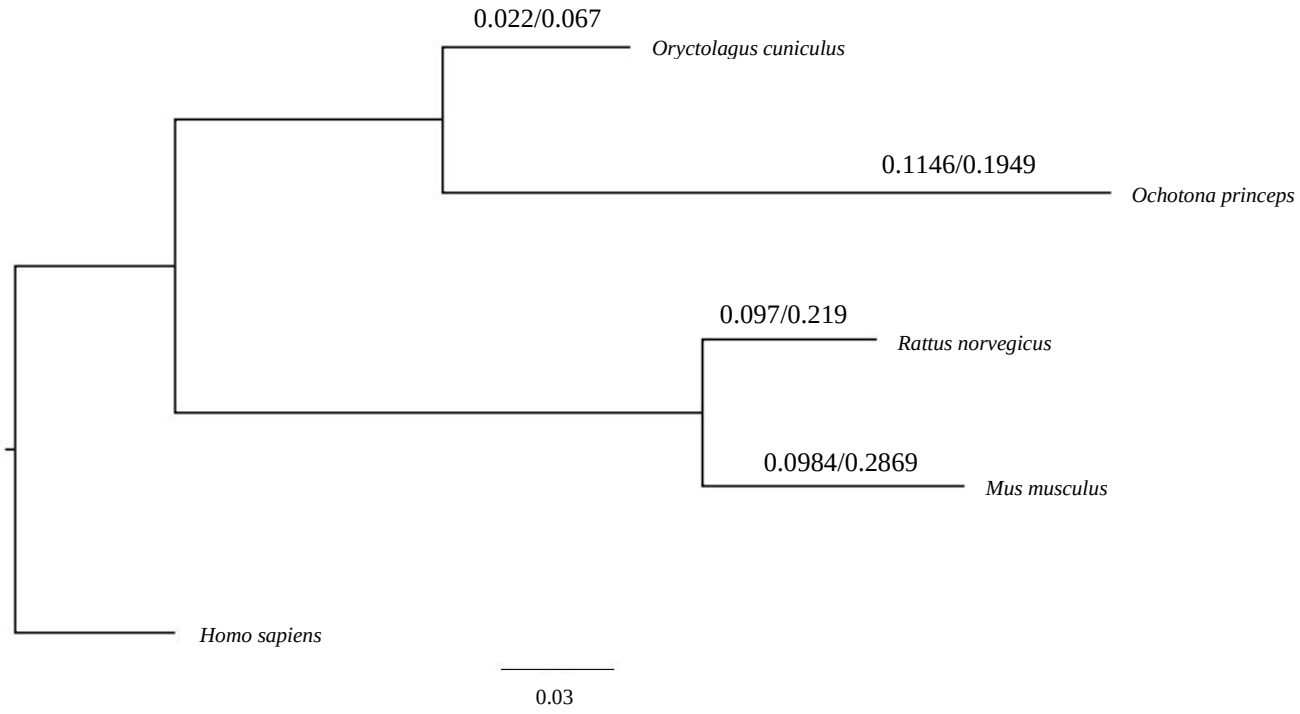

Figure 2.4. Maximum likelihood reconstruction of the *EHHADH* conserved domain; nucleotides 13-564. Species in this phylogeny include rabbit (*Or. cuniculus*), American pika (*O. princeps*), rat (*R. norvegicus*), mouse (*M. musculus*), and human (*H. sapiens*). The estimated d_N/d_S ratio since divergence from a common ancestor is listed above each branch.

	Or. cuniculus	*M. musculus*	*R. norvegicus*	*H. sapiens*
O. princeps	44	55	59	53
Or. cuniculus		45	48	37
M. musculus			21	47
R. norvegicus				48

Table 2.9. The observed number of amino acid differences between *EHHADH* of five mammals: n_{aa} = 184, L = 5.

Species	***Xi***
O. princeps	38.93
Or. cuniculus	22.11
M. musculus	20.34
R. norvegicus	24.06
H. sapiens	27.07

Table 2.10. The estimated numbers of substitutions within the *EHHADH* conserved domain for five mammals. The mean number of amino acid substitutions is about 26.502 while the median is 24.06.

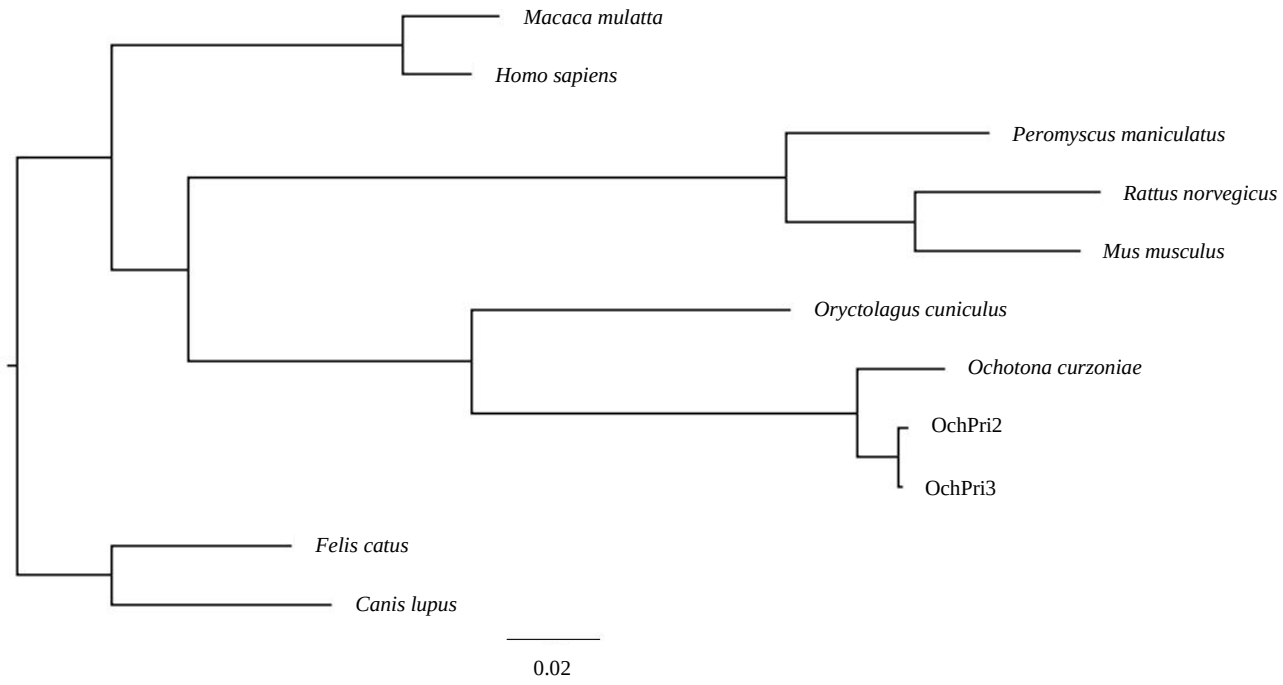

Figure 2.5. Maximum likelihood reconstruction of the *NOS2* gene. Species in this phylogeny include macaque (*M. mulatta*), human (*H. sapiens*), deer-mouse (*P. maniculatus*), rat (*R. norvegicus*), mouse (*M. musculus*), rabbit (*Or. cuniculus*), plateau pika (*O. curzoniae*), cat (*F. catus*), and wolf (*C. lupus*). American pika sequences are represented by the OchPri3 and OchPri2 genome assemblies.

Model	lnL	No. parameters	Parameter estimates				
Branch-specific							
1 ratio	-14440.271753	21	Background w	0.1509			
2 ratio/*Ochotona*	-14439.664527	22	Background w	0.1539			
			Foreground w	0.1289			
Branch-site							
Model A null/*Ochotona*	-14226.118946	23	site class	0	1	2a	2b
			proportion	0.84227	0.15415	0.00302	0.00055
			background w	0.07384	1	0.07384	1
			foreground w	0.07384	1	1	1
Model A/*Ochotona*	-14223.517314	24	site class	0	1	2a	2b
			proportion	0.84246	0.15383	0.00314	0.00057
			background w	0.07372	1	0.07372	1
			foreground w	0.07372	1	998.99967	998.99967

Table 2.11. Results of the CODEML branch-specific and branch-site analyses implemented for the *NOS2* alignment.

Seq 1	Seq 2	SynDif	SynPos	d_S	NSynDif	NSynPos	d_N	d_N/d_S
O. princeps	*Or. cuniculus*	103.50	289.33	0.4861	38.50	949.67	0.0417	0.0858
O. princeps	*R. norvegicus*	125.83	287.58	0.6567	78.17	951.42	0.0870	0.1325
O. princeps	*M. musculus*	136.67	287.58	0.7531	75.33	951.42	0.0837	0.1111
O. princeps	*H. sapiens*	130.33	289.25	0.6887	57.67	949.75	0.0633	0.0919
Or. cuniculus	*R. norvegicus*	120.17	286.58	0.6142	82.83	952.42	0.0924	0.1504
Or. cuniculus	*M. musculus*	128.17	286.58	0.6803	75.83	952.42	0.0842	0.1238
Or. cuniculus	*H. sapiens*	104.00	288.25	0.4920	50.00	950.75	0.0545	0.1108
R. norvegicus	*M. musculus*	52.50	284.83	0.2115	19.50	954.17	0.0207	0.0979
R. norvegicus	*H. sapiens*	112.50	286.50	0.5561	72.50	952.50	0.0803	0.1444
M. musculus	*H. sapiens*	116.83	286.50	0.5885	67.17	952.50	0.0741	0.1259

Table 2.12. Ratio of synonymous (d_S) and nonsynonymous (d_N) substitutions between pairwise comparisons of five mammals for the *NOS2* conserved domain.

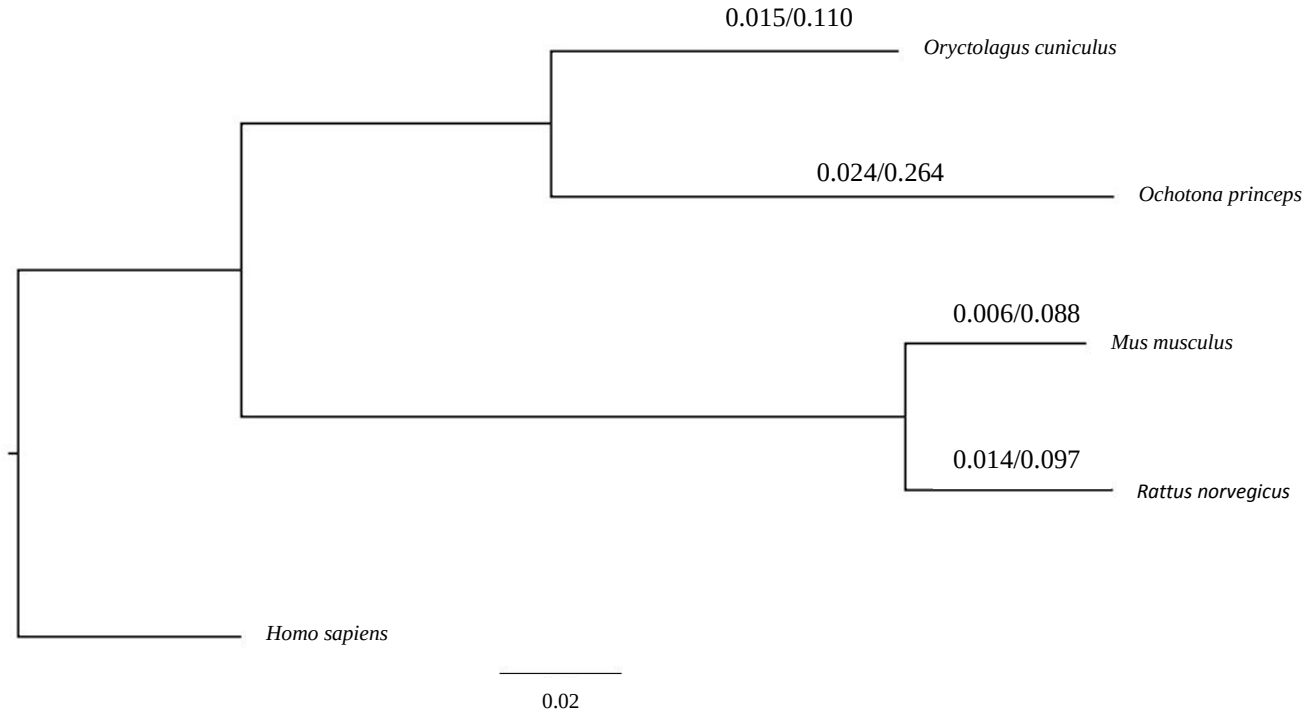

Figure 2.6. Maximum likelihood phylogeny of the *NOS2* conserved domain; nucleotides 251-1489. Species in this phylogeny include rabbit (*Or. cuniculus*), American pika (*O. princeps*), mouse (*M. musculus*), rat (*R. norvegicus*), and human (*H. sapiens*). The estimated d_N/d_S ratio since divergence from a common ancestor is listed above each branch.

	Or. cuniculus	*R. norvegicus*	*M. musculus*	*H. sapiens*
O. princeps	34	60	60	51
Or. cuniculus		61	57	43
R. norvegicus			18	53
M. musculus				51

Table 2.13. The observed number of amino acid differences between *NOS2* of five mammals: n_{aa} = 412, L = 5.

Species	***Xi***
O. princeps	29.711
Or. cuniculus	25.925
R. norvegicus	25.171
M. musculus	22.848
H. sapiens	26.841

Table 2.14. The estimated numbers of substitutions within the *NOS2* conserved domain for five mammals. The mean number of amino acid substitutions is about 26.099 while the median is 25.924.

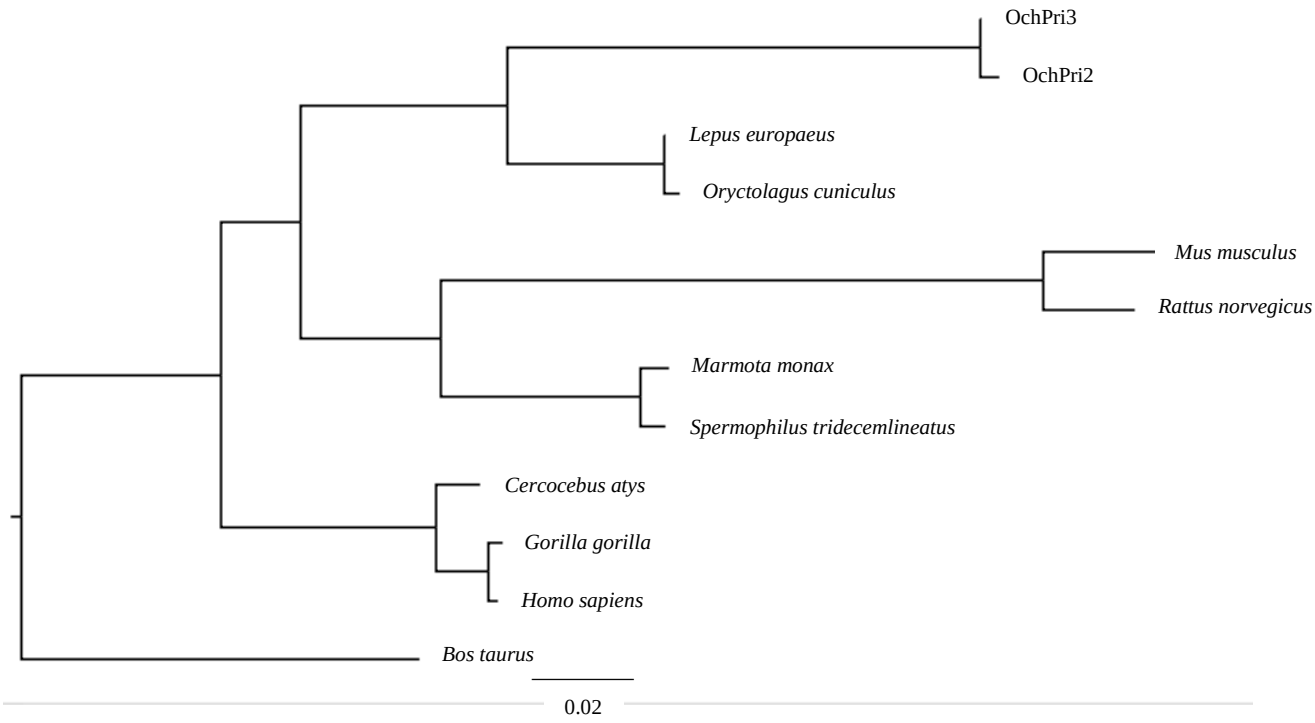

Figure 2.7. Maximum likelihood phylogeny of the *TGFB1* gene. Species in this phylogeny include hare (*L. europaeus*), rabbit (*Or. cuniculus*), mouse (*M. musculus*), rat (*R. norvegicus*), marmot (*M. monax*), ground-squirrel (*S. tridecemlineatus*), mangabey (*C. atys*), gorilla (*G. gorilla*), human (*H. sapiens*), and cattle (*B. taurus*). American pika sequences are represented by the OchPri3 and OchPri2 genome assemblies.

Model	lnL	No. parameters	Parameter estimates				
Branch-specific							
1 ratio	-3836.64123	23	Background w	0.058			
2 ratio/*Ochotona*	-3836.56767	24	Background w	0.057			
			Foreground w	0.064			
Branch-site							
Model A null/*Ochotona*	-3748.74516	25	site class	0	1	2a	2b
			proportion	0.903	0.07207	0.02309	0.00184
			background w	0.0187	1	0.0187	1
			foreground w	0.0187	1	1	1
Model A/*Ochotona*	-3742.81389	26	site class	0	1	2a	2b
			proportion	0.9096	0.07198	0.01709	0.00135
			background w	0.0183	1	0.01833	1
			foreground w	0.0183	1	22.62598	22.62598

Table 2.15. Results of the CODEML branch-specific and branch-site analyses implemented for the *TGFB1* alignment.

Seq 1	Seq 2	SynDif	SynPos	d_S	NSynDif	NSynPos	d_N	d_N/d_S
O. princeps	*Or. cuniculus*	45.00	171.75	0.3223	21.00	527.25	0.0409	0.1269
O. princeps	*R. norvegicus*	79.33	172.50	0.7124	51.67	526.50	0.1052	0.1477
O. princeps	*M. musculus*	82.67	171.75	0.7699	54.33	527.25	0.1109	0.1440
O. princeps	*H. sapiens*	55.50	173.83	0.4159	43.50	525.17	0.0878	0.2111
Or. cuniculus	*R. norvegicus*	63.33	171.25	0.5096	46.67	527.75	0.0941	0.1847
Or. cuniculus	*M. musculus*	70.67	170.50	0.6033	49.33	528.50	0.0997	0.1653
Or. cuniculus	*H. sapiens*	55.50	172.58	0.4200	29.50	526.42	0.0582	0.1386
R. norvegicus	*M. musculus*	24.00	171.25	0.1551	6.00	527.75	0.0115	0.0741
R. norvegicus	*H. sapiens*	71.83	173.33	0.6032	43.17	525.67	0.0870	0.1442
M. musculus	*H. sapiens*	78.83	172.58	0.7044	45.17	526.42	0.0911	0.1293

Table 2.16. Ratio of synonymous (d_S) and nonsynonymous (d_N) substitutions between pairwise comparisons of five mammals for the *TGFB1* conserved domain.

