MOLECULAR PHYLOGENETICS OF THE RHINOCEROS CLADE AND EVOLUTION OF *UCP1* TRANSCRIPTIONAL REGULATORY ELEMENTS ACROSS THE MAMMALIAN PHYLOGENY

By

Michael J. Gaudry

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Biological Sciences

University of Manitoba

Winnipeg

Canada

ABSTRACT

Aiming to resolve contentious phylogenetic relationships among rhinoceros subfamilies (Dicerorhininae, Rhinocerotinae, and Dicerotinae), I constructed a ~131 kilobase nuclear DNA dataset for the Malayan tapir and six rhinoceros species, including the extinct woolly rhinoceros. Phylogenetic analyses, possibly confounded by incomplete lineage sorting associated with quick ancestral speciation events, yielded opposing trees: ((Dicerorhininae, Rhinocerotinae) Dicerotinae) or (Rhinocerotinae (Dicerorhininae, Dicerotinae)), though five of six informative indels independently supported the latter relationship. Additionally, eye genes revealed no inactivating mutations that may underlie reputedly poor vision among rhinoceroses. Furthermore, I investigated thermogenic *UCP1* transcriptional regulators among 139 mammal species, expecting deleterious mutations in eutherians possessing *UCP1* pseudogenes and possibly even large-bodied species (e.g. rhinoceroses) that retain intact *UCP1*. Promoters and enhancers were conserved in all species with intact *UCP1*, suggesting that it remains functionally expressed in these species. However, these elements have been lost in some *UCP1*-less species, indicating the enhancer is non-pleotropic.

ACKNOWLEDGMENTS

I would especially like to thank my supervisor, Kevin Campbell, for his endless support over the course of my research. I will be forever appreciative for his advice and willingness to share his vast expertise. His energetic enthusiasm and curiosity have been increadibly motivational and he has shown me what it means to be a scientist. I was also lucky to receive valuable recommendations and guidance throughout my research from my committee members, Jason Treberg and Martin Jastroch.

I deeply appreciate the knowledgable suggestions offered by Mark Springer, as well as his generosity in providing DNA samples. I am grateful of Tom Gilbert, Eske Willerslev, and Rasumus Havmøller for suppling Malayan tapir, Sumatran rhinoceros and woolly rhinoceros tissue samples, which were used in the construction of DNA libraries thanks to the hard work of Tony Signore and Nathan Wales. Peter van Coeverden de Groot kindly provided me with extremely rare Javan rhinoceros bone samples. I also thank Margaret Docker for allowing me to use her Ion Torrent sequencer. Kai He took the time to teach me many of the molecular techniques that I utilized during this research, for which I am very grateful.

Finally, I would like to express my deep gratitude to my family and friends for always offering their unconditional love and support.

This research was funded by a Manitoba Graduate Scholarship, G.A. Lubinsky Memorial Scholarship, an NSERC Discovery Grant, and an NSERC Discovery Accelerator Supplement.

TABLE OF CONTENTS

MOLECULAR PHYLOGENETICS OF THE RHINOCEROS CLADE AND EVOLUTION OF UCP1 TRANSCRIPTIONAL REGULATORY ELEMENTS ACROSS THE MAMMALIAN PHYLOGENY.....i

ABSTRACT.....ii

ACKNOWLEDGMENTS.....iii

TABLE OF CONTENTS.....iv

LIST OF TABLES.....viii

LIST OF FIGURES.....ix

LIST OF ABBREVIATIONS.....xii

CHAPTER 1: GENERAL INTRODUCTION.....1

CHAPTER 2: MOLECULAR PHYLOGENETICS OF THE RHINOCEROS CLADE.....10

2.1. Abstract.....10

2.2. Introduction.....11

2.2.1. General introduction.....11

2.2.2. Perissodactyl evolution.....12

2.2.3. Rhinoceros evolution.....13

2.2.4. Previous studies of rhinoceros evolution.....19

2.2.5. Rhinoceros eyesight.....23

2.2.6. Objectives.....24

2.3. Materials and Methods.....25

2.3.1. Sampling.....25

2.3.2. Construction of DNA libraries..... 26
2.3.2.1. Black and Indian rhinoceros DNA library preparation..... 26
2.3.2.2. Javan rhinoceros DNA extraction and library preparation..... 28
2.3.2.3. Malayan tapir and Sumatran rhinoceros DNA extraction and library preparation..... 30
2.3.2.4. Woolly rhinoceros DNA extraction and library preparation..... 31
2.3.3. In-solution hybridization captures..... 34
2.3.4. Ion Torrent next-generation sequencing..... 37
2.3.5. Sequenced read assemblies 39
2.3.6. Genome mining..... 40
2.3.7. DNA Alignments..... 41
2.3.8. Phylogenetic analyses..... 41
2.3.8.1. Coalescence analyses..... 41
2.3.8.2. Concatenation analyses..... 43
2.3.8.3. Robinson-Foulds distances..... 45
2.3.9. Eye gene selection pressure using PAML..... 46
2.4. Results..... 47
2.4.1. Number of reads sequenced and sequence coverage..... 47
2.4.2. Coalescence phylogenetic trees..... 51
2.4.3. Concatenation phylogenetic trees..... 53
2.4.4. Phylogenetically informative indels..... 57
2.4.5. Eye gene selection pressure results..... 58
2.5. Discussion..... 60

2.5.1. Hybridization capture and next-generation sequencing..... 60

2.5.2. Rhinoceros phylogenetics..... 64

2.5.3. Rhinoceros eye gene selection pressure..... 72

2.6. Conclusions..... 73

CHAPTER 3: EVOLUTION OF *UCP1* TRANSCRIPTIONAL REGULATORY ELEMENTS ACROSS THE MAMMALIAN PHYLOGENY.....75

3.1. Abstract..... 75

3.2. Introduction..... 76

3.2.1. Evolution of uncoupling protein 1..... 76

3.2.2. Evolution of eutherian *UCP1* regulatory elements..... 83

3.3. Materials and methods..... 88

3.3.1. *UCP1* regulatory sequences..... 88

3.3.2. Phylogenetic trees..... 93

3.4. Results..... 94

3.4.1. *UCP1* coding sequences..... 94

3.4.2. *UCP1* basal promoter..... 96

3.4.3. CpG island..... 99

3.4.4. Putative regulatory region (PRR)..... 100

3.4.5. *UCP1* enhancer..... 101

3.5. Discussion..... 107

3.6. Conclusions..... 115

CHAPTER 4: FINAL DISCUSSION AND CONCLUSIONS.....116

4.1. Molecular phylogenetics of rhinoceroses..... 116

4.2. Molecular evolution of rhinoceros eye genes..... 117

4.3. Evolution of *UCP1* transcriptional regulatory elements across the mammalian phylogeny..... 118

LITERATURE CITED..... 120

APPENDICES..... 136

LIST OF TABLES

Table 2.1. PCR thermocycling regimens used to amplify DNA libraries..... 28

Table 2.2. Number of reads sequenced and average read length for the black, Indian, Javan, Sumatran and woolly rhinoceroses and the Malayan tapir..... 49

LIST OF FIGURES

Figure 2.1. Illustrations of the possible phylogenetic relationships among the three rhinoceros subfamilies, including the "two-horn" (A), "biogeographical" (B), and "separate lineage" (C) hypotheses. The hypothesis proposed by Fernando et al. (2006) linking dicerotines (black and white rhinoceroses) and rhinocerotines (Javan and Indian rhinoceroses) to the exclusion of dicerorhines is also depicted (D).....20

Figure 2.2. DNA library sequencing and assembly information for each species. Black bars represent the percentage of total reads per species that matched as blast hits to the white rhinoceros genome using the "discontinuous megablast" setting in Geneious. White bars denote the percent coverage of sequence data included in the phylogenetic analyses for each species relative to the complete coverage of the white rhinoceros (*Ceratotherium simum*). Grey bars represent the percentage of sequenced reads for each species that assembled using the "map to reference function" in Geneious to the white rhinoceros reference sequences that were targeted during hybridization capture experiments.....50

Figure 2.3. Percent sequence coverage of intron (black bars) and exon (white bars) regions relative to the complete coverage of the white rhinoceros (*Ceratotherium simum*).....50

Figure 2.4. ASTRAL-II species tree performed using the best-scoring GARLI maximum likelihood gene tree for each of the 199 gene segment alignments. Node values represent local posterior probability branch support and branch lengths are expressed in coalescence units.....52

Figure 2.5. SVDQuartets consensus tree performed in PAUP* 4.0a150 with 1000 bootstrap replicates. All possible quartet trees were evaluated in the analysis. Node values indicate bootstrap support percentages. Branch lengths are arbitrary in this tree.....53

Figure 2.6. Best-scoring maximum likelihood species tree generated in RAxML v 7.2.8 with the GTR GAMMA nucleotide substitution model being estimated for each of the 199 and 571 gene segment partitions of the 131,931 bp supermatrix. Node values denote bootstrap support percent values generated with 500 bootstrap replicates for 199 and 571 partition schemes, respectively. Branch lengths represent the number of nucleotide substitutions per site.....55

Figure 2.7. Bayesian concatenation tree for 131,931 bp supermatrix made using MrBayes with a 10,000,000 chain length sampled every 10,000 generations and a burn-in length of 1,000,000. The GTR substitution model with gamma rate variation was employed with *Camelus ferus* selected as the outgroup species. Branch lengths represent the number of nucleotide substitutions per site. Node values denote posterior probabilities.....56

Figure 2.8. Best-scoring maximum likelihood trees generated in RAxML v 7.2.8 with the GTR GAMMA nucleotide substitution model for 365 concatenated intron partitions

totaling 67,724 bp and 206 concatenated exon partitions totaling 64,207 bp. Blue branches indicate incongruent topologies between intron and exon trees. Node values denote bootstrap support percent values generated with 500 bootstrap replicates. Branch lengths represent the number of nucleotide substitutions per site.....57

Figure 2.9. Alignment segments showing phylogenetically informative insertions and deletions. Numbers correspond to the locations within the concatenated alignment. Indels linking two-horned rhinos or Asian rhinos are highlighted in red and green, respectively.....58

Figure 2.10. Omega (ω) values represented by colour for twelve eye genes acquired using the "free-ratio" model in CODEML. The 12 genes (*ARR*, *CNGB3*, *GNAT2*, *GNGT2*, *GRK7*, *GUCA1B*, *OPN1LW*, *OPN4*, *OPSD*, *PDE6C*, *PDE6H*, *SWS1*) are ordered in each rectangle as depicted in the legend. Purifying selection, neutral evolution and positive selection are characterized by $\omega < 1$, $\omega = 1$, and $\omega > 1$, respectively. The Javan rhinoceros, *Rhinoceros sondaicus*, lacked sequenced data for the *PDE6H* locus, which are denoted by blank boxes. Some omega values are infinite due to the denominator in the dN/dS ratio equaling zero.....60

Figure 3.1. Maximum likelihood gene tree of *UCP1*, *UCP2*, and *UCP3* coding sequences (N=448) modified from Gaudry et al. (2017) to include the sixteen additional species with recently available genome projects (see Appendix 5). The stem placental mammal branches are indicated in blue. Note that the *UCP1* stem placental branch is much longer than those of *UCP2* and *UCP3*, demonstrating a greater number of nucleotide substitutions per site. Placental mammal genes are highlighted with blue boxes. The tree was rooted with the western clawed frog (*Xenopus tropicalis*) *UCP3*.....80

Figure 3.2. Schematic of the murid *UCP1* enhancer with putative transcription factor binding motifs shown for the rat (green) and mouse (blue) based on a combination of previous studies (see text for details). Regions of overlap between adjacent transcription factor motifs are underlined.....86

Figure 3.3. Dot plot comparison of the gray short-tailed opossum *UCP1* exon 1 versus a section of the platypus *UCP1* gene occurring between *TBC1D9* and *ELMOD2* (accession number: NW_001794248.1). Sequence alignments of the platypus (top) and gray-short tailed opossum (bottom) are provided with the potential coding sequences indicated in bold; putative splice sites are underlined. Note that two regions within the platypus clearly display homology to the opossum exon 1 (199-226 and 400-520), suggesting the presence of a 186 bp insertion in the platypus exon 1 sequence. The blue shaded area represents the region where an automated predictor program, which created a seven exon *UCP1* gene for the platypus, placed a 30 bp 'exon 1' in order to obtain an open reading frame free from premature stop codons (accession number: XM_001512650), though this region shares no homology with exon 1 of the opossum. The original platypus start codon also appears to be mutated to AAG (red font), with the predicted platypus 'exon 2' occurring 6 bp downstream of the "ATG" start site in the opossum. Note that these

differences between the two species likely arise from a misassembly error in the platypus (see text for details).....95

Figure 3.4. *UCP1* basal promoter elements alignment for select mammalian species with putative protein binding motifs indicated. Highlighted sites indicate shared nucleotides to the species in which the motif was first described (mouse or rat) and the typical TATA box (5'-TATAAAA-3') sequence (Xu et al. 1991). The consensus sequence represents the simple majority based on species for which the *UCP1* gene is intact. Species with documented *UCP1* pseudogenes (Gaudry et al. 2017) are denoted in red font and were not included in the consensus calculations.....98

Figure 3.5. Sequence identity comparisons of the *UCP1* genes of the rat, cow, pangolin, armadillo, bottlenose dolphin, and killer whale versus the human. All DNA sequences are shown 5' (left) to 3' (right). *UCP1* exons 1-6 are denoted with orange rectangles while *UCP1* upstream transcriptional regulatory elements are denoted in light blue (enhancer box, putative regulatory region, CpG island; from left to right). Gaps in sequence coverage are represented by white rectangles. Notably, the putative regulatory region is absent in the rat, but conserved in the cow. Upstream regulatory elements also appear to have been deleted in the Javan pangolin and armadillo, which have deletions of *UCP1* exons 1-2, and 3-5, respectively. Deletion of the entire *UCP1* gene between *TBC1D9* (yellow arrows) and *ELMOD2* (green arrows) has occurred in bottlenose dolphin and killer whale ~8-15 MYA (Gaudry et al. 2017) and included the upstream regulatory elements. Sequence identity percentage is represented with a color scale.....100

Figure 3.6. Maximum likelihood *UCP1* gene tree illustrating substitution rates in several eutherian lineages (eulipotyphlans, canids, afroinsectiphilians, vesper bats, myomorph rodents; boxed in blue) that are comparable or higher than lineages with *UCP1* pseudogenes (denoted in red). Branch lengths represent the number of nucleotide substitutions per site.....104

Figure 3.7. A maximum likelihood species tree based on 41 gene segments (50,911 base pairs) that illustrates the gain and loss of known *UCP1* regulatory elements (CpG island, PRR, TATA box, enhancer) through the evolutionary history of Mammalia. Red branches indicate lineages with a non-functional *UCP1* gene (Gaudry et al. 2017).....109

LIST OF ABBREVIATIONS

aDNA	ancient DNA
a.k.a.	also known as
ASTRAL-II	accurate species tree algorithm version 2
BAT	brown adipose tissue
BLAST	Basic local alignment search tool
bp	base pair
BRE	brown adipocyte regulatory element
BSA	bovine serum albumin
C/EBP	CCAAT-enhancer-binding protein
cAMP	cyclic adenosine monphosphate
CDS	coding sequence
CpG	5'–cytosine–phosphate–guanine–3' dinucleotides
CRE	cAMP response element
CREB	cAMP response element binding protein
°C	degrees Celcius
DNA	deoxyribonucleic acid
dN	non-synonymous substitutions per non-synonymous site
dS	synonymous substitutions per synonymous site
dNTP	deoxynucleotide
dnTRE	down thyroid hormone response element
DR1	direct repeat with 1 spacer nucleotide
DR3	direct repeat with 3 spacer nucleotides
DR4	direct repeat with 4 spacer nucleotides
EDTA	ethylenediaminetetraacetic acid
GARLI	Genetic Algorithm for Rapid Likelihood Inference
gDNA	genomic DNA
GTR	general time reversible
HS	high sensitivity
ILS	incomplete lineage sorting
ISPs	Ion Sphere™ Particles
kb	kilobase (1000 base pairs)
kg	kilogram
M	molar
m	meter
mM	millimolar
ml	milliliter
MDA	multiple displacement amplification
mRNA	messenger ribonucleic acid
MUSCLE	multiple sequence comparison by log-expectation
MYA	million years ago
NBRE	nerve growth factor response element
NCBI	National Center for Biotechnology Information
NF-Y	nuclear a transcription factor Y

ng	nanogram
NGS	next-generation sequencing
NR4A	nuclear receptors 4A
NST	non-shivering thermogenesis
NUMT	nuclear mitochondrial DNA segment
ω	non-synonymous:synonymous nucleotide substitution ratio
PAUP*	Phylogenetic Analysis Using Parsimony
PCR	polymerase chain reaction
pg	picogram
PGM	Personal Genome Machine
pM	picomolar
PPARγ (PPARG)	peroxisome proliferator-activated receptor γ
PPRE	peroxisome proliferator response element
PRR	putative regulatory region
RAxML	Randomized Axelerated Maximum likelihood
RAR	retinoic acid receptor
RARE	retinoic acid response elements
RF	Robinson-Foulds
RNA	ribonucleic acid
rpm	revolutions per minute
RXR	retinoid X receptor
RXRα	retinoid X receptor α
SRA	sequence read archive
SVDQuartets	Singular Value Decomposition scores for species quartets
TBP	TATA-binding protein
TFIID	transcription factor IID
U	units
UCP1	uncoupling protein 1
µg	microgram
µl	microlitre
URE1	UCP regulatory element 1
WGA	whole genome amplification
WGS	whole genomoe shotgun

CHAPTER 1: GENERAL INTRODUCTION

In Chapter 2 of this thesis I construct a 131,931 base pair (bp) molecular data set composed of both protein-coding and non-coding nuclear DNA to examine the unresolved phylogenetic relationships among six species within the family Rhinocerotidea. The five extant rhinoceros species—white (*Ceratotherium simum*), black (*Diceros bicornis*), Indian (*Rhinoceros unicornis*), Javan (*Rhinoceros sondaicus*), and Sumatran (*Dicerorhinus sumatrensis*) rhinos—along with the extinct woolly rhinoceros (*Coelodonta antiquitatis*) were included in this study. With recent advances in molecular biology it is now feasible to recover and sequence ancient DNA from extinct species such as the woolly rhinoceros, providing a window into the past that was previously unattainable.

Ancient DNA (aDNA; i.e. DNA recovered from archeological, museum, or fossil specimens) research is fraught with challenges (Hofreiter et al. 2001a; Pääbo et al. 2004). Contamination is a major cause for concern as bacteria and fungi typically invade ancient tissue samples to play their role in the decomposition process, leaving behind traces of microbial DNA. Moreover, when ancient specimens are handled by mankind, they frequently become contaminated by modern human DNA (Hofreiter et al. 2001a). One recently developed molecular technique that can be used to help detect and minimize contamination artifacts is hybridization capture. This method utilizes single stranded RNA or DNA probes (a.k.a. "baits") designed to bind and enrich complementary target sequences within DNA libraries (Horn 2012). Bait molecules that have been hybridized to target strands can be immobilized using either a solid-phase or in-solution methodology and non-target DNA that failed to hybridize to a probe is then washed

away, increasing the proportion of targeted fragments in the library and reducing exogenous DNA contamination in the sequenced reads (Horn 2012). Another major obstacle with aDNA is that endogenous nucleases remain active following the death of the organism, acting to cleave and break down genomic DNA. Furthermore, micro-organisms within decaying tissue act to digest aDNA (Allentoft et al. 2012). These factors impose a molecular time limit of survival as the oldest aDNA recovered to date is from a ~560,000-780,000 year ago permafrost preserved horse bone (Orlando et al. 2013). Indeed, nuclease and microbial activity can be slowed under certain environmental conditions such as desiccation, high salinity, or cold temperatures (Allentoft et al. 2012; Hofreiter et al. 2001a), however, aDNA is still typically highly fragmented with the majority of endogenous strands being <100 bp (Poinar et al. 2006). This fragmented nature of aDNA reduces the feasibility of sequencing using the classical polymerase chain reaction (PCR) and Sanger sequencing approach due to both time and economic constraints. Fortunately, recent advances in next-generation sequencing (NGS) allow for extremely cost-effective sequencing of DNA fragments that can be as short as 35 bp. However, aDNA is also problematic in that it is typically damaged by chemical modifications of nucleotides that accumulate in the absence of cellular correction mechanisms that arrest with the death of the organism. The most common form of aDNA damage is hydrolytic deamination of cytosines, altering these nucleotides to uracils. Once fragments containing this form of damage are amplified using polymerases, C→T and G→A misincorporations are introduced (Pääbo et al. 2004). Thus, high sequence coverage, preferably from multiple specimens, is required to distinguish between true C→T and G→A transitions and those that are likely the result of DNA damage (Hofreiter

et al. 2001b). NGS is greatly beneficial in this regard as this technology makes it possible to acquire millions of sequencing reads from multiple DNA libraries in a single run.

Extant rhinoceroses are adapted to cope with warm climates of their tropical and sub-tropical habitats as their thick armour-like skin contains sweat glands (Hiley 1977) and highly vascularized skin folds (Endo 2009) that function to promote dissipation of excess heat. Numerous behavioural adaptations, such as reducing activity during the daytime heat, seeking shade, and wallowing in mud, also aid in thermoregulation (Hutchins and Kreger 2006). By contrast, the extinct woolly rhinoceros endured the extreme cold of the Pleistocene Ice Ages at northern Eurasian latitudes until the last members of this lineage died out ~10,000 years ago. With such polarizing ecological niches exploited by these species, the family Rhinocerotidea presents an excellent model system for investigating the molecular underpinnings of cold-tolerance, much like those previously explored among elephantids and sirenians (Campbell et al. 2010; Springer et al. 2015). However, when performing comparative analyses, having a well-established evolutionary history is key for deducing common ancestry underlying the origin of biological traits.

The field of phylogenetics aims to reconstruct evolutionary relationships among taxa using heritable characteristics as indicators of common ancestry. Early studies primarily utilized morphological features (e.g. bone/skull anatomy of modern and fossil specimens) as markers. A downside of retrieving morphological data is that it can be extremely labor-intensive and time-consuming (Hillis 1987). Moreover, this approach can be confounded by physical features that arise via convergent evolution (Gaubert et al. 2005). For example, feeding specializations arose independently multiple times among

rhinoceroses, some species being grazers (e.g. Indian, white, and woolly rhinoceroses) while others being browsers (e.g. Javan, black, and Sumatran rhinoceroses) (Prothero et al. 1993), thus anatomical features associated with these niche-dependent specializations (e.g. dental, skull, mouth/lip morphology) may not be reflective of common ancestry. With molecular sequence data now relatively easily accessible in large quantities, typically providing a high number of phylogenetically informative characters, many modern studies have shifted to using DNA or protein sequences as indicators of phylogenetic ancestry (Gaubert et al. 2005; Hillis 1987). Patterns of convergent evolution can still pose a problem for molecular phylogenies, however, they are typically less affected by environmental factors (e.g. exploitation of similar ecological niches) than morphological characteristics (Hillis 1987). Morphological studies greatly benefit from thier ability to incorporate extinct taxa as fossil specimens, however, as detailed above, recent advances in methods of recovering aDNA also allows for molecular data to be recovered from extinct species. Previous studies examining the evolutionary relationships among these six rhinoceros species have been plagued by contradictory findings, thus my primary goal in Chapter 2 of this thesis was to establish a robust phylogeny by assembling and analyzing the largest molecular data set of nuclear DNA sequence to date.

Several genes that I utilized as molecular markers for phylogenetic analyses play a key role in visual processes (e.g. phototransduction) and are expressed in the retina. This provided an excellent opportunity in Chapter 2 to also briefly examine the functionality of these genes among rhinoceroses, which are known to have poor vision (Nowak 1999; Skinner and Chimimba 2005). Indeed, it is possible that inactivations of

these loci could underlie nearsightedness among members of this clade as mutations to some of these same genes have already been discovered to be associated with the loss of eyesight in subterranean mammals (e.g. star-nosed mole [*Condylura cristata*], naked mole-rat [*Heterocephalus glaber*] and Cape golden mole [*Chrysochloris asiatica*]) (Emerling and Springer 2014). Using selection pressure analyses developed by Yang (2007) that estimate the non-synonymous (dN) to synonymous (dS) nucleotide substitution ratio (ω) it is possible to assess modes of evolution acting upon coding sequences of rhinoceros eye genes. Under purifying selection ($\omega<1$) the ratio of synonymous substitutions outweighs the ratio of non-synonymous substitutions, thus amino acid residues within the primary sequence of the protein are highly conserved. The opposite occurs under positive selection ($\omega>1$), where the ratio of non-synonymous substitutions is greater than the ratio of synonymous substitutions, promoting functional change of the protein as a high proportion of nucleotide mutations result in amino acid replacements. Signatures of neutral evolution ($\omega=1$) are indicative of non-expressed or inactivated genes where neither non-synonymous nor synonymous substitutions are favored, and the nucleotides are instead mutating randomly in the absence of natural selection (Yang 2007). This latter scenario would be expected for rhinoceros eye genes that are not expressed or inactivated and potentially underlie poor eyesight in this lineage.

In addition to poor eyesight, an interesting aspect of rhinoceroses is their massive body size. Indeed, the white rhinoceros is among the largest of all extant terrestrial animals (1,400 – 2,300 kg; Silva and Downing 1995), only surpassed by modern elephants (i.e. *Loxodonta africana* and *Elephas maximus*). Body size has substantial biological implications as it strongly correlates with generation time, longevity, and

metabolic rate (Bromham 2009; Martin and Palumbi 1993). Furthermore, the allometric link between surface area and body size dictate that large-bodied species retain heat more efficiently than smaller species, thus body size is inevitably tied to thermoregulation (McNab 1983). In the examination of 133 mammalian species, a previous study by colleagues and myself (Gaudry et al. 2017) demonstrated an example of this latter relationship by detailing independent inactivations of uncoupling protein 1 (UCP1), the integral effector protein of eutherian thermogenic brown adipose tissue (BAT), in several mammalian taxa (elephantids, sirenians, hyraxes, cetaceans, and equids), temporally coinciding with the evolution of increased body size in each of these lineages. Additionally, *UCP1* is pseudogenized in lineages that exhibit drastic reductions in metabolic intensity (Pilosa [sloths, anteaters], Cingulata [armadillos] and Pholidota [pangolins]) (Gaudry et al. 2017). BAT is the main contributor of non-shivering thermogenesis (NST) in eutherian mammals and its heat producing capabilities rely upon a mechanism known as mitochondrial proton leak catalyzed by high tissue-specific expression of UCP1. Thus, we concluded that while BAT-mediated NST was long-believed to be the single most important eutherian adaptation to cold environments, driving niche expansion and facilitating exploitation of northern habitats (Cannon and Nedergaard 2004), it became obsolete in various large-bodied lineages, resulting in *UCP1* gene inactivation. Indeed, these findings fit with the previous studies noting the absence of BAT in members of each of these lineages (Rowlatt et al. 1971) and a prediction by Heldmaier (1971) that BAT would provide negligible (if any) thermal benefits in species larger than 10 kg. Yet, several large-bodied lineages, including rhinoceroses, camels, pinnipeds, and the hippopotamus do not fit this pattern of

inactivation, instead retaining an intact *UCP1* coding sequence. However, with BAT expression levels largely unexplored in these species, it is conceivable that they have little or no use for BAT-mediated NST and thus, may not express UCP1. Therefore, in Chapter 3 of this thesis I hypothesized that while some large-bodied lineages retain an intact *UCP1* coding sequence, neutral evolution may have lead to the accumulation of mutations among regulatory elements postulated to modulate the transcription of this gene, potentially hindering or precluding *UCP1* expression.

Gene transcription, the process of synthesizing complementary precursor mRNA strands from DNA, can be controlled in a variety of ways. In eutherians mammals, *UCP1* transcription has been proposed to be modulated by multiple DNA elements including an upstream enhancer, promoter, and CpG island (Shore et al. 2013; Villaroya et al. 2017). Generally, enhancers contain DNA motifs that bind transcription factors and encourage gene expression by stabilizing promoter-bound proteins involved in the initiation of transcription (Pennachio et al. 2013). An enhancer can influence a single gene or multiple genes (the latter scenario is referred to as pleiotropism) and can be situated along a chromosome at highly variable distances relative to their target genes in either the upstream or downstream direction. Three-dimensional loop conformations within the DNA permit enhancer-promoter interactions despite these regions potentially being separated by up to 1,000,000 bp (Pennachio et al. 2013). In contrast to enhancers, promoters are required to be located in close proximity to the transcription initiation site (i.e. immediately upstream of the gene). These regions contain DNA sequence motifs functioning to bind proteins (transcription factors) that interact with RNA polymerase II to begin synthesis of a complementary RNA molecule (Smale and Kadonaga 2003). For

example, the transcription factor IID (TFIID) protein complex functions to recruit RNA polymerase II to the initiation site and is composed in part by the TATA-binding protein (TBP) subunit that recognizes the TATA box, a common motif within the promoter of many genes (Lee and Young 2000; Smale and Kadonaga 2003). Another method of modulating gene expression is through DNA methylation. CpG islands are present in several genes, typically located near the 5' promoter, these regions are GC-rich with frequent CpG dinucleotides (5'–cytosine–phosphate–guanine–3'), and act as DNA methylation sites that can block RNA synthesis by directly impeding binding of proteins that participate in transcription (Bird 2002). Also, these sites can interact with methyl–CpG–binding proteins, which repress transcription by promoting chromatin condensation by histones, causing genes to become inaccessible to the transcriptional machinery required for expression (Bird and Wolffe 1999).