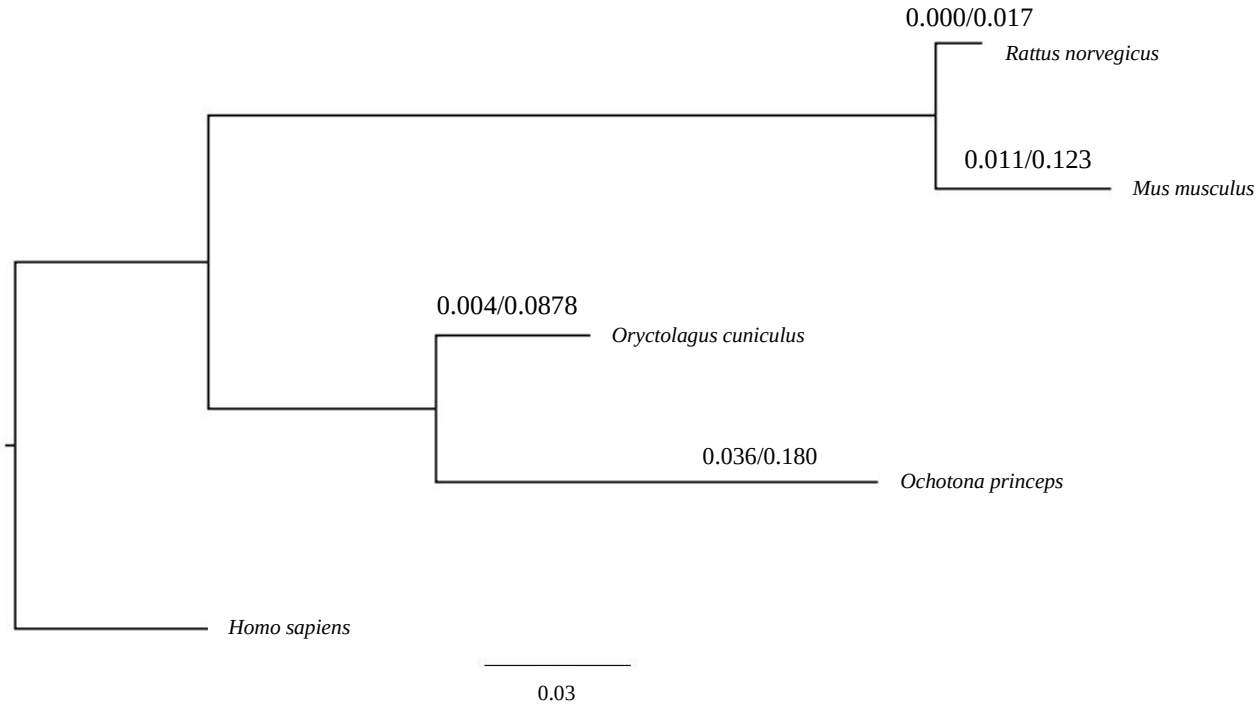

Figure 2.8. Maximum likelihood reconstruction of the *TGFB1* conserved domain; nucleotides 84-785. Species in this phylogeny include rat (*R. norvegicus*), mouse (*M. musculus*), rabbit (*Or. cuniculus*), American pika (O. *princeps*), and human (*H. sapiens*). The estimated d_N/d_S ratio since divergence from a common ancestor is listed above each branch.

	Or. cuniculus	*R. norvegicus*	*M. musculus*	*H. sapiens*
O. princeps	15	38	38	33
Or. cuniculus		36	36	26
R. norvegicus			5	35
M. musculus				36

Table 2.17. The observed number of amino acid differences between *TGFB1* of five mammals: n_{aa} = 232, L = 5.

Species	***Xi***
O. princeps	17.87
Or. cuniculus	13.61
M. musculus	14.76
R. norvegicus	14.37
H. sapiens	19.91

Table 2.18. The estimated numbers of substitutions within the *TGFB1* conserved domain for five mammals. The mean number of amino acid substitutions is about 16.104 while the median is 14.76.

Discussion

Many studies have reported on the physiological and ecological adaptive responses of vertebrates under chronic cold and hypoxia (Hayes and Chappell 1986; Chappell and Hammond 2003; Bouverot 2012; Cheviron et al. 2014), but molecular changes which have been the raw material on which phenotypic adaptation to hypoxia evolved have not been well described. A number of studies have used pikas, which have successfully colonized high elevation environments, to investigate the mechanisms involved in adaptation to hypoxia, and there is now a growing body of evidence that suggests pikas have adapted to hypoxia through regulation of neuroendocrine function (Du et al. 1992), pulmonary circulation (Ge et al. 1998), heart function (Xz et al. 2008), gene expression (Li et al. 2009), and genetic adaptation (Luo et al. 2008). From these studies, a number of genes have been identified as candidates for contributing to these adaptive responses.

Using *Ochotona curzoniae* as a model of hypoxic tolerance, it was demonstrated that this species maintains normal physiological conditions when exposed to hypoxia when compared to rodent species. For example, one study reported that *O. curzoniae* displayed reduced pulmonary vasoconstriction when compared to rats after both species were exposed to hypoxia in a hypobaric chamber (Ge et al. 1998). This was apparent in structural differences that had accrued in the pulmonary capillary beds of the two species. The researchers then tested if these structural differences were due, in part, to differences in the expression of growth factors; specifically, *TGFB1*, a mediator of vascular remodeling (Pepper 1997). A histological analysis showed that *TGFB1* was present in high amounts in the lungs of the rats but not in pikas, suggesting that the structural differences between the two species may be associated with the secretion of *TGFB1*.

However, the mechanism for *TGFB1* inhibition in pikas remained unknown (Ge et al. 1998). It was then suggested that an instance of positive selection acting on pikas' T-complex testis expressed protein 1 domain containing 4 (*TCTEX1D4*), which inhibits *TGFB1* signaling, has led to stronger inhibition of *TGFB1* in pikas (Korrodi-Gregório et al., 2013).

In a series of studies that combined observational and experimental approaches in physiology, it was reported that simulated hypoxia corresponding to an elevation of 5,000 meters did not alter the structure and function of *O. curzoniae* hepatocytes, whereas in lowland rats, mice, and guinea pigs, there was extensive hepatic cell damage via cytolysis (Li, et al. 1986; Li et al. 1987; Xiaoguang et al 1986). It was then hypothesized that the observed cytolysis was due to the loss of lysosomal membrane integrity and subsequent enzyme activity but that, in *O. curzoniae*, *IGF1* and *IGFBP1* acted to protect cell structure (Chen et al. 2007). Insulin-like growth factor I is synthesized primarily in the liver (Schmid 1995) and is mediated through receptor and binding proteins (*IGFBP*s) (Kelley et al. 2002; Duan and Xu 2005). In a follow up study, several animal species received injection of cobalt (II) chloride ($CoCl_2$), which induces endogenous hypoxia, and hepatic *IGF1* and *IGFBP1* mRNA and protein expression were measured (Chen et al. 2007). A significant increase in *IGFBP1* mRNA and protein expression was recorded in *O. curzoniae* at six hours post-injection and maintained up to 12 hours compared with controls. Because *IGFBP*s influence the action of *IGF1* (Duan and Xu 2005), these data support the hypothesis that *IGF1* has a role in the hypoxia response of liver cells.

In addition to secreted proteins like *IGF1* and *TGFB1*, key enzymes involved in energy metabolism pathways seem to be a critical component of the response to hypoxia

in pikas. For example, metabolic enzyme activity of Enoyl-CoA, Hydratase/3-Hydroxyacyl CoA Dehydrogenase (*EHHADH*) was shown to have a positive relationship with elevation in cardiac tissue for three species of pika, and *EHHADH* activity in the cardiac tissue of the American pika was significantly greater than that of its lowland sister species, the collared pika (*Ochotona collaris*) (Sheafor 2003). This is of interest because *EHHADH* activity indicates relative fatty acid oxidation capacity (Wakil et al. 1954). Additional differences between high elevation pikas and lowland taxa have also been reported for nitric oxide synthase (*NOS*) genes, which produce nitric oxide (NO) from L-arginine (Laubach et al. 1995). In sea level mammals, NO is a vasodilator and plays a role in preventing pulmonary hypertension due to hypoxia (Frostell et al. 1991; Pepke-Zaba et al. 1991), as well as facilitating acclimatization to hypoxia via the ventilator response (Haxhiu et al. 1995). Chronic hypoxia thus induces *NOS* mRNA expression and enhances NO production (Cras et al. 1996; Palmer et al. 1998).

When the role of neuronal nitric oxide synthase (*NOS1*) was assessed in *O. curzoniae*, *NOS1* levels seemed to have a significant effect on ventilation via tidal volume (Pichon et al. 2009). That is, with increasing *NOS1* concentration, ventilation was actually limited through a decrease in tidal volume. This effect was only noted during acute hypoxia, which led the authors to hypothesize that limiting ventilation during severe and sudden hypoxia perhaps represents one way in which pikas have optimized their ventilation pattern for life at high elevation (Pichon et al. 2009). In another study investigating whether inducible nitric oxide synthase (*NOS2*) expression and NO production are regulated by chronic hypoxia in *O. curzoniae*, *NOS2* mRNA and protein levels in lung tissues of high elevation pikas were significantly decreased when compared

to low elevation pikas (Xie et al. 2014). Increasing elevation (hypoxia) was correlated with decreased expression of *NOS2* in lungs as well as diminished NO plasma levels when compared to lowland conspecifics. Although the mechanism is unknown, the authors concluded that suppression of *NOS2* and lower NO levels at high elevation may facilitate the plateau pikas' survival in hypoxic environments (Xie et al. 2014).

From the above, because there have been reports of adaptive responses correlated with hypoxia (elevation) for *IGF*s, *TGFB1*, *BHCAD*, and *NOS*s, these genes may have been a target of selection during pikas' colonization and persistence in high elevation environments. Indeed, I performed evolutionary sequence analyses using publically available data from two *O. princeps* draft genomes, which indicated that the American pika lineage has experienced adaptive evolution at all four loci. In order to validate these results, I present evidence that *O. princeps* has accumulated comparably high levels of amino acid substitutions within the active/functional site of each gene. A comparably high rate of nonsynonymous substitutions in *O. princeps* could be explained by either a change in the mutation rate per year, or the alteration of the selective constraint of each molecular (Kimura 1987).

This is good evidence for adaptive evolution, and I suggest that positive selection was involved in producing these results. Thus, I hypothesize a recent selective sweep and, in order to test this hypothesis, I propose to use these four genes for a more intensive investigation of intra- and inter-population genetic variation by collecting additional data for more individuals from each gene and performing additional statistical tests for detecting deviations from neutral expectations of among-population variation.

CHAPTER THREE: USING NUCLEAR LOCI TO DETECT SELECTIVE SWEEPS IN THE AMERICAN PIKA (*OCHOTONA PRINCEPS*) GENOME

Introduction

Populations experience stabilizing selection in a relatively constant environment (Wright 1935). This is due to most genes being under functional constraints, which causes natural selection to remove deleterious alleles from populations and results in neutral evolutionary changes occurring more commonly at the molecular level (Kimura 1984). This interplay between mutation, genetic drift, and purifying selection are what drive molecular evolution. However, in rare instances, a new allele may arise that increases individual fitness. This new allele is considered advantageous, and positive selection may act to increase the frequency of this allele.

If genetic variability at the molecular level is maintained mainly by mutation-drift balance, then the neutral expectation can be treated as the null when examining the distribution of allele frequencies within a population (Nei et al. 2010). Accordingly, one may hypothesize that a given gene has been evolving in a strictly neutral manner, and then test neutral predictions using DNA sequence data and statistical methods. If we are able to reject the null of neutrality, then this may indicate purifying, balancing, or positive selection acting on a particular locus (Nei 2013). However, obtaining statistical evidence for selection is complicated by the fact that demographic fluctuations can also produce similar allele frequency patterns as some types of selection (Tajima 1989; Aris-Brosou and Excoffier 1996).

Polymorphism-based tests are one class of statistical tests for selective neutrality which investigate intrapopulation patterns of allele frequencies (Ewens 1972; Watterson 1977, 1978). These tests compare observed values of allele frequencies with the neutral expectation. For example, Tajima's *D* (Tajima 1989) and Fu's *Fs* (Fu 1997) tests statistics have been developed to assess the number of segregating sites and alleles within a population predicted to be in mutation-drift balance. An abundance of rare segregating sites or alleles is oftentimes interpreted to represent purifying or positive selection (Nei and Kumar 2000). These tests, however, require population level data, of which there is a limited amount for non-model organisms. Furthermore, focused studies require candidate genes that have been previously identified as a target of positive selection. Therefore, in the study of natural selection at the molecular level, it is useful to identify an organism with which a reference genome is available and which has previously been the subject of evolutionary and biological studies.

For this study I used polymorphism-based tests to detect positive selection on candidate genes in the American pika. American pika are alpine specialists with interesting biological traits; exhibiting a high resting metabolic rate (MacArthur and Wang 1973) and elevated Hemoglobin-Oxygen (Hb-O_2) affinities (Tufts et al. 2014). Both of these traits would appear to be adaptive, as many pika populations typically inhabit elevations above 3,000 meters. Cold, low oxygen environments could select for adaptations for efficient oxygen acquisition and use. Indeed, there is previous evidence to support this; two growth factors, *TGFB1* and *IGF1*, as well as the *NOS2* enzyme were hypothesized to have putative roles in the hypoxic response of *O. curzoniae* (Ge et al. 1998; Chen et al. 2007; Korrodi-Gregório et al. 2013; Xie et al. 2014) while *EHHADH*

enzyme activity levels were associated with elevation in *O. princeps* individuals (Sheafor 2003). In addition, I have previously described evidence which suggested an increased rate of amino acid substitution having occurred within the pika lineage at these four genes (Chapter 2). Lastly, the historical demography of American pika has been well documented (Hafner and Sullivan 1995; Galbreath et al. 2009, 2010), thereby making statistical tests for detecting selection easier to interpret. Therefore, I chose to further investigate if *O. princeps* exhibits signs of ongoing or recent selection at the *IGF1*, *EHHADH*, *NOS2*, and *TGFB1* loci using a population genetics approach.

Material and Methods

Sampling and laboratory procedures

I sequenced portions of four genes that were most significant from my previous analyses using specimens of *O. princeps* and its sister species *O. collaris*. The specimens used to create the new datasets were drawn from collections that were described in earlier studies (Galbreath et al. 2009, 2010). I chose a total of 30 individuals—five each from the five *O. princeps* mitochondrial lineages and five *O. collaris*—to use during sequencing. Mitochondrial lineages are abbreviated as follows: Cascade Range (CR), Central Utah (CU), Northern Rocky Mountain (NRM), Sierra-Nevada (SN), and Southern Rocky Mountain (SRM). The geographic distribution of the specimens analyzed are shown in Figure 3.1. Briefly, I plotted the geographic coordinates from each collection location onto a range map using DIVA-GIS v 7.5 (Higmans et al 2004). Spatial data was downloaded from IUCN website (http://www.iucnredlist.org/), while locality information was obtained from the Arctos museum database (http://arctosdb.org/). Of the 25 *O.*

princeps specimens used, there were 22 unique localities, while all five samples of *O. collaris* originated from a single locality.

Coding regions to be sequenced and primers for PCR amplification were identified using the OchPri3 genome assembly available from UCSC genome browser. In each case, primers were placed in non-coding intron regions. In other words, a primer pair was designed so that the forward and reverse primer annealed to the 5' and 3' intron regions flanking an exon. This method allowed for an entire exon to be sequenced in addition to 5' and 3' intron sequence. Primers were designed using Primer3 software available from IDT website (http://www.idtdna.com/primerquest/home/index).

All PCRs were carried out on a thermal cycler in 25 µl reactions containing 1 µl DNA, 18.4 µl water, 2.5 µl buffer, 1.5 $MgCl_2$, 0.5 µl dNTPs, 1 µl forward and reverse primer, and 0.1 µl of Taq polymerase. Bull's Eye Taq DNA polymerase was used for all PCRs (MIDSCI). An initial denaturation at 94^o for 4 minutes followed by 35 cycles at 94^o for 30 seconds, annealing (variable temperature) for 30 seconds, and 72^o for 30 seconds followed by a final extension at 72^o for 7 minutes. PCR products were electrophoresed in a 1% agarose gel and visualized on a gel imager system.

DNA was extracted from gels using a QIAquick Gel Extraction Kit according to the manufacturer's instructions (Qiagen, Valencia, CA). Sequencing of the sampled individuals was performed by GeneWiz, Inc. (http://www.genewiz.com/; North Brunswick, NJ). It was carried-out in 96-well plates using DNA extract normalized at 20 ng/µl for each sample. All DNA fragments were sequenced in both directions and electropherograms were obtained for both forward and reverse primers. Sequences were then determined by taking a consensus sequence. Because electropherograms were of

nuclear loci, two alleles per individual were logged into the final dataset. This included both homozygotes and a minority of heterozygotes which exhibited distinct overlapping peaks at heterozygous sites. Sequence screening was performed using Geneious v7.1 (Drummond 2011), and multiple alignments were constructed via MUSCLE using MEGA6 (Tamura et al. 2013) and default parameters.

Evolutionary change of amino acid and DNA sequences

The evolutionary pattern of each gene was examined more closely by calculating several measures of amino acid and DNA sequence polymorphisms. This was to compare the current datasets with previously published information to confirm the absence of sequencing errors and the use of functional genes rather than pseudogenes, both of which could significantly influence the outcome of downstream analyses.

Amino acid sequences: I calculated the rate of amino acid substitution, which can range from the highly conserved histone H4 (0.01×10^{-9}) to the rapidly evolving fibrinopeptides (9.00×10^{-9}) (Dayhoff and Schwartz 1978), by comparing the evolutionary distance between *Ochotona* and *Or. cuniculus*. Evolutionary distance was calculated using the Distance application in MEGA6 and was measured by the proportion (p-distance) of different amino acids between pika sequences and European rabbit. The p-distance gives an estimate of the number of amino acid substitutions per site. The probability that no amino acid change has occurred during T years at a site can be calculated (Nei 1987) by

$$\mathbf{Q = e^{-2rT}}$$

Therefore,

$$\mathbf{-\ln(Q) = 2rT}$$

Q can be estimated by 1 – p, so that

$$\mathbf{2rT = -\ln(1 - p)}$$

Fossil evidence suggests that the divergence between ochotonids and leporids occurred around 37 million years ago (MYA) (McKenna and Bell 2013). This corroborates with molecular divergence estimates within Glires (Asher et al. 2005) (but see Bininda-Emonds et al. 2007 and Matthee et al. 2004). This means that the rate of amino acid substitution can be calculated by

$$\mathbf{r = -\ln(1 - p) / (2*37\ MYA)}$$

DNA sequences: The previous analysis was repeated but using DNA sequence data and considering both the rate of synonymous and nonsynonymous substitutions separately. DNA divergence can be calculated in the same way as amino acid divergence (Nei 1987); however, DNA sequences are less conserved than amino acid sequences, and we therefore need to correct for multiple hits and reversions in order to obtain an accurate estimate of the number of nucleotide substitutions between lineages.

The simplest model of nucleotide substitution is the Jukes and Cantor model (1969). This model assumes that nucleotide substitution occurs at any site with equal frequency and that at each site a nucleotide changes to one of the three remaining nucleotides with equal probability per unit time. Under the Jukes and Cantor model, 2rT is calculated by,

$$\mathbf{2rT = -(3/4)\ln(1 - (4/3)p)}$$

where p is the proportion of different nucleotides. Therefore, rate can be calculated by

$$\mathbf{r = -(3/4)\ln(1 - (4/3)p) / (2*37\ MYA)}$$

The rate of nonsynonymous substitution is extremely variable among genes, and have been found to range from approximately zero to about 3.1×10^{-9}. The rate of synonymous substitution is also variable from gene to gene, but not as much as the nonsynonymous rate. Rates for synonymous substitutions have been reported to range from 1.57×10^{-9} to 6.39×10^{-9}. The mean synonymous and nonsynonymous rates for growth factors is about 3.49×10^{-9} and 0.804×10^{-9}, respectively, while the mean synonymous and nonsynonymous rates for enzymes is about 3.08×10^{-9} and 0.37×10^{-9}, respectively (Graur and Li 2000).

It is also useful to know the ratio of transitions to transversions, which typically ranges from 0.5 – 2 in nDNA, but can be higher in mDNA (Vigilant et al. 1991). If nucleotide substitution occurs at random, then transversions are expected to occur twice as much as transitions because, of the $4^2 = 16$ different types of nucleotide pairs, eight are transversions, four are transitions, and four are identical pairs. Thus, to determine the pattern of change between *Ochotona* and *O. cuniculus*, I calculated the ratio of transitions to transversions by

$$\mathbf{R = p(transitions) / p(transversions)}$$

When this ratio is greater than 1, more transitions and less transversions have occurred than expected by chance.

The patterns of nucleotide substitution at the first, second, and third codon positions are also of interest. Typically, the third positions are highly variable while the second positions are relatively invariable. This reflects the fact that no synonymous substitution occurs at the second positions, whereas many synonymous substitutions may occur at the third positions. At the first positions, a small proportion of nucleotide

substitutions are synonymous. To assess these patterns between *Ochotona* and *Or. cuniculus,* I calculated the number of base differences per site for each codon position separately using the Distance application in MEGA6. This same application was also used to calculate the proportion of transitional and transversional type substitutions between *Ochotona* and *Or. cuniculus*.

Phylogenetics

To assess the population structure present within *Ochotona*, I estimated the phylogeny of each gene using a Bayesian approach with the Metropolis-coupled Markov chain Monte Carlo method implemented in the program MrBayes 3.2.5 (Huelsenbeck and Ronquist 2001). The analysis was run twice with four chains per run for 10 million generation and the tree was saved every 100 generations. The first 25% of the trees (25000 trees) were discarded and a 50% majority-rule consensus tree was obtained from the remaining trees.

I used two different methods of phylogenetic reconstruction. First, I constructed gene trees using both coding and non-coding (genomic) datasets and specifying *O. collaris* as an outgroup. Then, I constructed trees using only coding (mRNA) sequence from each dataset and obtained homologous mRNA sequences from *Or. cuniculus* or *O. curzoniae* to serve as an outgroup. The appropriate nucleotide substitution model for each dataset was determined using the Akaike Information Criterion (AIC) (Akaike 1981) available in jModelTest (Darriba et al. 2012)

Significant population structure has previously been described across the range of *O. princeps*. Thus, I calculated the relative nucleotide differentiation among lineages for each gene when considering the five mitochondrial lineages. I did this by first estimating

nucleotide diversity for the entire species disregarding structure, and then calculated interpopulational nucleotide diversity (δ_{ST}) by

$$\boldsymbol{\delta_{ST} = \pi_T - \bar{\pi}_S}$$

where π_T is the nucleotide diversity calculated for the entire species and $\bar{\pi}_S$ is the average nucleotide diversity within lineages. The proportion of interpopulational diversity is then given by

$$\boldsymbol{N_{ST} = \delta_{ST} / \pi_T}$$

N_{ST} is the coefficient of nucleotide differentiation, and is equivalent to F_{ST} when using genotype data (Nei and Kumar 2000). It measures the degree of relative nucleotide differentiation among subpopulations or, in this case, lineages. Nucleotide diversity was calculated using the Distance application in MEGA6.

The American pika and collared pika are considered distinct sister species (Lissovsky et al. 2007; Lanier and Olson 2009; Galbreath and Hoberg 2012). Thus, to obtain an estimate of the time of divergence between *O. princeps* and *O. collaris*, I first obtained an estimate of the extent of DNA divergence between the two species by taking the average number of nucleotide differences and dividing by the total number of nucleotides examined for each gene separately. This gave the average number of nucleotide substitutions per site (*Dxy*). Then, to obtain an estimate of the number of net nucleotide substitutions per site (*Da*), I accounted for intraspecific diversity.

$$\boldsymbol{Da = Dxy - ((\pi_{O.\ princeps} + \pi_{O.\ collaris})/2)}$$

Because the expected value of *Dxy* is given by,

$$\boldsymbol{Dxy = \theta + 2rT}$$

where r is the mutation rate per nucleotide site, if both lineages are in mutation-drift balance, then $Dx = Dy = \theta$, so

$$\boldsymbol{Da = 2rT}$$

Using estimates of synonymous and nonsynonymous substitution rates obtained earlier from comparing *Ochotona* with *Or. cuniculus,* I calculated divergence time between *O. princeps* and *O. collaris* by

$$\mathbf{T = \mathit{Da}/(2*r)}$$

These calculations rely on the assumption that the DNA region being analyzed is evolving neutrally, which is likely not the case in reality. But, these estimates can still shed some light on the evolutionary history of a gene, and having multiple estimates of T are useful for comparisons.

Lastly, to examine possible differences in selection pressures between homologs of *O. princeps* and *O. collaris,* I isolated coding sequence from each dataset and calculated the average rate of nonsynonymous substitutions/rate of synonymous substitutions (d_N/d_S).

Homozygosity test for protein polymorphism

To compare the observed homozygosity for each locus with the theoretical homozygosity predicted by neutral expectations, I use the Ewens-Watterson test. This test is based on the sampling theory of neutral alleles (Ewens 1972), which is a way to determine the probability of observing a set of allele frequencies. Ewens (1972) showed that this probability is independent of population size and mutation rate, making it possible to estimate the distribution of expected homozygosity given the number of individuals sampled and number of alleles observed. This distribution can then be used for testing

whether or not observed homozygosity is higher or lower than the neutral expectation. Following Watterson (1977, 1978) homozygosity (*F*) is calculated by the following formula:

$$F = \text{sum of all } n_i^2/n^2$$

where *n* is the sample size, and *ni* are the numbers of representatives of these alleles in the sample. Observed homozygosity (*F*) is then compared to the expected homozygosity (*F*). Expected homozygosity (*F*) values were estimated using Arelquin v 3.5 (Excoffier and Lischer 2010) and a null distribution of expected homozygosity was generated by simulating 10,000 random neutral samples using the same number of individuals sampled and number of alleles observed. The probability of observing random samples with *F* values identical or smaller than the original sample is then reported. When the observed homozygosity matches the expected homozygosity, this probability will equal about 50%.

When observed homozygosity is too high, this suggests that there is purifying selection or the presence of an advantageous allele. If observed homozygosity is too low, this suggests that there is some type of balancing selection (Nei 1987). This test is based on the infinite-allele model, and is therefore applicable to protein polymorphisms (Nielsen 2001). Thus, alleles were called by considering only nonsynonymous substitutions.

Statistical tests for detecting selection

Since all morphological and physiological characters are controlled by genetic variation, it is of interest to study the changes in DNA sequences that are caused by natural selection. The neutral theory of molecular evolution makes several predictions which can be tested using DNA sequence data. To test the predictions under the neutral theory using intraspecific genetic variation, it is useful to use polymorphism-based methods, which

may identify purifying selection, balancing selection, or positive (directional) selection. Several test statistics have been developed, and all examine the relationship between the number of segregating sites or alleles and nucleotide diversity.

Any nucleotide site that shows two or more nucleotides among a set of sequences is called a segregating site. The total number of segregating sites in a data set is defined by *S*. The number of segregating sites per nucleotide site (*Ps*) is thus the total number of segregating sites divided by the total number of nucleotides examined. Note that *S* and *Ps* are dependent on sample size and increase as the number of sequences increases. However, we can use *Ps* to estimate theta (θ), which is the product of mutation rate and population size and is independent of sample size (i.e., $\theta = 4N_e\mu$). The expected value of *Ps* is given by

$$\boldsymbol{Ps = a_1\theta}$$

therefore,

$$\boldsymbol{\theta = Ps/a_1}$$

where

$$\boldsymbol{a_1 = 1 + 2^{-1} + 3^{-2} + \ldots\ldots + (n-1)^{-1}}$$

I first obtained a count of the number of segregating sites (*S*), using DnaSP v5 (Librado and Rozas 2009), and then calculated *Ps* to obtain an estimate of θ. I also used DnaSP to calculate the average number of nucleotide differences between sequences, *k*, which can be divided by sequence length in order to obtain an estimate of nucleotide diversity, π. If there are many deleterious alleles in the population, we would expect that Ps/a_1 is exaggerated, whereas π is not affected, because nucleotide diversity is mainly determined by high-frequency variants. By contrast, if there is balancing selection that

increases allele frequencies, we expect that π is enhanced, whereas Ps/a_1 is not. Therefore, by testing the difference between Ps/a_1 and π, we may be able to detect the presence of deleterious mutations or balancing selection.

Tajima's *D* statistic: Tajima's *D* statistic is based on the infinite-site model without recombination, and is appropriate for short DNA sequences (Excoffier and Lischer 2010). It compares estimates of θ and nucleotide diversity (π). The *D* statistic is calculated by

$$D = (n(\pi) - S/a_1) / (e_1 S + e_2 S(S\text{-}1))^{1/2}$$

where

$$e_1 = c_1/a_1$$

$$e_2 = c_2/(a_1^2 + a_2)$$

$$c_1 = b_1 - (1/a_1)$$

$$c_2 = b_2 - (n + 2)/(a_1 n) + a_2/a_1^2$$

$$b_1 = (n + 1) / 3(n - 1)$$

$$b_2 = 2(n^2 + n + 3) / 9n(n - 1)$$

$$a_1 = 1 + 2^{-1} + 3^{-2} + \ldots\ldots + (n - 1)^{-1}$$

$$a_2 = 1 + 2^{-2} + 3^{-2} + \ldots\ldots + (n - 1)^{-2}$$

When $D > 0$, this suggests balancing selection, when $D < 0$, this suggests purifying selection or positive (directional) selection.

Fu's *Fs* statistic: Fu (1997) described the *Fs* statistic, which is also based on the infinite-site model without recombination and appropriate for short DNA sequences (Excoffier and Lischer 2010). Here, the probability of observing *n* alleles from a random neutral sample is calculated, given the mean number of nucleotide differences between

two sequences. The higher the probability, the more likely it is that the number of sampled alleles is consistent with the neutral hypothesis. This probability is denoted by *S'*, which can give a good approximation if there are too many alleles than predicted by the neutral theory. The *S'* statistic can be used to calculate *Fs*, which gives a negative value when there is an abundance of rare alleles, and is usually interpreted as evidence against the neutrality of mutations. *Fs* is calculated by

$$\boldsymbol{Fs = \ln (S'/(1 - S'))}$$

The calculation of *S'* follows relatively complicated statistics, and I will refer to reader to see Fu (1997) for details. Here, I will obtain a measure of *Fs* and, from that, determine *S'*.

Significance of both Tajima's *D* and Fu's *Fs* is determined by computer simulation in which a set of allele sequences is generated according to the coalescent theory (Kingman 1982; Hudson 1983; Tajima 1983). I used Arlequin v 3.5 (Excoffier and Lischer 2010b) for both tests of selective neutrality, running 10,000 coalescent simulations to assess significance.

Results

IGF1 Locus

The *Ochotona IGF1* locus is composed of 4 exons and spans a genomic region of 65,675 nucleotides. I obtained two sequences of the *IGF1* locus, together totaling 881 nucleotides and containing exons 2 and 3. There is a gap >40,000 nucleotides separating these two sequences (Figure 3.2 Top).

The coding sequence of *IGF1* is 465 nucleotides which equates to 155 codons. Exons 2 and 3 were concatenated, resulting in 342 nucleotides, or 114 codons. This is approximately 74% of the gene (Figure 3.2 Bottom).

Evolutionary change of amino acid sequences

The number of amino acid differences per site between *O. princeps* and *Or. cuniculus* was calculated to be 0.04511. Therefore, the rate of amino acid substitution (Table 3.1) can be calculated by

$$\mathbf{r = -\ln(1 - 0.04511) / (2*37\ MYA)}$$

$$\mathbf{r = 0.04616 / (2*37\ MYA) = 0.6238 \times 10^{-9}}$$

Evolutionary change of DNA sequences

The number of synonymous differences per synonymous site between *O. princeps* and *Or. cuniculus* was calculated to be 0.1487. Therefore, the rate of synonymous substitution per synonymous site after Jukes and Cantor correction is equal to

$$\mathbf{r = -(3/4)\ln(1 - (4/3)\ 0.1487) / (2*37\ MYA)}$$

$$\mathbf{r = 0.1657 / (2*37\ MYA) = 2.239 \times 10^{-9}}$$

The rate of nonsynonymous substitution per synonymous site can be calculated in the same way. The number of nonsynonymous differences per synonymous site between *O. princeps* and *Or. cuniculus* was calculated to be 0.02592. Therefore, after Jukes and Cantor correction, the rate is equal to 0.3565×10^{-9} (Table 3.1).

When comparing the *IGF1* locus between *O. princeps* and *Or. cuniculus*, the proportion of transitional differences was equal to 0.02417 while that of transversions was equal to 0.03235. Thus, the ratio of transitions to transversions is 0.7471. The pattern of nucleotide substitution at the first, second, and third codon positions indicated the

greatest number of differences at the third position and the fewest at the first position (rather than the second position). The p-distance was 0.1154 at the third position, 0.03629 at the second position, and 0.01754 at the first position (Table 3.2).

Phylogenetics

For phylogenetic reconstruction of the *IGF1* gene, I used the HKY+G nucleotide substitution model for the genomic dataset while, for the mRNA dataset, I used the HKY model. Results from the genomic dataset show very limited structure, with only two clades evident: three SRM individuals grouped together while three CU individuals grouped together (Figure 3.3). Similarly, the coding dataset showed extremely low diversity within *Ochotona* (Figure 3.4).

Nucleotide diversity for *O. princeps* was calculated to be 0.0027, while the average within clade diversity was 0.00198. The interpopulational diversity (N_{ST}) at the *IGF1* locus (Table 3.3) is thus

$$\boldsymbol{N_{ST}} \textbf{ = (0.0027 – 0.00198) / 0.0027 = 0.2711}$$

The average number of nucleotide substitutions per site (*Dxy*) between *O. princeps* and *O. collaris* is equal to 0.005997. After accounting for intraspecific diversity within *O. princeps* (0.002703) and *O. collaris* (0.000), the net nucleotide substitutions per site (*Da*) was estimated to be 0.004646. Using the rate for both synonymous and nonsynonymous substitutions calculated for the *IGF1* locus, the estimated time of divergence (Table 3.4) between *O. princeps* and *O. collaris* is equal to

T = (((0.004645502/(2 * 2.23921 x 10-9)) + ((0.004645502/(2 * 0.356509 x 10-9))) / 2 = 3,780,000 MYA

The rate of nonsynonymous substitutions between *O. princeps* and *O. collaris* was calculated to be 0.0001553398 while the rate of synonymous substitutions was 0.0351351414. Therefore, the rate of nonsynonymous substitution/synonymous substitution is equal to 0.0044212089 (Table 3.5).

Homozygosity test for protein polymorphism

Of 44 sequences from the *IGF1* locus, there are two alleles when considering the primary protein structure (Table 3.6), with the most common allele having a haplotype frequencies of about 95.5%. The observed homozygosity is equal to

$$F = \text{sum of all } ni^2 / 44^2 = 0.91322$$

The expected homozygosity value, calculated in Arlequin, is approximately 0.77559. Of the 10,000 coalescent simulations, 76.55% of random neutral samples show homozygosity values identical to or smaller than the one observed (Table 3.7).

Statistical tests of neutrality

I identified six segregating sites resulting in seven haplotypes in 44 sequences after analyzing 342 sites. Therefore, the number of segregating sites per nucleotide site (*Ps*) is 0.017544. With 44 sequences, a_1 is equal to 4.353315049. Theta is thus estimated to be

$$\theta = 0.017544 / 4.353315049 = 0.00403$$

The average number of nucleotide differences, *k*, was equal to 0.760 after analysis of 342 sites. Therefore, nucleotide diversity is equal to

$$\pi = 0.760/342 = 0.00222$$

Tajima's *D*: From the above we see that Tajima's *D* can be calculated by

$$\frac{342(0.00222) - 6/4.353}{(e_1 S + e_2 S(S\text{-}1))^{1/2}}$$

$$\frac{\textbf{-0.619119752}}{(e_1 S + e_2 S(S\text{-}1))^{1/2}}$$

The variables for resolving this equation are listed in Table 3.6. The *D* statistic is equal to -1.17367. The *P*-value was reported to be 0.11 after running 10,000 coalescent simulations. Because the *P*-value was estimated by simulation, it can be interpreted that, of the 10,000 simulations, 1,100 simulations generated a *D* value equal to or more extreme than the one observed.

Fu's *Fs*: The *Fs* statistic was calculated to be -3.16082. Therefore,

$$S' = 0.042390966 / 1.042390966 = 0.040667051$$

This is the probability of observing seven alleles in a sample size of 46, with an average of 0.986 nucleotide differences between two sequences, in a random neural sample. The expected number of alleles, calculated in Arlequin, was 3.68674. The *Fs P*-value was reported to be 0.0557, meaning 557 computer simulations generated an *Fs* value equal to or more extreme than the one observed.

EHHADH Locus

The *Ochotona EHHADH* locus is composed of seven exons and spans a genomic region of 24,552 nucleotides. I sequenced two regions of the *EHHADH* locus totaling 1,335 nucleotides which included exon 6 and a portion of exon 7. There is a 1,232 nucleotide gap separating these two sequences (Figure 3.5 Top).

The coding sequence of *EHHADH* is 2,175 nucleotides which equates to 725 codons. Exon 6 and the portion of 7 were concatenated, resutling in 1,015 nucleotides, or 338 codons. This is approximately 47% of the gene (Figure 3.5 Bottom).

Evolutionary change of amino acid sequences

The number of amino acid differences per site between *O. princeps* and *Or. cuniculus* was calculated to be 0.1689. Therefore, the rate of amino acid substitution (Table 3.1) can be calculated by

$$r = -\ln(1 - 0.1689) / (2*37\text{MYA})$$

$$r = 0.18496 / (2*37\text{MYA}) = 2.499 \times 10^{-9}$$

Evolutionary change of DNA sequences

The number of synonymous differences per synonymous site between *O. princeps* and *Or. cuniculus* was calculated to be 0.2636. Therefore, the rate of synonymous substitution per synonymous site after Jukes and Cantor correction is equal to

$$r = -(3/4)\ln(1 - (4/3)\ 0.2636) / (2*37\ \text{MYA})$$

$$r = 0.3248 / (2*37\ \text{MYA}) = 4.389 \times 10^{-9}$$

The rate of nonsynonymous substitution per synonymous site can be calculated in the same way. The number of nonsynonymous differences per synonymous site between *O. princeps* and *Or. cuniculus* was calculated to be 0.08611. Therefore, after Jukes and Cantor correction, the rate is equal to 1.23598×10^{-9} (Table 3.1).

When comparing the *EHHADH* locus between *O. princeps* and *Or. cuniculus*, the proportion of transitional differences was equal to 0.088596 while that of transversions was equal to 0.039483. Thus, the ratio of transitions to transversions is 2.2439. The pattern of nucleotide substitution at the first, second, and third codon positions indicated the greatest number of differences at the third position and the fewest at the second position. The p-distance was 0.20096 at the third position, 0.0825444 at the second position, and 0.1008 at the first position (Table 3.2).

Phylogenetics

For phylogenetic reconstruction of the *EHHADH* gene, I used the HKY+G nucleotide substitution model for both the genomic and mRNA datasets. The topology produced from genomic data showed evidence of substructing within *O. princeps* corresponding to recognized subspecies. That is, there are three clades, including a SN clade (five individuals), a CR clade (five individuals), a SRM clade (2 individuals), and a CU clade (five individuals) (Figure 3.6). The topology produced from mRNA data showed *O. collaris* as basal and paraphyletic, with *O. princeps* being nested within *O. collaris* diversity (Figure 3.7).

Nucleotide diversity for *O. princeps* was calculated to be 0.0026, while the average within clade diversity was 0.00108. The interpopulational diversity (N_{ST}) at the *EHHADH* locus (Table 3.3) is thus

$$\boldsymbol{N_{ST} = (0.0026 - 0.00108) / 0.0026 = 0.58}$$

The average number of nucleotide substitutions per site (*Dxy*) between *O. princeps* and *O. collaris* is equal to 0.0076267274. After accounting for intraspecific diversity within *O. princeps* (0.0025617976) and *O. collaris* (0.0012699), the net nucleotide substitutions per site (*Da*) is estimated to be 0.005710879. Using the rate for both synonymous and nonsynonymous substitution calculated for the *EHHADH* locus, the estimated time of divergence (Table 3.4) between *O. princeps* and *O. collaris* is equal to

$$\mathbf{T = (((0.005710879/(2 * 4.39E\text{-}09)) + ((0.005710879/(2 * 1.24E\text{-}09))) / 2 = 1{,}480{,}000\ MYA}$$

The rate of nonsynonymous substitutions between *O. princeps* and *O. collaris* was calculated to be 0.0053942624 while the rate of synonymous substitutions was 0.0070509022. Therefore, the rate of nonsynonymous substitution/synonymous substitution is equal to 0.7650457041 (Table 3.5).

Homozygosity test for protein polymorphism

Of 40 sequences from the *EHHADH* locus, there are six alleles when considering primary protein structure (Table 3.6), with the most common allele having a haplotype frequencies of about 72.5%. The observed homozygosity is equal to

$$F = \text{sum of all } ni^2 / 40^2 = 0.54375$$

The expected homozygosity value, calculated in Arlequin, is approximately 0.35741. Of the 10,000 coalescent simulations, 91.70% of random neutral samples show homozygosity values identical to or smaller than the one observed (Table 3.7).