Chapter 3 of this thesis was intended as a continuation of my previous research (Gaudry et al. 2017) utilizing a comparative approach to examine the evolution of upstream elements that putatively modulate *UCP1* transcription among, not only rhinoceroses, but a total of 139 mammals. Previous studies have primarily focused on murid rodents (mice and rats) to describe possible protein-binding motifs within the *UCP1* enhancer and promoter regions (see Chapter 3). I aimed to examine whether or not these same motifs are universally conserved among eutherian mammals. Moreover, I aimed to determine if *UCP1* transcriptional regulatory regions have become deteriorated within lineages with *UCP1* pseudogenes and possibly even large-bodied species with an intact *UCP1* coding sequence (CDS). With UCP1 currently under extensive medical research as a possible method for combating obesity and diabetes (Feldmann et al. 2009;

Ishigaki et al. 2005), a deeper understanding of the evolution of transcriptional mechanisms controlling its expression may prove to be highly beneficial in achieving its full potential as a therapeutic treatment.

CHAPTER 2: MOLECULAR PHYLOGENETICS OF THE RHINOCEROS CLADE

Michael J. Gaudry, Anthony V. Signore, Nathan Wales, M. Thomas P. Gilbert, Eske Willerslev, Peter J. van Coeverden de Groot, Mark S. Springer, and Kevin L. Campbell

Author contributions:

MJG conceived of the project, performed experiments, interpreted the results, prepared figures, drafted the manuscript.

AVS constructed DNA libraries.

NW constructed DNA libraries.

MTPG provided laboratory space and tissue samples.

EW provided laboratory space and tissue samples.

PJVCDG provided tissue samples.

MS provided DNA samples and commented on the manuscript.

KLC assisted in designing the experiments, helped interpret the results and reviewed the manuscript.

2.1. Abstract

Evolutionary relationships among the three rhinoceros subfamilies (Dicerorhininae, Rhinocerotinae, and Dicerotinae) are unresolved despite numerous morphological and molecular phylogenetic studies. Here, I aimed to resolve this family tree utilizing hybridization capture techniques and next-generation sequencing to assemble a multi-locus 131 kb dataset of nuclear coding and non-coding regions from a

Malayan tapir and six rhinoceros species, including the extinct woolly rhinoceros. Estimated phylogenies from concatenation and coalescent analyses of the entire dataset, and concatenation of intron versus exon supermatrix subsets, yielded conflicting topologies linking either Asian versus African ((Dicerorhininae, Rhinocerotinae) Dicerotinae) or one- versus two-horned rhinoceroses (Rhinocerotinae (Dicerorhininae, Dicerotinae)). Incomplete lineage sorting and speciation events in quick succession likely account for these conflicting hypotheses. However, five of six discovered intronic indels support the latter relationship. Additionally, selection pressures were estimated for eye genes and revealed no evidence of pseudogenization despite notoriously poor vision among rhinoceroses; however, near-neutral evolution was found in the white rhinoceros for *GUCA1B* and *OPN4* genes.

2.2. Introduction

2.2.1. General introduction

Rhinoceroses are under severe threat from anthropogenic habitat destruction and relentless poaching for their highly valued horns used in traditional Asian medicines (Ferreira et al. 2015; Prothero 1992, Ripple et al. 2015). Population levels of the five remaining rhinoceros species have been decimated and for two species in particular, only dozens of individuals remain in the wild (Hariyadi et al. 2016). With these megaherbivores in peril, it is of utmost importance that we collect as much information as possible while extant members still exist.

One rhinoceros species that suffered recent extinction is the woolly rhinoceros, *Coelodonta antiquitatis*. In contrast to its tropically distributed cousins, the woolly rhinoceros was adapted to endure extreme cold temperatures permitting it to thrive in

high northern latitudes during the Pleistocene Ice Ages (Deng et al. 2011). Its success was reflected by its broad distribution across northern Eurasia, from Siberia to Spain, until the last individuals died off ~10,000 years ago (Kahlke and Lacombat 2008). By exploiting such a strikingly different ecological niche to modern rhinoceroses, this iconic species provides a fascinating opportunity to study adaptations to cold environments; however, before meaningful evolutionary comparisons can be made, the phylogenetic groundwork needs to be established.

Phylogenetics, the study of evolutionary affiliation between extant and/or extinct organisms, plays a fundamental role in our understanding of all aspects of biology. Remarkably, the evolutionary history of modern rhinoceroses is still unclear with contradictory topologies and quick radiations evidenced by previous molecular studies (Price and Bininda-Emonds 2009; Orlando et al. 2003; Steiner and Ryder 2011; Tougard et al. 2001; Willerslev et al. 2009). However, sister species relationships are strongly supported between the extinct woolly rhinoceros and the Sumatran rhinoceros (*Dicerorhinus sumatrensis*), between the black (*Diceros bicornis*) and white rhinoceroses (*Ceratotherium simum*) and lastly, between the Indian (*Rhinoceros unicornis*) and Javan rhinoceroses (*Rhinoceros sondaicus*) (Orlando et al. 2003; Willerslev et al. 2009). This study aims to elucidate the relationships between the five extant rhinoceroses and the extinct woolly rhinoceros using nuclear DNA sequence data.

2.2.2. Perissodactyl evolution

The family Rhinocerotidae (rhinoceroses), along with the families Tapiridae (tapirs) and Equidae (horses), make up the order Perissodactyla, characterized in part by their mesaxonic foot, odd number of toes, and hindgut fermenting digestive system

(Steiner and Ryder 2011; Prothero and Schoch 1989). Although modern artiodactyls are the dominant large-bodied grassland herbivores today, perissodactyls of the Eocene, Oligocene, and Miocene were more successful and prevalent than the even-toed ungulates (Prothero 1992) and ranged throughout North America, Eurasia, and Africa (Prothero et al. 1989). The first perissodactyl fossils originate from the early Eocene, though the earliest members are thought to date back to the late Paleocene (Radinsky 1969). Current phylogenetic hypotheses place rhinoceroses as a sister group to tapirs forming the suborder Ceratomorpha, that last shared a common ancestor some 51.8 million years ago (MYA), with horses (suborder Hippomorpha) diverging ca. 4.8 million years earlier based on molecular clock estimates from Meredith et al. (2011). However, a recently discovered fossil in India places the oldest known tapiromorph at ~53.7 MYA (Kapur and Bajpai 2015).

2.2.3. Rhinoceros evolution

An abundance of prehistoric rhinoceros fossils have been painstakingly uncovered, showing that rhinoceroses were once much more diverse and widespread than they are today (Cerdeño 1998; Prothero 1992). The most primitive known rhinocerotoid genus *Hyrachyus* (Superfamily Rhinocerotoidea), appears in the fossil record by the early Eocene and was tapir-like with teeth adapted to masticate leaves and long legs typical of an effective runner (Prothero 1992). *Hyrachyus* species were long thought to be ancestral tapirs (Cope 1873, Radinsky 1966); however, they are now placed as basal rhinoceroses substantiated mainly upon dental morphology (Prothero et al. 1986; Domning et al. 1997). These animals were approximately 1.3 m in length, roughly the size of an adult

sheep (Cope 1873), and lacked horns like many of its early rhinoceros descendants (Prothero et al 1993). *Hyrachyus* is thought to have utilized land bridges connecting North America, Europe, and Asia to disperse throughout northern latitudes of all three continents. Interestingly, fossils of this genus have been found on the Island of Jamaica (Domning et al. 1997) and as far north as Elsemere Island (Prothero 1992). However, by the middle-late Eocene, continental drift leading to geographic isolation (e.g. breaking of the European-North American land bridge) and climate change spurred these early rhinoceroses to diverge into three distinct families: Amynodontidae, Hyracodontidae, and Rhinocerotidae (Prothero et al. 1989, Prothero 1992).

The amynodonts resembled modern day hippopotamuses with some species having tapir-like proboscises for grasping food. They thrived in Asia and North America during the late-Eocene but diminished towards the Oligocene with one genus, *Cadurocotherium*, reaching Europe prior to the extinction of the entire amynodont lineage by the mid-Miocene (Wall 1980; Prothero 1992). Like the amynodonts, the hyracodonts were highly successful during the mid-late Eocene and into the Oligocene in North America and Asia. This family consisted primarily of species that were approximately the same size as *Hyrachyus spp.* with elongated legs well suited for running (Prothero 1992). However, one lineage within this family, the Indricotheres, reached enormous proportions with *Peraceratherium* being the largest terrestrial mammal to ever walk the earth (Prothero 1992). *Peraceratherium* occurred throughout Asia, weighed ~20,000 kg, and had a long neck similar to modern day giraffes that allowed it to forage on treetops and sustain such an enormous body size (Prothero et al. 1989;

Prothero 1992). The Indricothere subfamily was the last lineage of hyracodonts and thrived until the mid-Miocene before its extinction (Prothero 1992).

Of the three rhinoceros families, the Rhinocerotidae is the only one to persist to the present day. *Teletacera*s was the first member of the Rhinocerotidae appearing in the fossil record in the late Eocene and lacked horns, as did the amynodonts and hyracodonts (Cerdeño 1998; Prothero 1992). *Diceratherium*, a North American genus, was the first group of rhinoceroses to evolve horns, though these did not bear much resemblance to those of modern rhinoceroses as they occurred in pairs located side by side on the upper lateral portion of the snout (Prothero et al. 1989, Prothero 1992). As is reflected by modern grazing and browsing rhinos, early members of the Rhinocerotidae originated in the late Oligocene and had adaptations to suit their specific diets. Teleoceratines of North America were specialized for grazing on grass given their high-crowned continuously growing teeth and broad lips (similar to those of modern white rhinoceros) as well as large graviportal bodies (Prothero 1992). In contrast, many species of the aceratherine lineage appearing in Eurasia prior to spreading to North America were specialized browsers exhibiting low-crowned teeth and either prehensile lips or a proboscis specialized for plucking leaves, a feature common to the modern black rhinoceros and even modern tapirs. During the Miocene, the Rhinocerotidae attained peak diversity (Cerdeño 1998) but by the end of this epoch, both the teleoceratines and aceratherines went extinct presumably due to climatic shifts leading to the glaciation events and ultimately the Ice Ages (Prothero 1992). This marked the end of rhinoceroses in North America, however, some European species persisted, giving rise to modern rhinoceros

that are now classified into three subfamilies: Dicerorhininae, Rhinocerotinae, and Dicerotinae (Guérin 1980, Guérin 1982, Groves 1983, Prothero 1989).

The first definitive member of the Dicerorhininae lineage was *Dicerorhinus sansaniensis* (Prothero 1992) and primitive dicerorhine fossils dating to the late Oligocene to early Miocene have been found throughout Europe, eastern Africa, and southern Asia (Zin-Maung-Maung-Thein et al. 2008). While most dicerorhines were likely warm-adapted, forest-dwelling browsers such as the Sumatran rhinoceros (*Dicerorhinus sumatrensis*), one of the most fascinating members of this lineage was the iconic cold-adapted woolly rhinoceros (*Coelodonta antiquitatis*). In contrast to most members of this lineage, the woolly rhino possesses a relatively long and thin skull with eyes positioned more posteriorly, a wide snout, and a dental morphology indicative of an efficient grazer (Kahlke and Lacombat 2008). Fossils of the most ancestral members of the genus *Coelodonta* have been found in the mountains of Tibet dating back ~3.7 MYA (Deng et al. 2011). At least one of the woolly rhinoceros' cold-tolerant adaptations predates the Pleistocene, which include morphological features of the skull that facilitated sweeping motions. As suggested by wear patterns on the leading edge of fossilized ~1 m long bilaterally flattened anterior horns (Fortelius 1983; Kahlke and Lacombat 2008), the woolly rhino likely swept away snow to provide access to edible vegetation (Deng et al. 2011). Thus, the Tibetan environment may have spurred the pre-adaptation of woolly rhinos to cope with the extreme cold of the Ice Ages that began ~2.8 MYA (Deng et al. 2011). With a thick fur coat, large body size (~2000 kg), a stocky graviportal build with short limbs and ears to reduce heat loss during the cold climates, the aptly named woolly rhinoceros was once found throughout northern Eurasia (Boeskorov et al. 2011; Deng et

al. 2011; Prothero 1992). In fact in northern Siberia, several mummified woolly rhinoceros carcasses have been found preserved in the permafrost with soft tissues remaining intact (Boeskorov et al. 2001; Boeskorov et al. 2011). Unlike the woolly mammoths (*Mammuthus primigenius*), however, the woolly rhinoceros never crossed Beringia into North America (Prothero 1992). As exemplified by cave paintings throughout Europe (Guérin 1989; Loose 1975; Orlando et al. 2003), humans co-existed with woolly rhinos and hunted them, contributing to their eventual extinction ~10,000 years ago coupled with the effects of climate change (Lorenzen et al. 2011; Stuart and Lister 2012).

The Sumatran rhinoceros, is the only extant dicerorhine and the closest living relative to the extinct woolly rhinoceros. In contrast to the large woolly rhinoceros, the Sumatran rhino is the smallest extant rhinocerotid (900-1,000 kg; Silva and Downing 1995) and a browser. However, these two species share several morphological characteristics suggestive of their close ancestry. For instance, in addition to numerous ossification and dental features (Guérin 1980; Guérin 1989; Orlando et al. 2003), both species exhibit two horns, and the Sumatran rhino, with a coat of fur especially evident in newborns, is the hairiest extant rhinoceros (Willerslev et al. 2009). Once having a much larger distribution spreading throughout Southeast Asia, the Sumatran rhinoceros has been heavily poached for its horn and now occupies only 1% of its former territory, relegated to a handful of small populations on the islands of Sumatra in Indonesia and Borneo (Zin-Maung-Maung-Thein et al. 2008, Havmøller et al. 2016). With a 2009 population study estimating only 200 individuals remaining in the wild (Zafir et al. 2011), this species is in severe peril of becoming extinct.

Fossils of the earliest known rhinocerotinine genus, *Gaindatherium,* have been discovered in Portugal and date to the early Miocene (Prothero 1992). Modern rhinocerotinines, which include the Indian and the Javan rhinoceros, make up the genus *Rhinoceros.* Although the Indian rhinoceros is a grazer and the Javan rhinoceros is a browser, they are similar in body size, weighing 1,410 – 2,000 kg and 1,500 – 2,000 kg, respectively (Silva and Downing 1995). These two species share several morphological features including dental and ossification patterns, thick armour-like skin folds and, perhaps most prominently, a single horn, while all other extant rhinos possess two horns. The Indian rhinoceros inhabits regions of India and Nepal (Foose and van Strien 1997) with a population estimated at 2,800 individuals (Talukdar 2009). Despite continued poaching, the population numbers are slowly on the rise due to successful conservation efforts to protect this species within parks such as Kaziranga National Park in India (Cedric et al. 2016). The Javan rhinoceros, which suffered a recent extirpation from Vietnam in 2010 (Brook et al. 2011; Brook et al. 2014), is now restricted to only a small region on the Island of Java, Indonesia and is perhaps the most critically endangered mammal in the world with only ~50 surviving individuals (Hariyadi et al. 2016).

Fossil remains of the oldest known dicerotine, *Paradiceros mukirii*, date back to the mid-Miocene (~18 MYA) in Kenya and Morocco (Hooijer 1968; Hooijer 1978; Prothero et al 1993). The modern members of the Dicerotinae lineage are the black and white rhinoceroses of Africa. The black rhino weighs 816 – 1,300 kg (Silva and Downing 1995) and is a specialized browser ranging across South-east Africa. This species is also known as the hook-lipped rhinoceros for its specialized prehensile lip that is able to pluck vegetation from bushes and trees. In contrast, the white rhinoceros is much larger,

weighing 1,400 – 2,300 kg (Silva and Downing 1995), and is the third largest extant land mammal behind African and Asian elephants. This species is sometimes referred to as the square-lipped rhinoceros as its mouth is well adapted for grazing, and is distributed across Southern Africa. Despite their differences in diet, the black and white rhinos share several characteristics, the most obvious being their skull morphology (Groves 1983), possession of two horns, and their use of horny pads to masticate vegetation as they lack incisors (Groves 1983; Prothero et al. 1986). Despite continued poaching, conservation efforts have allowed white and black rhinoceros populations to respectively increase from ~6,000 and 2,500 in the early 1990s to ~17,475 and ~4,230 individuals by 2007 (Milliken et al. 2009).

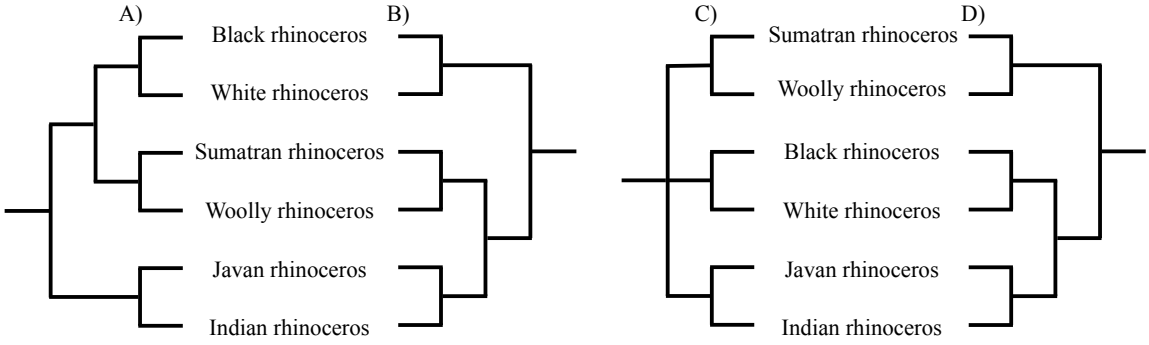
With rhinoceroses in peril of becoming extinct in the near future, it is imperative that we collect as much information as possible about these animals. Species-level phylogenetic studies bear a fundamental role in comparative biology because they provide the evolutionary framework needed to infer meaningful comparisons. While multiple previous studies, elaborated in the following section, have utilized molecular data in an effort to resolve the relationships among modern rhinoceroses, limited success has been achieved.

2.2.4. Previous studies of rhinoceros evolution

Three leading hypotheses prevailed from numerous early morphological studies attempting to resolve the evolutionary relationships among extant rhinoceroses (Morales and Melnick 1994; Price and Bininda-Emonds 2009; Steiner and Ryder 2011). The "two-horn" (a.k.a. "number of horns") hypothesis proposes that the dicerotines and dicerorhines form a taxonomic clade based on their shared morphological characteristic

of having two horns, to the exclusion of the one-horned rhinocerotines (Figure 2.1A) (Simpson 1945; Loose 1975). By contrast, the "biogeographical" (a.k.a. "geographic split") hypothesis suggests that Asian rhinoceroses (dicerorhines and rhinocerhines) and African rhinoceroses (dicerotines) form distinct clades based on geographical distribution and shared morphological traits of the teeth, skull, and skeleton (Figure 2.1B) (Groves 1983; Pocock 1945; Prothero et al. 1986). Finally, the "separate lineage" hypothesis proposes that the dicerorhines, rhinocerotines, and dicerotines represent individual lineages with neither being more closely related to one group than the other, forming a trichotomy at the base of the Rhinocerotidae phylogeny (Figure 2.1C) (Guérin 1982; Cerdeño 1995; Prothero and Schoch 1989).

Figure 2.1. Illustrations of the possible phylogenetic relationships among the three rhinoceros subfamilies, including the "two-horn" (A), "biogeographical" (B), and "separate lineage" (C) hypotheses. The hypothesis proposed by Fernando et al. (2006) linking dicerotines (black and white rhinoceroses) and rhinocerotines (Javan and Indian rhinoceroses) to the exclusion of dicerorhines is also depicted (D).



Recent molecular studies have similarly resulted in inconclusive or contradictory phylogenies and provide no definitive resolution of the rhinoceros family tree. Restriction site mapping of mitochondrial ribosomal genes performed by Morales and Melnick (1994) supported the "two-horn" hypothesis of rhinoceros classification. Conversely, results from Tougard et al. (2001) supported the "biogeographical" hypothesis, grouping

Asian and African rhinoceroses into a separate clades based on sequence data from mitochondrial cytochrome b (*cytb*) and *12S rRNA* genes. Orlando et al. (2003) successfully sequenced *12S rRNA* and partial *cytb* markers from >40,000 year old fossilized woolly rhinoceros tooth roots marking the first published study to sequence woolly rhinoceros DNA fragments. They confirmed that the closest extant relative of the woolly rhinoceros is the Sumatran rhino, solidifying its position in the dicerorhine lineage. Their phylogenetic analyses further supported the separation of Asian and African rhinoceroses, a finding in line with the "biogeographical" hypothesis, though the stem branch lengths were extremely short, suggestive of a quick radiation. Interestingly, Fernando et al. (2006) proposed a topology incongruent with the three prevailing morphological hypotheses described above, placing the Sumatran rhinoceros as the most basal extant rhinoceros with the rhinocerotinines being sister to the dicerotines based on mitochondrial *12S rRNA* sequences (Figure 2.1D). However, these studies utilizing partial or complete mitochondrial genes were all superseded by Willerslev et al. (2009) who examined the entire mitochondrial genomes (~16,000 bp) of all five extant rhinos and the extinct woolly rhinoceros. The sequence data strongly supported the dicerorhines, rhinocerotinines, and dicerotines sister pairings. However, the mitochondrial genomes provided no statistical support to resolve the phylogenetic relationships between these three lineages, instead arguing that a trichotomy best characterizes the relationships between the three lineages of modern rhinoceroses, supporting the "separate lineage" hypothesis. Willerslev et al. (2009) further advised the use of nuclear markers to increase the resolution power in future analyses.

Clear differences have been reported by Springer et al. (2001) comparing the efficacy of mitochondrial vs. nuclear data in resolving mammalian phylogenetic relationships. Among protein-coding regions in mammals, mitochondrial DNA has a quicker substitution rate than nuclear DNA (Vawter and Brown 1986). Furthermore, in contrast to nuclear DNA, the mitochondrial genome is inherited in a haploid, maternal fashion without recombination between parental sources (Moore 1995). Springer et al. (2001) found that nuclear exonic markers were less obscured by multiple hits (also known as superimposed substitutions) where secondary mutations of a nucleotide revert the base to its ancestral state, masking a previously occurring single nucleotide polymorphism. Thus, nuclear coding regions reliably outperformed mitochondrial markers when resolving phylogenetic relationships that predate the Eocene (deep divergences; Springer et al. 2001).

To this end some nuclear markers were utilized by Price and Bininda-Emonds (2009), who performed genome mining techniques to assemble a 19,260 bp supermatrix comprised from 33 mitochondrial genes and 6 nuclear genes, making up about ~82% and ~18% of the complete dataset, respectively. Their phylogenetic results supported the "biogeographical" hypothesis, grouping Asian rhinoceroses separately from the African rhinoceroses.

Steiner and Ryder (2011) used only two mitochondrial markers (*12S rRNA* and *cytb*) and both coding and non-coding sequence data from nine nuclear markers (*BRCA1*, *EDNRB*, *Kit*, *MC1R*, *MITF*, *SNAI2*, *SOX10*, *TBX15*, *TYR*; see Appendix 1 for full gene names) to examine the phylogeny of perissodactyls including the Sumatran, white, black, and Indian rhinoceroses. Phylogenetic results from their mitochondrial dataset supported

the Sumatran rhino being a sister species to African dicerotines (concordant with the "two-horn" hypothesis), with the Indian rhinoceros diverging earlier. By contrast, the coding nuclear markers suggested that the Indian rhinoceros is the sister species to dicerotines, however this relationship was not strongly supported by bootstrap values. When mitochondrial and nuclear markers were concatenated, the Sumatran rhinoceros shared a common ancestor with dicerotines some 25 MYA, with the Indian rhino diverging ~1 million years earlier. Despite the seemingly contradictory findings between the mitochondrial and nuclear datasets, the authors concluded that the results from the concatenated supermatrix accurately reflected the true relationships among rhinoceroses.

Interestingly, recent research by Welker et al. (2017) analyzed ancient protein sequence retrieved from *Stephanorhinus sp.* (a rhinoceros lineage that went extinct during the Late Pleistocene), woolly rhinoceros, as well as modern rhinoceros specimens, which place *Stephanorhinus* sp. within the dicerorhininae. These amino acid sequences were added to the translated protein-coding DNA previously analysed by Steiner and Ryder (2011) and achieved a phylogeny supporting the "biogeographical" hypothesis.

The phylogeny of rhinoceroses remains debated with further research needed to elucidate the relationships among dicerorhines, rhinocerotinines, and dicerotines. Studies involving ancient DNA typically focus on mitochondrial DNA because of its higher abundance within fossils (Orlando et al. 2003; Binladen et al. 2006). Using hybridization capture techniques and next-generation sequencing (NGS) the current study sought to finally resolve the phylogeny of modern rhinoceroses and the extinct woolly rhinoceros using nuclear markers.

2.2.5. Rhinoceros eyesight

Many loci included in my phylogenetic analyses produce gene products that are involved in vision. All extant rhinoceroses have been long believed to have poor vision and are thought to instead rely mostly upon their strong auditory and olfactory senses (Skinner and Chimimba 2005). Providing evidence of their near-sightedness, Nowak (1999) showed that motionless humans and objects as large as vehicles did not elicit a response from the black rhinoceroses until within 20-30 meters. Interestingly, Emerling and Springer (2014) discovered that several retinal genes (e.g. *SWS1*, *GUCA1B*, *PDE6H*, *ARR3*, *PDE6C*, and *CNGB3*) are inactivated for subterranean mammals like the star-nosed mole (*Condylura cristata*), naked mole-rat (*Heterocephalus glaber*) and Cape golden mole (*Chrysochloris asiatica*), as they have drastically reduced or no need for eyesight. In contrast to functional genes that evolve under natural selection, inactivated genes (pseudogenized) evolve under a lack of selection pressure termed neutral evolution where mutations are accrued at random. Given the reputedly poor eyesight of rhinoceroses I hypothesized that genes involved in vision may have been inactivated or are evolving under neutral evolution in this lineage. Thus, coding regions of eye genes were examined for frameshift, nonsense, and splice site mutations and the modes of evolution acting upon these loci were characterized using selection pressure analyses.

2.2.6. Objectives

The objective of this study was to first sequence >100 kb of nuclear exonic and intronic targets using hybridization capture and NGS techniques from the Malayan tapir (*Tapirus indicus*) and four of the five extant rhinoceroses: the black, Indian, Sumatran, and Javan rhinoceroses. Secondly, I aimed to sequence the same genomic targets from

ancient DNA of the extinct woolly rhinoceros. My third goal was to acquire orthologous sequence data from publically available sources for the closely related white rhinoceros, four horses (the donkey [*Equus asinus*], Przewalski's horse [*Equus ferus przewalskii*], thoroughbred horse [*Equus ferus caballus* – thoroughbred breed] and Mongolian horse [*Equus ferus caballus* – Mongolian breed]), cow (*Bos taurus*), pig (*Sus scrofa*) and camel (*Camelus ferus*), with intentions of using these nucleotide sequences to resolve the phylogenetic relationships of dicerorhines, rhinocerotinines, and dicerotines. Lastly, I performed selection pressure analyses on eyesight genes to determine if these loci were evolving under neutral evolution, which would be consistent with pseudogenization in rhinoceroses with reputedly poor vision.

2.3. Materials and Methods

2.3.1. Sampling

The starting material used to create a DNA library for each individual tissue/DNA sample is summarized in Appendix 2. Previously extracted genomic DNA (gDNA) samples from one black and one Indian rhinoceros were acquired from Dr. Mark Springer (University of California, Riverside, California, USA). Two Javan rhinoceros bone samples were acquired from Dr. Peter van Coeverden de Groot (Queens University, Kingston, Ontario, Canada). Dr. Tom Gilbert (Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark) provided two blood samples from a single Malayan tapir, the blood meal of one leech that fed upon a Sumatran rhinoceros, and permafrost-preserved bone, tooth, or skin samples from five woolly rhinoceros individuals.

2.3.2. Construction of DNA libraries

2.3.2.1. Black and Indian rhinoceros DNA library preparation

To augment the quantity of DNA for the black and Indian rhinoceroses (~3 µl gDNA samples), whole-genome amplifications (WGAs) were completed using a REPLI-g Mini kit (Qiagen, Toronto, Ontario, Canada). These reactions were performed using 2.5 µl DNA template and a 16 hour incubation period at 30°C. The reactions were then purified using 3 volumes of homemade serapure magnetic beads (Rohland and Reich 2012) and eluted in 25 µl nuclease-free water. The gDNA samples were then enzymatically fragmented in reactions consisting of 15 µl DNA (~80 and ~230 ng DNA input for Indian and black rhino, respectively), 2 µl 10x buffer, 0.2 µl 100x bovine serum albumin (BSA), 2.0 µl NEBNext dsDNA Fragmentase (New England Biolabs, Whitby, Ontario, Canada), and 0.8 µl nuclease-free water. Fragmentation reactions were incubated in a MJ Mini Gradient Thermal Cycler (Bio-Rad Laboratories, Mississauga, Ontario, Canada) at 37°C for 18 minutes. Immediately afterwards, the fragmentase enzyme was inactivated with the addition of 5 µl 0.5 M ethylenediaminetetraacetic acid (EDTA). A 2 µl subsample of the reaction was subsequently electrophoresed on a 1.5% agarose gel to ensure that DNA fragments ranged ~100-600 bp in length. The reactions were then purified using 3 volumes of homemade serapure magnetic beads and eluted in 54 µl nuclease-free water.