Statistical tests of neutrality

I identified 12 segregating sites resulting in 10 haplotypes in 40 sequences after analyzing 1,015 sites. Therefore, the number of segregating sites per nucleotide site (*Ps*) is 0.011823. With 40 sequences, a_1 is equal to 4.394948116. Theta is thus estimated to be

$$\theta = 0.011823 / 4.394948116 = 0.00278$$

The average number of nucleotide differences, *k*, was equal to 1.405 after analysis 1,015 sites. Therefore, nucleotide diversity is equal to

$$\pi = 1.405 / 1015 = 0.00138$$

Tajima's D: From the above we see that Tajima's *D* is equal to

$$\frac{1015\,(0.00138) - 12/4.394948116}{(e_1 S + e_2 S(S\text{-}1))^{1/2}}$$

$$\frac{\textbf{-1.32970766}}{(e_1 S + e_2 S(S\text{-}1))^{1/2}}$$

The variables for resolving this equation are listed in Table 3.6. The *D* statistic is equal to -1.54591. The *P*-value was reported to be 0.0388 after running 10,000 coalescent simulations. Because the *P*-value was estimated by simulation, it can be interpreted that of the 10,000 simulations, 388 simulations generated a *D* value equal to or more extreme than the one observed.

Fu's *Fs*: The *Fs* statistic was calculated to be -4.10219. Therefore,

$$S' = 0.01654 / 1.01654 = 0.01627$$

This is the probability of observing 10 alleles in a sample size of 40, with an average of 1.405 nucleotide differences between two sequences, in a random neural sample. The expected number of alleles, calculated in Arlequin, is 5.29371. The *Fs P*-value was reported to be 0.014, meaning 140 computer simulations generated an *Fs* value equal to or more extreme than the one observed.

NOS2 Locus

The *Ochotona NOS2* locus is composed of 26 exons and spans a genomic region of 10,980 nucleotides. I targeted a 2,322 bp region of this locus and obtained two raw sequences from three primer pairs. These two sequences are a combined length of 1,887 nucleotides and include exons 14, 15, 16, and 17. There is a gap of 9,860 nucleotides separating the two sequenced regions (Figure 3.8 Top).

The coding sequence of *NOS2* is 3,354 nucleotides which equates to 1,118 codons. Exons 14, 15, 16, and 17 were concatenated, resulting in 456 nucleotides, or 152 codons. This equates to approximately 14% of the gene (Figure 3.8 Bottom)

Evolutionary change of amino acid sequences

The number of amino acid differences per site between *Ochotona* and *Or. cuniculus* was calculated to be 0.1163. Therefore, the rate of amino acid substitution (Table 3.1) can be calculated by

$$\mathbf{r = -\ln(1 - 0.1163) / (2*37\ MYA)}$$

$$\mathbf{r = 0.1236 / (2*37\ MYA) = 1.6703 \times 10^{-9}}$$

Evolutionary change of DNA sequences

The number of synonymous differences per synonymous site between *Ochotona* and *Or. cuniculus* was calculated to be 0.2963 after analysis of 152 codons. Therefore, the rate of synonymous substitution per synonymous site after Jukes and Cantor correction is equal to

$$\mathbf{r = -(3/4)\ln(1 – (4/3)\ 0.2963) / (2*37\ MYA)}$$

$$\mathbf{r = 0.3769 / (2*37\ MYA) = 5.09377 \times 10^{-9}}$$

The rate of nonsynonymous substitution per synonymous site can be calculated in the same way. The number of nonsynonymous differences per synonymous site between *Ochotona* and *Or. cuniculus* was calculated to be 0.05837. Therefore, after Jukes and Cantor correction, the rate is equal to 0.82111×10^{-9} (Table 3.1).

When comparing 456 sites of the *NOS2* locus between *Ochotona* and *Or. cuniculus*, the proportion of transitional differences was equal to 0.07664 while that of transversions was equal to 0.03766. Thus, the ratio of transitions to transversions is 2.03478. The pattern of nucleotide substitution at the first, second, and third codon positions indicated the greatest number of differences at the third position and the fewest at the second position. The p-distance was 0.2007 at the third position, 0.06567 at the second position, and 0.07641 at the first position (Table 3.2).

Phylogenetics

For phylogenetic reconstruction of the *NOS2* gene, I used the GTR+G nucleotide substitution model for the genomic dataset, and the GTR+I model for the mRNA dataset. Results from both genomic and mRNA data show *O. collaris* nested within *O. princeps* diversity. An unrooted phylogeny is shown for the genomic dataset, while a single sequence from *O. curzoniae* was used to root the mRNA phylogeny. Interestingly, O. *collaris* is grouped within Northern Rocky Mountain diversity (Figures 3.8 and 3.9).

Nucleotide diversity for the entire *O. princeps* population was calculated to be 0.0141, while the average within clade diversity was 0.00558. The interpopulational diversity (N_{ST}) at the *NOS2* locus (Table 3.3) is thus

$$\boldsymbol{N_{ST}} \textbf{ = (0.0141 – 0.00558) / 0.0141 = 0.6050}$$

The average number of nucleotide substitutions per site (*Dxy*) between *O. princeps* and *O. collaris* is equal to 0.0202472872. After accounting for intraspecific diversity within *O. princeps* (0.014099655) and *O. collaris* (0.006275144), the net nucleotide substitutions per site (*Da*) is estimated to be 0.010059888. Using the rate for both synonymous and nonsynonymous substitution calculated for the *NOS2* locus, the estimated time of divergence (Table 3.4) between *O. princeps* and *O. collaris* is equal to

T = (((0.010059888/(2 * 5.09377 x 10-9)) + ((0.010059888/(2 * 0.821107))) / 2 =

3,560,000 MYA

The rate of nonsynonymous substitutions between *O. princeps* and *O. collaris* was calculated to be 0.0123811372 while the rate of synonymous substitutions was 0.0175708681. Therefore, the rate of nonsynonymous substitution/synonymous substitution is equal to 0.7046400391 (Table 3.5).

Homozygosity test for protein polymorphism

Of 34 sequences from the *NOS2* locus, there are seven alleles when considering primary protein structure (Table 3.6), with two allelic variants having a haplotype frequencies of about 35.3% each. The observed homozygosity is equal to

$$F = \text{sum of all } ni^2 / 34^2 = 0.26644$$

The expected homozygosity value, calculated in Arlequin, is approximately 0.29307. Of the 10,000 coalescent simulations, 47% of random neutral samples show homozygosity values identical to or smaller than the one observed (Table 3.7).

Statistical tests of neutrality

I identified 12 segregating sites resulting in 11 haplotypes in 34 sequences after analyzing 433 sites. Therefore, the number of segregating sites per nucleotide site (*Ps*) is 0.027714. With 34 sequences, *a1* is equal to 4.088798226. Theta is thus estimated to be

$$\theta = 0.027714 / 4.088798226 = 0.00678$$

The average number of nucleotide differences, *k*, was equal to 3.342 after analysis 433 sites. Therefore, nucleotide diversity is equal to

$$\pi = 3.342/433 = 0.00772$$

Tajima's *D*: From the above we see that Tajima's *D* is equal to

$$\frac{433\ (0.00772) - 12/(4.088798226)}{(e_1 S + e_2 S(S\text{-}1))^{1/2}}$$

$$\frac{0.407912317}{(e_1 S + e_2 S(S\text{-}1))^{1/2}}$$

The variables for resolving this equation are listed in Table 3.6. The *D* statistic is equal to 0.4413. The *P*-value was reported to be 0.273 after running 10,000 coalescent

simulations, meaning that 2,730 of the 10,000 simulations generated a *D* value equal to or more extreme than the one observed.

Fu's *Fs:* The *Fs* statistic was calculated to be -1.4689. Therefore,

$$S' = 0.230178542 / 1.230178542 = 0.18711$$

This is the probability of observing 11 alleles in a sample size of 34, with a mean of 3.342 nucleotide differences between two sequences, in a random neural sample. The expected number of alleles (calculated in Arlequin) is 8.54622. The *Fs P*-value was reported to be 0.307.

TGFB1 Locus

The *Ochotona TGFB1* locus is composed of seven exons and spans a genomic region of 13,650 nucleotides. I sequenced three regions of the *TGFB1* locus totaling 1,468 nucleotides which include exons 1, 2 and 3. There is a 3,387 nucleotide gap separating sequence one and sequence two while there is a 1,299 nucleotide gap separating sequence two and sequence three (Figure 3.11 Top).

The coding sequence of *TGFB1* is 1,176 nucleotides which equates to 392 codons. Exons 1, 2 and 3 were concatenated, resulting in 639 nucleotides, or 213 codons. This is approximately 54% of the gene (Figure 3.11 Bottom).

Evolutionary change of amino acid sequences

The number of amino acid differences per site between *Ochotona* and *Or. cuniculus* was calculated to be 0.06651. Therefore, the rate of amino acid substitution (Table 3.1) can be calculated by

$$r = -\ln(1 - 0.0665094340) / (2*37 \text{ MYA})$$

$$r = 0.068824422 / (2*37 \text{ MYA}) = 0.9301 \times 10^{-9}$$

Evolutionary change of DNA sequences

The number of synonymous differences per synonymous site between *Ochotona* and *Or. cuniculus* was calculated to be 0.25481. Therefore, the rate of synonymous substitution per synonymous site after Jukes and Cantor correction is equal to

$$\mathbf{r = -(3/4)\ln(1 - (4/3)\ 0.2548075601) / (2*37\ MYA)}$$

$$\mathbf{r = 0.311345064 / (2*37\ MYA) = 4.2074 \times 10^{-9}}$$

The rate of nonsynonymous substitution per synonymous site can be calculated in the same way. The number of nonsynonymous differences per synonymous site between Ochotona and *Or. cuniculus* was calculated to be 0.0428660345. Therefore, after Jukes and Cantor correction, the rate is equal to 0.5965×10^{-9}.

When comparing the *TGFB1* locus between *O. princeps* and *Or. cuniculus*, the proportion of transitional differences was equal to 0.0485579937 while that of transversions was equal to 0.0509404389. Thus, the ratio of transitions to transversions is 0.953231. The pattern of nucleotide substitution at the first, second, and third codon positions indicated the greatest number of differences at the third position and the fewest at the second position. The p-distance was 0.1942452830 at the third position, 0.0330516432 at the second position, and 0.0716431925 at the first position (Table 3.2).

Phylogenetics

For phylogenetic reconstruction of the *TGFB1* gene, I used the GTR+I nucleotide substitution model for genomic dataset, and the HKY model for the mRNA dataset. The genomic dataset showed evidence of substructuring within *O. princeps*. That is, there are three monophyletic clades present within *O. princeps* diversity: a CR clade (five individuals), a SN clade (four individuals), a SRM clade (three individuals), and a CU

clade (three individuals) (Figure 3.12). The mRNA dataset shows a split between three CU individuals and all other individuals, including *O. collaris* (Figure 3.13).

Nucleotide diversity for *O. princeps* was calculated to be 0.0086, while the average within clade diversity was 0.00428. The interpopulational diversity (N_{ST}) at the *TGFB1* locus (Table 3.3) is thus

$$\boldsymbol{N_{ST} = (0.0086 - 0.00428) / 0.0086 = 0.5026}$$

The average number of nucleotide substitutions per site (*Dxy*) between *O. princeps* and *O. collaris* is equal to 0.0131427401. After accounting for intraspecific diversity within *O. princeps* (0.008590909) and *O. collaris* (0.002061323), the net nucleotide substitutions per site (*Da*) is estimated to be 0.007816624. Using the rate for both synonymous and nonsynonymous substitution calculated for the *TGFB1* locus, the estimated time of divergence (Table 3.4) between *O. princeps* and *O. collaris* is equal to

T = (((0.007816624/(2 * 4.20737 x 10-9)) + ((0.007816624/(2 * 0.596484 x 10-9))) / 2 = 3,740,000 MYA

The rate of nonsynonymous substitutions between *O. princeps* and *O. collaris* was calculated to be 0.0029897858 while the rate of synonymous substitutions was 0.0027602906. Therefore, the rate of nonsynonymous substitution/synonymous substitution is equal to 1.083141681 (Table 3.5).

Homozygosity test for protein polymorphism

Of 50 sequences from the *TGFB1* locus, there are four alleles when considering primary protein structure (Table 3.6), with the most common allele having a haplotype frequency of about 90%. The observed homozygosity is equal to

$$\boldsymbol{F = \text{sum of all } ni^2 / 50^2 = 0.8136}$$

The expected homozygosity value, calculated in Arlequin, is approximately 0.52088. Of the 10,000 coalescent simulations, 94.6% of random neutral samples show homozygosity values identical to or smaller than the one observed (Table 3.7).

Statistical tests of neutrality

I identified 10 segregating sites resulting in 10 haplotypes in 50 sequences after analyzing 638 sites. Therefore, the number of segregating sites per nucleotide site (*Ps*) is 0.015674. With 50 sequences, a_1 is equal to 4.479205338. Theta is thus estimated to be

$$\theta = 0.015674/4.479205338 = 0.003499281$$

The average number of nucleotide differences, *k*, was equal 1.376 after analysis 638 sites. Therefore, nucleotide diversity is equal to

$$\pi = 1.376 / 638 = 0.00215674$$

Tajima's *D*: From the above we see that Tajima's *D* can be calculated by

$$\frac{638(0.00215674) - 10/4.479205338}{(e_1 S + e_2 S(S\text{-}1))^{1/2}}$$

$$\frac{-0.856538745}{(e_1 S + e_2 S(S\text{-}1))^{1/2}}$$

The variables for resolving this equation are listed in Table 3.6. The *D* statistic is equal to -1.09662. The *P*-value was reported to be 0.14 after running 10,000 coalescent simulations. Because the *P*-value was estimated by simulation, it can be interpreted that of the 10,000 simulations, 1,400 simulations generated a *D* value equal to or more extreme than the one observed.

Fu's *Fs*: The *Fs* statistic was calculated to be -3.67471. Therefore,

$$S' = 0.025356758 / 1. 0.025356758 = 0.024729693$$

This is the probability of observing 10 alleles in a sample size of 50, with an average of 1.376 nucleotide differences between two sequences, in a random neural

sample. The expected number of alleles, calculated in Arlequin, was 5.52644. The *P*-value was reported to be 0.0738, meaning 738 computer simulations generated an *Fs* value equal to or more extreme than the one observed.

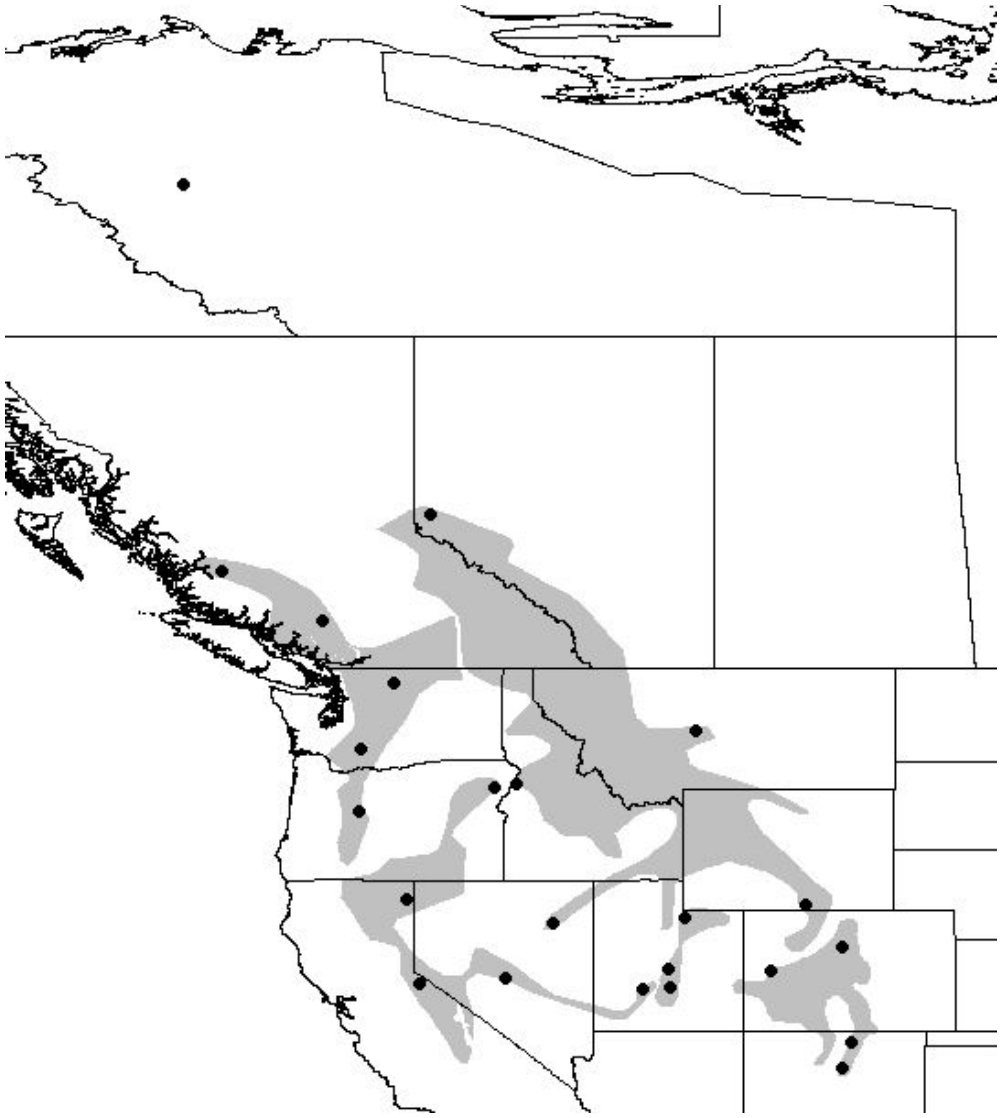

Figure 3.1. *Ochotona princeps* range (gray area) and samples used in this study (marked with dots). The single locality in the Northwest Territories marks the area in which *O. collaris* samples originate.

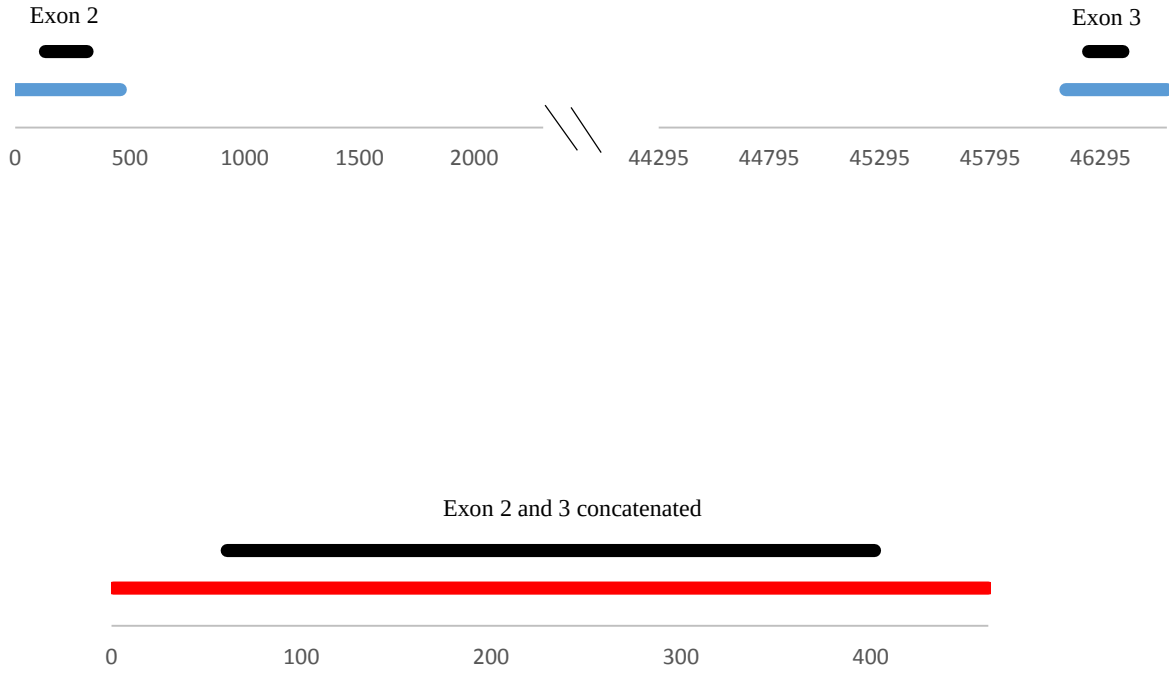

Figure 3.2. (Top) Axis represents the targeted *IGF1* genomic region measured in numbers of nucleotides. Blue lines represent raw sequence data while black lines represent the location of exons within raw sequence. (Bottom) Axis represents the *Ochotona IGF1* coding sequence measured in numbers of nucleotides. Black line represents exons 2 and 3 concatenated.

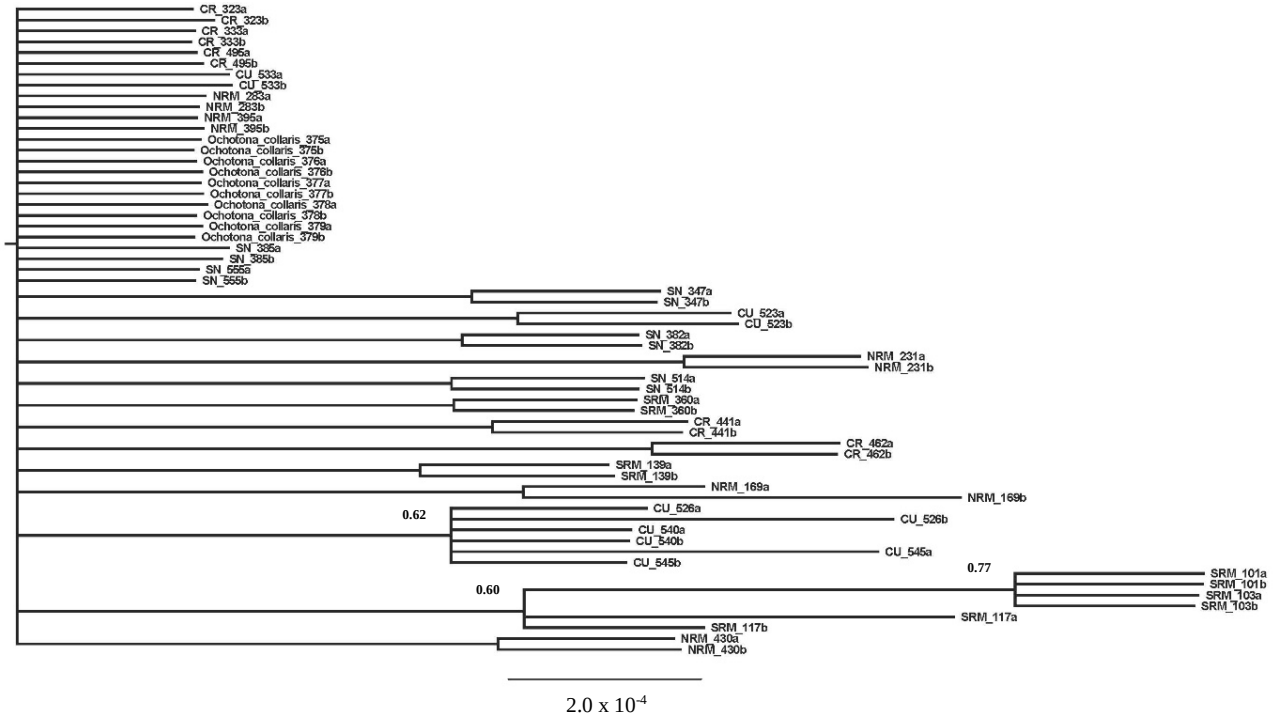

Figure 3.3. Bayesian phylogeny constructed using the *IGF1* genomic dataset.

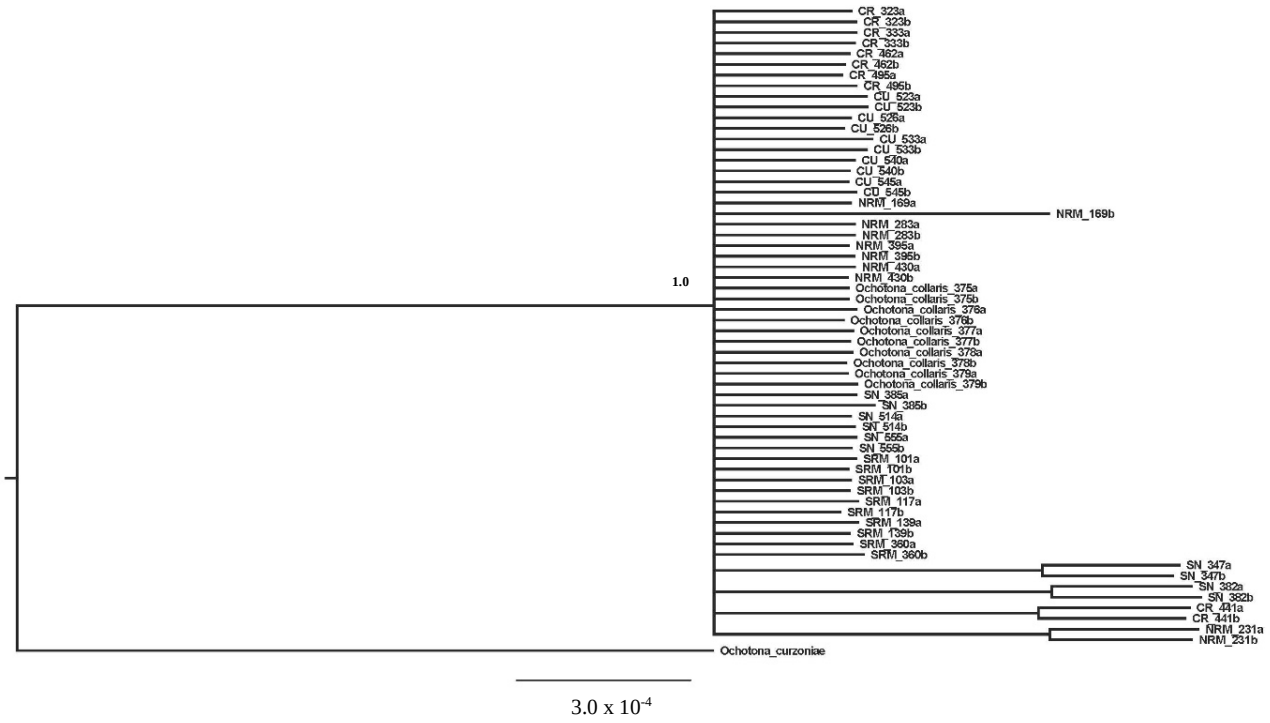

Figure 3.4. Bayesian phylogeny constructed using *IGF1* mRNA dataset.

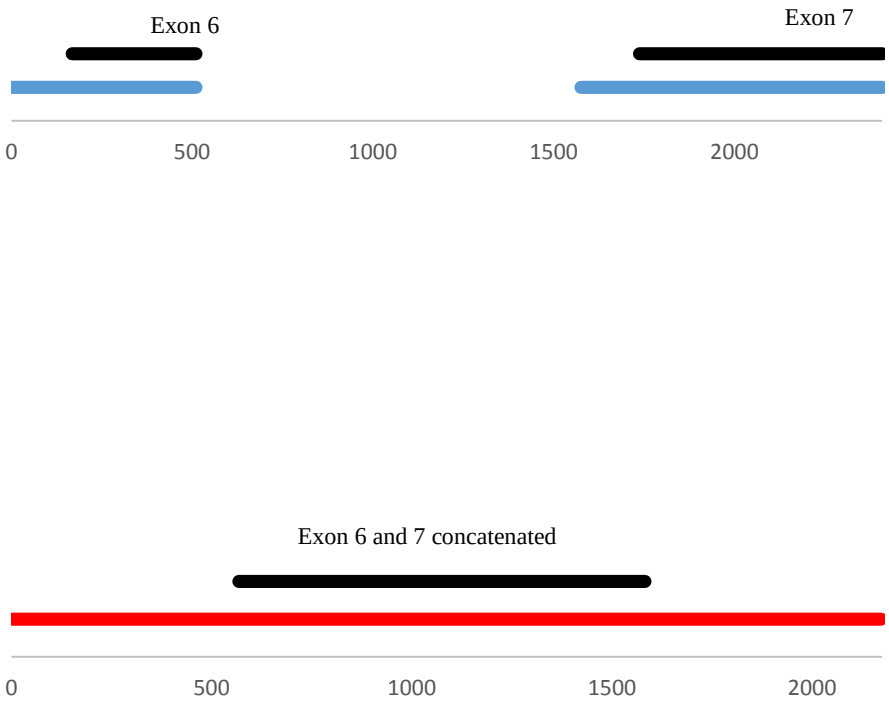

Figure 3.5. (Top) Axis represents the targeted *EHHADH* genomic region measured in numbers of nucleotides. Blue lines represent raw sequence data while black lines represent the location of exons within raw sequence. (Bottom) Axis represents the *Ochotona EHHADH* coding sequence measured in numbers of nucleotides. Black line represents exons 6 and 7 concatenated.

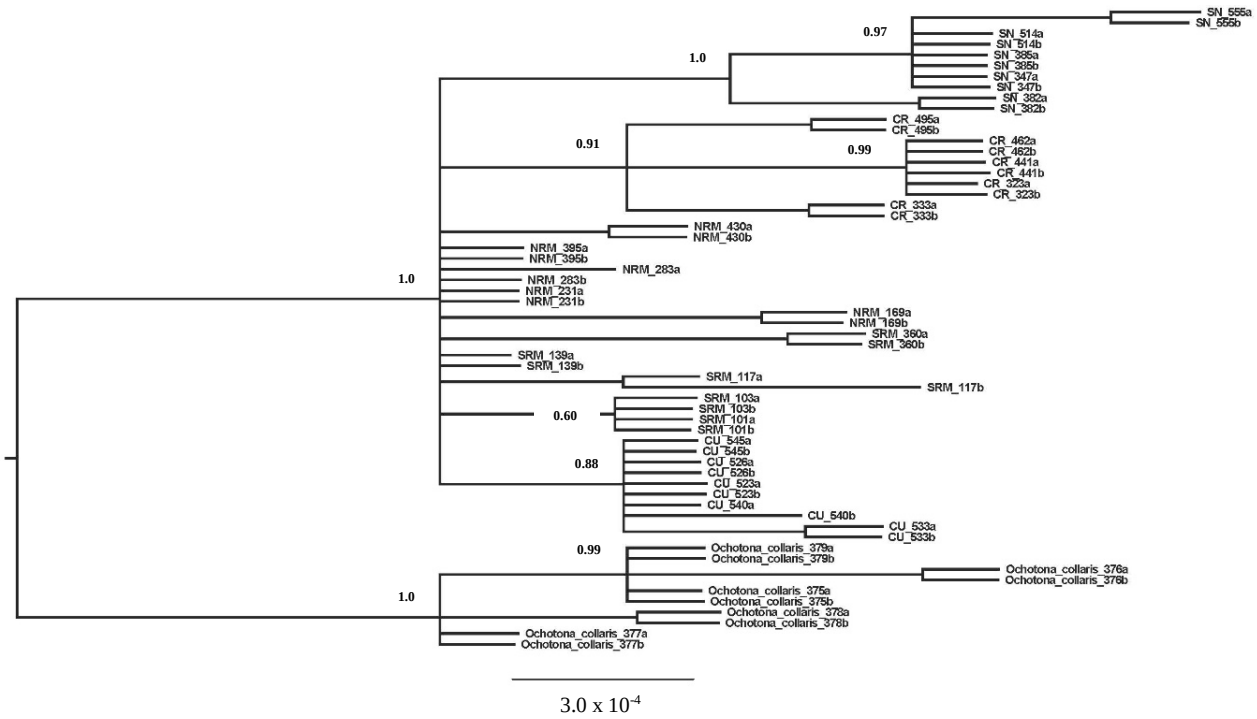

Figure 3.6. Bayesian phylogeny constructed using the *EHHADH* genomic dataset.

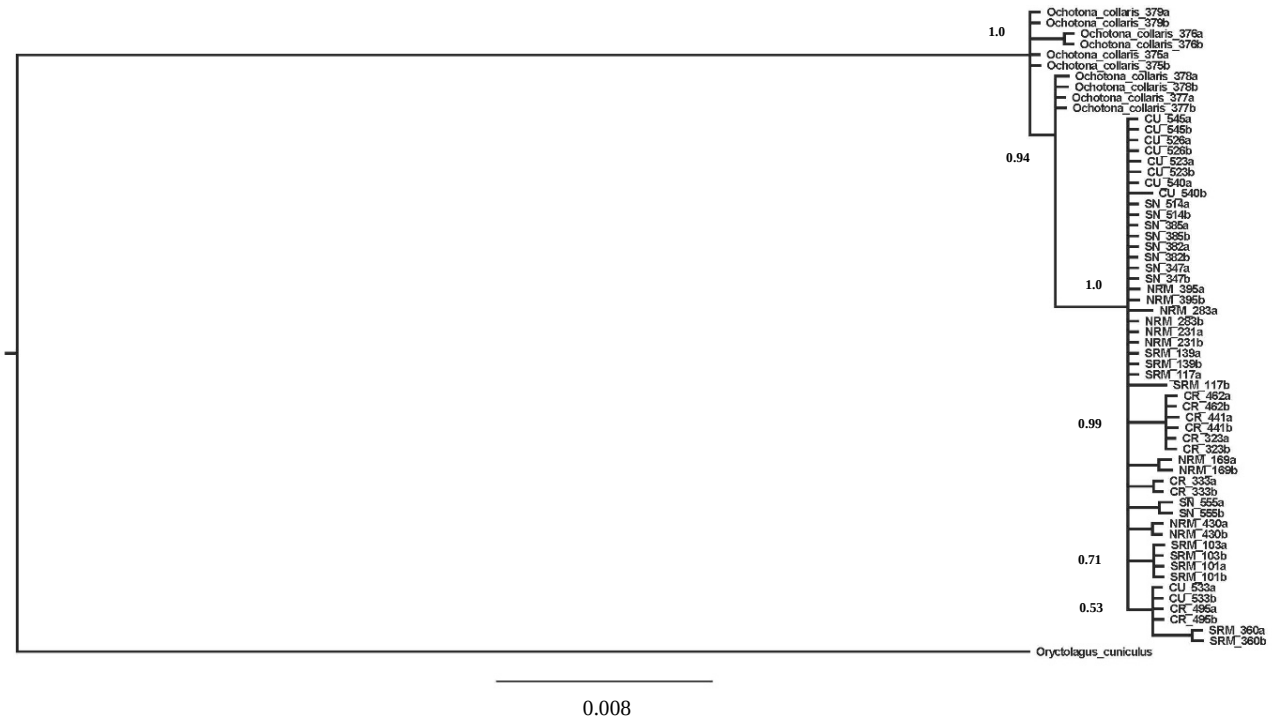

Figure 3.7. Bayesian phylogeny constructed using the *EHHADH* mRNA dataset.

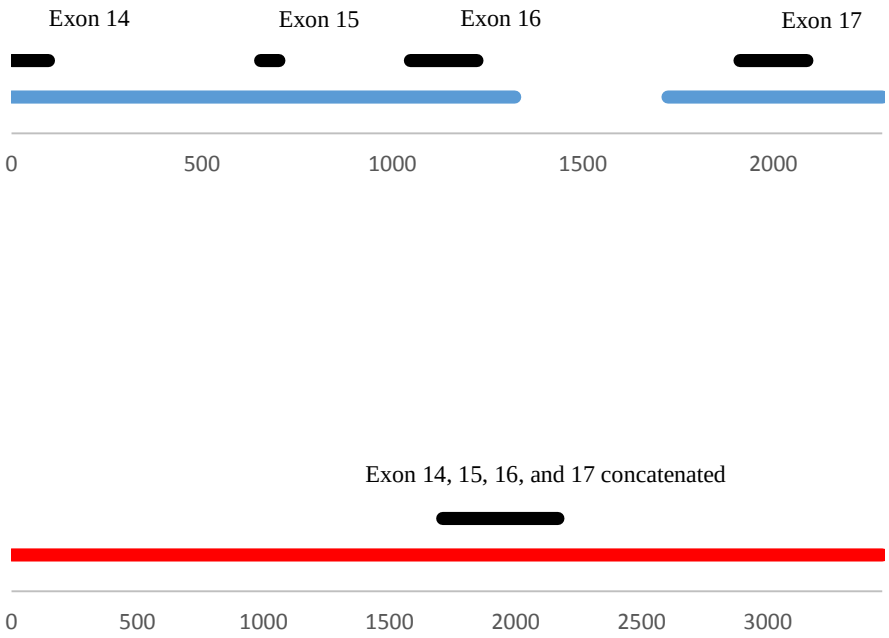

Figure 3.8. (Top) Axis represents the targeted *NOS2* genomic region measured in numbers of nucleotides. Blue lines represent raw sequence data while black lines represent the location of exons within raw sequence. (Bottom) Axis represents the *Ochotona NOS2* coding sequence measured in numbers of nucleotides. Black line represents exons 14, 15, 16 and 17 concatenated.

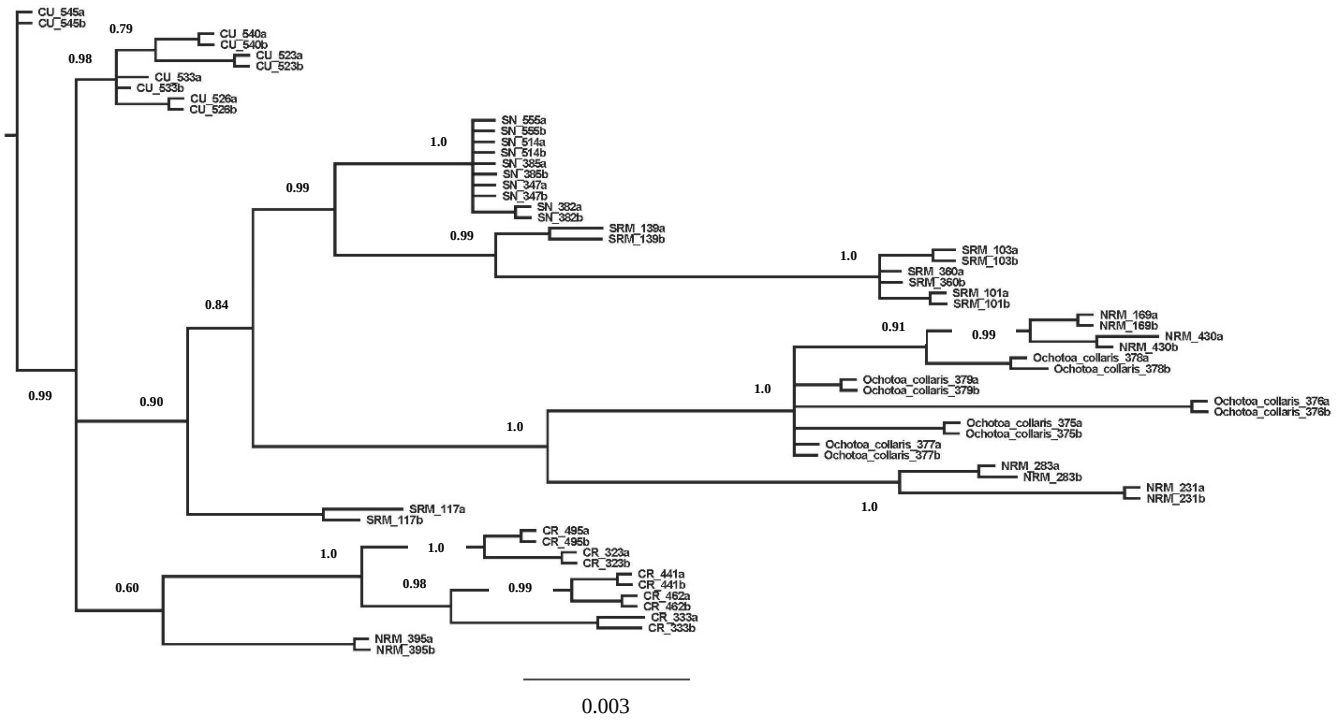

Figure 3.9. Bayesian phylogeny constructed using the *NOS2* genomic dataset.