A NEBNext Fast DNA Library Prep Set for Ion Torrent (New England Biolabs, Whitby, Ontario, Canada) was used to create DNA libraries. To prepare the DNA fragments for blunt-end adaptor ligation, overhanging nucleotides on 5’ and 3’ ends of the DNA fragments were removed and the 5’ ends were phosphorylated. Reactions were

prepared consisting of 6 µl NEBNext end repair reaction buffer, 3 µl NEBNext end repair enzyme mix, and 51 µl fragmented DNA, which were then incubated at 25°C for 20 minutes, 70°C for 10 minutes, and subsequently held at 4°C. Barcoded NEXTFlex™ adaptors (Appendix 2) were then ligated to the DNA fragments by adding 10 µl T4 DNA ligase buffer for Ion Torrent, 11.5 µl P1 adaptor, 11.5 µl barcoded adaptor, 1 µl Bst 2.0 WarmStart DNA polymerase, and 6 µl T4 DNA ligase (see Appendix 2 for barcode sequences). These reactions were incubated for 30 minutes at 25°C and 5 minutes at 65°C, then purified using 3 volumes of serapure magnetic beads and eluted in 25 µl nuclease free water. DNA fragments were then isolated according to their length using an E-gel iBase electrophoresis system (Invitrogen, Carlsbad, California, USA) and E-Gel SizeSelect SYBR Safe 2.0% agarose gels (Invitrogen, Carlsbad, California, USA). Briefly, 20 µl (totaling 100 ng) DNA was added to the input wells and electrophoresed alongside a 50 bp DNA ladder (Invitrogen, Carlsbad, California, USA). To account for the length of the adaptors when retrieving 100, 200, 400, and 600 bp fragments, the samples were withdrawn from the extraction wells when the DNA ladder indicated that the fragments were actually 150, 330, 550, and 790 bp, respectively. While libraries equal to or less than 200 bp were sequenced in this study, larger fragment sizes were stored at -20°C as they could be used for future studies.

Black and Indian rhinoceros 200 bp libraries were amplified in polymerase chain reactions (PCRs) consisting of 1.0 µl adaptor ligated DNA, 12.5 µl NEB High-Fidelity 2x PCR Master Mix, 1.0 µl primers for Ion Torrent (forward primer sequence: 5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG-3'; reverse primer sequence: 5'-CCTCTCTATGGGCAGTCGGTGAT-3'), and 10.5 µl nuclease-free water, using 13-15

cycles of thermocycling regimen #1 (Table 2.1). A 2 µl subsample of the reactions were electrophoresed on 1.5% agarose gels and amplicons were visualized using a Molecular Imager VersaDoc MP 4000 system (Bio-Rad Laboratories, Mississauga, Ontario, Canada). Since hybridization reactions required high concentrations of DNA, four identical 20 µl PCRs were performed and pooled together before purifying the reactions using 3 volumes of serapure beads and eluting in 25 µl nuclease free water. This method of pooling multiple PCRs was used instead of increasing the cycle number in an attempt to reduce PCR amplification bias and maintain library complexity. The amplified libraries were then quantified using a NanoDrop 2000C spectrophotometer (Fisher Scientific, Ottawa, Ontario, Canada) and subsequently concentrated to ~50 ng/µl using a SPD1010 integrated SpeedVac™ vacuum centrifuge (Thermo Fisher Scientific, Waltham, Massachusetts, USA).

Table 2.1. PCR thermocycling regimens used to amplify DNA libraries.

PCR regimen #	**1**	**2**	**3**	**4**	**5**
Initial denaturation	98°C - 30 sec	95°C - 12 min	95°C - 12 min	98°C - 30 sec	98°C - 30 sec
Cycles	98°C - 10 sec 58°C - 30 sec 72°C - 30 sec	95°C - 20 sec 58°C - 30 sec 72°C - 1 min	95°C - 20 sec 60°C - 30 sec 72°C - 40 sec	98°C - 20 sec 62°C - 30 sec 72°C - 1 min	98°C - 20 sec 58°C - 30 sec 72°C - 30 sec
Final extension	72°C - 5 min	72°C - 5 min	72°C - 5 min	72°C - 5 min	72°C - 5 min
Hold	4°C	4°C	4°C	4°C	4°C

2.3.2.2. Javan rhinoceros DNA extraction and library preparation

Two Javan rhinoceros bone samples acquired from Dr. Peter van Coeverden de Groot (Queens University, Kingston, Ontario, Canada) were used to prepare two DNA libraries. The surfaces of the bones were rinsed clean with nuclease-free water. DNA extractions were performed on the bones following Dabney et al. (2013). Briefly, a Dremel rotary tool was used to drill into the bone several times, reducing it to a fine

powder. Approximately 100 mg of bone powder was used for each DNA extraction and added to 1 ml of extraction buffer (0.45 M EDTA and 0.25 mg/ml proteinase K). The mixture was vortexed to ensure equal mixing and incubated overnight at 37°C in a thermomixer set at ~900 rpm. After ~18 hours of incubation, the mixture was centrifuged at 16,000 x g for 2 minutes to pellet the undissolved bone material. The supernatant was then transferred to a 15 ml tube with 13 ml of binding buffer composed of 5 M guanidine hydrochloride, 40% isopropynol (by volume), 0.05% Tween-20, and 90 mM sodium acetate (pH 5.2). The entire supernatant-binding buffer mixture was then mixed and spun through a MinElute purification column (Qiagen, Toronto, Ontario, Canada). Since the MinElute spin column has a maximum capacity of ~750 µl, several successive centrifugations at 1,500 x g for 4 minutes were required to spin the entire volume through the silica column. Following each spin, the waste flow-through was removed from the collection tube. Upon completing the series of centrifugations, the MinElute spin column was placed in a fresh collection tube and dry-spun at 3,300 x g for 4 minutes. The column was then washed twice using ~750 ul PE buffer from the MinElute DNA purification kit (Qiagen, Toronto, Ontario, Canada) and centrifuged at 3,300 x g for 2 minutes, and the flow-through was later removed. The column was then dry-spun once at 16,100 x g for 1 minute to remove any residual PE buffer. With the column in a fresh 1.5 ml collection tube, the DNA was then eluted in 25 µl nuclease-free water and quantified using a NanoDrop 2000C spectrophotometer (Fisher Scientific, Ottawa, Ontario, Canada).

The extracted DNA was electrophoresed on a 1.5% agarose gel, confirming that it was already highly fragmented (<400 bp); therefore, I immediately proceeded to the "end-repair" DNA library preparation stage using the NEBNext Fast DNA Library Prep

Set for Ion Torrent (New England Biolabs, Whitby, Ontario, Canada). First, 15 µl of DNA was diluted with 36 µl nuclease-free water and the resulting 51 µl diluted DNA was transferred into the end-repair reaction. The remainder of the DNA library preparation followed the methods described above for the black and Indian rhinos. The E-gel iBase (Invitrogen, Carlsbad, California, USA) electrophoresis system was used to isolate 80, 150, and 200 bp fragments (excluding adaptors). The 150 bp libraries were then amplified in PCRs akin to those detailed for the black and Indian rhinoceroses and the purified PCR products were concentrated to ~50 ng/µl using a SPD1010 integrated SpeedVac^{TM} vacuum centrifuge (Thermo Fisher Scientific, Waltham, Massachusetts, USA).

2.3.2.3. Malayan tapir and Sumatran rhinoceros DNA extraction and library preparation

Malayan tapir and Sumatran rhinoceros libraries were created by Nathan Wales and Anthony Signore using workspace and samples provided by Tom Gilbert at the Natural History Museum of Denmark (University of Copenhagen, Copenhagen, Denmark). DNA extractions were performed on two 100 µl liquid blood samples from the Malayan tapir and an excised blood meal from a leech that had fed upon a Sumatran rhinoceros. The leech was first frozen at -20°C and a ~100 µl volume of solid blood was dissected from its gastrointestinal tract. All blood samples were incubated at 55°C overnight and DNA extractions were processed the following day using the DNeasy blood and tissue kit (Qiagen, Toronto, Ontario, Canada).

A NEBNext DNA Library Prep Master Mix Set for 454 (New England Biolabs, Whitby, Ontario, Canada) was used to create DNA libraries for the Malayan tapir and

Sumatran rhinoceros. Unique barcode A adaptors (Appendix 2) were ligated to the DNA fragments and the libraries were purified using a MinElute PCR purification kit (Qiagen, (Qiagen, Toronto, Ontario, Canada), then eluted in 40 µl EB. Each library was then divided into two PCR reactions, each consisting of 17.1 µl nuclease-free water, 5.0 µl 10x AmpliTaq Gold buffer, 5.0 $MgCl_2$, 0.2 µl dNTPs, 2.5 µl primers for Ion Torrent, 0.2 µl AmpliTaq Gold polymerase (Applied Biosystems, Foster City, California, USA), and 20 µl template DNA using 10 cycles of thermocycling regimen #2 (Table 2.1). A 2.0 µl subsample from each reaction was electrophoresed on 1.5% agarose gels to ensure that the PCRs were successfully amplified and reactions for the same DNA library were combined, purified using a MinElute PCR purification kit (Qiagen, Toronto, Ontario, Canada), and eluted in 25 µl EB. These PCR products were then shipped to the University of Manitoba where I prepared them for hybridization capture experiments by transferring 100 ng DNA into the E-gel iBase (Invitrogen, Carlsbad, California, USA) and performing size selection to isolate 100, 200, 400, and 600 bp fragments (excluding adaptors). These fragments were then used as a template for PCR amplifications, subsequently purified, quantified and vacuum centrifuged as previously described for the black and Indian rhinoceroses.

2.3.2.4. Woolly rhinoceros DNA extraction and library preparation

Ancient DNA extractions were performed by Nathan Wales and Anthony Signore on permafrost-preserved samples of four bones and one piece of skin from five woolly rhinoceros individuals that originated from China and Siberia (see Appendix 2) in a dedicated ancient DNA laboratory workspace provided by Tom Gilbert at the University of Copenhagen, who also provided the tissue samples. The DNA extractions were

performed according to Dabney et al. (2013) using 112-176 mg of woolly rhinoceros tissue as outlined above for the Javan rhinoceros. An extraction blanks serving as a negative control was subjected to the same DNA extractions procedures to control for exogenous DNA contamination. The extraction blank revealed minimal contamination with a DNA concentration of 328.71 pg/µl for fragment sizes between 100 – 1,000 bp, whereas woolly rhinoceros positive DNA extractions provided concentrations of ~3,200 – 88,900 pg/µl within the same size range. A NEBNext DNA Library Prep Master Mix Set for 454 was used to create the ancient DNA libraries. First, end repair reactions were performed using 85 µl DNA 10 µl NEBNext 10x end repair buffer and 5 µl NEBNext end repair enzyme mix. The reaction was then incubated at 12°C for 20 minutes followed by 37°C for 15 minutes in a thermocycler. The end-repaired DNA was then purified with a MinElute PCR Purification kit (Qiagen, Toronto, Ontario, Canada) where the reaction was first mixed with a 5x volume of PB buffer, applied to a MinElute spin column (Qiagen, Toronto, Ontario, Canada), and centrifuged at 8000 x g for 1 minute. The column was then washed by adding 750 µl PE buffer and centrifuged at 8000 x g for 1 minute. The flow-through was discarded and the column was placed into an empty collection tube and dry-spun at 8000 x g for 1 minute to remove any residual ethanol. To elute the DNA, the spin column was placed in a fresh 1.5 ml tube, 30 µl EB buffer was added, and the solution incubated at 37°C for 15 minutes before spinning at 8000 x g for 1 minute.

While each of the modern DNA libraries were single indexed, with only the A adaptor having a barcode sequence, the ancient woolly rhino libraries were dual indexed where both the A and P adaptors contain unique barcode identifiers (Appendix 2). This

step helps to reduce the possibility of chimeric reads being incorporated into the dataset and ensures the authenticity of both ends of the sequenced reads. Adaptor sequences with specialized phosphorothioate bonds were acquired from Dr. James Haile (University of Oxford, UK) that resist digestion from endonuclease contamination commonly associated with ancient DNA samples. To perform the adaptor ligation, 10 µl NEBNext 5x quick ligation buffer, 5 µl adaptor mix, and 50 µl Quick T4 ligase was added to a 1.5 ml tube containing the 30 µl purified end repaired DNA from the previous step. The reaction was then incubated at 20°C for 20 minutes in a thermomixer and purified using a MinElute spin column (Qiagen, Toronto, Ontario, Canada) in the same manner detailed above but eluted in 42 µl EB. The 42 µl of purified adaptor ligated DNA was then mixed in a 0.2 ml tube with 5 µl NEBNext adaptor fill-in reaction buffer and 3 µl Bst DNA polymerase (large fragment), and incubated at 65°C for 20 minutes followed by 80°C for 20 minutes. The libraries were then purified as described earlier using a MinElute PCR Purification kit (Qiagen, Toronto, Ontario, Canada) and eluted in 20 µl EB.

The DNA libraries were then amplified in a 100 µl PCR containing 20 µl template DNA, 0.8 µl dNTPs, 4 µl primers for Ion Torrent, 2 µl AmpliTaq Gold polymerase (Applied Biosystems, Foster City, California, USA), 10 µl of 10x AmpliTaq Gold buffer, 10 µl $MgCl_2$ (2.5 mM), 4 µl (0.4 mg/ml) BSA, and 49.2 µl nuclease-free water using 10, 12, or 14 cycles of thermocycling regimen #3 (Table 2.1). The PCR products were then purified using Agencourt AMPure XP Beads (Beckman Coulter, Brea, California, USA), quantified on an Agilent 2100 Bioanalyzer (Agilent Technologies) in a high sensitivity DNA assay, and shipped to the University of Manitoba.

As described for the black and Indian rhinoceroses, size selection of the woolly rhinoceros libraries was performed to isolate 80, 150, and 200 bp fragments (excluding adaptors). Woolly rhinoceros 80 and 150 bp libraries were then re-amplified in PCRs containing 1 µl template DNA, 1 µl primers, 12.5 µl 2X NEBNext Master Mix, and 10.5 µl nuclease-free water with 15 cycles of thermocycling regimen #4 (Table 2.1). Four identical PCRs were performed for each library, then pooled together and purified with serapure magnetic beads and eluted in ~25 µl nuclease-free water. The amplified libraries were then quantified and vacuum centrifuged as described for the black and Indian rhinoceroses.

2.3.3. In-solution hybridization captures

Biotinylated 120mer MyBaits RNA probes (Mycroarray, Ann Arbor, Michigan, USA) were designed to target 199 genomic regions from the publically available draft genome of the white rhinoceros on GenBank (Di Palma et al. 2012) (see Appendix 3 for accession numbers), and synthesized with a 4x tiling pattern (i.e. 90 bp of overlap with each 120 bp bait). Probes were designed to target exons of all 54 genes in Appendix 1 plus 30 bp of upstream and downstream intronic flanking sequence. Repeat regions were replaced with "NNNs" using the program RepeatMasker 4.0.5 (Smit et al. 2013). Genes targeted in this study (listed in Appendix 1) include 26 nuclear "Assembling the Tree of Life" genes used in previous mammalian phylogenetic analyses of Meredith et al. (2011), 8 nuclear markers previously used by Steiner and Ryder (2011), and 21 nuclear genes (divided into 165 gene segments) unique to this study.

Vacuum centrifuge concentrated libraries were used for in-solution hybridization capture experiments with MyBaits RNA probes (Mycroarray Ann Arbor, Michigan, USA) following the manufacturers version 2.3.1 protocol. To summarize, a library master mix was prepared at room temperature by mixing 5.9 µl (~300 ng) of DNA template, 2.5 µl of 1 µg/µl human Cot-1 DNA, 2.5 µl of 1 µg/µl salmon sperm DNA, and 0.6 µl proprietary blocking agent. Next, a hybridization master mix was prepared at room temperature by mixing 20 µl 20X saline-sodium phosphate-EDTA hybridization buffer, 0.8 µl 500 mM EDTA, 8 µl 50X Denhardt's solution, and 0.8 µl 10% sodium dodecyl sulfate. Finally, a capture baits master mix was prepared on ice by mixing 5 µl RNA probes and 1 µl 20 U/µl SUPERase-In RNase block (Invitrogen, Carlsbad, California, USA). The library master mix was then incubated in a MJ Mini Gradient Thermal Cycler (Bio-Rad Laboratories, Mississauga, Ontario, Canada) at 95°C for 5 minutes to denature the DNA. The tube containing the hybridization master mix was transferred to the thermocycler after the temperature had been reduced to 65°C and incubated for 3 minutes. Next, the capture baits master mix tube was added to the tubes within the thermocycler and incubated for 2 minutes at 65°C. All three tubes were held in the thermocycler at 65°C while transferring 9.5 µl of the library master mix and 10.5 µl of the hybridization master mix to the tube containing the capture baits master mix. The solution was mixed by pipetting and incubated at 65°C for 8 hours, 64°C for 4 hours, 63°C for 4 hours, and 62°C for 4 hours.

DNA library fragments that had annealed to the biotinylated RNA probes were then isolated and purified by preparing 50 µl Dynabeads MyOne Streptavidin C1 magnetic beads (Thermo Fisher, Waltham, Massachusetts, USA) by pelleting them on a

magnetic particle stand in a 1.5 ml tube and removing the supernatant. The beads were washed with 200 µl binding buffer, vortexed for 10 seconds, and pelleted on the magnetic tube rack while the supernatant was discarded. This washing sequence was repeated twice and afterwards the beads were resuspended in 20 µl binding buffer, transferred to a fresh 0.2 ml tube and incubated at 65°C for 2 minutes in the thermocycler. The streptavidin-coated beads were then added to the hybridization capture mixture, incubated at 65°C for 45 minutes and mixed by pipetting every 10 minutes. The beads were pipetted into a 1.5 ml tube, pelleted, and the supernatant removed, at which point 500 µl of wash buffer 2 (preheated to 65°C) was added to the beads and mixed by pipetting. The tube was incubated at 65°C for 5 minutes in a ThermoMixer (Eppendorf, Mississauga, Ontario, Canada), pelleted, and the supernatant again removed. The wash sequence was repeated an additional 2 times before the beads were finally resuspended in 30 µl nuclease-free water.

Post-hybridization PCR amplifications were performed in 25 µl reactions using 1 µl resuspended beads, 12.5 µl NEB High-Fidelity 2x PCR Master Mix, 1 µl primers for Ion Torrent and 10.5 µl nuclease-free water (New England Biolabs, Whitby, Ontario, Canada) and 14 - 20 cycles of thermocycling regimen #5 (Table 2.1). A 2 µl subsample of each reaction was electrophoresed on 1.5% agarose gels where a visualized band of the expected length confirmed successful PCR amplification of the enriched libraries. Successful PCRs were purified with serapure magnetic beads and DNA was eluted in 20 µl of nuclease-free water. DNA quantification assays were then performed using high sensitivity (HS) assays and a Qubit 2.0 fluorometer (Invitrogen, Carlsbad, California, USA). DNA concentrations (ng/µl) were converted to picomolar (pM) units and diluted

to 100 pM for sequencing. Note that while five woolly rhinoceros DNA libraries were prepared, two libraries (WR1 and WR5; Appendix 2) were not sequenced as bioanalyzer results revealed lower DNA quantities than expected and post-hybridization PCR amplifications did not produce a band during electrophoresis even after 20 cycles.

2.3.4. Ion Torrent next-generation sequencing

The DNA libraries were sequenced in-house using an Ion Torrent Personal Genome Machine (PGM; Applied Biosystems, Foster City, California, USA) next-generation sequencer. Briefly, the DNA libraries were first amplified on Ion Sphere™ Particles (ISPs) using the Ion One Touch 2 System and an Ion PGM Template OT2 Hi-Q Kit (Applied Biosystems, Foster City, California, USA). Briefly, 2.0 µl of each of the 100 pM target-enriched DNA libraries was added to 23 µl of nuclease free water to create a diluted DNA library. Next, the following reagents were added to a 2 ml tube containing 800 µl Ion PGM Hi-Q Reagent Mix: 15 µl nuclease-free water, 50 µl Ion PGM Hi-Q Enzyme Mix, 25 µl diluted DNA library, 100 µl Ion PGM Hi-Q ISPs, and 10 µl Ion PGM Calibration Standard. This mixture was pipetted into a filter assembly with 1.5 ml reaction oil, loaded into the Ion One Touch 2 System (Applied Biosystems, Foster City, California, USA), and the "Ion PGM Hi-Q OT2 Kit -200" program was run for ~ 16 hours. The two recovery tubes were then removed from the machine and all except 100 µl of the reaction solution was removed from each tube. The remaining solution was pipetted up and down to resuspend the pelleted ISPs, and 500 µl Ion One Touch wash solution was added to each recovery tube. Following this step, the solutions were pooled together in a 1.5 ml tube and centrifuged at 15,500 x g for 2.5 minutes and once again, all except 100 µl of liquid was removed from the tube. The template-positive ISPs were then

enriched using the Ion One Touch ES system (Applied Biosystems, Foster City, California, USA). An eight-well strip was prepared: 100 µl suspended ISPs in one well, 130 µl washed Dynabeads MyOne Streptavidin C1 beads (Thermo Fisher, Waltham, Massachusetts, USA) in another well, 300 µl Ion One Touch Wash Solution in three wells, 300 µl of freshly-prepared melt-off solution (125 mM NaOH and 0.1% Tween 20) in one well, and the last two wells were left empty as described in the manufacturer's protocol. A 2 µl sample was withdrawn from the well containing the ISPs for a quality control assay. The Ion One Touch ES system was then initiated and after the run, the machine deposited the enriched ISPs into a 0.2 ml tube pre-loaded with 10 µl neutralization solution.

The quality control assays were then performed on the 2 µl ISP sample using an Ion Sphere Quality Control kit (Applied Biosystems, Foster City, California, USA) and Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, California, USA) with the Ion_PluginV3.10 firmware file installed. The Qubit 2.0 flurometer calibration factor and AF488 and AF647 values were inputted into the Qubit Easy Calculator Microsoft Excel Spreadsheet (http://ioncommunity.lifetechnologies.com/community/products/pgm/) to estimate the percent of template-positive ISPs. Values between 10 and 30% were considered optimal; however, libraries were still sequenced if the percent of templated ISPs fell outside of that range.

The template-positive ISPs were then sequenced in one direction using an Ion PGM Hi-Q Sequencing Kit and Ion Torrent PGM (Applied Biosystems, Foster City, California, USA). First, the sequencer was washed with a filtered chlorite solution and rinsed with 18.2 MΩ cm^{-1} water. The Ion PGM was initialized according to the user

manual with one modification that 100 µl of 100 mM NaOH was added to the wash 2 bottle instead of the recommended 70 µl to account for the site-specific water quality. Enriched template-positive ISPs were loaded into Ion 314v2 and Ion 318v2 barcoded chips (Applied Biosystems, Foster City, California, USA). "TargetSeq" sequencing runs were planned using the Torrent Suite 4.4 Software with 330 and 500 dNTP flows for 150 and 200 bp libraries, respectively. On occasion, multiple libraries were pooled together on the same chip and sequenced simultaneously (see Appendix 4). In total 11 sequencing runs were performed: 9 using Ion 314 v2 BC chips and 2 using Ion 318 v2 BC chips (Applied Biosystems, Foster City, California, USA). The 314 v2 BC chips are capable of sequencing 400,000-550,000 reads while the 318 v2 BC chips can sequence up to 10 times as many reads.

2.3.5. Sequenced read assemblies

Sequenced reads were binned automatically using the Torrent Suite 4.4 software system according to the 'A' adaptor indexes sequence and imported into Geneious 7.1.9 (Biomatters Limited, Auckland, New Zealand). Woolly rhinoceros reads contained 'P' adaptor indexes that were not trimmed automatically from the 3' ends by the Torrent Suite software, thus requiring additional trimming performed in Geneious. This was accomplished using the "Trim Ends" function that identified and removed the index sequences (Appendix 2), allowing 1 mismatch per index.

Assemblies of sequenced reads in Geneious 7.1.9 (Biomatters Limited, Auckland, New Zealand) required reference sequences to be used as a template; therefore, the same white rhinoceros sequences previously used to design the RNA capture probes were chosen. Assemblies were performed against each reference exon plus 30 bp of upstream

and downstream flanking sequence using the "assemble to reference" function and "medium/low" sensitivity with three iterations and 20% maximum mismatches per read. Using multiple iterations helped to build out the flanking sequences of the targets, which were incorporated into the phylogenetic analyses. Assembled reads were examined by eye, especially for the woolly rhinoceros that contained frequent DNA damage (G→A, C→T transitions) commonly associated with ancient DNA due to cytosine deamination to uracil (Binladen et al. 2006; Brotherton et al. 2007, Hofreiter et al. 2001b). Ancient DNA damage is often identifiable from the associated mutations occuring in some, but not all, of the sequenced reads. In cases where these mutations were suspected to result from cytosine deamination, corrections to the consensus sequences were implemented. Multiple libraries were sequenced for the tapir, Javan rhinoceros, and woolly rhinoceros, but the reads were combined when performing the assemblies to acquire consensus sequences for each species. In total, assemblies were performed for 199 gene segments, which included flanking intron segments that provided regions of mostly neutral evolution for phylogenetic analyses.

2.3.6. Genome mining

Genome mining was used to retrieve orthologous sequence data for the white rhinoceros (*C. simum*), four equines (donkey [*E. asinus*], Przewalski's horse [*E. ferus przewalskii*], thoroughbred horse [*E. ferus caballus*], and Mongolian horse [*E. ferus caballus*]), as well as the cow (*B. taurus*), pig (*S. scrofa*), and camel (*C. ferus*) (see Appendix 3 for accession numbers). The non-rhinoceros species included in the phylogenetic analyses served as outgroup species. Briefly, nucleotide blasts (blastn) were performed on the NCBI web server using white rhinoceros DNA sequence as queries

using the “megablast” and “discontinuous megablast” functions. The species were selected from the whole-genome shotgun sequence database. Top scoring contigs were imported into Geneious 7.1.9 to be trimmed to the correct length and the coding regions were annotated with the “transfer annotations” function using the white rhinoceros exons as reference sequences.

2.3.7. DNA Alignments

Alignments were created for each of the 199 gene segments and all 14 species using the MUSCLE (Edgar 2004) plugin in Geneious 7.1.9 (Biomatters Limited, Auckland, New Zealand). Species were excluded from gene segment alignments if they completely lacked coverage for that gene segment. Alignments were then examined by eye to correct occasional misalignments. DNA insertions were removed if they were common to only a single species.

Many genes included in this study are made up of multiple exons (Appendix 1); however, assemblies and alignments were performed on an exon-by-exon basis, such that a gene segment would consist of a single exon plus the upstream and downstream flanking intron sequence. In some cases, where two exons were in very close proximity (<200 bp), the two exons plus the surrounding intron sequence were grouped into the same gene segment. This methodology avoided a common problem made in many studies that have used coalescence analyses, which is to artificially concatenate gene-coding sequences from multiple exons and analyze it as a single coalescence gene or “c-gene” (Gatesy et al. 2016; Scornavacca and Galtier 2017). This methodology can lead to misleading results as it does not account for differential genetic recombination that may

occur between multiple exons of a gene with large interspersed introns (Gatesy et al. 2016).

2.3.8. Phylogenetic analyses

Two main phylogenetic approaches were used to estimate species trees in this study: coalescence and concatenation analyses. Coalescence models generally create multiple phylogenetic trees for subsets of the total sequence data, which are then summarized to form a species tree. In contrast, to perform more commonly used concatenation analyses, several gene segment alignments are linked together to form a single large alignment (referred to as a supermatrix), which is then analyzed to estimate a species tree. The DNA alignments were also examine by eye for synapomorphic indels (insertion/deletions) that had potential to be phylogenetically informative within the family Rhinocerotidae. In all, 199 gene segments (totaling 131,931) were analyzed in this study from 54 genes, including 365 intron segments (67,724 bp) and 206 exons (64,207 bp).

2.3.8.1. Coalescence analyses

As summary coalescent analyses can be susceptible to gene tree estimation errors (Chou et al. 2015, Gatesy et al. 2016), efforts were made to optimize gene tree accuracy. The program GARLI 2.0 (Genetic Algorithm for Rapid Likelihood Inference; Zwickl 2006) was used to create gene trees that avoid confounding, indiscriminate relationships. When too few informative sites are recognized between taxa to produce clear dichotomies with a branch length greater than 1 e-8 nucleotide substitutions per site, GARLI 2.0 collapses branches into polytomies. This effectively eliminates artifactual

relationships that would have otherwise been arbitrarily resolved with programs like RAxML that require all relationships to be fully resolved as bifurcations even in the absence of sufficient evidence to support such a relationship. The GARLI 2.0 program was run to create a best scoring tree for each of the 199 gene segment alignments, which were then analyzed using ASTRAL-II (Mirarab and Warnow 2015), a summary coalescence algorithm that combines the gene trees to form a consensus species tree. The length of each gene segment is listed in Appendix 1. ASTRAL-II accommodates gene trees with species that have been excluded due to missing data, allows for polytomies and treats all gene trees as being unrooted.

The program SVDQuartets (Singular Value Decomposition scores for species quartets; Chifman and Kubatko 2014) is another coalescence method utilized to estimate a species tree. In contrast to summary coalescence, SVDQuartets does not rely on the construction of gene trees, but instead analyzes single nucleotide polymorphisms (SNPs) to produce the best scoring topological arrangement for all possible groups of four species. The resulting trees, termed "quartet trees", are then coalesced in PAUP* 4.0a150 to construct an estimated species tree (Chifman and Kubatko 2014; Chou et al. 2015). To run this program, all 199 gene segment alignments were first concatenated in Geneious 7.1.9. to form a 131,931 bp supermatrix that was then converted to a nexus file format. The boundaries for the each of the 199 gene segments were then added to the nexus file, which was then imported into PAUP* 4.0a150 (Swofford 2002) to run SVDQuartets with 1000 bootstrap replicates evaluating all possible quartet trees. The program was performed with the "multispecies coalescence tree" model.

2.3.8.2. Concatenation analyses

To perform concatenation analyses, all 199 gene segment alignments were concatenated in Geneious 7.1.9 to form a 131,931 bp supermatrix. The supermatrix was then analyzed using RAxML (Randomized Axelerated Maximum likelihood) version 7.2.8 (Stamatakis 2006) using the following parameters: "GTR Gamma" nucleotide model with "rapid bootstrapping and search for best scoring ML tree" algorithm, "start with complete random tree" and 500 bootstrap replicates. Multiple partitioning schemes were performed, where independent GTR Gamma nucleotide models are estimated for each partition. The first scheme made each of the 199 gene segments individual partitions while the second scheme made each intron or exon region as a partition with a total of 571 partitions.