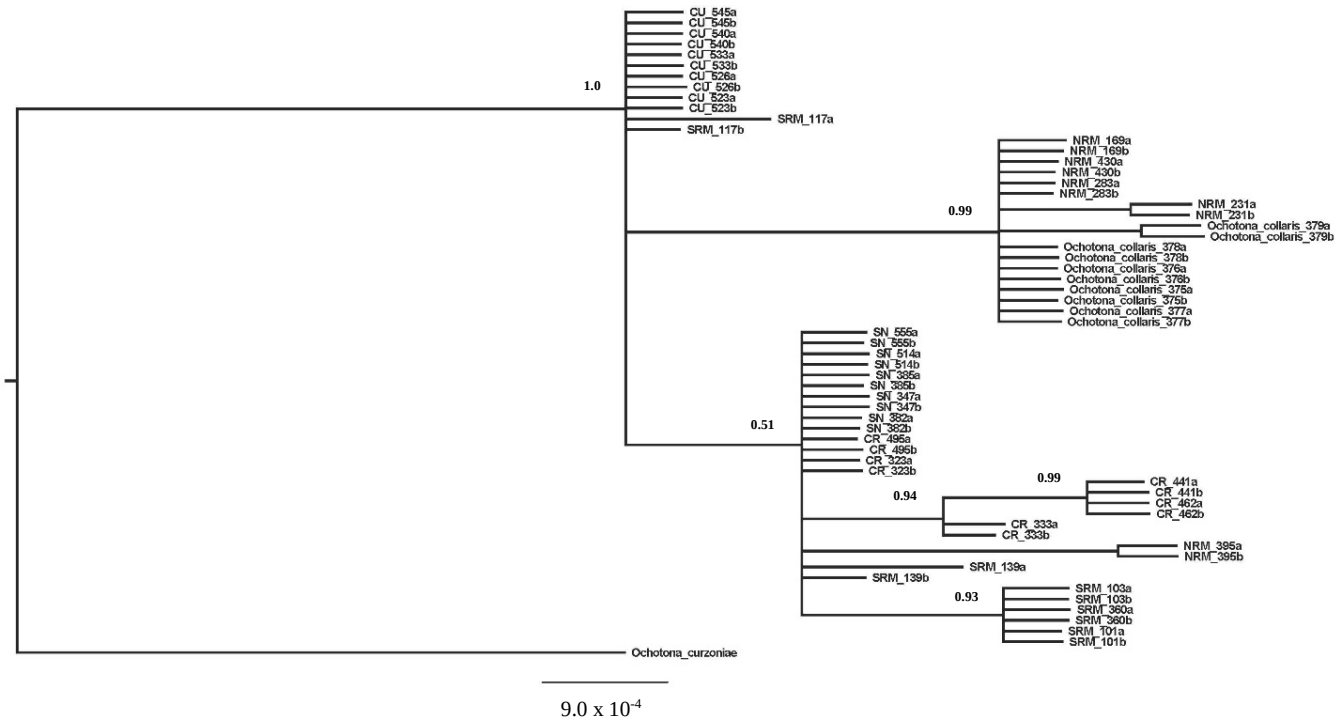

Figure 3.10. Bayesian phylogeny constructed using *NOS2* mRNA dataset.

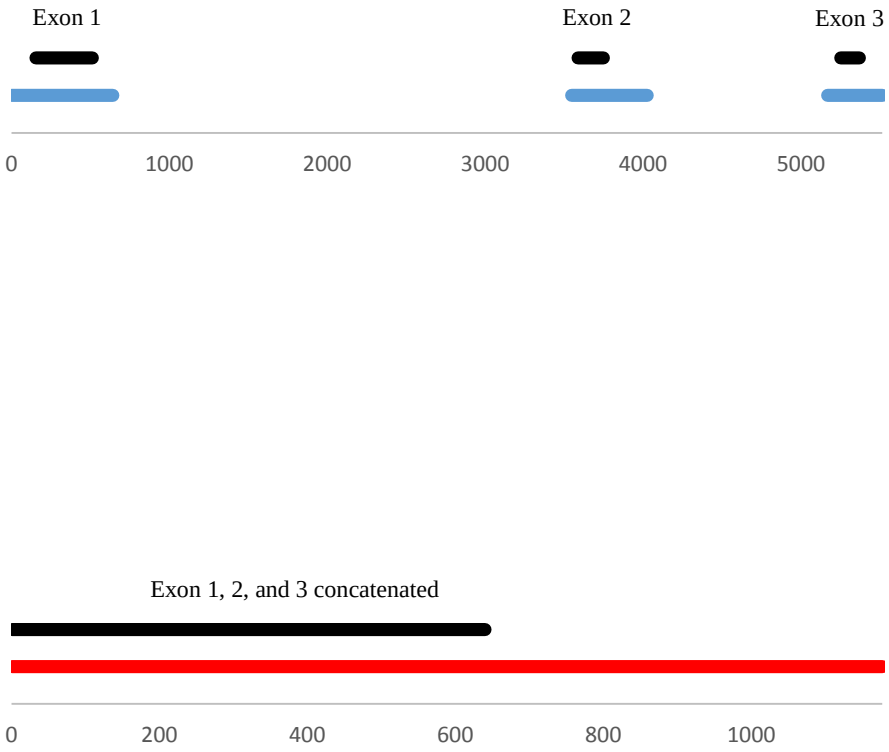

Figure 3.11. (Top) Axis represents the targeted *TGFB1* genomic region measured in numbers of nucleotides. Blue lines represent raw sequence data while black lines represent the location of exons within raw sequence. (Bottom) Axis represents the *Ochotona TGFB1* coding sequence measured in numbers of nucleotides. Black line represents exons 1, 2 and 3 concatenated.

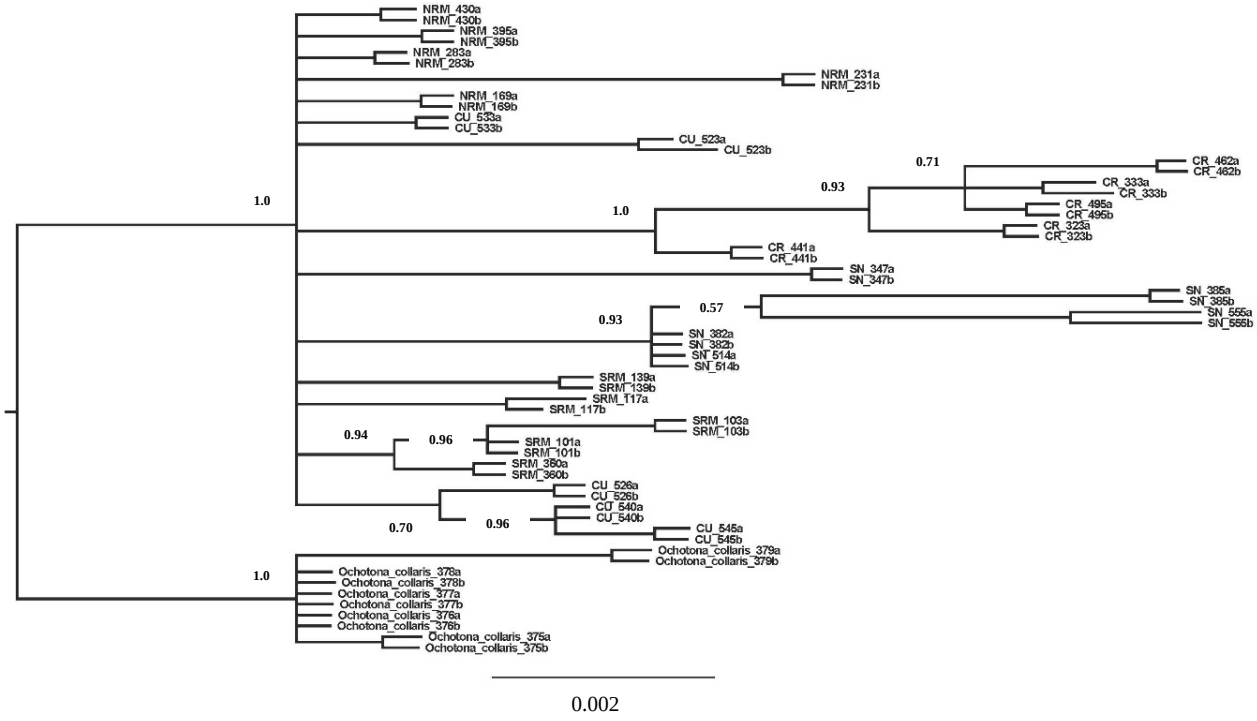

Figure 3.12. Bayesian phylogeny constructed using the *TGFB1* genomic dataset.

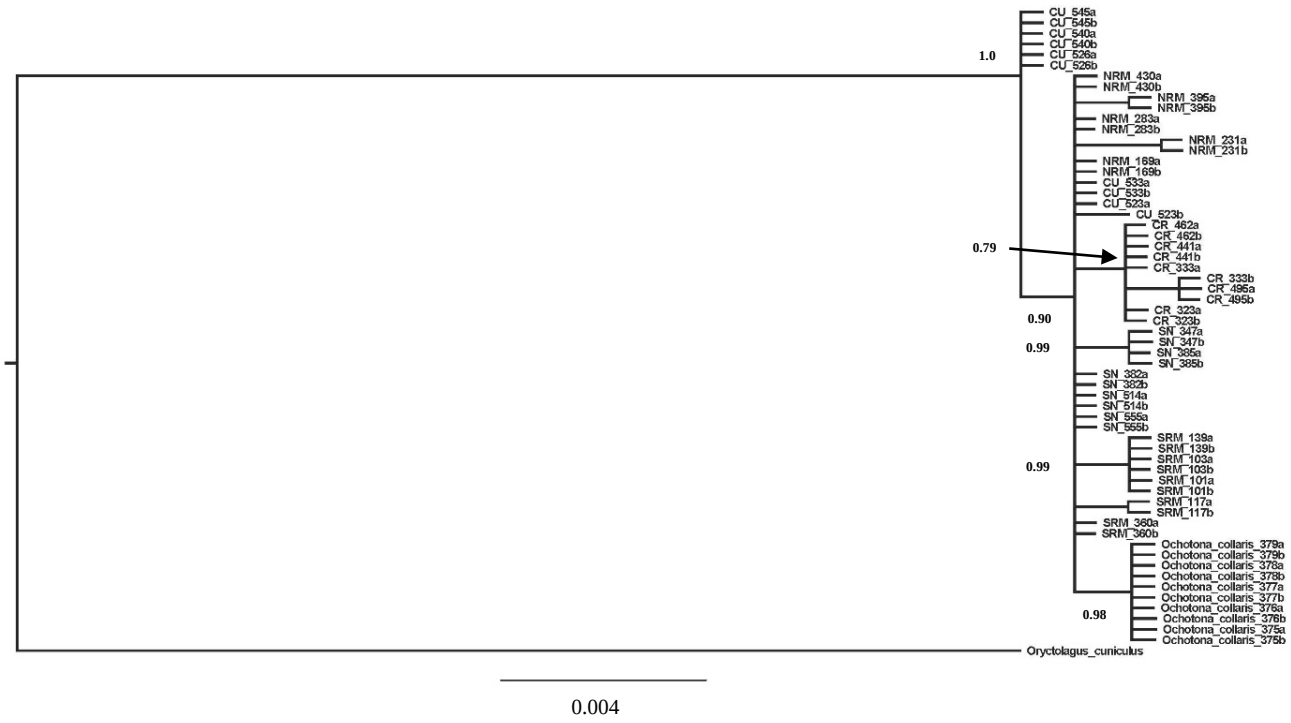

Figure 3.13. Bayesian phylogeny constructed using *TGFB1* mRNA dataset.

Ochotona/Or. cuniculus	Codons	Amino acid	Synonymous	Nonsynonymous
IGF1	113	0.62384×10^{-9}	2.23921×10^{-9}	0.356509×10^{-9}
EHHADH	338	2.49943×10^{-9}	4.38863×10^{-9}	1.23598×10^{-9}
NOS2	152	1.67027×10^{-9}	5.09377×10^{-9}	0.821107×10^{-9}
TGFB1	212	0.93006×10^{-9}	4.20737×10^{-9}	0.596484×10^{-9}

Table 3.1. Amino acid and DNA substitution rates. All rates are based on comparisons between pika and rabbit genes. The time of divergence was set at 37 million years ago. Rates are in units of substitutions per site per 10^9 years.

Ochotona/ *Or. cuniculus*	Nucleotides	p-distance	Transistions	Transversions	Ratio	First codon position	Second codon position	Third codon position
IGF1	342	0.0565	0.0242	0.0324	0.7471	0.0175	0.0363	0.1154
EHHADH	1015	0.1281	0.0886	0.0395	2.2439	0.1008	0.0825	0.2010
NOS2	456	0.1143	0.0766	0.0377	2.0348	0.0764	0.0657	0.2007
TGFB1	638	0.0995	0.0486	0.0509	0.9532	0.0716	0.0331	0.1942

Table 3.2. Evolutionary change of DNA sequences between pika and rabbit

***IGF1* (0.0027)**		***EHHADH* (0.0026)**		***NOS2* (0.0141)**		***TGFB1* (0.0086)**	
CU	0.0013	CU	0.0005	CU	0.0020	CU	0.0039
CR	0.0020	CR	0.0012	CR	0.0052	CR	0.0030
NRM	0.0021	NRM	0.0009	NRM	0.0127	NRM	0.0050
SN	0.0020	SN	0.0010	SN	0.0005	SN	0.0057
SRM	0.0025	SRM	0.0018	SRM	0.0075	SRM	0.0038
N_{ST} **= 0.2711**		N_{ST} **= 0.5775**		N_{ST} **= 0.6050**		N_{ST} **= 0.5026**	

Table 3.3 Degree of relative nucleotide differentiation among lineages of *Ochotona princeps*. The coefficient of nucleotide differentiation (N_{ST}) is the proportion of interpopulational diversity and is equivalent to F_{ST}.

Gene	*Dxy*	*Da*	T
IGF1	0.0059968677	0.004645502	3,780,000
EHHADH	0.0076267274	0.005710879	1,480,000
NOS2	0.0202472872	0.010059888	3,560,000
TGFB1	0.0131427401	0.007816624	3,740,000

Table 3.4. The extent of DNA divergence between *Ochotona princeps* and *Ochotona collaris*. The average number of nucleotide substitutions per site (Dxy). The number of net nucleotide substitutions per site (Da). Divergence time (years) between *O. princeps* and *O. collaris*.

Gene	d_S	d_N	d_N/d_S
IGF1	0.0351351414	0.0001553398	0.0044212089
EHHADH	0.0053942624	0.0070509022	0.7650457041
NOS2	0.0123811372	0.0175708681	0.7046400391
TGFB1	0.0029897858	0.0027602906	1.083141681

Table 3.5 Ratio of synonymous (d_S) and nonsynonymous (d_N) substitutions between pairwise comparisons of *Ochotona princeps* and *Ochotona collaris*.

Gene	n	k	n_1	n_2	n_3	n_4	n_5	n_6	n_7
IGF1	44	2	42	2	-	-	-	-	-
EHHADH	40	6	29	4	2	2	2	1	-
NOS2	34	7	12	12	2	2	2	2	2
TGFB1	50	4	45	2	2	1	-	-	-

Table 3.6. Pika sample data showing number of individuals (*n*), number of alleles observed (*k*), and the numbers of representatives of these alleles in the sample (n_i).

Gene	n	k	observed F	expected F	F <= *obs* F
IGF1	44	2	0.91322	0.77559	0.76550
EHHADH	40	6	0.54375	0.35741	0.91680
NOS2	34	7	0.26644	0.29307	0.46980
TGFB1	50	4	0.8136	0.52088	0.94580

Table 3.7. Pika sample statistics comparing the observed homozygosity (*F*) and expected homozygosity (*F*) using the Ewens-Watterson test homozygosity test for protein polymorphism.

Ochotona	*n*	*S*	*H*	Sites	*Ps*	*θ*	*k*	*π*	*D*	*Fs*
IGF1	44	6	7	342	0.0175	0.00403	0.760	0.00222	-1.1737 P = 0.11	-3.1608 P = 0.0557
EHHADH	40	12	10	1015	0.0118	0.00278	1.405	0.00138	-1.5459 P = 0.0388	-4.1022 P = 0.014
NOS2	34	12	11	433	0.0277	0.00678	3.342	0.00772	0.4413 P = 0.273	-1.4689 P = 0.307
TGFB1	50	10	10	638	0.0157	0.00350	1.376	0.00216	-1.09662 P = 0.14	-3.67471 P = 0.0738

Table 3.8. Molecular diversity statistics. Number of sequences (N), number of segregating sites (S), number of haplotypes (H), number of nucleotide sites analyzed (Sites), proportion of segregating sites (Ps), diversity statistic theta (θ), mean number of nucleotide differences (k), and nucleotide diversity (π), Tajima's *D* (*D*), and Fu's *Fs* (*Fs*).

	IGF1	***EHHADH***	***NOS2***	***TGFB1***
e_1	0.027345314	0.027113684	0.02664957	0.027613132
c_1	0.118952079	0.115329224	0.108964713	0.123684889
b_1	0.348837209	0.35042735	0.353535354	0.346938776
a_1	4.349998621	4.253543039	4.088798226	4.479205338
e_2	0.003810829	0.003891526	0.004032719	0.003705319
c_2	0.078291337	0.07671072	0.073933279	0.080360997
b_2	0.2329105	0.234045584	0.236284413	0.231564626
a_2	1.621946573	1.619618963	1.615085536	1.624732734

Table 3.9. Variables for calculating Tajima's *D* test statistic

Discussion

Evolutionary change of amino acid and DNA sequences

I calculated several measures of amino acid and DNA sequence polymorphisms by comparing *O. princeps* with *Or. cuniculus*. The evolutionary rates of amino acid and DNA substitutions at each locus are in general agreement with previously published rates for a variety of protein coding genes (Dayhoff and Schwartz 1978) (Table 3.1). However, when considering only the *IGF1*, *NOS2* and *TGFB1* coding sequences, the synonymous rate for each locus is about 6.5127 times faster than the nonsynonymous rate, while for *EHHADH* the synonymous rate is only about 3.551 times faster than the nonsynonymous rate (Table 3.1). This could be due to a decrease in the synonymous rate, an increase in nonsynonymous rate, or both. However, when comparing the predicted synonymous rate and nonsynonymous rate for *EHHADH* (4.38863×10^{-9} and 1.23598×10^{-9}, respectively) to the average synonymous rate and nonsynonymous rate reported for enzymes (3.08×10^{-9} and 0.37×10^{-9}, respectively), in addition to the observation that *EHHADH* is predicted to have the fastest rate of change for amino acid sequences (Table 3.1), it is possible that this discrepancy is caused solely by an increased rate of nonsynonymous substitutions at the *EHHADH* locus.

The ratio of transitions to transversions, and the proportional differences at the first, second and third codon positions for all four genes are also in general agreement with what is known of the evolution of protein-coding genes (Table 3.2). But, I note that *EHHADH*, *NOS2* and *TGFB1* exhibit a moderate level of divergence—when comparing the p-distance between pika and rabbit—but that *IGF1* shows very low divergence (Table 3.2). In fact, pika and rabbit *EHHADH*, *NOS2* and *TGFB1* have diverged, by proportion,

about 2.0162 times more so than *IGF1*. This indicates a higher similarity between pika and rabbit *IGF1,* and that this particular gene is likely evolving under strict functional constraints.

It is also interesting to note a transitional bias for both enzymes. Transitions are less likely to result in amino acid substitutions, and transversions result in mispairings which cause the double helix to distort at the site of mutation and makes the mispairing easier for DNA repair enzymes to recognize (Jukes 1987; Collins and Jukes 1994). Thus, many potential transversions are repaired and transitions are more likely to become fixed. It is difficult to speculate why, or if it is even meaningful that both enzymes show a transitional bias while both growth factors do not. It may, indeed, be related to functional constraint. That is, for a molecule with strict functional constraint, the larger is the chance that mutations are deleterious and will be eliminated by natural selection (Kimura 1984).

Phylogenetics

When a phylogenetic tree is constructed from a single locus, the tree may not necessarily agree with the species tree (Pamilo and Nei 1988), and for closely related species, this chance is quite high due to incomplete lineage sorting. Indeed, none of the gene trees constructed from mRNA datasets show *O. princeps* and *O. collaris* as reciprocally monophyletic. These results are not necessarily unexpected, because polymorphic genes have likely been diverging longer than the time of species divergence (Maddison 1997; Edwards and Beerli 2000; Nichols 2001) and, for closely related species in which insufficient time has passed for drift to erase shared variation, we expect there to be pooled diversity. On the other hand, phylogenies constructed from genomic datasets

revealed emerging structure within *O. princeps* corresponding to the five previously described subspecies (Hafner and Smith 2010).

The proportion of interpopulational diversity (N_{ST}) was exceedingly high for each gene (Table 3.3). The N_{ST} index ranges from 0 to 1, with values close to zero suggesting panmixis, while values on the higher end of this range indicate that populations are not currently interbreeding (Wright 1969). Hafner and Sullivan (1995) reported an F_{ST} value of 0.856, averaged across 26 allozymes, which, in conjunction with the high N_{ST} values reported in this study, strongly suggest isolation across the American pikas' range. It is interesting, however, that *IGF1* shows a relatively much lower value when compared to *EHHADH*, *NOS2*, and *TGFB1*. Mathematically, *IGF1* shows a lower value for nucleotide differentiation because the average nucleotide diversity within lineages is similar to the nucleotide diversity calculated for the entire dataset. However, the biological interpretation is much less clear.

Using substitution rates obtained earlier by comparing the evolutionary distance between pika with rabbit (Table 3.1), I estimated the temporal split between *O. princeps* and *O. collaris*. The results showed similar estimates for *IGF1*, *NOS2*, and *TGFB1*, but a much lower estimate for *EHHADH* (Table 3.4). This is likely due to the fact that the majority of data collected from the *EHHADH* locus is represented by coding sequence (Figure 3.5), while, for the other three loci, the majority of data was non-coding (intron) sequence (Figures 3.2, 3.8, and 3.11). Two other estimates for the *O. princeps*/*O. collaris* split are reported to be 2.562 MYA (Galbreath et al. 2010) and 4.621 MYA (Lanier and Olson 2009). Both these estimates were based on mDNA. Thus, taking the average of all

six estimates, these two species are predicted to have begun diverging about 3.2905 MYA.

I also obtained a level of divergence between *O. princeps* and *O. collaris* by calculating *dN*/*dS* for each gene (Table 3.5). Notably, *dN*/*dS* for *EHHADH*, *NOS2*, and *TGFB1* are higher than expected if the strength of purifying selection has been strong since divergence. However, these higher values could be the stochastic result of comparing two closely related species (Rocha et al. 2006). Although, from a biological perspective, this could have been caused by a change in the selective constraint in either of the two lineages after splitting. On the other hand, *dN*/*dS* calculated from *IGF1* data suggests a strong level of purifying selection at this locus. Indeed, it is interesting to note that the *IGF1* gene also shows a reduced evolutionary distance between pika and rabbit (Table 3.2) and a low level of population differentiation among lineages of *O. princeps* (Table 3.3).

Statistical tests of neutrality

Observed homozygosity was higher than expected for the *IGF1*, *EHHADH*, and *TGFB1* loci (Table 3.7), which suggests that there is purifying selection or the presence of an advantageous allele at each locus. This is further supported by the results of both Tajima's *D* and Fu's *Fs* test statistics, which were negative, as expected when low-frequency polymorphisms are in excess (Table 3.8). Although these tests can also be interpreted as being due to rapid population expansion, there is evidence of pika population declines since the Last Glacial Maximum (Galbreath et al. 2009), thereby suggesting that positive selection may have been involved in producing these results.

Indeed, there are several lines of evidence that support the interpretation of a selective sweep affecting the *IGF1*, *EHHADH,* and *TGFB1* loci.

First, previous studies have suggested that the action of *IGF1* has a protective effect in *O. curzoniae* hepatic cells in response to hypoxia (Chen et al. 2007), while *TGFB1* inhibition was cited to be important in the pulmonary adaption in *O. curzoniae* to high altitude (Ge et al. 1998; Korrodi-Gregorio et al. 2013). In addition, levels of *EHHADH* activity were shown to be dramatically higher in high elevation *O. princeps* when compared to their relatively low elevation congeners, *O. collaris* (Sheafor 2003). Although these three genes could instead be under very strong purifying selection, this is not supported by the predicted increase in d_N/d_S along the branch leading to *Ochotona* in the PAML analyses (Chapter 2). Also, *O. princeps* is predicted to have accumulated more amino acid substitutions within the functional region of these three genes in both the statistical and ancestral state reconstruction tests (Chapter 2). It is possible, and likely, that the majority of coding sequence in these three genes is under purifying selection, but the evidence presented here also points to the possibility that there were a minority of amino acid changes occurring along the *Ochotona* branch which experienced fixation at an increased rate. Indeed, the elevated level of homozygosity, and the observation that θ is inflated relative to π for *IGF1*, *EHHADH*, and *TGFB1*, validate the possibility of a selective sweep.

It is unknown, however, if the molecular diversity observed in *O. princeps* is due to ancient or recent positive selection. Because *IGF1* and *TGFB1* were initially associated with adaptation in *O. curzoniae,* it is possible that positive selection acted on the these two loci before the pika radiation, which is predicted to have occurred after 8 –

11 MYA (Lanier and Olson 2009; Melo-Ferreira et al. 2015). Given that these two growth factors are predicted to have slower rates of evolutionary change at the amino acid and DNA sequence levels (when compared to both enzymes used in this study) (Table 3.1), it may be feasible to detect such an ancient selective sweep. An ancient selective sweep affecting the *IGF1* locus is further supported by a relatively even distribution of nucleotide diversity among the five lineages of *O. princeps* (Table 3.3), and the strong level of purifying selection indicated between *O. princeps*/*O. collaris* (Table 3.5). These same patterns are not seen, however, at the *TGFB1* locus. Most notable is the elevated *dN*/*dS* ratio between *O. princeps* and *O. collaris* (Table 3.5). While this result may be due to stochastic error, it is possible that there was a change in selective pressure along one lineage after the *O. princeps* and *O. collaris* split. It is reasonable that, because *TGFB1* was initially indicated for *O. curzoniae*—another high elevation pika species—*O. collaris* may have experienced relaxed selection at this locus after colonizing low elevation habitats, thereby increasing the proportion of neutral changes allowed at this locus. This would predict that, if a selective sweep at the *TGFB1* locus occurred prior to the *Ochotona* radiation, then the adaptive allele has been maintained in both *O. curzoniae* and *O. princeps* due to subsequent purifying selection maintaining it at high frequency, and that *O. collaris* has lost the selective constraint at the *TGFB1* locus upon colonizing low elevation environments. This scenario is highly contingent on *Ochotona* history, and would require additional data from *O. collaris*, as well as individual Eurasian pika species to test this hypothesis.

There is good evidence that positive selection has only recently occurred at the *EHHADH* locus in only the American pika lineage. Sheafor (2003) reported a significant

increase in *EHHADH* activity when comparing high elevation *O. princeps* individuals with low elevation *O. collaris* individuals, and the Bayesian phylogeny constructed from *EHHADH* mRNA dataset shows *O. princeps* possessing the more derived state (Figure 3.7). Initially, this result was unexpected because it is predicted that *O. collaris* represents a mere pocket of *O. princeps'* diversity (Galbreath and Hoberg 2012). However, if a selective sweep occurred after these two lineages split, then it is expected that *O. collaris* would have retained the ancient *EHHADH* variant while *O. princeps* accrued a new, derived allele at the *EHHADH* locus. This idea is also supported by the high d_N/d_S calculated for the *EHHADH* gene when comparing *O. princeps* and *O. collaris* individuals.

Observed homozygosity was nearly identical to the expected value for the *NOS2* locus (Table 3.4). Although my earlier analyses hinted at an increased rate of nonsynonymous substitutions occurring along the branch leading to *Ochotona* for this locus, this evidence was initially weak. Because several studies had identified *Ochotona NOSs* as a target of selection (Pichon et al. 2009; Ge et al. 2013; Xie et al. 2014), investigation of this gene was justified. However, subsequent evidence of non-neutral evolution was essentially non-existent for the *NOS2* locus; all three polymorphism-based tests of neutrality suggest that the *NOS2* locus is evolving neutrally. This is in contrast to previous studies suggesting that both *NOS1* and *NOS2* may have been a target of selection in the closely related plateau pika (Pichon et al. 2009; Xie et al. 2014). Thus, adaptive effects of *NOS* enzymes may be due to phenotypic plasticity. However, if this is so, one might predict *NOS* enzymes to experience strict functional constraints. Therefore,

it is more likely that, as a whole, the *NOS2* gene is evolving under purifying selection and that the small region of this gene which I have analyzed is evolving neutrally.

Conclusion

I have garnered a body of evidence which suggests an interesting pattern of evolution at the *IGF1*, *EHHADH,* and *TGFB1* loci of *Ochotona*. These three genes thus represent candidates for further investigation into functional effects, which may provide insight into the adaptive evolution of these genes within the *Ochotona* genus. Indeed, my examination of the variation within partial sequences of four nuclear genes from a large sample of *O. princeps* suggests a pattern consistent with selection at the *IGF1*, *EHHADH,* and *TGFB1* loci that could not be explained by neutral evolution.

SUMMARY AND CONCLUSION

I used tools of molecular biology and mathematical theory of population genetics to study the American pika from an evolutionary perspective. I chose to study pikas from an evolutionary perspective because there is evidence that this species has adapted phenotypically to life in high elevation environments. Therefore, I expected to find evidence of positive selection in the American pika genome.

This project consisted of two separate, but complementary, studies. First, I investigated selection and interspecific divergence, comparing pikas to closely related mammal species, and then investigated selection and intraspecific divergence, comparing the distribution of genetic diversity within *O. princeps* using population-level data. To accomplish this, I initially collected data for 54 different loci from public databases for interspecific tests. Loci were chosen based on previous studies of pikas, and *O. princeps* sequence data were available from two draft genomes. Multiple alignments and phylogenies were constructed for each of the 54 genes, and the results were subjected to the branch-site and branch-specific statistical tests of the CODEML module of PAML. These analyses indicated four genes, which were then resequenced from 30 different pika specimens: 25 *O. princeps* and five *O. collaris*.

Data collection consisted of designing multiple primers for each gene, DNA amplification via polymerase chain reaction and Sanger sequencing. These data were then used to calculate molecular diversity and phylogenetic statistics and, specifically, test the null hypothesis of each gene evolving neutrally. Using three separate tests of neutrality, I obtained substantial evidence for non-neutral evolution of the *EHHADH* gene. Enoyl-

CoA, hydratase/3-hydroxyacyl CoA dehydrogenase is involved in beta-oxidation within the mitochondria to produce acetyl-CoA, which enters the citric acid cycle, and NADH, which donates electrons to the electron transport chain. Activity of this enzyme can therefore be used to indicate relative fatty acid oxidation capacity (Wakil et al. 1954). Sheafor (2003) reported that high elevation pikas had greater *EHHADH* enzyme activity when compared with low elevation pikas, and that this increase in oxidative enzyme activity may enhance oxygen utilization in hypoxic environments. Given that this gene has shown both physiological evidence of adaptation at the organismal level and statistical evidence of positive selection at the DNA level, it thus represents a candidate for further investigation of the adaptive effect of this gene within *O. princeps*.

WORKS CITED

AKAIKE, H. 1981. Likelihood of a model and information criteria. Journal of Econometrics 16:3–14.

ARIS-BROSOU, S. AND L. EXCOFFIER. 1996. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Molecular Biology and Evolution 13:494–504.

ASHER, R. J. ET AL. 2005. Stem Lagomorpha and the Antiquity of Glires. Science 307:1091–1094.

AVIVI, A. ET AL. 2010. Neuroglobin, cytoglobin, and myoglobin contribute to hypoxia adaptation of the subterranean mole rat Spalax. Proceedings of the National Academy of Sciences 107:21570–21575.

AWAD, A. S. ET AL. 2006. Selective sphingosine 1-phosphate 1 receptor activation reduces ischemia-reperfusion injury in mouse kidney. American Journal of Physiology-Renal Physiology 290:F1516–F1524.

BAI, Z. ET AL. 2015a. Intermittent cold exposure results in visceral adipose tissue "browning" in the plateau pika (*Ochotona curzoniae*). Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology 184:171–178.

Bininda-Emonds, O. R., Cardillo, M., Jones, K. E., MacPhee, R. D., Beck, R. M., Grenyer, R., ... & Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, *446*(7135), 507-512.

BOUVEROT, P. 2012. Adaptation to Altitude-Hypoxia in Vertebrates. Springer Science & Business Media.

BROWN, J. H. 1971. Mammals on Mountaintops: Nonequilibrium Insular Biogeography. The American Naturalist 105:467–478.

BRUNSON, J. T., R. N. SEAMAN AND D. J. NASH. 1977. INTERPOPULATIONAL VARIATION OF BLOOD PROTEINS IN PIKA (*OCHOTONA PRINCEPS SAXATILIS*). The Great Basin Naturalist 37:77–80.

CASTILLO, J. A., C. W. EPPS, A. R. DAVIS AND S. A. CUSHMAN. 2014. Landscape effects on gene flow for a climate-sensitive montane species, the American pika. Molecular Ecology 23:843–856.

CHAPPELL, M. A. AND K. A. HAMMOND. 2003. Maximal aerobic performance of deer mice in combined cold and exercise challenges. Journal of Comparative Physiology B 174:41–48.

CHEN, X.-Q., S.-J. WANG, J.-Z. DU AND X.-C. CHEN. 2007a. Diversities in hepatic HIF-1, IGF-I/IGFBP-1, LDH/ICD, and their mRNA expressions induced by CoCl2 in Qinghai-Tibetan plateau mammals and sea level mice. American Journal of Physiology - Regulatory, Integrative and Comparative Physiology 292:R516–R526.

CHEVIRON, Z. A., A. D. CONNATY, G. B. MCCLELLAND AND J. F. STORZ. 2014. Functional Genomics of Adaptation to Hypoxic Cold-Stress in High-Altitude Deer Mice: Transcriptomic Plasticity and Thermogenic Performance. Evolution 68:48–62.

COLLINS, D. W. AND T. H. JUKES. 1994. Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics 20:386–396.

CONNER, D. A. 1983. Seasonal changes in activity patterns and the adaptive value of haying in pikas (*Ochotona princeps*). Canadian Journal of Zoology 61:411–416.

CONNER, D. A. 1985. Analysis of the vocal repertoire of adult pikas: ecological and evolutionary perspectives. Animal Behaviour 33:124–134.

CONNER, J. K. AND D. L. HARTL. 2004. A Primer of Ecological Genetics. Sinauer Associates.

CRAS, T. D. L., C. XUE, A. RENGASAMY AND R. A. JOHNS. 1996. Chronic hypoxia upregulates endothelial and inducible NO synthase gene and protein expression in rat lung. American Journal of Physiology - Lung Cellular and Molecular Physiology 270:L164–L170.

DARRIBA, D., R. DOALLO, D. POSADA AND G. L. TABOADA. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9:772.

DAWKINS, R. 2006. The Selfish Gene: 30th Anniversary Edition. OUP Oxford.

DAYHOFF, M. O. AND R. M. SCHWARTZ. 1978. Chapter 22: A model of evolutionary change in proteins. in in Atlas of Protein Sequence and Structure.

DEARING, M. D. 1996. Disparate determinants of summer and winter diet selection of a generalist herbivore, Ochotona princeps. Oecologia 108:467–478.

DUAN, C. AND Q. XU. 2005. Roles of insulin-like growth factor (IGF) binding proteins in regulating IGF actions. General and Comparative Endocrinology 142:44–52.

Du, J. Z., & You, Z. B. (1992). A radioimmunoassay of corticotrophin releasing factor of hypothalamus in *Ochotona curzoniae*. *Acta Theriol. Sin*, *12*(3), 223-229.

DYKHUIZEN, D. AND D. L. HARTL. 1980. Selective Neutrality of 6pgd Allozymes in *E. Coli* and the Effects of Genetic Background. Genetics 96:801–817.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792–1797.

EDWARDS, S. AND P. BEERLI. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54:1839–1854.

EWENS, W. J. 1972. The sampling theory of selectively neutral alleles. Theoretical Population Biology 3:87–112.

EXCOFFIER, L. AND H. LISCHER. 2010a. ARLEQUIN VER 3.5. 1.2 USER MANUAL (2010) Computational and molecular population genetics lab (CMPG). Institute of Ecology and Evolution University of Berne, Baltzerstrasse 6:3012.

EXCOFFIER, L. AND H. E. L. LISCHER. 2010b. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10:564–567.

FISHER, R. A. 1930. The Genetical Theory of Natural Selection: A Complete Variorum Edition. OUP Oxford.

FROSTELL, C., M. D. FRATACCI, J. C. WAIN, R. JONES AND W. M. ZAPOL. 1991. Inhaled nitric oxide. A selective pulmonary vasodilator reversing hypoxic pulmonary vasoconstriction. Circulation 83:2038–2047.

FU, Y. X. 1997. Statistical Tests of Neutrality of Mutations against Population Growth, Hitchhiking and Background Selection. Genetics 147:915–925.

GALBREATH, K. E., D. J. HAFNER AND K. R. ZAMUDIO. 2009. When Cold Is Better: Climate-Driven Elevation Shifts Yield Complex Patterns of Diversification and Demography in an Alpine Specialist (american Pika, *Ochotona Princeps*). Evolution 63:2848–2863.

GALBREATH, K. E., D. J. HAFNER, K. R. ZAMUDIO AND K. AGNEW. 2010. Isolation and introgression in the Intermountain West: contrasting gene genealogies reveal the complex biogeographic history of the American pika (*Ochotona princeps*). Journal of Biogeography 37:344–362.

GALBREATH, K. E. AND E. P. HOBERG. 2012. Return to Beringia: parasites reveal cryptic biogeographic history of North American pikas. Proceedings of the Royal Society B: Biological Sciences 279:371–378.

GE, R.-L. ET AL. 2013. Draft genome sequence of the Tibetan antelope. Nature Communications 4:1858.

GE, R.-L., K. KUBO, T. KOBAYASHI, M. SEKIGUCHI AND T. HONDA. 1998. Blunted hypoxic pulmonary vasoconstrictive response in the rodent *Ochotona curzoniae* (pika) at high altitude. American Journal of Physiology - Heart and Circulatory Physiology 274:H1792–H1799.

GLOVER, D. G., M. H. SMITH, L. AMES, J. JOULE AND J. M. DUBACH. 1977. Genetic variation in pika populations. Canadian Journal of Zoology 55:1841–1845.

GRAUR, D. AND W.-H. LI. 2000. Fundamentals of Molecular Evolution. Sinauer.

GRAYSON, D. K. 1987. The Biogeographic History of Small Mammals in the Great Basin: Observations on the Last 20,000 Years. Journal of Mammalogy 68:359–375.

GRAYSON, D. K. 2005. A brief history of Great Basin pikas. Journal of Biogeography 32:2103–2111.

GRAYSON, P. AND A. CIVETTA. 2012. Positive Selection and the Evolution of *izumo* Genes in Mammals. International Journal of Evolutionary Biology 2012:e958164.

GRAYSON, P. AND A. CIVETTA. 2013. Positive selection in the adhesion domain of *Mus* sperm Adam genes through gene duplications and function-driven gene complex formations. BMC Evolutionary Biology 13:217.

GUTHRIE, R. D. 1973. Mummified Pika (*Ochotona*) Carcass and Dung Pellets from Pleistocene Deposits in Interior Alaska. Journal of Mammalogy 54:970–971.

Hall, B. G. (2004). *Phylogenetic trees made easy: a how-to manual* (Vol. 547). Sunderland: Sinauer Associates.

HAFNER, D. J. AND A. T. SMITH. 2010. Revision of the subspecies of the American pika , *Ochotona princeps* (Lagomorpha: Ochotonidae). Journal of Mammalogy 91:401–417.

HAFNER, D. J. AND R. M. SULLIVAN. 1995. Historical and Ecological Biogeography of Nearctic Pikas (Lagomorpha: Ochotonidae). Journal of Mammalogy 76:302–321.

Hartl, Daniel L., Andrew G. Clark, and Andrew G. Clark. *Principles of population genetics*. Vol. 116. Sunderland: Sinauer associates, 1997.

HARTL, D. L. 2000. A primer of Population Genetics. in A primer of Population Genetics. Sinauer.

HAXHIU, M. A., C. H. CHANG, I. A. DRESHAJ, B. EROKWU, N. R. PRABHAKAR AND N. S. CHERNIACK. 1995. Nitric oxide and ventilatory response to hypoxia. Respiration Physiology 101:257–266.

HAYES, J. P. AND M. A. CHAPPELL. 1986. Effects of Cold Acclimation on Maximum Oxygen Consumption during Cold Exposure and Treadmill Exercise in Deer Mice, Peromyscus maniculatus. Physiological Zoology 59:473–481.