Bayesian analyses were performed on this dataset with MrBayes 3.2.6 (Huelsenbeck and Ronquist 2001) plug-in in Geneious using the following parameters: "GTR" substitution model, "gamma" rate variation, *Camelus ferus* as the outgroup, 4 gamma categories and 10,000,000 chain length sampled every 10,000 trees with a burn-in of 1,000,000 generations.

Rhinoceros relationships were further assessed by subdividing the 131,931 bp supermatrix into 365 intron partitions totaling 67,724 bp and 206 exon partitions totaling 64,207 bp. RAxML concatenation trees were generated for each of these supermatrices using the same settings described above. This step should provide insights as to whether or not evolutionary inferences differ between non-coding gene segments assumed to be evolving largely under lack of selection pressure and coding sequences that are assumed to be evolving under either purifying or positive selection pressures.

2.3.8.3. Robinson-Foulds distances

The phylogenetic analyses used in this study provided contradictory topological arrangements of the three main rhinocerotid subfamilies (Dicerorhininae, Rhinocerotinae, and Dicerotinae) supporting either the "two-horned" or the "biogeographical" hypothesis described above. One major factor known to confound molecular phylogenies is referred to as incomplete lineage sorting (ILS). This evolutionary phenomenon arises when multiple alleles of a single locus (containing distinct polymorphisms) are present in an ancestral species and inherited by descendant lineages following speciation events occurring in quick succession. In the descendant species, redundant alleles are eventually lost and genetic polymorphisms in the remaining locus are not representative of the true monophyletic relationships between descendants, leading to incongruence during molecular phylogenetic analyses between gene and species trees (Galtier and Daubin 2008; Mirarab and Warnow 2015; Suh et al. 2015). High levels of topological discordance were noticed within GARLI gene trees, thus to provide an indication of ILS within my molecular data set, I calculated pairwise Robinson-Foulds (RF) distances (Robinson and Foulds 1981) between gene trees and each of the two prevailing species trees using the program IQ-TREE (Nguyen et al. 2015). However, RF distances calculations require fully bifurcating trees, so instead of using GARLI gene trees (which contain polytomies), RAxML was used to create best-scoring gene trees suitable for these analyses employing the same settings described above for concatenation analyses. Non-ceratomorph species were excluded from the 199 gene segment alignment as I aimed to only receive an indication of ILS within the rhinocerotid lineage and not horses or even-toed ungulates. The Malayan tapir was retained in the alignments to serve as the outgroup

species. Species trees were pruned to reflect the same species with sequence data for each of the gene segment alignments. The gene segment NGB 1 was excluded from the RF analyses as only three species (white, black and woolly rhinos) had coverage for this region and RAxML requires ≥4 species to create a tree. Thus, a total of 198 gene trees were compared to 2 species trees. RF distances were scaled using the following equation: RF distance/(2x(number of shared species between the two trees – 3)) (Rosenberg and Kumar 2001). Scaled RF distances range from 0 (representing absolute congruence between in branching configurations between gene and species trees) to 1 (representing total incongruence between the trees).

2.3.9. Eye gene selection pressure using PAML

To investigate whether the 12 eyesight genes (*ARR*, *CNGB3*, *GNAT2*, *GNGT2*, *GRK7*, *GUCA1B*, *OPN1LW*, *OPN4*, *OPSD*, *PDE6C*, *PDE6H*, and *SWS1*; see Appendix 1 for full gene names) included in the non-coding versus coding phylogenetic analyses of the notoriously near-sighted rhinoceroses were not undergoing neutral evolution consistent with gene inactivation, selection pressure analyses were performed using the program CODEML in the PAML 4.8 software package (Yang 2007). First, the coding exons for each gene were combined to form coding sequences and alignments were built for each species. These loci were then examined for mutations that would make the gene non-functional, including splice site mutations violating the GT-AG rule (introns start with 5' guanine followed by a thymine and end with 3' adenine followed by a guanine; Burset et al. 2000), nonsense mutations and frameshift mutations. In the event that these mutations were found, thus providing evidence of pseudogenization, the inactivated gene would be excluded from the phylogenetic analyses. The species tree inputted into the

CODEML program was based on the topologies from the maximum likelihood and Bayesian concatenation phylogenetic results obtained from this study. All termination codons were removed from the alignments, as CODEML is not designed to accommodate sequences with stop codons. The "free-ratio" model was performed for each gene alignment, which allows the selection pressure to be estimated for every branch tip and internal node of the unrooted species tree. This model is designed to provide an overall picture of selection pressure acting upon a gene by calculating the ratio of non-synonymous mutations per non-synonymous site (dN) versus the number of synonymous mutations per synonymous site (dS), denoted as the "dN/dS" ratio or omega value (ω). Calculating this ratio can provide insight as to whether the gene is under purifying selection ($\omega < 1$) where functional conservation is being selected, neutral evolution ($\omega = 1$) where selection is not acting upon the gene and nucleotide substitutions are occurring at random, or positive selection ($\omega > 1$) where functional change is being selected.

2.4. Results

2.4.1. Number of reads sequenced and sequence coverage

Hybridization capture and next-generation sequencing techniques proved to be very effective, resulting in the total number of sequenced reads and average read length summarized in table 2.2 Of any species examined in this study, the most reads were sequenced for the ancient woolly rhinoceros libraries, totaling ~2.3 million (Table 2.2). About 35% of the woolly rhino reads matched as BLAST hits to the white rhinoceros genome and 7% assembled to loci targeted with the hybridization capture experiments. In total, ~68% sequence coverage of all nuclear markers was attained for this extinct species (Figure 2.2). This relatively high level of coverage stemming from three individual

libraries allowed many sites of DNA damage, commonly associated with ancient DNA libraries (G→A, C→T transitions) (Binladen et al. 2006; Brotherton et al. 2007; Hofreiter et al. 2001 b), to be identified and corrected before generating gene segment consensus sequences. Modern DNA libraries of the black, Indian, and Sumatran rhinoceroses as well as the Malayan tapir provided even higher coverage of phylogenetic markers, all exceeding 80% (Figure 2.2). Surprisingly, the greatest coverage was achieved for the Malayan tapir (~93%; Figure 2.2) with only ~404,000 reads sequenced (Table 2.2). For each species, a larger proportion of sequence was retrieved for exons in comparison to introns included in the dataset (Figure 2.3).

Most modern DNA libraries had a relatively high (>84%) BLAST hit percentage of all sequenced reads matching to the white rhinoceros genome (Figure 2.2). However, the Javan rhinoceros libraries provided very poor coverage (~10%; Figure 2.2) despite ~841,000 reads sequenced (Table 2.2). Several of the Javan rhino reads also contained sites of likely DNA damage (G→A, C→T transitions relative to other rhinoceros species), which could not be confirmed or corrected due to the low level of sequence coverage. Only ~1% of all reads sequenced for the Javan rhinoceros assembled to targeted loci and only ~5% matched as BLAST hits to the white rhinoceros genome (Figure 2.2). Furthermore, BLASTs against the NCBI nucleotide database revealed that many of the sequenced reads from these libraries were from bacterial and fungal sources (data not shown).

Occasional chimeric reads (where a single read is composed of at least two artificially joined, non-contiguous segments of DNA) were discovered upon assembly of

the black rhinoceros sequence data. In these cases shortest chimeric region of the read was trimmed away and excluded from the assembly.

No signs of cross-contamination between rhinoceros and/or tapir DNA libraries were evident during the Torrent Suite binning process, which sorts reads according to their A adaptor barcodes. Authenticities of the sequenced libraries were further confirmed by performing assemblies against the published mitochondrial genomes of every species (Accession numbers: Black rhinoceros: FJ905814.1; Indian rhinoceros: NC_001779; Javan rhinoceros: FJ905815.1; Sumatran rhinoceros: FJ905816.1; woolly rhinoceros: FJ905813.1; Malayan tapir: NC_023838), as some mitochondrial reads were captured as bycatch even though they were not specifically targeted during the probe design. In each case, bycatch reads from each DNA library provided identical matches to mitochondrial sequences of their respective species (data not shown).

Table 2.2 Number of reads sequenced and average read length for the black, Indian, Javan, Sumatran and woolly rhinoceroses and the Malayan tapir.

Species	**Number of reads sequenced**	**Average read length (bp)**
Diceros bicornis	993110	173
Rhinoceros unicornis	820313	167
Rhinoceros sondaicus	840972	114
Dicerorhinus sumatrensis	386074	186
Coelodonta antiquitatis	2350901	92
Tapirus indicus	404284	182

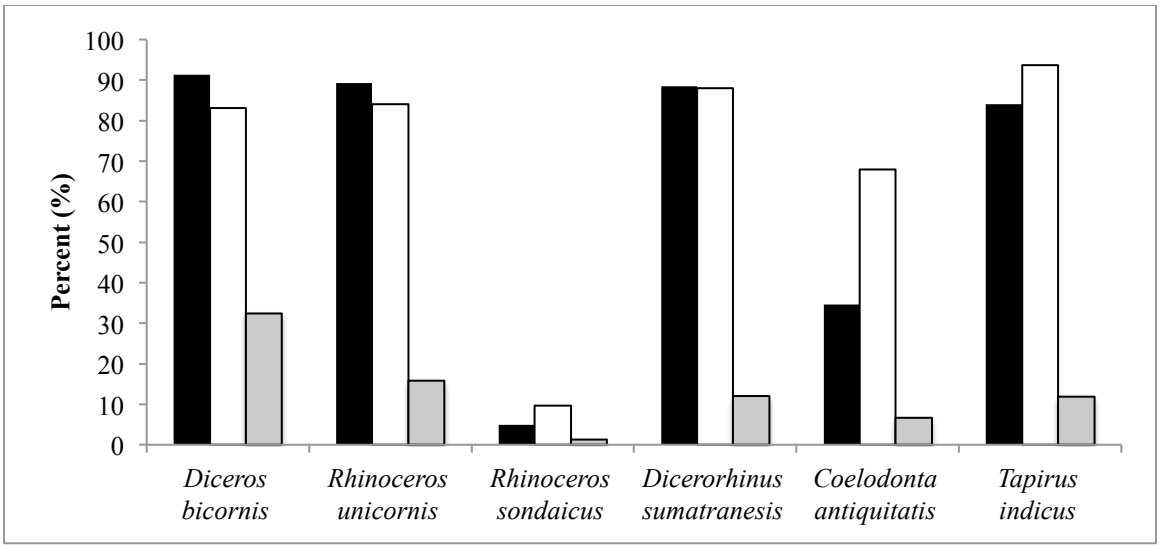


Figure 2.2. DNA library sequencing and assembly information for each species. Black bars represent the percentage of total reads per species that matched as blast hits to the white rhinoceros genome using the "discontinuous megablast" setting in Geneious. White bars denote the percent coverage of sequence data included in the phylogenetic analyses for each species relative to the complete coverage of the white rhinoceros (*Ceratotherium simum*). Grey bars represent the percentage of sequenced reads for each species that assembled using the "map to reference function" in Geneious to the white rhinoceros reference sequences that were targeted during hybridization capture experiments.

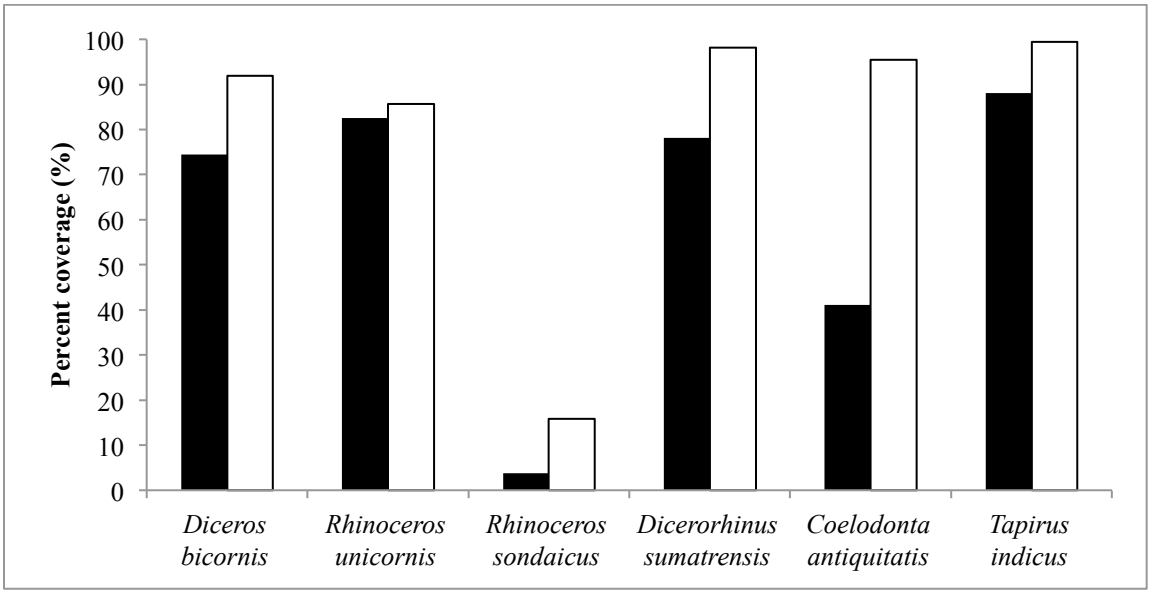


Figure 2.3. Percent sequence coverage of intron (black bars) and exon (white bars) regions relative to the complete coverage of the white rhinoceros (*Ceratotherium simum*).

2.4.2. Coalescence phylogenetic trees

The cumulative length of the 199 gene segment alignments examined in this study totaled 131,931 bp, of which 64,207 bp were from protein- coding regions while 67,724 bp was non-coding intron sequence. The length of each gene segment is listed in Appendix 1.

The ASTRAL-II phylogeny results using best-scoring GARLI 2.0 trees for 199 gene segments are displayed in figure 2.4. As expected, all rhinoceros species group into their respective dicerorhines, rhinocerotinines, and dicerotines sister pairings. The Malayan tapir was placed sister to rhinoceroses with high support. The Asian rhinoceroses (rhinocerotines and dicerorhines) were grouped together and sister to the African rhinoceroses (dicerotines). This relationship is supported with relatively high posterior probability branch support of 0.83, but a very short branch length (0.0791 coalescence units). Of the 199 GARLI gene trees constructed, 51 displayed unresolved polytomies at the base of the Rhinocerotidae family, 13 inconclusive branching arrangements (not any phylogenetic arrangments depicted in Figure 2.1), 52 supported the ((dicerorhines,rhinocerotines),dicerotines) hypothesis, 42 supported the ((dicerorhines, dicerotines),rhinocerotines) hypothesis, and 41 supported the ((dicerotines,rhinocerotines),dicerorhines) hypothesis.

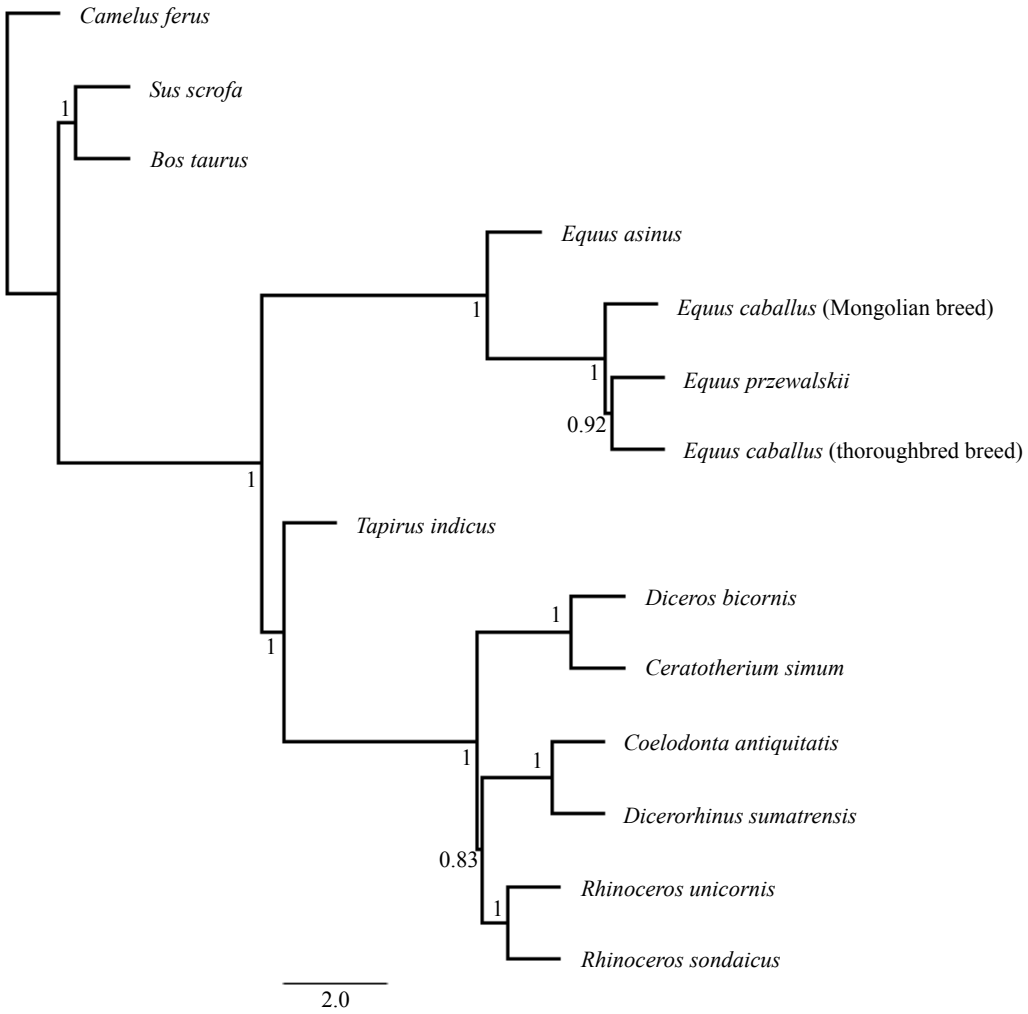


Figure 2.4. ASTRAL-II species tree performed using the best-scoring GARLI maximum likelihood gene tree for each of the 199 gene segment alignments. Node values represent local posterior probability branch support and branch lengths are expressed in coalescence units.

The species tree results from the SVDQuartets single-site analyses are presented in figure 2.5. As with the results attained using the summary coalescent ASTRAL-II method, all rhinoceroses grouped into their respective sister pairings. The topologies of all outgroup species were identical between coalescence methods. However, in contrast to the ASTRAL-II results, the one-horned dicerorhines and dicerotines were strongly supported (94% bootstrap value) as sister taxa (Figure 2.5).

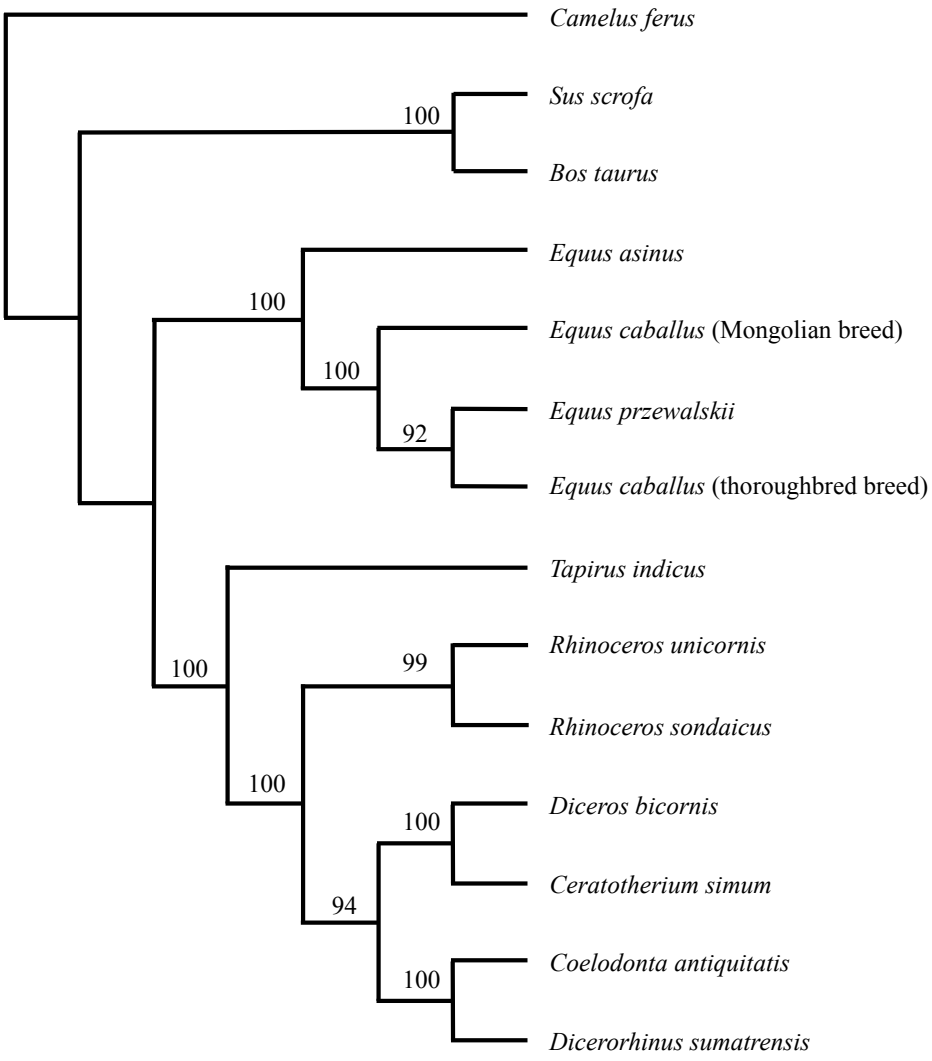


Figure 2.5. SVDQuartets consensus tree performed in PAUP* 4.0a150 with 1000 bootstrap replicates. All possible quartet trees were evaluated in the analysis. Node values indicate bootstrap support percentages. Branch lengths are arbitrary in this tree.

2.4.3. Concatenation phylogenetic trees

The two partitioning schemes (199 gene segments versus 571 introns and exons) used to create RAxML concatenation species trees yielded identical topologies with nearly identical bootstrap support values (Figure 2.6). The RAxML tree relationships are concordant with those of the SVDQuartets (Figure 2.5) and Bayesian trees (Figure 2.7),

placing the rhinocerotines as a sister group to dicerorhines and dicerotines. However, this relationship is weakly supported with only 59% and 63% bootstrap support for the respective 199 and 571 gene segments. The weak support for this node is further reflected in the very short branch length ($6.5x10^{-4}$) leading from the rhinocerotines to all other rhinoceroses, indicating that few nucleotide substitutions separate these three lineages. In contrast, the MrBayes Bayesian concatenation tree (Figure 2.7) supports this relationship with a 0.99 posterior probability despite a short branch length ($6.9x10^{-4}$).

Interestingly, the RAxML concatenation trees constructed for the 67,724 bp intron and 64,207 bp exon supermatrices (Figure 2.8) show clear incongruences in topological arrangements among the rhinoceros subfamilies. The intron tree recovers the relationships found for the complete dataset in SVDQuartets (Figure 2.5), RAxML (Figure 2.6), and MrBayes (Figure 2.7), placing the rhinocerotines as the sister group to the dicerorhine and dicerotine pairing with 85% bootstrap support. In contrast, the RAxML exon tree agrees with the ASTRAL-II results (Figure 2.4), placing the dicerotines as the sister group to the more closely related dicerorhines and rhinocerotines pairing with 83% bootstrap support.

Scaled pairwise RF distances comparing the 198 RAxML gene trees to the pruned species tree reflecting the topology acquired from the SVDQuartets coalescent analyses (Figure 2.5) and the maximum likelihood and Bayesian concatenation analyses (Figures. 2.6 and 2.7) averaged 0.45, while those compared to the species tree topology acquired from ASTRAL-II (Figure 2.4) yielded a nearly identical scaled average RF distance of 0.46. This indicates that on average ~45% of branches are in conflict between gene and species trees.

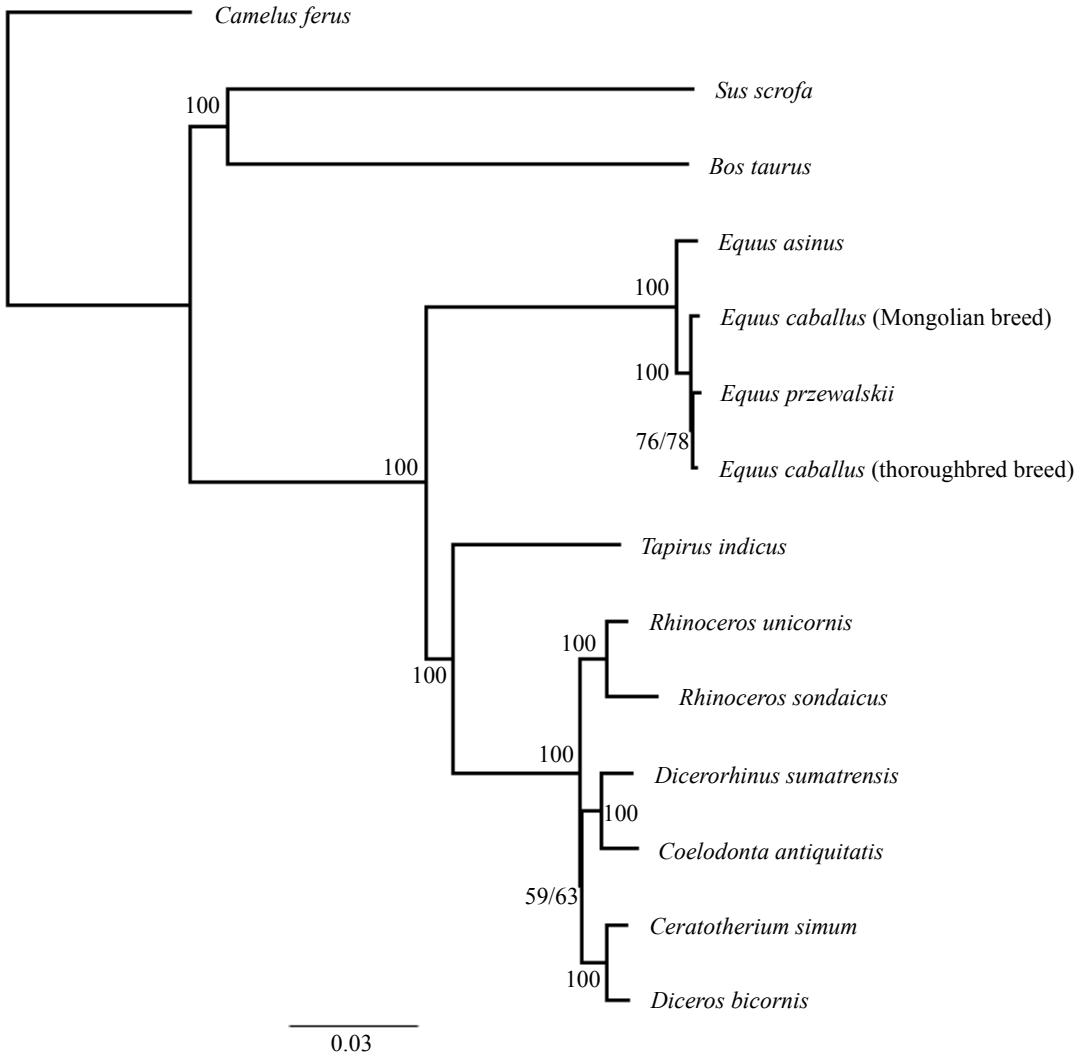


Figure 2.6. Best-scoring maximum likelihood species tree generated in RAxML v 7.2.8 with the GTR GAMMA nucleotide substitution model being estimated for each of the 199 and 571 gene segment partitions of the 131,931 bp supermatrix. Node values denote bootstrap support percent values generated with 500 bootstrap replicates for 199 and 571 partition schemes, respectively. Branch lengths represent the number of nucleotide substitutions per site.