HENRY, P. AND M. A. RUSSELLO. 2013. Adaptive divergence along environmental gradients in a climate-change-sensitive mammal. Ecology and Evolution 3:3906–3917.

Hijmans, R. J., Guarino, L., & Mathur, P. (2004). DIVA-GIS. Version 7.5. A geographic information system for the analysis of species distribution data.

HOLMES, W. G. 1991. Predator risk affects foraging behaviour of pikas: observational and experimental evidence. Animal behaviour 42:111–119.

HUDSON, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology 23:183–201.

HUELSENBECK, J. P. AND F. RONQUIST. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Hughes, A. L., & Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, *335*(6186), 167-170.

Hughes, A. L., & Nei, M. (1989). Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proceedings of the National Academy of Sciences*, *86*(3), 958-962.

IVINS, B. L. AND A. T. SMITH. 1983. Responses of pikas (*Ochotona princeps*, Lagomorpha) to naturally occurring terrestrial predators. Behavioral Ecology and Sociobiology 13:277–285.

JIZENG, D., W. YAN AND Y. ZHIBING. 1996. Norepinephrine stimulates the release of corticotropin-realeasing factor from median eminence of *Ochotona curzoniae*. Acta theriologica sinica 17:204–207.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, *3*(21), 132.

JUKES, T. H. 1987. Transitions, transversions, and the molecular evolutionary clock. Journal of molecular evolution 26:87–98.

KELLEY, K. M. ET AL. 2002. Comparative endocrinology of the insulin-like growth factor-binding protein. Journal of Endocrinology 175:3–18.

KIETZMANN, T., A. KRONES-HERZIG AND K. JUNGERMANN. 2002. Signaling cross-talk between hypoxia and glucose via hypoxia-inducible factor 1 and glucose response elements. Biochemical pharmacology 64:903–911.

KIMURA, M. 1962. On the Probability of Fixation of Mutant Genes in a Population. Genetics 47:713–719.

KIMURA, M. 1984. The Neutral Theory of Molecular Evolution. Cambridge University Press.

KIMURA, M. 1986. DNA and the Neutral Theory. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 312:343–354.

KIMURA, M. 1987. Molecular evolutionary clock and the neutral theory. Journal of Molecular Evolution 26:24–33.

KIMURA, M. AND T. OHTA. 1971. Theoretical Aspects of Population Genetics. Princeton University Press.

KING, J. L. AND T. H. JUKES. 1969. Non-Darwinian Evolution. Science 164:788–798.

KINGMAN, J. F. C. 1982. The coalescent. Stochastic Processes and their Applications 13:235–248.

KORRODI-GREGÓRIO, L. ET AL. 2013. An Intriguing Shift Occurs in the Novel Protein Phosphatase 1 Binding Partner, TCTEX1D4: Evidence of Positive Selection in a Pika Model. PLoS ONE 8:e77236.

LANIER, H. C. AND L. E. OLSON. 2009. Inferring divergence times within pikas (*Ochotona* spp.) using mtDNA and relaxed molecular dating techniques. Molecular Phylogenetics and Evolution 53:1–12.

Lanier, H. C., Massatti, R., He, Q., Olson, L. E., & Knowles, L. L. (2015). Colonization from divergent ancestors: glaciation signatures on contemporary patterns of genomic variation in Collared Pikas (*Ochotona collaris*). *Molecular ecology*, *24*(14), 3688-3705.

LAUBACH, V. E., E. G. SHESELY, O. SMITHIES AND P. A. SHERMAN. 1995. Mice lacking inducible nitric oxide synthase are not resistant to lipopolysaccharide-induced death. Proceedings of the National Academy of Sciences of the United States of America 92:10688–10692.

LEE, P. ET AL. 2014. Irisin and FGF21 are cold-induced endocrine activators of brown fat function in humans. Cell metabolism 19:302–309.

LEMAY, M. A., P. HENRY, C. T. LAMB, K. M. ROBSON AND M. A. RUSSELLO. 2013. Novel genomic resources for a climate change sensitive mammal: characterization of the American pika transcriptome. BMC Genomics 14:311.

Li, Q. F., You, Z. B., Chen, X. G., & Du, J. Z. (1986). Effect of chronic hypoxia on the liver of *Ochotona curzoniae* and rat. *Acta Theriologica Sinica*, *6*(4), 261-266.

Li, Q. F., Chen, X. G., You, Z. B., & Du, J. Z. (1987). A comparative study on effects of acute hypoxia upon livers of three small mammals. *Acta Theriologica Sinica*, *7*(1), 51-57.

LI, H. ET AL. 2013. VEGF189 Expression Is Highly Related to Adaptation of the Plateau Pika (*Ochotona curzoniae*) Inhabiting High Altitudes. High Altitude Medicine & Biology 14:395–404.

LI, H.-G. ET AL. 2009. The protein level of hypoxia-inducible factor-1α is increased in the plateau pika (*Ochotona curzoniae*) inhabiting high altitudes. Journal of Experimental Zoology Part A: Ecological Genetics and Physiology 311A:134–141.

LI, Q. ET AL. 2001. Cold adaptive thermogenesis in small mammals from different geographical zones of China. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology 129:949–961.

Librado, P., & Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, *25*(11), 1451-1452.

LISSOVSKY, A. A., N. V. IVANOVA AND A. V. BORISENKO. 2007. Molecular Phylogenetics and Taxonomy of the Subgenus Pika (*Ochotona*, Lagomorpha). Journal of Mammalogy 88:1195–1204.

LUO, Y. ET AL. 2008. Mitochondrial genome analysis of *Ochotona curzoniae* and implication of cytochrome c oxidase in hypoxic adaptation. Mitochondrion 8:352–357.

MACARTHUR, R. A. AND L. C. H. WANG. 1973. Physiology of thermoregulation in the pika, *Ochotona princeps*. Canadian Journal of Zoology 51:11–16.

MACARTHUR, R. A. AND L. C. H. WANG. 1974. Behavioral thermoregulation in the pika Ochotona princeps: a field study using radiotelemetry. Canadian Journal of Zoology 52:353–358.

MADDISON, W. P. 1997. Gene trees in species trees. Systematic biology 46:523–536.

MATTHEE, C. A., B. J. V. VUUREN, D. BELL AND T. J. ROBINSON. 2004. A Molecular Supermatrix of the Rabbits and Hares (Leporidae) Allows for the Identification of Five Intercontinental Exchanges During the Miocene. Systematic Biology 53:433–447.

MCKENNA, M. C. AND S. K. BELL. 2013. Classification of Mammals: Above the Species Level. Columbia University Press.

MEAD, J. I. AND F. GRADY. 1996. *Ochotona* (Lagomorpha) from Late Quaternary Cave Deposits in Eastern North America. Quaternary Research 45:93–101.

MELO-FERREIRA, J., A. LEMOS DE MATOS, H. AREAL, A. A. LISSOVSKY, M. CARNEIRO AND P. J. ESTEVES. 2015. The phylogeny of pikas (*Ochotona*) inferred from a multilocus coalescent approach. Molecular Phylogenetics and Evolution 84:240–244.

NEI, M. 1987. Molecular Evolutionary Genetics. Columbia University Press.

NEI, M. 2013. Mutation-Driven Evolution. OUP Oxford.

NEI, M. AND T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3:418–426.

NEI, M. AND S. KUMAR. 2000. Molecular Evolution and Phylogenetics. Oxford University Press.

NEI, M., Y. SUZUKI AND M. NOZAWA. 2010. The Neutral Theory of Molecular Evolution in the Genomic Era. Annual Review of Genomics and Human Genetics 11:265–289.

NICHOLS, R. 2001. Gene trees and species trees are not the same. Trends in Ecology & Evolution 16:358–364.

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity*, *86*(6), 641-647.

PALMER, L. A., G. L. SEMENZA, M. H. STOLER AND R. A. JOHNS. 1998. Hypoxia induces type II NOS gene expression in pulmonary artery endothelial cells via HIF-1. American Journal of Physiology - Lung Cellular and Molecular Physiology 274:L212–L219.

PAMILO, P. AND M. NEI. 1988. Relationships between gene trees and species trees. Molecular biology and evolution 5:568–583.

PEPKE-ZABA, J., T. W. HIGENBOTTAM, A. T. DINH-XUAN, D. STONE AND J. WALLWORK. 1991. Inhaled nitric oxide as a cause of selective pulmonary vasodilatation in pulmonary hypertension. The Lancet 338:1173–1174.

PEPPER, M. S. 1997. Transforming growth factor-beta: Vasculogenesis, angiogenesis, and vessel wall integrity. Cytokine & Growth Factor Reviews 8:21–43.

PICHON, A. ET AL. 2009. Long-term ventilatory adaptation and ventilatory response to hypoxia in plateau pika (*Ochotona curzoniae*): role of nNOS and dopamine. American Journal of Physiology - Regulatory, Integrative and Comparative Physiology 297:R978–R987.

PICHON, A. ET AL. 2013. Cardiac adaptation to high altitude in the plateau pika (*Ochotona curzoniae*). Physiological Reports 1.

Rambaut, A. (2007). FigTree, a graphical viewer of phylogenetic trees. *See http://tree. bio. ed. ac. uk/software/figtree*.

RJ, Z., R. XF, W. DB, W. DW, W. L AND S. SZ. 2012. [Functional difference of malate-aspartate shuttle system in liver between plateau zokor (*Myospalax baileyi*) and plateau pika (*Ochotona curzoniae*)]. Sheng li xue bao : [Acta physiologica Sinica] 64:177–186.

ROBSON, K. M., C. T. LAMB AND M. A. RUSSELLO. 2015. Low genetic diversity, restricted dispersal, and elevation-specific patterns of population decline in American pikas in an atypical environment. Journal of Mammalogy:gyv191.

Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of theoretical biology*, *239*(2), 226-235.

SCHMID, C. 1995. Insulin-like growth factors. Cell Biology International 19:445–457.

SELANDER, R. K. AND W. E. JOHNSON. 1973. Genetic Variation Among Vertebrate Species. Annual Review of Ecology and Systematics 4:75–91.

SHEAFOR, B. A. 2003. Metabolic enzyme activities across an altitudinal gradient: an examination of pikas (genus *Ochotona*). Journal of Experimental Biology 206:1241–1249.

SMITH, A. T. 1974. The Distribution and Dispersal of Pikas: Influences of Behavior and Climate. Ecology 55:1368–1376.

SMITH, A. T. AND M. L. WESTON. 1990. Ochotona princeps. Mammalian Species:1–8.

SPRINGER, M. S. ET AL. 2007. The Adequacy of Morphology for Reconstructing the Early History of Placental Mammals. Systematic Biology 56:673–684.

STORZ, J. F. 2007. Hemoglobin Function and Physiological Adaptation to Hypoxia in High-Altitude Mammals. Journal of Mammalogy 88:24–31.

TAJIMA, F. 1983. Evolutionary Relationship of Dna Sequences in Finite Populations. Genetics 105:437–460.

TAJIMA, F. 1989a. The effect of change in population size on DNA polymorphism. Genetics 123:597–601.

TAJIMA, F. 1989b. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

TAMURA, K., G. STECHER, D. PETERSON, A. FILIPSKI AND S. KUMAR. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution 30:2725–2729.

TUFTS, D. M. ET AL. 2014. Epistasis Constrains Mutational Pathways of Hemoglobin Adaptation in High-Altitude Pikas. Molecular Biology and Evolution:msu311.

VARNER, J. AND M. D. DEARING. 2014. Dietary plasticity in pikas as a strategy for atypical resource landscapes. Journal of Mammalogy 95:72–81.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES AND A. C. WILSON. 1991. African Populations and the Evolution of Human Mitochondrial DNA. Science 253:1503–1507.

WAKIL, S. J., D. E. GREEN, S. MII AND H. R. MAHLER. 1954. STUDIES ON THE FATTY ACID OXIDIZING SYSTEM OF ANIMAL TISSUES VI. β-HYDROXYACYL COENZYME A DEHYDROGENASE. Journal of Biological Chemistry 207:631–638.

WANG, D. ET AL. 2013. Testis-specific lactate dehydrogenase is expressed in somatic tissues of plateau pikas. FEBS Open Bio 3:118–123.

WANG, D., R. SUN, Z. WANG AND J. LIU. 1999. Effects of temperature and photoperiod on thermogenesis in plateau pikas (*Ochotona curzoniae*) and root voles (*Microtus oeconomus*). Journal of Comparative Physiology B 169:77–83.

WANG, J.-M., Y.-M. ZHANG AND D.-H. WANG. 2006. Seasonal Thermogenesis and Body Mass Regulation in Plateau Pikas (*Ochotona curzoniae*). Oecologia 149:373–382.

WATTERSON, G. A. 1977. Heterosis or Neutrality? Genetics 85:789–814.

WATTERSON, G. A. 1978. The Homozygosity Test of Neutrality. Genetics 88:405–417.

WOOTEN, M. C. AND M. H. SMITH. 1985. Large Mammals are Genetically Less Variable? Evolution 39:210–212.

WRIGHT, S. 1935. The analysis of variance and the correlations between relatives with respect to deviations from an optimum. Journal of Genetics 30:243–256.

Wright, S. (1969). The theory of gene frequencies. Vol. 2. *Evolution and the Genetics of Populations*, 294-324.

Xiaoguang, L. Q. Y. Z. C., & Jizeng, D. U. (1986). EFFECTS OF CHRONIC HYPOXIA ON THE LIVER OF *OCHOTONA CURZONIAE* AND RATS [J]. *Acta Theriologica Sinica*, *4*, 004.

XIE, L. ET AL. 2014. Inhibition of inducible nitric oxide synthase expression and nitric oxide production in plateau pika (*Ochotona curzoniae*) at high altitude on Qinghai-Tibet Plateau. Nitric Oxide 38:38–44.

XZ, Q., W. XJ, Z. SH, R. XF, W. L AND W. DB. 2008. [Hypoxic adaptation of the hearts of plateau zokor (Myospalax baileyi) and plateau pika (*Ochotona curzoniae*)]. Sheng li xue bao : [Acta physiologica Sinica] 60:348–354.

YANG, J. ET AL. 2008. Natural Selection and Adaptive Evolution of Leptin in the Ochotona Family Driven by the Cold Environmental Stress. PLoS ONE 3:e1472.

YANG, J. ET AL. 2011. Functional Evolution of Leptin of Ochotona curzoniae in Adaptive Thermogenesis Driven by Cold Environmental Stress. PLoS ONE 6.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS 13:555–556.

YANG, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution 15:568–573.

YANG, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24:1586–1591.

ZHANG, Q., Q. HE, F. XUE AND J. MA. 2014a. Inhibitor screening of lactate dehydrogenase C4 from black-lipped pika in the Western Sichuan Plateau. Bioscience, Biotechnology, and Biochemistry 78:651–654.

ZHANG, Q.-L. ET AL. 2014b. Cloning and characterization of lactate dehydrogenase C4 from *Ochotona curzoniae*. Molecular Biology 48:105–112.

ZHAO, T. B. ET AL. 2004. Cloning of hypoxia-inducible factor 1α cDNA from a high hypoxia tolerant mammal—plateau pika (*Ochotona curzoniae*). Biochemical and Biophysical Research Communications 316:565–572.

APPENDIX A

Primer Pairs

Locus	Primer Name	Sequence (5' -> 3')	Annealing Temperature
IGF1 exon 3	OpIGF1_F1	GAGACACACAGGAATCCAGAAG	51º
	OpIGF1_R1	AACACTGGTCCAACAGTTACTT	
IGF1 exon 2	OpIGF1_F2	GATTTGGAAACCCGGGAAATG	52º
	OpIGF1_R2	AGCCTCGCTGAAACTCTATG	
EHHADH exon 7	OpBHCAD_F2	CCACTGTTCCTGTTCATCCATA	51.5º
	OpBHCAD_R2	GGAAGAAGCAATCTCATCAACATC	
EHHADH exon 7	OpBHCAD_F3	CGTCCTGGATGTTGATGAGATT	52-53º
	OpBHCAD_R3	CCAACCCAGAAAGGTCAGATAC	
EHHADH exon 6	OpBHCAD_F4	CTCCTTGCCCTCTCAAAGAAATA	52º
	OpBHCAD_F4	CCCACAACACCAACAGAAGA	
NOS2 exon 17	OpNOS2_F1	CCTGTGCCTGTCTATCTTCAC	52º
	OpNOS2_R1	GAACTGCAATCTCTCCATCCA	
NOS2 exon 16	OpNOS2_F2	CATACCACAGAAGTCCAGCTT	52-53º
	OpNOS2_R2	CACATGCACAGAGGAGAGTT	
NOS2 exons 14 and 15	OpNOS2_F3	AAGGGCTGAAGTCCATGTC	52º
	OpNOS2_R3	CCTCAGCGCCAAGTCTC	
TGFB1 exon 2	OpTGFB1_F1	TGACGCCTCTCTCTCATCTC	52-53º
	OpTGFB1_R1	CATTTCCTGCTCATCCATTTCC	
TGFB1 exon 1	OpTGFB1_F3	CCCTGCTTCTTTCTTTCTGATAAAC	51º
	OpTGFB1_R3	GACATCCAAATCCCTCCTCAC	
TGFB1 exon 3	OpTGFB1_F4	TGTGTGTGTGTGTGTGTAAGA	52º
	OpTGFB1_R4	AAGGGCTGAAGTCCATGTC	

Table A.1. Primers used in PCR to amplify *O. princeps* DNA. The name of the gene, and the associated exon targeted with the primers, are listed in the first column. The name of the primers, primer sequences (5' -> 3'), and annealing temperatures follow.

APPENDIX B

Locality Information

Abbreviation	Catalog ID	State	County	Latitude	Longitude
SRM 101	CUMV:Mamm:20114	New Mexico	Taos	36.6122	-105.5012
SRM 103	CUMV:Mamm:20116	New Mexico	Taos	36.6033325195	-105.5015029907
SRM 117	CUMV:Mamm:20128	Colorado	Clear Creek	39.7999000549	-105.7880020142
SRM 139	CUMV:Mamm:20145	Colorado	Mesa	39.018236	-108.168282
SRM 360	CUMV:Mamm:20329	New Mexico	Santa Fe	35.7952003479	-105.7753982544
NRM 169	CUMV:Mamm:20166	Wyoming	Carbon	41.1777992249	-107.0139007568
NRM 231	CUMV:Mamm:20211	Montana	Judith Basin	46.9664	-110.6292
NRM 283	CUMV:Mamm:20255	Idaho	Adams	45.1870002747	-116.5752029419
NRM 395	CUMV:Mamm:20692	Nevada	Elko	40.5876998901	-115.3896026611
NRM 430	CUMV:Mamm:20589	Alberta	-	54.0773010254	-119.418800354
CR 323	CUMV:Mamm:20293	Washington	Okanogan	48.5105018616	-120.6398010254
CR 333	CUMV:Mamm:20302	Washington	Skamania	46.3302993774	-121.7480010986
CR 441	CUMV:Mamm:20729	British Columbia	-	52.2142982483	-126.3666000366
CR 462	CUMV:Mamm:20584	British Columbia	Lillooet Land	50.5817985535	-123.0280990601
CR 495	CUMV:Mamm:20632	Oregon	-	44.2608985901	-121.8089981079
SN 347	CUMV:Mamm:20316	Nevada	Nye	38.7252006531	-116.9261016846
SN 382	CUMV:Mamm:20682	California	Modoc	41.3698997498	-120.2388000488
SN 385	CUMV:Mamm:20684	California	Alpine	38.54582	-119.8162
SN 514	CUMV:Mamm:20575	Oregon	Baker	45.0577011108	-117.2830963135
SN 555	CUMV:Mamm:21157	Utah	Millard	38.3821	-112.4005
CU 523	CUMV:Mamm:20654	Utah	Summit	40.7647018433	-111.0160980225
CU 526	CUMV:Mamm:20572	Utah	Summit	38.3821	-112.4005
CU 533	CUMV:Mamm:20570	Utah	Sanpete	39.1118011475	-111.4710998535
CU 540	CUMV:Mamm:20663	Utah	Wayne	38.4402999878	-111.4829025269
CU 545	CUMV:Mamm:20669	Utah	Wayne	38.4402999878	-111.4829025269

Table B.1. Cornell Museum of Vertebrates (CUMV) catalog IDs of samples used in analyses and locations where samples were collected.