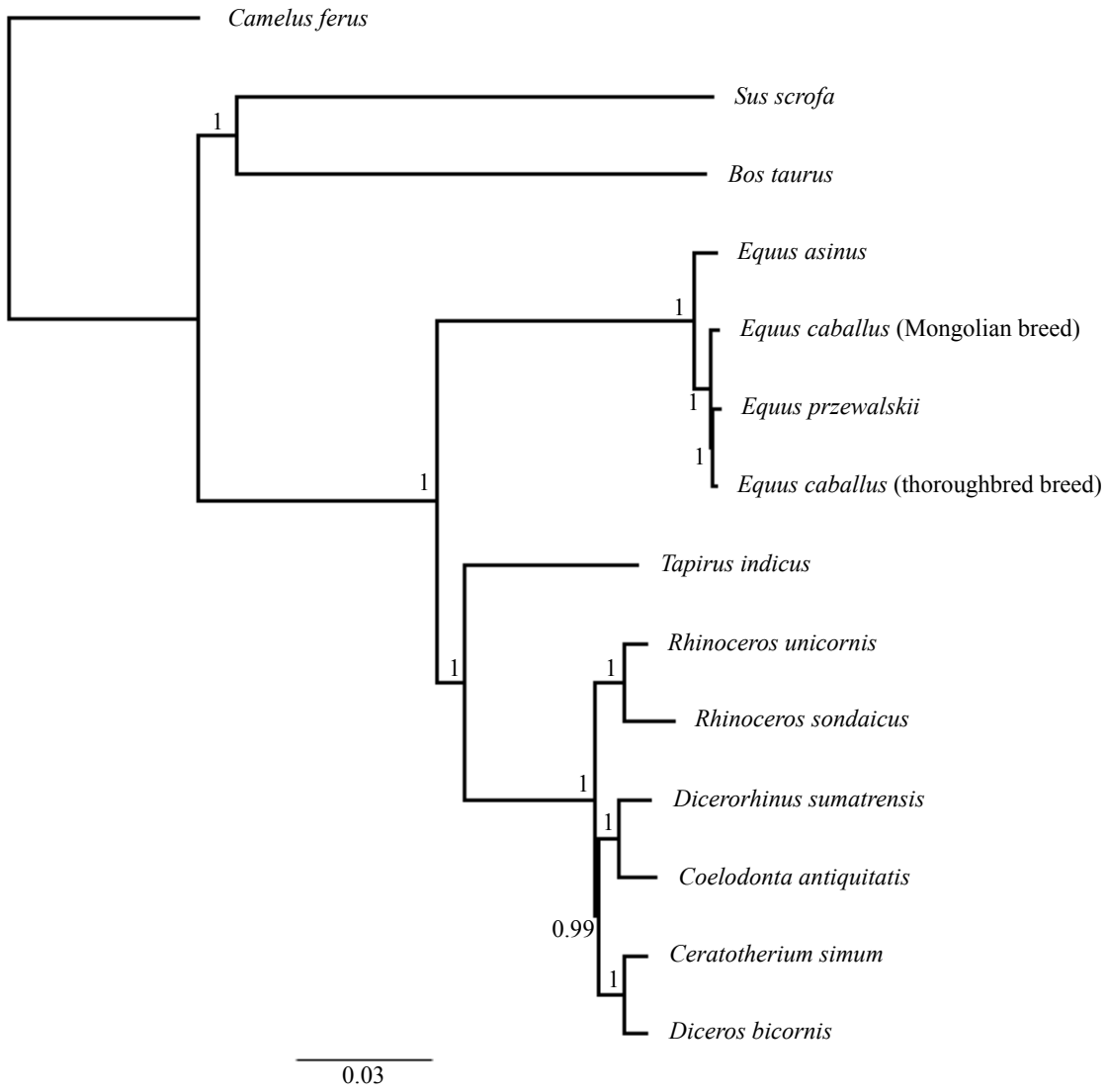


Figure 2.7. Bayesian concatenation tree for 131,931 bp supermatrix made using MrBayes with a 10,000,000 chain length sampled every 10,000 generations and a burn-in length of 1,000,000. The GTR substitution model with gamma rate variation was employed with *Camelus ferus* selected as the outgroup species. Branch lengths represent the number of nucleotide substitutions per site. Node values denote posterior probabilities.

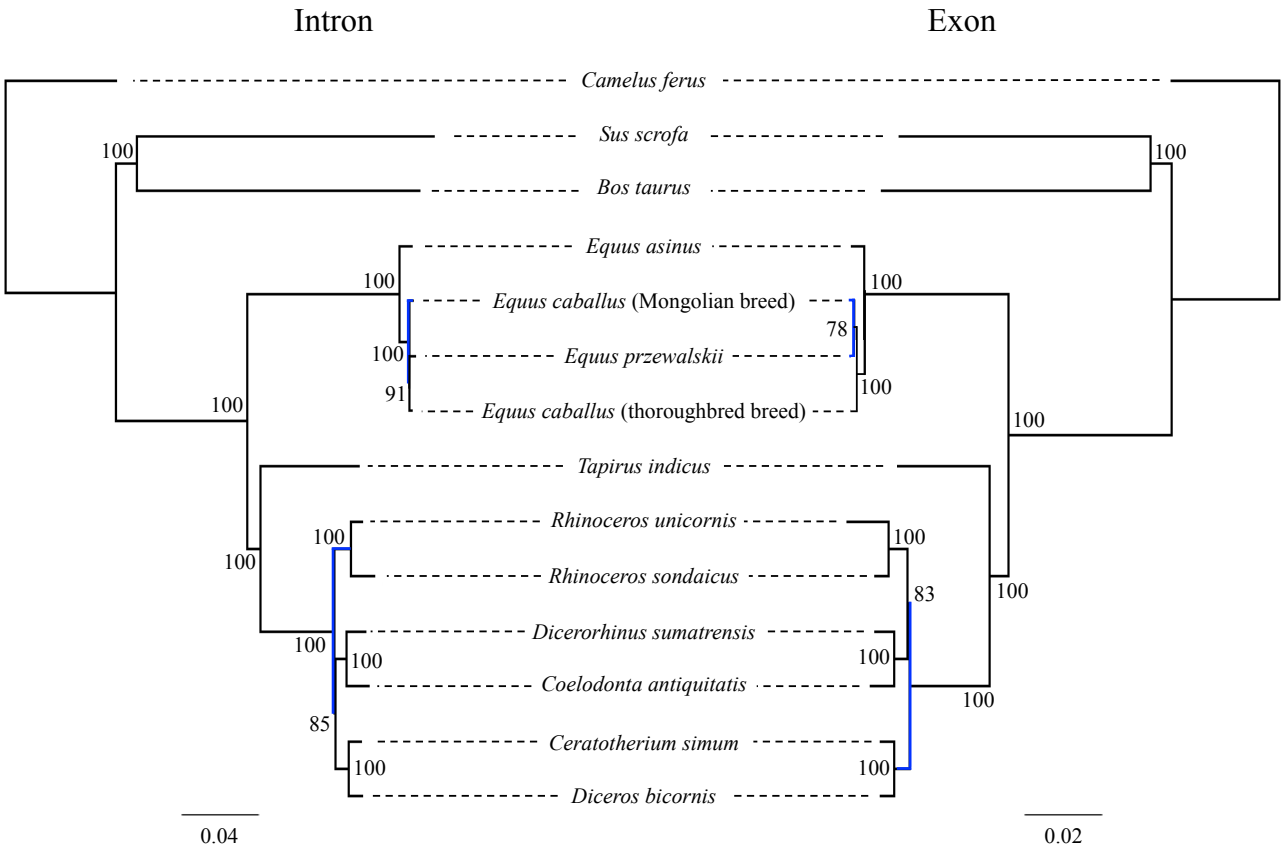


Figure 2.8. Best-scoring maximum likelihood trees generated in RAxML v 7.2.8 with the GTR GAMMA nucleotide substitution model for 365 concatenated intron partitions totaling 67,724 bp and 206 concatenated exon partitions totaling 64,207 bp. Blue branches indicate incongruent topologies between intron and exon trees. Node values denote bootstrap support percent values generated with 500 bootstrap replicates. Branch lengths represent the number of nucleotide substitutions per site.

2.4.4. Phylogenetically informative indels

Manual examination of the concatenated dataset revealed six phylogenetically informative indels (Figure 2.9), all of which were found within intron regions. Four indels found in *AMBN* intron 2, *LEPR* intron 6, *PDE6C* intron 10, and *UCP1* intron 2 were shared among two-horned rhinoceroses, but not the one-horned Indian rhinoceros or Malayan tapir. Similarly, a deletion in intron 3 of the *PDE6C* locus only occurs in the

one-horned Indian rhinoceros and Malayan tapir, but not the two-horned rhinoceroses. Only one deletion found within the dataset, in *GRK7* intron 2, supports the phylogenetic segregation of rhinoceroses according to their geographic locations as it is shared only amongst Asian species and is not found in the African rhinoceroses or the Malayan tapir.

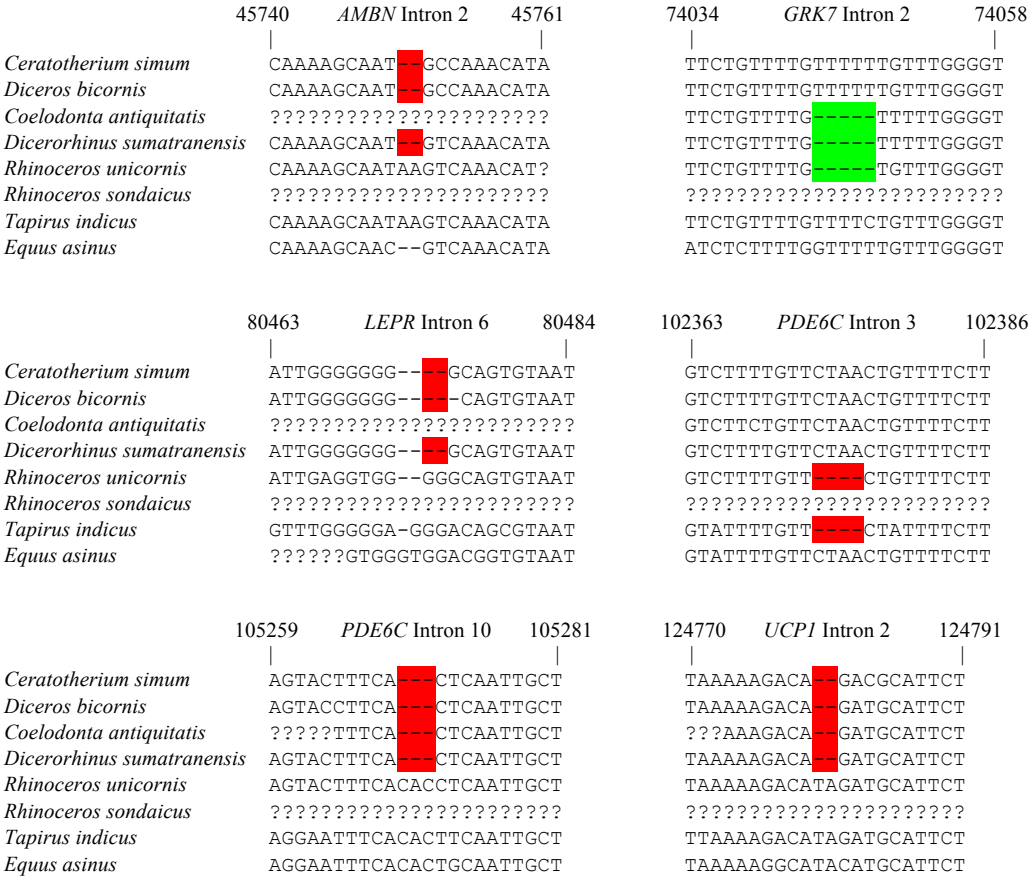


Figure 2.9. Alignment segments showing phylogenetically informative insertions and deletions. Numbers correspond to the locations within the concatenated alignment. Indels linking two-horned rhinos or Asian rhinos are highlighted in red and green, respectively.

2.4.5. Eye gene selection pressure results

All rhinoceros eye genes were intact with no apparent nonsense, splice site, or frameshift mutations; thus, were treated in the same manner as all other genes for use as phylogenetic markers. Omega (ω) values from selection pressure analyses acquired using

"free-ratio" models in CODEML are summarized in figure 2.10. The twelve eye genes examined appear to be largely under strong purifying selection pressures (ω<1) for *Camelus ferus*, *Sus scrofa,* and *Bos taurus,* as well as the stem equine ancestor, and the Malayan tapir. The nearest omega values to the neutral value of ω = 1 were found for the white rhinoceros *GUCA1B* (ω = 0.73) and *OPN4* loci (ω = 1.29) and for the Javan rhinoceros *GNAT2* (ω = 1.23), however, the latter species had low sequence coverage for the coding sequence of this gene (~27%). Also the stem ceratomorph had an omega value for *CNGB3* of 0.69 while the stem perissodactyl *CNGB3* and *GNGT2* genes had respective omega values of 0.71 and 1.39. Several genes had infinite omega values due to the denominator in the dN/dS ratio equaling zero. However, all genes examined appear to be under strong purifying selection in the stem rhinoceros ancestor.

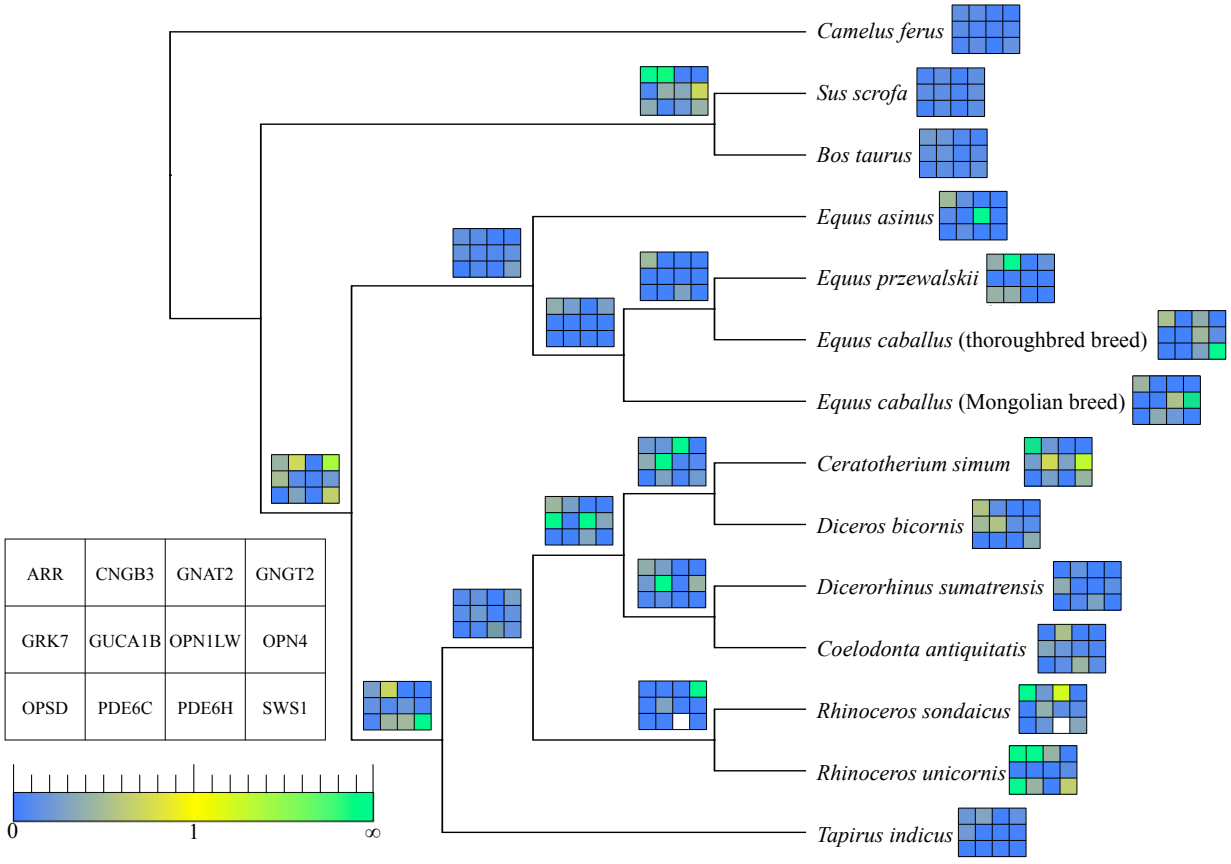


Figure 2.10. Omega (ω) values represented by colour for twelve eye genes acquired using the "free-ratio" model in CODEML. The 12 genes (*ARR*, *CNGB3*, *GNAT2*, *GNGT2*, *GRK7*, *GUCA1B*, *OPN1LW*, *OPN4*, *OPSD*, *PDE6C*, *PDE6H*, *SWS1*) are ordered in each rectangle as depicted in the legend. Purifying selection, neutral evolution and positive selection are characterized by $\omega < 1$, $\omega = 1$, and $\omega > 1$, respectively. The Javan rhinoceros, *Rhinoceros sondaicus*, lacked sequenced data for the *PDE6H* locus, which are denoted by blank boxes. Some omega values are infinite due to the denominator in the dN/dS ratio equaling zero.

2.5. Discussion

2.5.1. Hybridization capture and next-generation sequencing

With the exception of the Javan rhinoceros, modern DNA libraries of all species provided relatively high coverage of nuclear markers targeted for phylogenetic analyses. The highest sequence coverage was achieved for the Malayan tapir at 93.7% (Figure 2.2), which is surprising considering that hybridization capture experiments were performed

using probes designed to match white rhinoceros genomic targets. Thus, despite over 53 million years of divergence between tapirids and rhinocerotids (Kapur and Bajpai 2015), this study exemplifies that in-solution hybridization techniques can be highly effective for inter-familial gene capture. Springer et al. (2015) reported similar levels of success with chip-based hybridization capture experiments of sirenian genomic DNA (including the extinct Steller's sea cow, *Hydrodamalis gigas*) using probes designed from the African elephant (*Loxodonta africana*) and Cape rock hyrax (*Procavia capensis*).

The relatively high sequence coverage (88.1%) for the Sumatran rhinoceros (Figure 2.1) was also noteworthy bearing in mind the source of the tissue sample for this species was the blood meal of a leech. Although contamination from the leech itself and bacteria present within the digestive system of the leech would be expected, the sequenced Sumatran rhinoceros library was largely free from contamination as indicated by the relatively high percentage of reads matching as BLAST hits to the white rhinoceros genome (Figure 2.2). Consequently, the target enrichment process using hybridization capture techniques appears to have been highly effective for this DNA library.

Slightly lower sequence coverage was attained for the black and Indian rhinoceroses (each at ~83%; Figure 2.2) and may stem from the necessity to perform whole-genome amplifications of the previously extracted DNA samples in order to have adequate concentration of DNA for subsequent library construction procedures. Although, the REPLI-g Mini Kit (Qiagen, Toronto, Ontario, Canada) utilizes the multiple displacement amplification (MDA) approach, which allows for a more uniform amplification than PCR-based WGA methods, amplification bias is still known to occur

(Dean et al. 2002; Pinard et al. 2006). This happens when various areas of the genome are preferentially amplified over others, leading to over- or under-sampling of some genomic regions and a less diverse DNA library. An additional shortcoming of the MDA method WGA is that a small proportion (~0.5%) of chimeric DNA sequence can be formed due to occurrences of mispriming (Lasken and Stockwell 2007; Sabina and Laemon 2015). Indeed a small number of chimeric reads were identified in the black rhinoceros library, likely originating from this process. Sequenced reads for the black and Indian rhinoceroses were also slightly shorter on average than those of the Malayan tapir and Sumatran rhinoceros (Table 2.2), which likely contributed to the slightly lower coverage (Figures 2.2 and 2.3).

The Javan rhinoceros libraries provided comparatively low sequence coverage (~10%) with only ~1% of the reads assembling to the targeted regions (Figure 2.2). Furthermore, only ~5% of sequenced reads matched as BLAST hits to the white rhinoceros genome (Figure 2.2) indicating that these results stem from extensive exogenous DNA contamination, likely from bacterial and fungal sources. This level of contamination was much higher than that of all other modern libraries and even the ancient woolly rhinoceros libraries. Since both exogenous and endogenous DNA reads contained A-adaptor barcode indexes, microbial contamination of the Javan rhinoceros bone samples must have preceded the construction of the two DNA libraries and may be attributed to storage of the bone samples as museum specimens after they were collected from deceased individuals (Brook et al. 2011). This likely also explains the deterioration of the DNA as several sites of probable DNA damage were noticed in some of the Javan rhinoceros reads. As noted by Hagelberg and Clegg (1991), improper storage of animal

bones can result in fungal growth and an increased rate of DNA decay. Unfortunately, Javan rhinoceros tissue types other than bone are not allowed outside of Indonesia; thus, tissue samples suitable for use in DNA extractions were extremely limited. Nevertheless, to my knowledge, this was the first study to partially sequence these nuclear markers (Appendix 1) for this exceedingly rare species.

Sufficient quantities of endogenous DNA can be challenging to recover from ancient tissue specimens >10,000 years old as exemplified by the failure to amplify the two woolly rhinoceros libraries originating from bone sample WR1 and skin sample WR5 (Appendix 2) with post-hybridization capture PCRs. Nonetheless, sequencing results for ancient DNA libraries derived from three permafrost-preserved tooth and bone samples (WR2, WR3, and WR4; Appendix 2) proved to be highly successful. During the degradation process, ancient DNA becomes increasingly fragmented with age (Dabney et al. 2013) therefore; shorter fragment lengths (averaging 92 bp; Table 2.2) were targeted during size selection of the woolly rhinoceros libraries compared to most modern libraries. This produced shorter assemblies of the 5' and 3' intronic regions flanking each exon and contributed to the lower overall coverage the woolly rhinoceros libraries compared to most modern rhinoceros libraries (Figures 2.2 and 2.3). BLAST hits of the sequenced reads to the white rhinoceros genome were also substantially lower for the ancient woolly rhinoceros (~35%) compared to most modern rhinoceroses (>84% excluding the Javan rhinoceros), signifying that exogenous DNA contamination was much more prevalent in ancient samples, as expected. The authenticity of these woolly rhinoceros ancient DNA sequences was bolstered by the addition of P-adaptor barcodes to the 3' ends of the reads and matches of bycatch reads to the previously sequenced

mitochondrial genome (Willerslev et al. 2009) of this extinct species. By contrast, Springer et al. (2015) reported only ~8% of reads sequenced from ~1000 year old ancient DNA extracted from extinct Steller's sea cow museum bone samples aligned as BLAST hits to the Florida manatee (*Trichechus manatus latirostris*) genome.

Two previous studies recovered relatively small regions of the woolly rhinoceros nuclear genome and in both instances they were nuclear mitochondrial insertions (termed NUMTs; i.e. segments of mitochondrial DNA that have been transposed in the nuclear genome). Orlando et al. (2003) sequenced NUMTs while targeting mitochondrial *12S rRNA* and *cytb* genes from two woolly rhinoceros specimens. Similarly, Binladen et al. (2006) sequenced woolly rhinoceros NUMTs from two samples to evaluate the effects of postmortem nucleotide damage in ancient DNA. Thus, to my knowledge, the current study is the first to sequence non-NUMT nuclear targets from this extinct species.

2.5.2. Rhinoceros phylogenetics

Size matters when it comes to molecular phylogenetics as large datasets provide greater opportunity to resolve contentious relationships (Chou et al. 2015; Gatesy and Springer 2014). This dataset surpasses all other previous molecular studies attempting to elucidate the rhinoceros family tree, being nearly 16 times larger than that of Steiner and Ryder (2011). Furthermore, this dataset is comprised entirely of nuclear markers, which should provide greater resolving power for phylogenetic analyses than mitochondrial genes (Springer et al. 2001; Willerslev et al. 2009). To my knowledge, all previous molecular studies aiming to resolve the evolutionary history of rhinoceroses have focused on the most common analytical approach; concatenation of gene coding regions. In contrast, my study utilized both conventional concatenation and more recently developed

coalescence methods to analyze both coding and non-coding markers. Both methods have limitations; concatenation does not account for the effects of incomplete lineage sorting (ILS) while coalescence can overlook the effects of recombination (Gatesy et al. 2016; Chou et al. 2015); thus, the use of both analytical forms provided a broader scope of results to assess the rhinoceros phylogeny.

All inferred phylogenetic trees provided robust support for the suborders Hippomorpha and Ceratomorpha, confirming that tapirs are more closely related to rhinoceroses than to horses. The three rhinoceros sister pairings were also strongly supported in all analyses with ~100% bootstrap support or a posterior probability value of 1. However, a consensus was not reached regarding the phylogenetic relationships between the three subfamilies (Dicerorhininae, Rhinocerotinae, and Dicerotinae).

The two coalescence approaches used in this study yielded conflicting results, signifying that at least one set of results are incorrect. The ASTRAL-II summary coalescence method (Figure 2.4) grouped Asian rhinoceroses together (rhinocerotines and dicerorhines) and distinct from African rhinoceroses (dicerotines), supporting the "biogeographical" hypothesis proposed by Groves (1983), Pocock (1945), and Prothero et al. (1986). Indeeed, the majority of GARLI gene trees supported the ((dicerorhines,rhinocerotines),dicerotines) relationship with the two remaining non-trichotomy phylogenies [((dicerorhines, dicerotines),rhinocerotines) and ((dicerotines,rhinocerotines),dicerorhines)] being less, but nearly equally, frequent, generally fitting with multispecies coalescent predictions of gene tree distribution (Degnan and Rosenberg 2009). This topology agrees with the mitochondrial *12S rRNA* and *cyt b* concatenation results of Orlando et al. (2003) and Tougard et al. (2001), as well

as those of Price and Bininda-Emonds (2009) using a dataset heavily biased towards mitochondrial loci (33 mitochondrial and 6 nuclear genes). Maximum likelihood and maximum parsimony bootstrap support values were below 70% for phylogenies performed by Orlando et al. (2003), while they were 97% and 77%, respectively, for those performed by Tougard et al. (2001). Phylogenies performed by Price and Bininda-Emonds (2009) support this relationship with 84 and 92% bootstrap values from maximum parsimony and maximum likelihood analyses, respectively, and 0.99 posterior probability from Bayesian analyses. By contrast, the single-site coalescence method, SVDQuartets (Figure 2.5), grouped the two-horned rhinoceroses (dicerorhine and dicerotine) together to the exclusion of the single-horn rhinoceroses (rhinocerotines), supporting the "two-horn" hypothesis proposed by Simpson (1945) and Loose (1975). This topological arrangement is also in line with the findings of Steiner and Ryder (2011) based on their combined dataset of 9 nuclear and 2 mitochondrial loci, supported by bootstrap values of 61, and 92 for maximum parsimony and maximum likelihood analyses, respectively, and a posterior probability of 0.71 for Bayesian analyses.

Several issues can cause discrepancies when using summary coalescent analyses, most of which stem from gene tree estimation error (Chou et al. 2015). However, in this case, measures were taken to increase gene tree accuracy. For instance, gene trees were created with GARLI 2.0 instead of RAxML in order to allow for polytomies and avoid particularly weak or unsupported branches below 1e-8 in length, an approach also used by He et al. (2016) with talpid moles. Approximately one quarter (51/199) of the gene trees produced in GARLI resulted in polytomies at the base of Rhinocerotidae, showing insufficient resolution between the three rhinoceros subfamilies, which reduced the

effective size of the dataset but avoided potentially arbitrary bifurcations. Another factor to consider with summary coalescence is that each gene alignment is represented by only a single gene tree. Thus, markers composed of long coding sequence by artificially concatenating multiple smaller exons separated by sizeable introns, termed "concatalescence", as performed by Song et al. (2012), should be avoided as this can ignore the effects of recombination within exons of a gene and cause misleading results (Gatesy and Springer 2014; Gatesy et al. 2016; Scornavacca and Galtier 2017). Therefore, further efforts were made to avoid the introduction of erroneous gene trees by creating 199 relatively short gene segments (averaging ~660 bp), typically consisting of only a single exon and its flanking 5' and 3' intron sequence. Taken together, these efforts to optimize gene tree accuracy should bolster support for the ASTRAL-II results, however, they are in direct contradiction to the SVDQuartets SNP approach, which is unsusceptible to gene tree estimation error, though may also fail to identify the optimal phylogeny quartet trees are still required to be agglomerated.

Unlike the conflicting results obtained from the two coalescent methods, the maximum likelihood and Bayesian concatenation analyses of the complete dataset only differed in their level of support for the node placing one-horned rhinoceroses as a sister group to all two-horned rhinoceroses in agreement with conclusions reached by Steiner and Ryder (2011). The branch lengths separating the two- and one-horned rhinoceroses were very short ($6.5x10^{-4}$ for both RAxML 199 and 591 partition trees) indicating that overall the nucleotide sequences of all rhinoceroses were very similar and that the three modern rhinoceros lineages all arose within a short time period. Indeed, Steiner and Ryder (2011) noted a similarly short branch length and estimated the one-horned Indian

rhinoceros diverged from the most recent common ancestor of all modern rhinoceroses 26 MYA while the Sumatran rhinoceros split from African rhinoceroses only ~1 million years later. The RAxML tree provided only 59 and 63% bootstrap support for this relationship depending on the partitioning scheme (Figure 2.6), whereas the MrBayes tree provided a much stronger 0.99 posterior probability (Figure 2.7). The low bootstrap support values for this node in the RAxML analyses are unsurprising considering that opposing topologies were attained when the dataset was split into intron and exon sequences (Figure 2.8). Concatenation analyses of intron sequences supported the "two-horn" hypothesis with an 85% bootstrap value, agreeing with the branching arrangement acquired from SVDQuartets and concatenation results of the entire 131 kb supermatrix. While a small fraction of the intron sequences included in my dataset are expected to contain transcriptional regulatory elements (e.g. promoters) that would likely be evolving under natural selection, the majority of these non-coding markers are presumed to be accruing random nucleotide substitutions at a fairly constant rate predominantly under neutral evolution and should be largely free from any potentially confounding effects of natural selection. Conversely, exon sequences, evolving under the constraints of natural selection, produced a topology in support of the "biogeographical" hypothesis with an 83% bootstrap value, matching the topology recovered with ASTRAL-II. The lack of selection pressures in non-coding DNA sequences is expected to be beneficial as shared substitutions are more likely to reflect true phylogenetic relationships, whereas mutations within exons may affect protein function and therefore have a greater chance of arising due to parallel evolution (Bailey et al. 1991). Indeed, manual examination of the sequence data revealed that the vast majority of coding regions are highly conserved

rhinoceroses with very few informative sites, hence the importance of assembling large datasets. It is also worth noting that, although the intron-concatenated supermatrix was slightly longer than that of exons (67,724 versus 64,207), higher coverage (i.e. fewer gaps) was attained for exons in each species (Figure 2.3), however, the number of informative sites within introns may be higher due to the level of conservation within coding regions and a higher rate of mutation expected for non-coding sequences (Gojobori et al. 1982; Li et al. 1984).

One added benefit of including introns as markers in this study was that shared indels were discovered within these regions that can offer clues to the phylogenetic relationships among rhinoceroses. No indels were found within exons, a testament to the level of conservation within these protein-coding regions. Similarly, Steiner and Ryder (2011) used this method within one intron of the *Kit* gene to show shared deletions linking ceratomorphs (numerous indels supporting this clade were also found in the present study; data not shown). These molecular synapomorphies, unlike nucleotide substitutions, are not taken into account by most phylogenetic programs such as RAxML and Mr.Bayes that treat all gaps as unknown nucleotides (i.e. Ns). Five out of six indels supported the pairing of dicerotines and dicerorhines (two-horned species) and showed that the one-horned Indian rhinoceros shared the same (likely ancestral ceratomorph) characteristics as the Malayan tapir (Figure 2.9), providing independent support for the "two-horn" hypothesis as opposed to the "biogeographical" hypothesis.

Chou et al. (2015) tested the accuracy of ASTRAL-II, SVDQuartets, and RAxML concatenation with short gene segments (<200 bp). They found that the accuracy of the

analyses varied with the level of ILS in the dataset. ILS is known to be especially prevalent as a result of quick radiations as appears to have occurred in rhinoceroses where all three modern rhinoceros lineages are estimated to have arisen within ~1 million years (Steiner and Ryder 2011). Indeed, this phenomenon is known to even confuse the phylogeny of great apes where approximately 30 percent of genomic markers support relationships in contradiction to the correct phylogeny closely linking humans and chimpanzees to the exclusion of the more distantly related gorilla (Galtier and Daubin 2008). Chou et al. (2015) found that under low ILS, RAxML concatenation analyses were the most reliable. However, under high levels of ILS, ASTRAL-II outperformed both RAxML concatenation, corroborating the findings of Mirarab and Warnow 2015, and SVDQuartets. The latter coalescent method was found to be comparable to ASTRAL-II under low ILS but does not perform with the same accuracy under high ILS. Given the level of branching variation among GARLI gene trees and the average scaled RF distance for RAxML gene trees compared to species trees of ~0.45 indicating that nearly half of the ceratomorph relationships between gene and species trees are in conflict, it does appear that high levels of ILS are a confounding factor for rhinoceros nuclear markers included in this study.

Although the majority of the phylogenetic results inferred from molecular data presented in this study support the evolutionary arrangement of rhinoceroses according to their horn-number, this characteristic should not be the only morphological trait used to compare these species. Indeed, Groves (1983) examined 42 derived characteristics of the dentition, skull and skeleton in the Sumatran, Indian, white, and black rhinoceroses. This study reported 14 synapomorphies between the two Asian species, while 7 were present

between the Indian and African rhinoceroses, and only one synapomorphy (the loss of I_1 incisors) linked the two-horned species (Sumatran and African rhinoceroses). This lead Groves (1983) to conclude that the Asian rhinoceroses formed a distinct evolutionary taxon, placing the Indian and Sumatran rhinoceroses in a tribe termed "Rhinocerotini" and the African rhinoceroses in tribe "Dicerotini". While this may be the most convincing morphological examination of modern rhinoceroses to date, it is possible that convergent evolution among Asiatic species, spurred by overlapping geographical ranges of ancestral forms and similarity between ecological niches, obscures the cladistics, as has been well described for numerous mammals (e.g. pangolins and xenarthrans; Reiss 2001). It could also be the case that the African rhinoceroses evolved more derived morphological traits suited to the African ecosystems which may have been mischaracterized as being primitive features due to the selection of an extinct *Aceratherium* as an outgroup species. Indeed, members of this genus are not necessarily representative of ancestral rhinoceros characters, as they have been recently shown by Deng et al. (2013) to have also evolved numerous derived features of their own that suited the ecological niches they exploited.

In short, the contradictory phylogenies resulting from the nuclear sequence data may be indicative of the confounding factors that plague the rhinocerotid lineage, such as ILS. Given the extremely short branch lengths separating the three rhinoceros lineages acquired from both coalescent and concatenation methods, it is possible that the Dicerorhininae-Rhinocerotinae-Dicerotinae split occurred in such a short time frame that the molecular phylogeny between these lineages is unresolvable, using existing analyses. In fact, Willerslev et al. (2009) drew the same conclusion upon examination of the entire mitochondrial genomes of the same rhinoceros species. While the majority of the

evidence provided here supports the division of rhinoceroses according to their horn number, further research is needed to confirm this relationship. Perhaps future studies could include more nuclear loci or even the complete nuclear genome.

2.5.3. Rhinoceros eye gene selection pressure

Examination of all eye genes revealed that the coding sequences were fully intact and free from deleterious (nonsense, frameshift and splice site) mutations that would be expected for pseudogenes. Furthermore, PAML selection pressure results indicated that all eye genes in the stem rhinoceros ancestor were under strong purifying selection (Figure 2.10). While, the Javan rhinoceros did display a near neutral ω value for the *GNAT2* locus (ω = 1.23; Figure 2.10) this result should be taken with caution as very little coverage was attained for this species and its sequenced reads do shows several sites with G→A, C→T transitions, consistent with DNA damage (Binladen et al. 2006; Brotherton et al. 2007, Hofreiter et al. 2001 b). The white rhinoceros *GUCA1B* and *OPN4* genes also had omega values approaching neutrality (ω = 0.73 and 1.29, respectively) suggesting that these genes may be evolving under relaxed selection pressures in this species.

The *GUCA1B* (guanylate cyclase activator 1B) protein is abundantly expressed in both photoreceptor types of the retina—rod and cone cells—and functions to activate membrane-bound guanylate cyclase to return stimulated photoreceptors to their inactivated state (Sato et al. 2005; Payne et al. 1999). In humans, mutations in this gene have been known to result in a neurodegenerative disorder called retinitis pigmentosa, which reduces peripheral vision as well as visual acuity in low light levels (Sato et al.

2005). Interestingly, Emerling and Springer (2015) reported the inactivation of this gene in the nine-banded armadillo (*Dasypus novemcinctus*). Further investigation would be needed to determine if *GUCA1B* mutations in rhinoceroses contribute to an impairment of vision.

The *OPN4* gene encodes for the melanopsin protein, a photopigment found in photosensitive retinal ganglion cells. While this photosensitive protein does not play a role in the formation of visual images, it does function to regulate important processes such as the circadian rhythm and melatonin production (Borges et al. 2012). A study that estimated the selection pressure of the OPN4 genes of 26 vertebrates, including 6 mammals, found that this locus evolved under strong purifying selection pressure as indicated by global ω values <0.17. Similarly, Dong et al. (2012) reported a global ω value of 0.07 for the *OPN4* gene in a dataset that included 20 mammals; thus a value of 1.29 in the white rhinoceros is rather interesting. Further experiments may focus on characterizing potential functional effects of amino acid substitutions in the white rhinoceros melanopsin protein and with respect to other rhinoceroses.

Rhinoceroses have been long-thought to be near-sighted (myopic) and have very poor vision based mainly upon behavioural studies (Nowak 1999; Skinner and Chimimba 2005). However, research by Pettigrew and Manager (2008), examining the cellular make up of a black rhinoceros retina, showed that the ganglion cell density is higher than expected for a mammal with myopia and similar to that of many other mammals including rabbits, rodents, seals and dolphins. Thus, the visual acuity of rhinoceroses is still debatable and further research is needed to reveal the complete picture.

2.6. Conclusions

Despite assembling the largest molecular dataset of rhinocerotid DNA sequence to date, this study resulted in conflicting phylogenetic relationships among rhinoceros subfamilies, which have also plagued previous studies. Estimated phylogenies using concatenation and coalescent analyses of the entire ~131 kb dataset, yielded topologies linking either Asian versus African ((Dicerorhininae, Rhinocerotinae) Dicerotinae) or one- versus two-horned rhinoceroses (Rhinocerotinae (Dicerorhininae, Dicerotinae)). The same conflicting relationships were also obtained when concatenation analyses were performed on exon versus intron subsets of the 131 kb supermatrix. Furthermore, the phylogenies produced extremely short branch lengths separating the three rhinoceros subfamilies, indicating low molecular divergence between these lineages. Several factors may be to blame for the conflicting topologies including ILS, which is associated with rapid speciation events. However, interestingly, phylogenetically informative indels predominantly support the one- versus two-horned hypothesis. Future studies examining the molecular phylogenetics of rhinoceroses may benefit from performing whole-genome analyses, as the resulting dataset may be sufficiently large to provide a consistent resolution. This study also highlights the importance of performing multiple types of phylogenetic analyses to ensure congruence between various tree-building methods.

While rhinoceros eye genes first retrieved as phylogenetic markers provided an interesting opportunity to investigate these loci for inactivations that may in part explain nearsightedness among this clade, the genes are intact and generally evolving under purifying selection pressures. However, two loci (*GUCA1B* and *OPN4*) in the white rhinoceros appear to be evolving under relaxed selection pressures and may provide targets for future studies.

CHAPTER 3: EVOLUTION OF *UCP1* TRANSCRIPTIONAL REGULATORY ELEMENTS ACROSS THE MAMMALIAN PHYLOGENY

Authors: Michael J. Gaudry and Kevin L. Campbell

Author contributions:

> MJG conceived of the project, designed research, prepared DNA libraries, performed hybridization capture experiments, conducted sequencing and genome-mining, performed comparative bioinformatic analyses, prepared the figures, interpreted the results, and drafted the manuscript.
>
> KC conceived of the project, designed research, interpreted the results, and revised the manuscript.

Please note: a version of this article has been submitted to Frontiers in Physiology for consideration for publication.

3.1. Abstract

Uncoupling protein 1 (UCP1) permits non-shivering thermogenesis (NST) when highly expressed in brown adipose tissue (BAT) mitochondria. Exclusive to placental mammals, BAT has commonly been regarded to be advantageous for thermoregulation in hibernators, small-bodied species, and the neonates of larger species. While numerous regulatory control motifs associated with *UCP1* transcription have been proposed for murid rodents, it remains unclear whether these are conserved across the eutherian mammal phylogeny and hence essential for UCP1 expression. To address this shortcoming, we conducted a broad comparative survey of putative *UCP1* transcriptional regulatory elements in 139 mammals (135 eutherians). We find no evidence for presence of a *UCP1* enhancer in monotremes and marsupials, supporting the hypothesis that this

control region evolved in a stem eutherian ancestor. We additionally reveal that several putative promoter elements (e.g. CRE-4, CCAAT) identified in murid rodents are not conserved among BAT-expressing eutherians, and together with the putative regulatory region (PRR) and CpG island do not appear to be crucial for UCP1 expression. The specificity and importance of the upTRE, dnTRE, URE1, CRE-2, RARE-2, NBRE, BRE-1, and BRE-2 enhancer elements first described from rats and mice are moreover uncertain as these motifs differ substantially—but generally remain highly conserved—in other BAT-expressing eutherians. Other *UCP1* enhancer motifs (CRE-3, PPRE, and RARE-3) as well as the TATA box are also highly conserved in nearly all eutherian lineages with an intact *UCP1*. While these transcriptional regulatory motifs are generally also maintained in species where this gene is pseudogenized, the loss or degeneration of key basal promoter (e.g. TATA box) and enhancer elements in other *UCP1*-lacking lineages make it unlikely that the enhancer region is pleiotropic (i.e. co-regulates additional genes).

3.2. Introduction

3.2.1. Evolution of uncoupling protein 1

Uncoupling protein 1 (UCP1) expression is a defining characteristic of brown adipose tissue (BAT), allowing this specialized eutherian heater organ to function in non-shivering thermogenesis (NST). UCP1 spans the mitochondrial inner-membrane of brown adipocytes, acting to promote mitochondrial proton leak, which dissipates the electrochemical gradient that typically drives ATP synthase. In an effort to defend the mitochondrial protonmotive force, the electron transport chain thus pumps protons into

the inter-membrane space at an elevated rate via an increased level of substrate oxidation, thereby resulting in substantial heat production in the form of NST (Cannon and Nedergaard 2004; Klingenspor and Fromme 2012).

In addition to UCP1 expression, BAT is also characterized by the presence of multilocular lipid droplets that, due to their amplified surface area, can be readily oxidized by an abundance of mitochondria (Cannon and Nedergaard 2004; Klingenspor and Fromme 2012). In fact, mitochondria occur in such high quantities within BAT that they impart the tissue with its distinctive brown coloration due to their high iron content. Vital to its function, BAT is highly vascularized and localized primarily to the thoracic region, lying adjacent to major blood vessels of the heart (e.g. the Sulzer's vein) permitting effective transfer of NST heat to the rest of the body via the circulatory system (Klingenspor and Fromme 2012; Oelkrug et al. 2015). Overall, BAT-mediated NST is regarded as a more efficient means of heat production than shivering thermogenesis, which has major drawbacks as it impedes locomotion and produces heat in large muscle groups of the limbs that are prone to heat loss due to their high surface area to volume ratios (Oelkrug et al. 2015). For these reasons, UCP1 is widely considered to have provided a key thermoregulatory and evolutionary advantage to the eutherian lineage, particularly for small-bodied and hibernating species, and, while BAT in larger-bodied species (e.g. humans) is typically lost with the onset of adulthood, it has been generally understood to play vital role in their neonates (Cannon and Nedergaard 2004).

The *UCP1* gene predates the divergence of ray- and lobe-finned fishes (420 million years ago [MYA]) and can be distinguished from *UCP2* and *UCP3* paralogs by its conserved synteny among vertebrates, as *UCP1* is flanked by the upstream *TBC1D9*

and downstream *ELMOD2* loci (Jastroch et al. 2008; Klingenspor et al. 2008). UCP2 and UCP3 have been long-believed to play non-thermogenic roles, and are instead hypothesized to perform a multitude of functions including the reduction of reactive oxygen species by promoting a low level of mitochondrial proton leak when activated by fatty acids (Brand and Esteves 2005; Echtay 2007; Mailloux and Harper 2011). However, a recent study by Lin et al. (2017) suggests that proton uncoupling by UCP3 permits heat production in beige adipose tissue of pigs, compensating for the loss of UCP1 in this lineage (Berg et al. 2006). Nevertheless, the functional roles of both UCP2 and UCP3 remain hotly debated. Similarly, the ancestral function of UCP1 in non-eutherians is currently unclear (Klingenspor et al. 2008). UCP1 expression has been shown to increase with cold exposure in common carp (*Cyprinus carpio*) brain tissue, suggesting a possible a role in local thermogenesis (Jastroch et al. 2007). Interestingly, thirteen-lined ground squirrels (*Spermophilus tridecemlineatus*) have been suggested to promote UCP1 expression in neuronal tissue during hibernation to serve a similar purpose (Laursen et al. 2015). However, to date, this protein has not been definitively linked to heat production in ectothermic vertebrates (Jastroch et al. 2007). While the fat-tailed dunnart (*Sminthopsis crassicaudata*), a marsupial, displays a primitive "brownish" interscapular adipose depot that up-regulates UCP1 expression in response to cold exposure (Jastroch et al. 2008), this tissue is incapable of adaptive NST (Polymeropoulos et al. 2012) with no study demonstrating that UCP1 contributes to NST in marsupials. Although *UCP1* appears to have been inactivated early in the evolution of the eutherian superorder Xenarthra (Gaudry et al. 2017), BAT-mediated adaptive thermogenesis is widely known to occur in small-bodied members of the superorders Laurasiatheria and Euarchontoglires

(Oelkrug et al. 2015), and has been documented in the rock elephant shrew (*Elephantulus myurus*; Mzilikazi et al. 2007) and the lesser hedgehog tenrec (*Echinops telfairi*; Oelkrug et al. 2013), both members of the eutherian superorder Afrotheria. These observations strongly suggest that UCP1 was recruited for BAT-mediated NST in a common eutherian ancestor by gain of function mutations in the amino acid sequence of the protein and/or greater control over gene transcription that allowed highly concentrated UCP1 expression within BAT mitochondria (Klingespor et al. 2008).

Consistent with the gain of function hypothesis, comparative phylogenetic analyses reveal that the stem eutherian branch is highly elongated in *UCP1* gene trees relative to that of *UCP2* and *UCP3* paralogs (Saito et al. 2008; Hughes et al. 2009; Gaudry et al. 2017; Figure 3.1). It is thus likely that an elevated rate of non-synonymous *UCP1* nucleotide substitutions in the stem eutherian branch conferred this protein with the ability to facilitate proton leak at physiologically significant levels (Jastroch et al. 2008; Klingenspor et al. 2008). While Saito et al. (2008) first proposed *UCP1* evolved under positive selection in basal eutherians, more recent selection pressure analyses reveal non-synonymous to synonymous substitution ratios (dN/dS or ω) of ~0.5-0.6 that are more consistent with relaxed purifying selection (Hughes et al. 2009; Gaudry et al. 2017). However, given that UCP1 of placental mammals possess several unique amino acids relative to non-eutherians, it is possible that directional selection was limited to certain codons along the stem eutherian branch, though, so far this hypothesis remains statistically unsupported (Hughes et al. 2009; Gaudry et al. 2017).

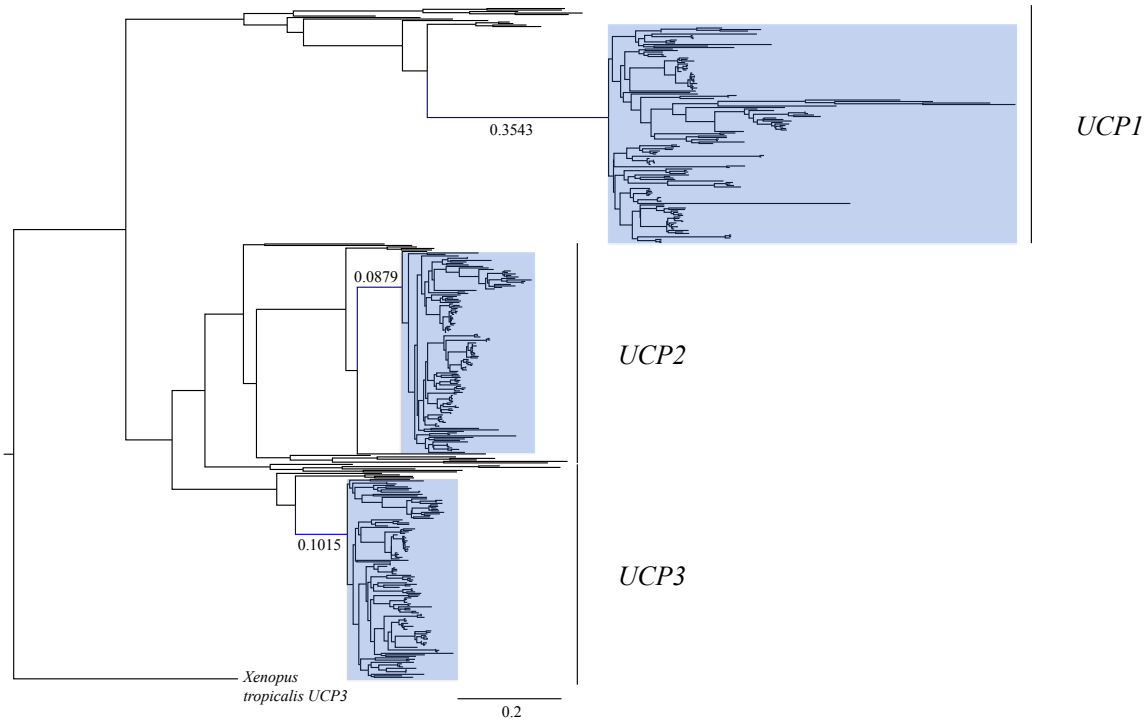


Figure 3.1. Maximum likelihood gene tree of *UCP1*, *UCP2*, and *UCP3* coding sequences (N=448) modified from Gaudry et al. (2017) to include the sixteen additional species with recently available genome projects (see Table 1). The stem placental mammal branches are indicated in blue. Note that the *UCP1* stem placental branch is much longer than those of *UCP2* and *UCP3*, demonstrating a greater number of nucleotide substitutions per site. Placental mammal genes are highlighted with blue boxes. The tree was rooted with the western clawed frog (*Xenopus tropicalis*) *UCP3*.

Along with the increased rate of UCP1 evolution in stem eutherians, expression of this protein also became highly tissue-specific during the rise of BAT (Cannon and Nedergaard 2004). In contrast to the seemingly constitutive presence of UCP1 in common carp brain, liver, and kidney tissues (Jastroch et al. 2007), eutherian UCP1 expression is tightly regulated, occurring predominantly in BAT (Cannon and Nedergaard 2004). One notable exception, however, is the recently discovered "beige or brite (brown in white)" adipocytes in rodents (mice and rats) and humans. These are derived from white adipose cells that, upon cold exposure, become BAT-like by expressing UCP1 and by having multilocular lipid droplets and an elevated mitochondrial

concentration (Harms and Seale 2013). An important distinction in BAT (and *UCP1*) evolution is that BAT-dependent NST relies upon exceptionally high levels of UCP1 expression, constituting up to 10% of the mitochondrial membrane proteins, whereas UCP2 and UCP3 expression is several orders of magnitude lower (0.01-0.1%) in other tissues (Brand and Esteves 2005). Interestingly, an enhancer box has been well documented to play a major role in eutherian *UCP1* gene transcription, but is absent in the gray short-tailed opossum (*Monodelphis domestica*; Jastroch et al. 2008), suggesting that it originated with the advent of eutherian UCP1-mediated NST, thus highlighting the importance that gene regulation likely played in the rise of eutherian BAT-mediated thermogenesis.

Given the thermoregulatory advantages conferred by BAT, it is believed that this tissue was fundamental to the evolutionary success of eutherian mammals, and it has even been hypothesized to underlie their colonization of cold ecological niches (Cannon and Nedergaard 2004). The documented inactivation of the *UCP1* gene in suids (pigs) (Berg et al. 2006) initially emphasized the importance of BAT-mediated thermogenesis, as this inactivation appears to have had detrimental consequences as newborn piglets are widely known to have meager thermoregulatory abilities, suffering from high infant mortality when cold-stressed and relying upon shivering thermogenesis and maternal nest-building in order to maintain homeothermy (Herpin et al. 2002; Berg et al. 2006). By contrast, two recent studies (Gaudry et al. 2017; McGaugh and Schwartz 2017) contested the conventional belief regarding the importance of BAT-mediated NST throughout the course of placental evolution. Indeed, Gaudry et al. (2017) not only detailed ancient pseudogenization events of *UCP1* in eight additional eutherian lineages: Equidae

(horses), Cetacea (whales and dolphins), Proboscidea (elephants and mammoths), Sirenia (sea cows), Hyracoidea (hyraxes), Pholidota (pangolins), Pilosa (sloths and anteaters), and Cingulata (armadillos), but concluded that extreme cold tolerance evolved in many of these groups in the absence of UCP1-mediated thermogenesis.

With the exception of xenarthrans and pangolins, who have adopted a strategy of reduced metabolic rates and body temperatures associated with their low energy diets, and pigs, for which no credible explanation for *UCP1* inactivation has yet been put forward, Gaudry et al. (2017) proposed that *UCP1* inactivations date back to a period of substantial planetary cooling ~55 to 22 MYA that triggered pronounced increases in body size in other UCP1-lacking lineages (Gaudry et al. 2017). The inverse relationship between the surface-area-to-volume ratio and size imparts greater retention of heat in larger bodied mammals, thus larger mammals have proportionally lower rates of heat production per gram of body mass (McNab 1983). This linkage is reflected in the diminishing fraction of eutherian body mass constituted by BAT, as well as a reduced NST capacity, with increasing body size (Heldmaier 1971; Oelkrug et al. 2015). Heldmaier (1971) further suggested that BAT-mediated NST is negligible for mammals >10 kg. Nonetheless, several large-bodied taxa retain an intact *UCP1* gene (e.g. rhinoceroses, pinnipeds, hippopotamus, and camel; Gaudry et al. 2017). Despite this finding, it remains conceivable that members of these groups do not express *UCP1* in BAT, even as neonates. For example, Rowlatt et al. (1971) noted the absence of BAT upon examination of a single newborn hippopotamus (*Hippopotamus amphibious*), while both UCP1 expression and discernable BAT was not detected in either Weddell seal (*Leptonychotes weddellii*) or hooded seal neonates (*Cystophora cristata*) (Pearson et al.

2014). Additionally, the Bactrian camel (*Camelus ferus*) *UCP1* gene displays a 12 base pair nucleotide deletion in exon 5 that would impart the loss of 4 amino acids in close proximity to a site that putatively binds purine nucleosides (e.g. guanosine diphosphate) to act as a regulator (inhibitor) of protein activity (Gaudry et al. 2017). Consequently, disruptions to *UCP1* regulatory regions may preclude expression of this protein in BAT of these lineages.

3.2.2. Evolution of eutherian *UCP1* regulatory elements

In eutherian mammals, the neuro-hormonal modulation and tissue-specific expression of *UCP1* is under the control of two regulatory regions in the 5' non-coding region of the gene—a complex distal enhancer region and a proximal promoter—through their interactions with a broad assemblage of transcription factors (Villarroya et al. 2017). Based primarily on murid rodent studies, several putative transcription factor binding motifs (see Figure 3.2) have been proposed within a conserved ~200 bp *UCP1* enhancer box located ~2–5 kb upstream of the transcriptional start site in eutherians (Cannon and Nedergaard 2004; Jastroch et al. 2008; Shore et al. 2012). For instance, two cAMP response elements (CREs) were discovered in mice and termed "CRE-3" and "CRE-2" (Kozak et al. 1994). CRE sites typically have a palindromic consensus sequence of 5'-T(G/T)ACGTCA-3' (Bokar et al. 1988; Kozak et al. 1994). While the first three nucleotides of the two mouse CREs deviate from the typical consensus sequence (Figure 3.2), the 5'-CGTCA-3' nucleotides remain conserved and are believed to be key for *UCP1* expression. Indeed, site-directed mutagenesis of these nucleotides within the enhancer CRE of glycoprotein hormone and phosphoenolpyruvate carboxykinase genes

has been shown to drastically reduce transcription factor (i.e. cAMP response element binding protein [CREB]) binding and expression in human and rat cells (Bokar et al. 1988). Two "brown adipocyte regulatory element" (BRE) protein-binding motifs (Kozak et al. 1994) also occur in the mouse *UCP1* enhancer box (Figure 3.2). Again, site directed mutagenesis of the "TTCC" nucleotides within the BREs to a "GTAC" sequence drastically reduces *UCP1* enhancer activity measured using transient expression assays (Kozak et al. 1994). This study further proposed that the CRE and BRE binding sites act in a cooperative and synergistic fashion to promote transcription. In addition, Sears et al. (1996) found a stretch of nucleotides they termed "UCP regulatory element 1" (URE1), though this is referred to as the peroxisome proliferator response element (PPRE) by Jastroch et al. (2008); Jastroch also predicted a second possible PPRE motif downstream of the URE1 (PPRE) site. The URE1 motif displays high similarity to DR-1 elements (Sears et al. 1996), which are known to comprise of two direct repeats of the "AGGTCA" half-site consensus sequence separated by a single nucleotide (hence the term DR-1; i.e. direct repeats separated by 1 spacer nucleotide). In mice this sequence occurs in the reverse and complement orientation of the first DNA strand (5'- TCACCCTTGACCA-3'), and although it is not an exact match to the consensus sequence, it has been shown to bind the peroxisome proliferator-activated receptor γ and retinoid X receptor α (PPARγ-RXRα) heterodimer transcription factor (Sears et al. 1996). Conversely, mutant variants of the URE1 sequence (i.e. 5'-TCACAATTGACCA-3' or 5'-TCACCCTAGACCA-3') failed to bind the PPARγ-RXRα transcription factor, suggesting a key role in the functionality of the *UCP1* enhancer (Sears et al. 1996). Additionally, in light of the requirement of triiodothyronine (T3) for proper BAT expression (Bianco and Silva 1987),

Rabelo et al. (1995) described two putative thyroid hormone response elements (TREs) in the rat *UCP1* enhancer termed "upTRE" and "dnTRE" (Figure 3.2). TREs typically include two or more variations of the "AGGT(C/A)A" half-site consensus sequence separated by four nucleotides (Brent et al. 1991; Umesono et al. 1991). This same half-site sequence was mentioned above for URE1 and is indeed recognized by multiple transcription factors (Brent et al. 1991). Mutations of the 3' portion of the upTRE (5'-AGGCAA-3') and the dnTRE (5'-AGGTCA-3') to "5'-ATTTAA-3'" and "5'-ATATTA-3'", respectively, eliminate T3 receptor interactions with the rat *UCP1* enhancer (Rabelo et al. 1995). Three putative retinoic acid response elements (RAREs) within the rat *UCP1* enhancer have also been described by Rabelo et al. (1996), though both RARE-1 and RARE-2 overlap with other binding motifs (see Figure 3.2). Nonetheless, mutations increasing the AT-richness of these former regulatory elements were shown to significantly disrupt retinoic acid receptor (RAR) and retinoid X receptor (RXR) transcription factor binding (Rabelo et al. 1996). Finally, Kumar et al. (2008) noted a putative nerve growth factor response element (NBRE) within the *UCP1* enhancer of mice (Figure 3.2) that binds nuclear receptors 4A (NR4A), which acts to promote gene transcription. In addition to the enhancer box, Shore et al. (2012) described a 678 bp putative regulatory region (PRR) located 2095 bp upstream of the transcriptional start site in humans that was conserved in fourteen of twenty-five of the eutherian species they examined. While Shore et al. (2012) found no evidence that this conserved region plays a role in *UCP1* expression, they did note that it encompassed several possible transcription factor binding motifs, including DR1, DR3, DR4, C/EBP (CCAAT-enhancer-binding proteins), CREB, and PPAR.

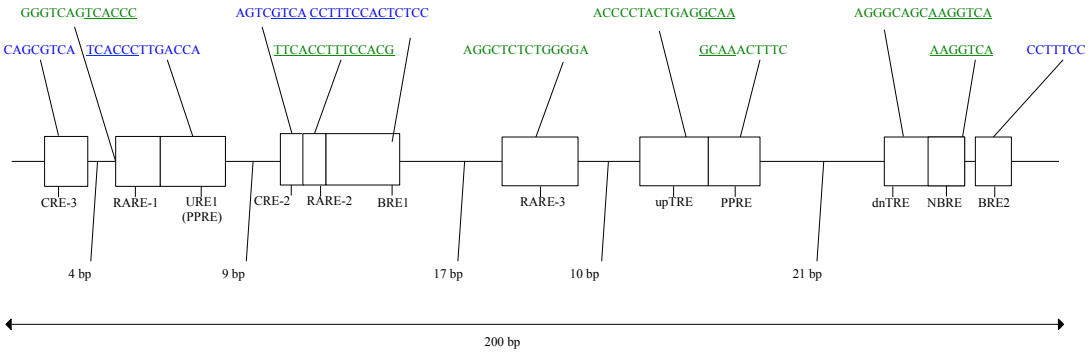


Figure 3.2. Schematic of the murid *UCP1* enhancer with putative transcription factor binding motifs shown for the rat (green) and mouse (blue) based on a combination of previous studies (see text for details). Regions of overlap between adjacent transcription factor motifs are underlined.

Transcriptional control of the *UCP1* gene has also been hypothesized to be regulated by a basal promoter occurring within ~250 bp upstream of the transcription start site (Shore et al. 2010). Within this region, Bouillaud et al. (1988) identified a putative TATA box and a CCAAT binding site located ~20 and ~30 bp upstream of the transcriptional start site of the rat *UCP1* gene, respectively. Generally, the TATA box consists of an A/T-rich consensus sequence (5'-TATAAAA-3'; Xu et al. 1991) that interacts with the TATA binding protein (TBP), one of the components of the transcription factor IID (TFIID) that initiates transcription via RNA polymerase II (Nakajima et al. 1988; Patikoglou et al. 1999). The promoters of some mammalian genes (e.g. globins) also contain a CCAAT box typically situated -60 to -100 bp upstream of the transcription start site that binds nuclear a transcription factor Y (NF-Y) subunit or C/EBP, which then aids in the initiation of transcription via RNA polymerase II (Mantovani 1999). Additionally, a putative CRE site (termed CRE-4) occurs ~130 bp upstream of the mouse *UCP1* transcriptional start site in a reverse and compliment orientation (5'-TGACGCGC-3'), with mutations to this sequence eliminating 90-95% of

reporter gene expression (Kozak et al. 1994). Yubero et al. (1994) further noted three GCCCCT sequences occurring within ~210 bp of the transcriptional start site of the rat, which DNAse 1 footprinting analyses suggest interact with nuclear proteins found within BAT cells, but these have not been defined as protein binding motifs.

Finally, a CpG island surrounding the *UCP1* proximal promoter and extending into exon 1 has been described in several eutherian species (Kiskinis et al. 2007; Shore et al. 2010; Shore et al. 2012). CpG islands contain high densities of cytosine (C) and guanine (G) nucleotide pairs occurring in the 5' to 3' direction and linked by a phosphate (i.e. 5'–C–phosphate–G–3'). These CpG dinucleotides are uncommon in vertebrate genomes, typically occurring at only 20-25% of the frequency anticipated by random chance and act as DNA methylation sites that can modulate gene transcription (Gardiner-Garden and Frommer 1987). Located immediately upstream of many housekeeping genes, CpG islands are believed to play a major role in their transcriptional control (Gardiner-Garden and Frommer 1987). Indeed, methylation of CpG dinucleotides immediately upstream of the *UCP1* gene have been shown to modulate gene activity by blocking transcription, whereas demethylation promotes transcription (Shore et al. 2010). Thus, this CpG island has been postulated to be important for *UCP1* gene regulation and, potentially, tissue specific expression within BAT (Kiskinis et al. 2007; Shore et al. 2010).

Because the majority of studies investigating the transcriptional control of *UCP1* have focused on rodents, the status of these transcription factor binding motifs in other eutherian species remain largely unexplored. Here we use genome mining and hybridization-capture techniques coupled with next-generation sequencing to identify and

examine *UCP1* transcriptional regulatory elements in 139 mammals (135 eutherians). Briefly, putative transcription factor binding motifs and CpG islands were evaluated using a comparative approach to first determine if they are universally conserved among eutherian superorders with functional BAT, and second to test if they are mutated or lost in large-bodied species that presumably have little or no need for NST. We further anticipated that crucial DNA motifs involved in *UCP1* transcription would have deteriorated via millions of years of neutral evolution in the nine lineages for which *UCP1* has been inactivated.

3.3. Materials and methods

3.3.1. *UCP1* regulatory sequences

In total, *UCP1* upstream regions of 139 mammals (1 monotreme, 3 marsupials, 3 xenarthrans, 11 afrotherians, 65 laursiatherians, and 56 euarchontoglires) were examined for transcriptional regulatory elements (see Appendix 5 for species list). This data set employed 116 species whose *UCP1* loci were previously annotated by Gaudry et al. (2017) together with sixteen additional species whose genomes have recently been sequenced (denoted by asterisks in Appendix 5). Regulatory elements of seven additional eutherians were also retrieved by hybridization capture and next-generation sequencing techniques. Briefly, *UCP1* enhancers, PRRs, and basal promoters of four rhinoceroses (black rhinoceros: *Diceros bicornis*, Indian rhinoceros: *Rhinoceros unicornis,* Sumatran rhinoceros; *Dicerorhinus sumatrensis,* and woolly rhinoceros; *Coelodonta antiquitatis*), one tapir (Malayan tapir; *Tapirus indicus*), and two sirenians (dugong; *Dugong dugon*, and Steller's sea cow; *Hydrodamalis gigas*), were targeted using hybridization capture

and next-generation sequencing techniques (Springer et al. 2015; Gaudry et al. 2017). Barcoded rhinoceros DNA libraries were constructed using NEBNext Fast DNA Library Prep Set for Ion Torrent and NEBNext DNA Library Prep Master Mix Set for 454 kits (New England Biolabs; Ipswich, Massachusetts, USA) and target-enriched using MyBaits (Mycroarray; Ann Arbor, Michigan, USA) 120mer RNA probes designed to capture *UCP1* exons and regulatory elements based on the orthologous sequences of the white rhinoceros (*Ceratotherium simum*) genome. The captured rhinoceros reads were sequenced on an Ion Torrent PGM platform using Ion 314 v2 and Ion 318 v2 barcoded chips and an Ion PGM Hi-Q sequencing kit (Applied Biosystems; Foster City, California, USA). Sirenian DNA libraries prepared following the methods of Meyer and Kircher (2010) were enriched using an Agilent SureSelect Capture array with probes designed from African elephant (*Loxodonta africana*) *UCP1* upstream sequences. Sirenian DNA reads were sequenced on Illumina GAIIx and HiSeq2500 (Illumina Inc.; San Diego, California, USA) platforms. Sequenced reads were assembled to reference sequences of the white rhinoceros or manatee (*Trichechus manatus*) using the "map to reference" feature in Geneious R9.1 (Biomatters Ltd.; Auckland, New Zealand) at 20% maximum mismatch per read and consensus sequences were generated.

For publically available genomes, *UCP1* regulatory sequences were acquired using genome-mining techniques of sequences available on the National Center for Biotechnology Information web server. *UCP1*-containing contigs were first acquired by performing nucleotide BLAST searches employing the "discontinuous megablast" option against whole genome shotgun (WGS) contigs of mammalian genome projects using human *UCP1* CDS (NM_021833.4) as a query. If the contigs did not extend ~5 kb

upstream of the *UCP1* transcriptional start site to include the enhancer box, an additional nucleotide BLAST was performed using the human *UCP1* enhancer sequence as a query. For several species with genome projects that have not yet been fully assembled (e.g. *Sus cebifrons*, *Sus verrucosus*, *Elephas maximus*, *Mammuthus primigenius, Balaena mysticetus, Balaenoptera physalus, Mylodon darwinii, Panthera unica*), short read archive (SRA) BLASTs were performed in order to obtain the *UCP1* regulatory elements. Contigs from top BLAST hits were then imported into Sequencher v5.1 (Gene Codes Corporation; Ann Arbor, Michigan, USA) and the exons and regulatory regions annotated by aligning orthologous human *UCP1* sequences (exons 1-6 and enhancer), initially at a 85% minimum match percentage. If the sequences were too divergent to assemble at that stringency, the minimum match percentage was progressively decreased to 60% or until the sequences successfully assembled. *UCP1* coding regions for the sixteen species not included in the Gaudry et al. (2017) study were also examined for the presence of inactivating (e.g. splice site, frameshift, and nonsense) mutations.

The PRR proposed by Shore et al. (2012) was generally less conserved than the enhancer, often with large insertions or deletions, therefore the same annotation methods described above could not be effectively applied to this region. Instead, dot plots were performed in Geneious R9.1 (Biomatters Ltd.) which uses the EMBOSS 6.5.7 dotmatcher tool to compare sequence identities of the human PRR versus the upstream sequence of other mammalian species using a window size of 25, a threshold of 45, and the high sensitivity setting with a probabilistic scoring matrix. The PRR was determined to be present if a conserved region >100 bp relative to the human sequence was discernible from the dot plots. The boundaries of the PRRs were estimated using the dot plot and

annotated. The PRRs of species listed in Appendix 6 were then screened in rVista 2.0 (Loots and Ovcharenko 2004) for the presence of putative transcription factor binding motifs (DR1, DR3, DR4, CEBP (CCAAT-enhancer-binding proteins), CREB, and PPAR) shared with humans, as performed by Shore et al. (2012). Insertions larger than 100 bp relative to the human PRR were removed prior to screening in rVista using the vertebrate TRANSFAC professional V10.2 library with the "matrix similarity optimized for function" setting.

Basal promoter regions were identified by performing alignments of 600 bp upstream of the ATG start codon for each species with available sequence data. The rat and mouse upstream sequences contain several putative promoter motifs (e.g. TATA box, CCAAT site, CRE-4, and GCCCCT sites) and thus were used as reference sequences. CpG islands within the 5' region of *UCP1* were identified using the EMBOSS CpGplot tool (http://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/). Kiskinis et al. (2007) noted that the *UCP1* CpG island occurs immediately upstream of the *UCP1* open reading frame but may also extend into exon 1, therefore, 1 kb upstream of exon 2 was screened for the presence of CpG islands. EMBOSS CpGplot positively identifies CpG islands if a sequence >200 bp contains an observed/expected ratio of CpGs exceeding 0.6, with a GC content >50%, meeting the criteria proposed by Gardiner-Garden and Frommer (1987). The default window size of 100 bp was used for these runs.

The *UCP1* genes of non-eutherian mammals were also examined for the presence or absence of regulatory elements. Contigs of the Tasmanian devil (*Sarcophilus harrisii*) and Tammar wallaby (*Macropus eugenii*) were too short to encompass a potential enhancer occurring ~5 kb upstream of the transcriptional start site. However, contigs of

the platypus (*Ornithorhynchus anatinus*) and gray short-tailed opossum were sufficiently long to create dot plots of the upstream sequence in order to screen for homologous regulatory elements occurring in the human. Some eutherian species displayed inactivated *UCP1* genes with deletions of whole exons (e.g. Chinese pangolin; *Manis pentacatyla*, Javan pangolin; *Manis javanica*, nine-banded armadillo; *Dasypus novemcinctus*), or deletion of the entire gene (killer whale and bottlenose dolphin). The annotation techniques described above did not reveal the presence of a *UCP1* enhancer in these species; thus, sequence identity comparisons against human *UCP1* were performed using Easyfig 2.1 (Sullivan et al. 2011). This analysis was also performed for the rat and cow (*Bos taurus*) since these were species are known to display *UCP1* enhancers while the cow also contains a PRR region (Shore et al. 2012).

Finally, regions containing enhancer and basal promoter sequences for each species were imported into Geneious 9.1 and multispecies nucleotide alignments were generated using the MUSCLE alignment tool (Edgar 2004) with default settings. A consensus eutherian sequence representing the simple majority (>50%) was generated from this dataset based only on species for which the *UCP1* gene is intact (i.e. species with documented *UCP1* pseudogenes (Gaudry et al. 2017) were not included in the consensus calculations). For some eutherian species, pairwise alignments were also created against the human enhancer to obtain the percent sequence identity values. Conserved motifs and putative transcription factor binding sites were annotated. Recognized transcription factor binding motifs within the *UCP1* enhancer (illustrated in Figure 3.2) were examined by eye in each eutherian species and scrutinized for mutations that potentially affect DNA-protein interactions based on previous site directed

mutagenesis studies. Additionally, the consensus enhancer region sequence (see above), together of those of seven species spanning the three mammalian superorders for which *UCP1* is intact, were screened for the presence of all vertebrate transcription factors in the TRANSFAC professional V10.2 library using rVista with the "matrix similarity optimized for function" setting.

3.3.2. Phylogenetic trees

To generate a combined *UCP1*, *UCP2*, and *UCP3* coding sequence phylogenetic tree, the data set of Gaudry et al. (2017) was updated to include coding sequences of the sixteen additional species with recently published genomes (Appendix 5). The resulting 448 *UCP* genes were aligned using MUSCLE (Edgar 2004), and a maximum likelihood tree constructed using RAxML (Randomized Axelerated Maximum likelihood) version 7.2.8 (Stamatakis 2006) with the "GTR Gamma" nucleotide model and "rapid bootstrapping and search for best scoring tree" setting. The program was performed for 500 bootstrap replicates.

In order to trace the evolutionary gain and loss of *UCP1* transcriptional regulatory elements, we also constructed a 41-gene species tree for the 139 mammals included in this study following the methods of Gaudry et al. (2017). Briefly, this data set included 30 nuclear (*A2AB, ADRB2, APP, ATP7A, ADORA3, APOB, BCHE, BDNF, BMI1, BRCA1, BRCA2, CHRNA1, CMYC, CNR1, CREM, DMP1, ENAM, EDG1, FBN1, GHR, IRBP, MC1R, PLCB4, PNOC, RAG1, RAG2, SWS1, TTN, TYR1, VWF*) and 11 mitochondrial loci (*12S rRNA, 16S rRNA, CYTB, COI, COII, COIII, ND1, ND2, ND3, ND4, ND5*). A 50,911 bp concatenated supermatrix was aligned in MUSCLE. The

supermatrix was divided into 32 partitions. Each nuclear gene was assigned an individual partition, while *12S rRNA* and *16S rRNA* were combined to create one partition, and the nine remaining mitochondrial genes were also combined into a single partition. An independent GTR Gamma model was estimated for of these partitions and a maximum likelihood tree was generated in RAxML 7.2.8 using the same settings described above with 100 bootstrap replicates.

3.4. Results

3.4.1. *UCP1* coding sequences

All of the sixteen newly acquired *UCP1* CDSs were intact with the exception of the Javan pangolin, which displays the same mutations as the Chinese pangolin pseudogene (i.e. frameshift, splice site and nonsense mutations, deletion of exons 1 and 2) documented by Gaudry et al. (2017). Similarly, the 12 bp deletion that calls into question the functionality of the Bactrian camel *UCP1* gene (Gaudry et al. 2017) is also present in the dromedary camel (*Camelus dromedarius*). Conversely, the *UCP1* CDS of the giraffe (*Giraffa camelopardalis*) is intact, despite its large body size.

The predicted platypus *UCP1* CDS available on GenBank (accession number: XM_001512650) is unique in that it creates a hypothetical open reading frame composed of seven exons; the usual 126 bp exon 1 is divided into two separate exons of 30 and 120 bp in length. The placement of these putative exons are displayed in a dot plot comparison with the 5' region of the gray short-tailed opossum *UCP1* locus (Figure 3.3). Notably, two separate regions within the platypus read display homology to the opossum *UCP1* exon 1 sequence, revealing what appears to be a 186 bp insertion in the platypus

exon 1 sequence. The original platypus start codon also appears to be mutated to "AAG" thus translocating the predicted 30 bp 'exon 1' of the platypus 176 bp upstream of the gray short-tailed opossum start codon (Figure 3.3). By contrast, BLAST searches of platypus RNA sequencing projects (SRX182802, SRX17144, SRX17145, SRX081892, SRX081881, SRX081882, SRX328084, SRX328085, SRX081887-SRX081890) reveal an intact *UCP1* mRNA sequence (data not shown) that differs from the predicted coding sequence. Briefly, the platypus mRNA coding sequence indicates that the predicted 30 bp 'exon 1' coding sequence is not translated, that there is no insertion in exon 1 of the platypus, and that the ATG start codon found in other mammals is indeed intact at the expected position (i.e. there is a misassembly error in the predicted GenBank sequence).

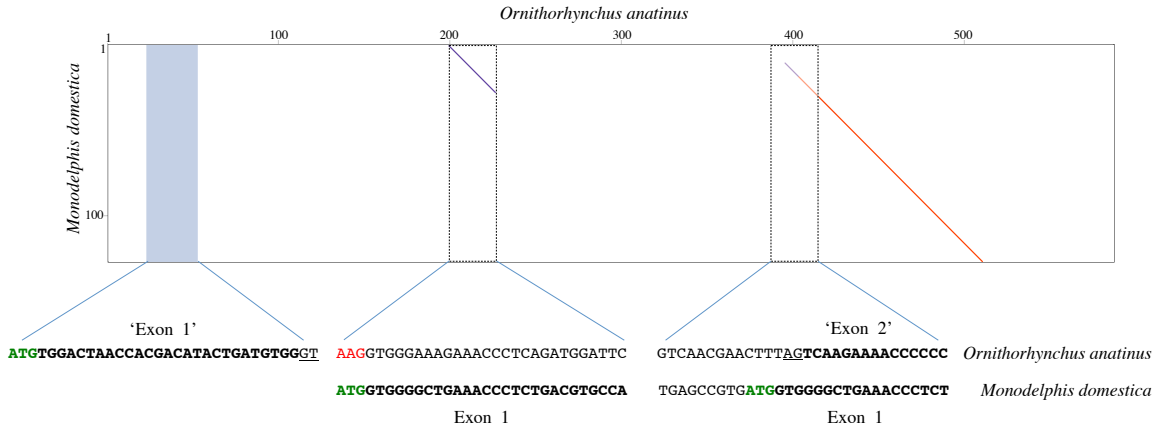


Figure 3.3. Dot plot comparison of the gray short-tailed opossum *UCP1* exon 1 versus a section of the platypus *UCP1* gene occurring between *TBC1D9* and *ELMOD2* (accession number: NW_001794248.1). Sequence alignments of the platypus (top) and gray-short tailed opossum (bottom) are provided with the potential coding sequences indicated in bold; putative splice sites are underlined. Note that two regions within the platypus clearly display homology to the opossum exon 1 (199-226 and 400-520), suggesting the presence of a 186 bp insertion in the platypus exon 1 sequence. The blue shaded area represents the region where an automated predictor program, which created a seven exon *UCP1* gene for the platypus, placed a 30 bp 'exon 1' in order to obtain an open reading frame free from premature stop codons (accession number: XM_001512650), though this region shares no homology with exon 1 of the opossum. The original platypus start codon also appears to be mutated to AAG (red font), with the predicted platypus 'exon 2' occurring 6 bp downstream of the "ATG" start site in the opossum. Note that these

differences between the two species likely arise from a misassembly error in the platypus (see text for details).

3.4.2. *UCP1* basal promoter

An alignment of the basal *UCP1* promoter for representative species is displayed in figure 3.4. Notably, the most upstream GCCCCT motif (nucleotides 1-6 of the promoter alignment; Figure 3.4) described in the rat by Yubero et al. (1994) is not present in any non-murid species. While the CRE-4 consensus sequence (5'-TGAAGGGC-3') is similar to that described by Kozak et al. (1994) in mice (5'-TGACGCGC-3'), this site does differ substantially in many species (e.g. common shrew [*Sorex araneus*], human, etc.) and is absent in the gray short-tailed opossum, walrus, cow, and giraffe (Figure 3.4). The second and third GCCCCT sites, respectively occurring at 242-248 and 308-315 of the alignment, are relatively well conserved (Figure 3.4). By contrast, the putative CCAAT site in the rat (Bouillaud et al. 1988) is highly variable in other mammals. The TATA box described by Bouillaud et al. (1988) is intact in the majority of species including all marsupials where it occurs as a 5'-TATAARR-3' sequence 260-280 upstream of the ATG start codon of exon 1. While a 5'-TATAAGG-3' sequence is found ~200 bp upstream of the platypus *UCP1* coding sequence, the validity of this site is uncertain due to a misassembly in this region of the GenBank sequence (see above). Interestingly, the walrus motif contains a T→A mutation causing a 5'-TAAATAA-3' sequence, while the panda, white rhinoceros, horse, and bats share a 5'-TACAWAA-3' sequence. Among species that possess pseudogenized *UCP1* genes, an intact TATA box still remains ~290 bp upstream of the the African elephant (*Loxodonta africana*) and manatee (*Trichechus manatus*) coding sequence while the closely related Cape rock

hyrax (*Procavia capensis*) deviates from the consensus (5'-TACGTGA-3'). Similarly, the pig retains a TATA box identical to that of the cow, camel, and giraffe (5'-GATATAA-3'), though a number of mutations in cetaceans have resulted in a sequence (5'-GACGTCAA-3') that is virtually unrecognizable as a TATA box (Figure 3.4).

Figure 3.4. *UCP1* basal promoter elements alignment for select mammalian species with putative protein binding motifs indicated. Highlighted sites indicate shared nucleotides to the species in which the motif was first described (mouse or rat) and the typical TATA box (5'-TATAAAA-3') sequence (Xu et al. 1991). The consensus sequence represents the simple majority based on species for which the *UCP1* gene is intact. Species with documented *UCP1* pseudogenes (Gaudry et al. 2017) are denoted in red font and were not included in the consensus calculations.

3.4.3. CpG island

CpG islands meeting the criteria of Gardiner-Garden and Frommer (1987) were not detected in the monotreme or marsupial assemblies. Conversely, a CpG island within or immediately upstream of exon 1 was identified in 91 of 113 eutherian species with available sequence coverage for this region (Appendix 5). The presence of the CpG island was found to vary extensively among small-bodied species as it was detected in the common shrew, but is absent from the European hedgehog (*Erinaceus europaeus*) and star-nosed mole (*Condylura cristata*; Appendix 5). Many rodent species (e.g. mouse, rat), known to express functional BAT, also lack a CpG island (Appendix 5). Similarly, among the four afroinsectiphilians examined, a CpG island was only identified in the lesser hedgehog tenrec (containing 39 CpG dinucleotides), despite a relatively high number of CpG sites (37-41) located between 600 bp upstream and 200 bp downstream of the start codon in the other three species. Conversely, CpG islands were identified in closely related paenungulates (elephants, sirenians, and hyraxes), which have >50 CpG dinucleotides in the same region, and armadillos—despite both of these groups having a non-functional *UCP1*. Among artiodactyls, CpG islands were detected in camels, the okapi (*Okapia johnstoni*), and all whale *UCP1* pseudogenes (except for the killer whale and bottlenose dolphin for which the entire gene is deleted; Figure 3.5), but not the giraffe or the pig (*Sus scrofa*). This element is also missing in the pangolin pseudogenes, which is likely due to deletion of a portion of the gene upstream of exon 3 (Figure 3.5).

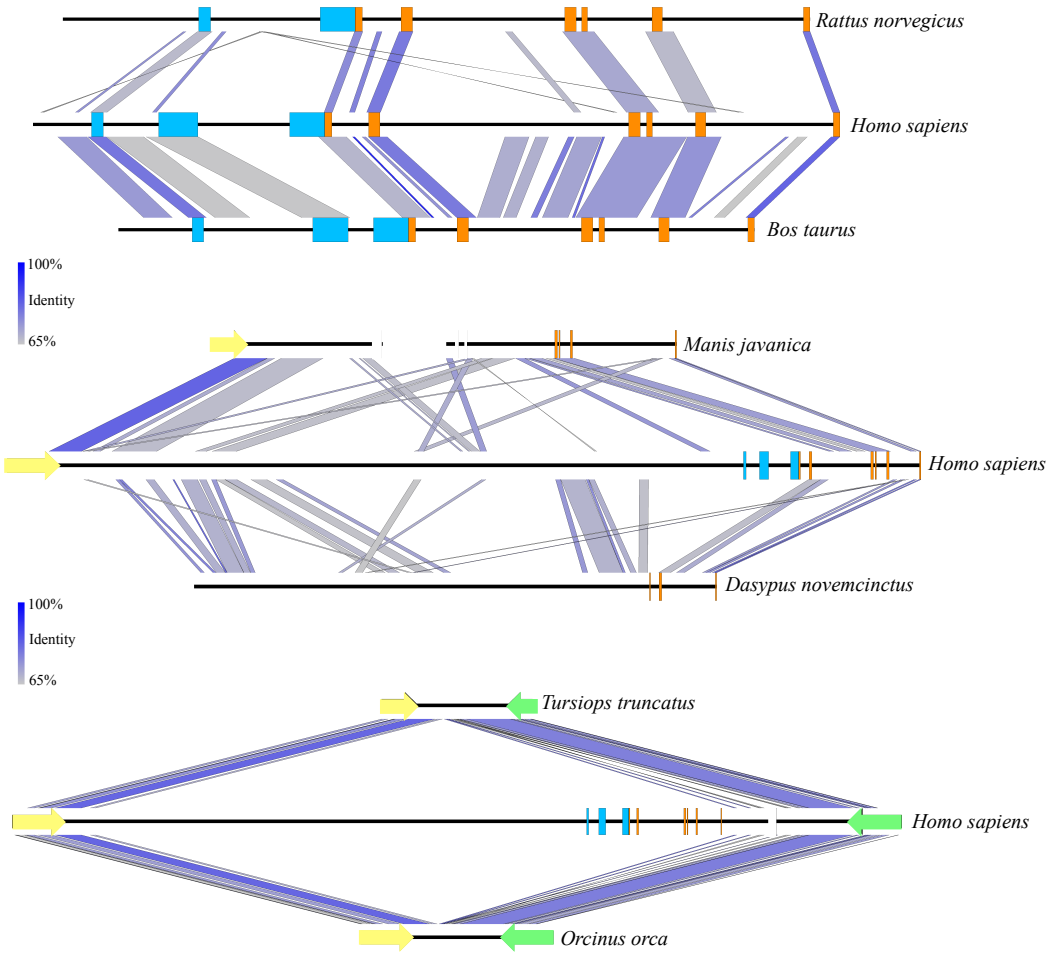


Figure 3.5. Sequence identity comparisons of the *UCP1* genes of the rat, cow, pangolin, armadillo, bottlenose dolphin, and killer whale versus the human. All DNA sequences are shown 5' (left) to 3' (right). *UCP1* exons 1-6 are denoted with orange rectangles while *UCP1* upstream transcriptional regulatory elements are denoted in light blue (enhancer box, putative regulatory region, CpG island; from left to right). Gaps in sequence coverage are represented by white rectangles. Notably, the putative regulatory region is absent in the rat, but conserved in the cow. Upstream regulatory elements also appear to have been deleted in the Javan pangolin and armadillo, which have deletions of *UCP1* exons 1-2, and 3-5, respectively. Deletion of the entire *UCP1* gene between *TBC1D9* (yellow arrows) and *ELMOD2* (green arrows) has occurred in bottlenose dolphin and killer whale ~8-15 MYA (Gaudry et al. 2017) and included the upstream regulatory elements. Sequence identity percentage is represented with a color scale.

3.4.4. Putative regulatory region (PRR)

A distinct PRR was found to be present in 97 of the 125 eutherian mammals examined for which sequence is available (Appendix 5), though this element was not

observed in the platypus or gray short-tailed opossum (Appendix 7). PRRs were observed from all afrotherians, but not the armadillo, a xenarthran (Appendix 5), though insertions within this region are prevalent in the elephant shrew, lesser hedgehog tenrec, and aardvark (Appendix 7). By contrast, the dot plots of the elephant and manatee—for which *UCP1* is pseudogenized—reveal a high conservation of the PRR with virtually no indels, though only the 3' half of the PRR is present in the hyrax (Appendix 7). As seen for the cow (Figure 3.5), giraffe, camel, and several whales (Appendix 7), the PRR is conserved among most artiodactyls, but is missing in the pig *UCP1* pseudogene (Appendix 7) and deleted in the bottlenose dolphin, killer whale, and Javan pangolin (Figure 3.5). A PRR is also absent in several species known to express functional BAT, including the shrew and star-nosed mole, several bats (*Myotis spp*. and *Eptesicus fuscus,* etc.), and many rodents (Appendix 5), including the mouse and rat (Figs. 5 and Appendix 7). Similarly, both *Canis familiaris* and *Lycaon pictus* lack a PRR, despite this feature being present in all other carnivores (Appendix 5). The transcription factor binding sites identified within PRRs of selected species using rVista 2.0 are listed in Appendix 6. PPAR, DR1, DR3, DR4, CREB, and C/EBP sites are relatively common within this region in species with and without a functional *UCP1* locus.

3.4.5. *UCP1* enhancer

UCP1 enhancer sequences were retrieved for 121 eutherian species (Appendix 5). Enhancer boxes were typically found within 5 kb upstream of exon 1, however, for some members of the afroinsectiphilia (i.e. aardvark and elephant shrew), the enhancer occurs at ~ -7.5 kb (Appendix 7). Dot plots of the upstream regions of the platypus and the gray

short-tailed opossum reveal no evidence for a *UCP1* enhancer (Appendix 7), suggesting it is absent within both monotremes and marsupials.

Contrary to the findings of Shore et al. (2012), who noted the absence of an enhancer in the upstream region of the common marmoset (*Callithrix jacchus*), American pika (*Ochotona princeps*), thirteen-lined ground squirrel (*Spermophilus tridecemlineatus*), common shrew, and European hedgehog, we identified this element in each of these species except the hedgehog. The contig encompassing hedgehog *UCP1* CDS (accession number: AMUD01193160.1), however, only extends 1126 bp upstream of exon 1 and BLAST searches failed to provide hits of a *UCP1* enhancer located on other contigs, thus its presence or absence from the genome remains inconclusive. Similarly, low sequencing coverage likely explains the apparent lack of a *UCP1* enhancer in the zebu (*Bos indicus*), Brazilian guinea pig (*Cavia apera*), and desert woodrat (*Neotoma lepida*), as enhancers have been recovered from their close phylogenetic relatives (Appendix 5).

The enhancer is highly conserved in large-bodied species with intact *UCP1* loci (i.e. rhinoceroses, camels, giraffe, and pinnipeds) as well as several species with *UCP1* pseudogenes (e.g. elephantids, sirenians, suids, equids, and some cetaceans; Appendix 5). However, seven species lack both a *UCP1* enhancer and an intact *UCP1*. For instance, the entire *UCP1* gene including the enhancer has been deleted in the killer whale and bottlenose dolphin (Figure 3.5). The enhancer has also been deleted in the sperm whale (*Physeter macrocephalus*; Appendix 7), yet it remains present in the baiji (*Lipotes vexillifer*) and all baleen whales, indicating an independent loss in both the sperm whale and delphinids. The dot plots also fail to provide evidence for an *UCP1* enhancer in the

Cape rock hyrax, though this element is present in other paenungulates for which this gene is also pseudogenized (Appendix 7). Sequence identity comparisons also suggest the enhancer is lost in pangolins and the nine-banded armadillo (Figure 5; Appendix 5). Interestingly, BLAST searches failed to identify this regulator in the WGS contigs or SRA of the two-toed sloth (*Choleopus hoffmanni*), although partial coverage was recovered for the extinct giant ground sloth (*Mylodon darwinii*) from a pair of SRA reads (Appendix 8).

Dot plots of the murid (rat and mouse) upstream sequence (Appendix 7) illustrate marked divergence from humans with the exception of a small region encompassing the *UCP1* enhancer. By contrast, the upstream sequence of the many laurasiatherians and even paenungulates lacking an intact *UCP1* (e.g. elephants and manatees) is surprisingly similar to that of humans (Appendix 7). In fact, pairwise sequence comparisons of the enhancers versus that of the human reveal that this region is more highly conserved (>80%) in large-bodied species that both possess and lack an intact *UCP1* than relative to the mouse (74%) and rat (69%) *UCP1* (data not show), despite the latter sharing a more recent common ancestor with humans. This pattern is mirrored in the *UCP1* gene tree (Figure 3.6) as many small-bodied lineages (i.e. afroinsectiphlians, myomorph rodents, vesper bats, and most notably, eulipotyphlans) display long branch lengths indicative of high rates of molecular evolution that are comparable to those of many species with *UCP1* pseudogenes (e.g. pangolins, pigs, armadillo, and hyrax). Canines are also worth noting, as their branch is highly elongated compared to other carnivores. By contrast, short branches found for most large-bodied species, even among those with non-

functional *UCP1* (e.g. paenungulates, cetaceans, and equids), reflect low nucleotide substitution rates.

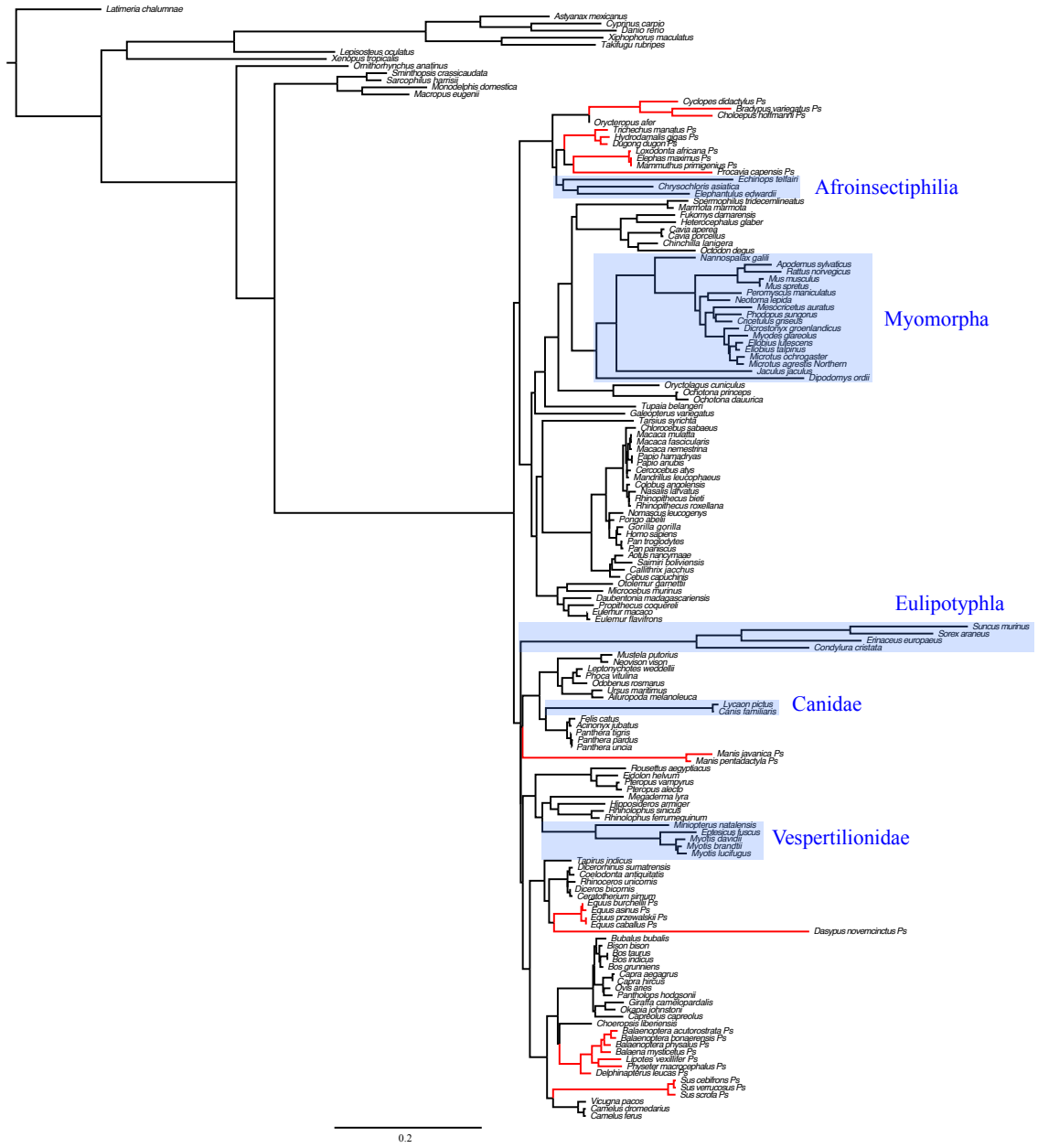


Figure 3.6. Maximum likelihood *UCP1* gene tree illustrating substitution rates in several eutherian lineages (eulipotyphlans, canids, afroinsectiphilians, vesper bats, myomorph rodents; boxed in blue) that are comparable or higher than lineages with *UCP1* pseudogenes (denoted in red). Branch lengths represent the number of nucleotide substitutions per site.

Enhancer region alignments revealed a number of marked differences within transcription factor binding motifs among species (Appendix 8). For instance, while the CRE-3 site contains a set of core nucleotides (5'-CGTCA-3') that are highly conserved in most eutherians, mutations to one or two nucleotides within this region are observed in a number of species (e.g. *Condylura cristata*, *Dipodomys ordi*, *Cricetulus griseus*), while the 5' portion of this site appears to be deleted in the Philippine tarsier (*Tarsius syrichta*). Notably, the CRE-3 motif was detected in the each species for which the enhancer was screened in rVista except for *Condylura cristata* (Appendix 9). Various mutations to this motif are also found in species with a pseudogenized *UCP1* (e.g. elephants, pigs, whales, and horses; Appendix 8). The RARE-1 site is especially conserved in the section that overlaps with the URE1 motif, where the consensus sequence (5'-TTACCCTTGCTCA-3') closely resembles the mouse URE1 site proposed by Sears et al. (1996). However, mutations at sites (e.g. nucleotide positions 32-33 of the alignment in Appendix 8) shown to block transcription binding in mice (Sears et al. 1996) are observed in several species with intact *UCP1* (e.g. rabbit; *Oryctolagus cunculus,* Philippine tarsier; *Tarsius syrichta*, white rhino; *Ceratotherium simum*, and tapir; *Tapirus indicus*). The aardvark displays a 4 bp insertion occurring within the URE1 that results in a single nucleotide (C→A) substitution to this motif. Notably, among species lacking a functional *UCP1*, the Javan warty pig (*Sus verrucosus*) exhibits a marked disruption to the URE1 site.

The CRE-2 motif is well conserved among most eutherians, however, the consensus eutherian sequence (5'-ATTCTTTA-3'; Appendix 8) is a poor match to the mouse 5'-AGTCGTCA-3' sequence (Kozak et al. 1994). Indeed, of seven species for with the enhancer region was screened using rVista, this site was identified as a cAMP

response element only within the mouse (Appendix 9). Notably, several species with an intact *UCP1* display deletions within the CRE-2 motif (e.g. black capped squirrel monkey; *Simiri boliviensis,* thirteen-lined ground squirrel; *Spermophilus tridecemlineatus*, and natal long-fingered bat; *Miniopterus natalensis*). Similarly, the two TTCC motifs described for the mouse BRE-1 site (Kozak et al. 1994) are not found in any non-murid eutherians. This region, however, is TC-rich in nearly all species with a single convergent TTCC site found in the dog and natal long-fingered bat (Appendix 8). In contrast, the AT-richness of the BRE-1/RARE-2 region is substantially increased in horses, whales, and pigs—all of which lack a functional UCP1—relative to species with an intact gene.

The RARE-3 site consensus sequence (5'-TGACCCTTTGGGGAT-3'; Appendix 8) is strongly conserved among eutherians with the exception of a 2-bp deletion in the tiger (*Panthera tigris*). The PPRE motif predicted by Jastroch et al. (2008) is also a highly conserved element within the *UCP1* enhancer, with a consensus sequence of 5'-GCAAACTTTC-3'. Of note, a PPARG (or PPARγ) site with a consensus sequence of 5'-CAAACTTTCTCCTACTT-3' was identified to overlap with this PPRE motif in six of the seven species (all except for the mouse) for which the enhancer was screened using rVista (Appendix 9). Conversely, the rat upTRE motif (Rabelo et al. 1995) appears to have arisen from a 14 bp deletion in this species, and is therefore not present in other lineages (Appendix 8). Additionally, the white-headed capuchin (*Cebus capuchinis*) and polar bear (*Ursus maritimus*), both of which likely express functional BAT, have deletions within the putative upTRE region. The 5' portion of the dnTRE motif (5'-AGGGCAGCAAGGTCA-3') described by Rabelo et al. (1995) is also exclusive to the

rat, as the consensus sequence (5'-AG**A**AGGGGTGAGGTCA-3') has numerous differences and an insertion [bold]; deletions to this region are also found in the Damaraland mole-rat (*Fukomys damarensis*), *Myotis spp.* bats, and the lesser hedgehog tenrec (Appendix 8). The NBRE site, which overlaps with the 3' region of the dnTRE, is not strongly conserved in all species, with nucleotide deletions in artiodactyls, the Damaraland mole-rat, great roundleaf bat (*Hipposideros armiger*), David's myotis, and natal long-fingered bat, and insertions in both the tiger and the giant ground sloth (Appendix 8). The most crucial nucleotides of the BRE-2 motif (5'-TTCC-3'; bases 219-222 of the enhancer alignment; Appendix 8) described by Kozak et al. (1994) are only found in mice (the species in which it was first described) and the Chinese rufous horseshoe bat (*Rhinolophus sinicus*).

3.5. Discussion

No traces of an enhancer, PRR, or CpG island were detected in the upstream region of the platypus or gray short-tailed opossum loci, though both appear to possess a TATA box within the proximal promoter. By contrast, each of these elements were observed in afrotherians, euarchontoglirans, and laurasiatherians, while a portion of the *UCP1* enhancer was also obtained in a single xenarthran, the giant ground sloth, a species that went extinct during the late Pleistocene ~12,000 years ago (Moore 1978). We can thus deduce that the *UCP1* gene of stem mammals contained a TATA box, while the other transcriptional regulatory elements evolved in a common ancestor of eutherians as proposed by Jastroch et al. (2008). However, despite functioning as a hypothetical methylation site (CpG island) or encompassing putative transcription factor binding sites

in some species (PRR) these motifs are not required for BAT transcription, as exemplified by high *UCP1* expression within the BAT of mice and rats (Pedersen et al. 2001; Wu et al. 2012), which lack both of these elements. Indeed, these elements have repeatedly been lost in eutherian mammals (Figure 3.7). Shore et al. (2012) reached a similar conclusion as roughly half of the eutherian species they examined lacked a PRR and a CpG island. Given the proposed function of the CpG island as a regulator of UCP1 tissue-specific expression (Kiskinis et al. 2007), a lower level of methylation in BAT as opposed to other tissues would be expected, however, Shore et al. (2012) discovered that the *UCP1* CpG island remains virtually un-methylated in BAT, white adipose tissue, and liver despite greatly reduced UCP1 expression levels in the latter two tissues. Therefore the function of this region remains unclear, however, Shore et al. (2012) did characterize a CpG island in the zebrafish suggesting its presence could be an ancestral condition of the *UCP1* gene that was lost in non-eutherian mammals, but retained (and again lost) in some eutherians (see Figure 3.7).

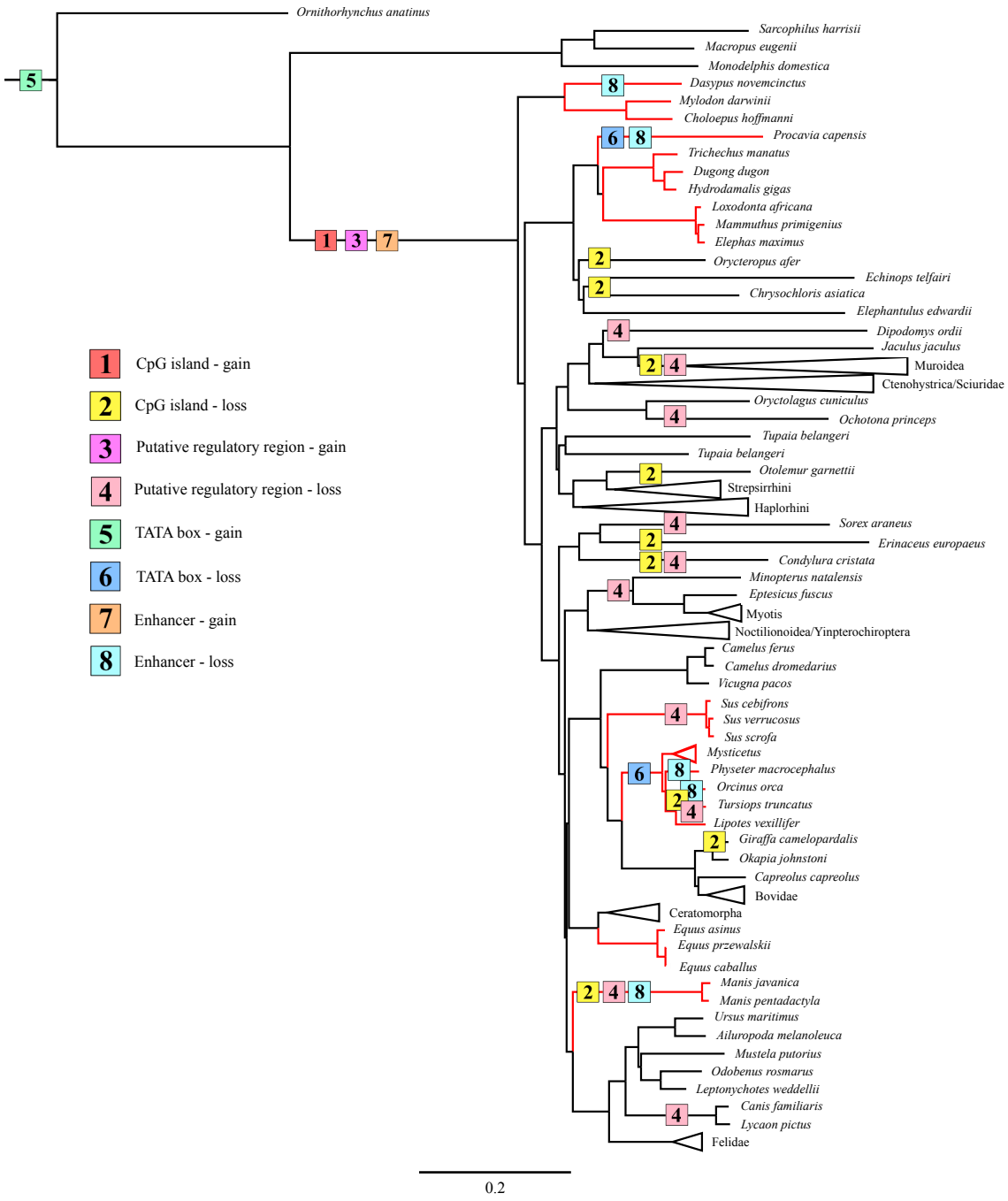


Figure 3.7. A maximum likelihood species tree based on 41 gene segments (50,911 base pairs) that illustrates the gain and loss of known *UCP1* regulatory elements (CpG island, PRR, TATA box, enhancer) through the evolutionary history of Mammalia. Red branches indicate lineages with a non-functional *UCP1* gene (Gaudry et al. 2017).

Alignment of the proximal promoter CRE-4 site among representative eutherians reveals that the 5'-TGACGCGC-3' sequence proposed by Kozak et al. (1994) is conserved in the rat, but deviates considerably in the shrew, cow, and human, which are known to express functional BAT (Przełecka 1981; Heaton 1972; Alexander et al. 1975). Thus, while the CRE-4 site may play an important role within the murid lineage, it likely does not apply to other eutherians. Similarly, the CCAAT box proposed by Bouillaud et al. (1988) in the rat is highly variable among eutherians (and even among rodents), thus is also unlikely to be a key site for promoter activity. Of the three GCCCCT sites proposed by Yubero et al. (1994), only the two located proximal to exon 1 are conserved, however, to our knowledge transcription factors that bind to these nucleotides have not yet been identified. Overall, the TATA box of the *UCP1* promoter is highly conserved in most eutherians, but does vary in some species. For instance, the shared TACA box variant among the horse, rhino, bats, and panda is interesting given that bats and bears possess discernible BAT (Rowlatt et al. 1971; Thomas et al. 1990). While TATA box variants of the flowering plant *Arabidopsis thaliana*, including the 5'-TACAAAAG-3' sequence, can still bind the TATA binding protein (TBP) without any structural modifications to the protein, transcription activity levels are substantially (76-85%) lower compared to the 5'-TATAAAAG-3' sequence (Patikoglou et al. 1999). Considering the high level of TBP conservation among eukaryotes (Peterson et al. 1990), its ability to bind TATA box variants may also apply to mammals. The same T→C transition at the third nucleotide position has been described in the TATA (TACA) box of rabbit uteroglobin with respect to the rat and human, causing a 7-fold reduction in activity when binding to TBP (Klug et al. 1994). However, two other proteins (TATA core factor and TATA palindrome factor)

present in uteroglobin-expressing cells bind the TACA box with high efficiency to promote cell specific-expression of the protein (Klug et al. 1994), thus the same possibility may apply to bears, bats, and rhinos. The mutated 5'-TAAATAA-3' site of the walrus retains a high A/T richness and can thus likely still efficiently bind the TBP (Patikoglou et al. 1999). Notably, the TATA boxes of the hyrax and cetacean *UCP1* pseudogenes are poorly conserved, likely due to mutations accumulating under neutral evolution (Figure 3.7).

In general, the *UCP1* enhancer appears to be among the most crucial elements of transcriptional regulation as it is one of the few highly conserved regions in the upstream sequence between humans and rodents (Appendix 7). Indeed, excluding four species with low sequence coverage (see below), the enhancer was recovered from all eutherians with an intact *UCP1* gene, and therefore is likely essential for *UCP1* expression in BAT. This conclusion is at odds with that of Shore et al. (2012), who incorrectly deduced that this region was deleted in a number of species. While we were unable to retrieve an enhancer in four species (i.e. European hedgehog, zebu, Brazilian guinea pig, and desert woodrat), contigs of these species either do not extend ~5 kb upstream of *UCP1* exon 1 or contain large sequencing gaps.

In concert with our prediction that large body size may be associated with relaxed selection pressures for UCP1 expression, several anomalies among putative transcription factor binding motifs exist that could be indicative of degradation of these elements were observed. For instance, rhinoceroses display a deletion within the BRE-2 site, and multiple mutations occur within the dnTRE and NBRE regions of camels and the alpaca (*Vicugna pacos*). However, deletions also occur within these regions of some small-

bodied species (Damaraland mole-rat, lesser hedgehog tenrec, and *Myotis spp*. bats) that also have an intact *UCP1*, while felids display a highly divergent nucleotide sequence within this 3' region of the enhancer box. Overall, it thus seems unlikely that these transcriptional regulatory element mutations would substantively impact *UCP1* expression in the large-bodied species. Notably, *UCP1* regulatory regions (enhancer, PRR, CpG island, promoter) are also present in all large-bodied species (e.g. rhinoceroses, pinnipeds, camel), except the giraffe where a CpG island was not detected (Appendix 5). Again, this finding suggests that the UCP1 protein may be present in BAT and/or beige tissue of these lineages, highlighting the need for future investigation of UCP1 expression in these species.

In support of our hypothesis that transcriptional regulators would be deteriorated or lost in eutherians with *UCP1* pseudogenes, at least five independent lineages (sperm whale, hyrax, pangolins, armadillo, and the family delphinidae [killer whale and bottlenose dolphin]) lack an *UCP1* enhancer (Figure 3.7); notably the TATA box is also lost/mutated in these lineages. By contrast, we identified several lineages (elephantids, sirenians, suids, equids, and some cetaceans) that retain a highly conserved enhancer despite inactivation of their *UCP1* genes >20 MYA (Gaudry et al. 2017). The presence of a conserved enhancer upstream of the pig *UCP1* pseudogene was also noted by Shore et al. (2012), who suggested that an added function might explain its high degree of sequence identity to that of humans. One such added function could be pleiotropy; the regulation multiple genes (He and Zhang 2006). Indeed, evolutionary constraint increases (i.e. a higher degree of purifying selection) in mammalian enhancers with increasing pleiotropy (Hiller et al. 2012). Considering that pleiotropic enhancers are not uncommon

among mammals (Hiller et al. 2012), this hypothesis cannot be entirely discounted. However, the loss of an *UCP1* enhancer in the sperm whale, killer whale, bottlenose dolphin, hyrax, armadillo, and pangolins implies that this enhancer is non-pleiotropic. The apparent conservation of most enhancer elements in the other species for which *UCP1* is pseudogenized (e.g. baleen whales, elephants, sirenians, horses) is presumably in part due to an inherently slow rate of molecular evolution arising from their large body size. Indeed, other pseudogenized genes (e.g. *AMBN*, *AMEL*, *ENAM*, and *MMP20*) in baleen whales and the Steller's sea cow (*Hydrodamalis gigas*) show exceptionally low rates of molecular decay (Meredith et al. 2011; Springer et al. 2015). Consequently the high (>80%) enhancer sequence identity shared between *UCP1*-pseudogenized species (horse, minke whale, pig, baiji, bowhead whale, African elephant, and manatee) and humans is not surprising. It thus also remains possible that slow rates of DNA evolution may explain the retention and conservation of these regulatory elements in some large-bodied species with intact *UCP1* CDS. By contrast, the higher sequence divergence in rats and mice, which share only 69 and 74% of *UCP1* enhancer similarity with humans, respectively, can likely be attributed to a relatively fast mutation rate.

Surprisingly, an elevated mutation rate is also evident in the *UCP1* coding sequence of canids as well as the small-bodied lesser hedgehog tenrec, myomorph rodents, vesper bats, and, particularly within members of the order eulipotyphla (Figure 3.6). While selection pressure analyses indicate that the *UCP1* coding sequences of these species display relatively low dN/dS ratios (<0.22; Gaudry et al. 2017), associated with functional conservation of the protein, the very high substitution rates in these groups equate to a substantively elevated number of nonsynonymous amino acid substitutions

relative to other eutherian lineages. Notably, these high substitution rates are not found for *UCP2* or *UCP3* sequences of these species (cf. Figure 3.1), suggesting that this is not solely a size-dependent phenomenon. Consequently these lineages provide intriguing comparative opportunities to study functional UCP1 attributes, as BAT-mediated NST is likely crucial for thermoregulation in these lineages.

A key finding of this study is that several transcription factor binding motifs first described in either mice or rats (BRE-1, BRE-2, upTRE, dnTRE) appear to be restricted to this clade of mammals. Other enhancer motifs (URE1, CRE-2, RARE-2, NBRE) presumed to be key for transcription factor binding in murid rodents (Kozak et al. 1994: Kumar et al. 2008; Rabelo et al. 1996; Sears et al. 1996) are also mutated in other eutherian lineages (Appendix 8). Although both single point mutations (Bokar et al. 1988) or combination of mutations (Rabelo et al. 1996) have been shown to alter transcription factor binding to some of these motifs in murid rodents, the effect of the observed differences to these motifs in other eutherians needs to be assessed. Nonetheless, the rVista enhancer screening (Appendix 9) demonstrates that a number of putative transcription factor binding elements (e.g. CRE-2, PPARG) are not shared between murid rodents and the consensus sequence. This analysis also suggests that components of the transcriptional control of *UCP1* expression may be differentially regulated among eutherian mammals. For example, the CRE-3 element was identified in each species selected for screening except for the star-nosed mole (Appendix 9). By contrast, the high level of sequence identity of the PPRE and RARE-3 elements across Placentalia (Appendix 8) indicates that their function has remained strongly constrained

throughout eutherian evolution, and is suggestive that they are universally required for the regulation and specificity of *UCP1* transcription.

3.6. Conclusions

To our knowledge, this study represents the broadest comparative analysis of *UCP1* transcriptional regulatory elements among mammals. Our results demonstrate that the CpG island and PRR are not universally conserved among BAT-expressing eutherians and thus are likely not required for *UCP1* transcription. In contrast, the TATA box and two of the three GCCCCT sites in the promoter are highly conserved and presumably play a transcriptional role, while the CRE-4 and CCAAT sites differ substantially among eutherians and likely are unimportant. While a *UCP1* enhancer was found to be present in every eutherian superorder (Xenarthra [partial], Afrotheria, Laurasiatheria, Euarchontoglires), its absence among non-eutherian mammals supports the hypothesis that it originated with the rise of BAT in a stem placental ancestor. Within this region, however, the specificity and importance of the upTRE, dnTRE, URE1, CRE-2, RARE-2, NBRE, BRE-1, and BRE-2 enhancer elements first described from rats and mice are uncertain as these motifs differ substantially—but generally remain highly conserved—in other BAT-expressing eutherians. Conversely, the RARE-3 and PPRE motifs are among the most highly conserved putative transcription factor binding elements and are likely functional across the eutherian phylogeny. Finally, while some *UCP1*-less species still retain a *UCP1* enhancer, this sequence conservation is presumably due to a slow rate of neutral evolution. Nonetheless, lack of an enhancer in seven *UCP1*-less species strongly suggests this element is non-pleiotropic.

CHAPTER 4: FINAL DISCUSSION AND CONCLUSIONS

4.1. Molecular phylogenetics of rhinoceroses

Hybridization capture and NGS techniques proved to be highly successful in Chapter 2 where I examined the molecular phylogenetics among six rhinoceros species using the largest molecular dataset analyzed for this purpose to date. Despite rhinoceros and tapir lineages being separated by >53 million years of evolution (Kapur and Bajpai 2015), inter-familial gene capture techniques of the Malayan tapir were clearly effective and provided an outgroup for phylogenetic analyses. Interestingly, much higher sequence coverage was achieved from the >10,000 year old aDNA of the woolly rhinoceros relative to DNA libraries of the Javan rhinoceros (Figure 2.3), likely stemming from less than ideal storage conditions and a high degree of microbial contamination of the Javan rhinoceros bones. Nevertheless, the previously documented sister species relationships between Javan and Indian rhinoceroses (Rhinocerotines) were upheld in all phylogenetic analyses, as were those within both Dicerotines and Dicerorhines. However, evolutionary histories among these three subfamilies remain contested as contradictory relationships resulted from the various phylogenetic analyses. The "biogeographical" hypothesis, suggesting that phylogenetic relationships mirror geographical distributions [((Dicerorhininae, Rhinocerotinae) Dicerotinae)], is supported by the ASTRAL-II coalescent analysis (Figure 2.4), as well as the RAxML concatenation analysis of protein-coding regions totaling ~64 kb (Figure 2.8). Conversly, the "number of horns" hypothesis, suggesting that two-horned rhinoceroses form a monophyletic taxon that is sister to more distantly related one-horned rhinoceroses [(Rhinocerotinae (Dicerorhininae, Dicerotinae))], is supported by SVDQuartet coalescence phylogenies

(Figure 2.5), Bayseian and RAxML concatenation analyses of the entire 131 kb supermatrix (Figure 2.6 and 2.7), as well as a RAxML tree of the intronic markers totaling ~68 kb (Figure 2.8). A high level of gene tree discordance was reflected by an average scaled RF value of ~0.45, which may be indicative of high ILS coinciding with quick speciation events. Exon versus intron disagreement may stem from confusion brought on by natural selection acting upon protein-coding regions of the markers, resulting in patterns of parallel evolution. The intron data should provide more accurate results as these regions were largely evolving under a lack of natural selection and are expected to show more phylogenetically informative nucleotide substitutions (Bailey et al. 1991). Six phylogenetically informative indels were discovered within intron sequences, five of which supported the "number of horns" hypothesis (Figure 2.9). Thus, while two of the phylogenetic analyses performed on this molecular dataset suggest the African versus Asian grouping of rhinoceroses, the majority of the results favor the one- versus two-horned relationship, though it is clear the Rhinocerotidae family tree remains somewhat uncertain due to confounding factors.

4.2. Molecular evolution of rhinoceros eye genes

Examination of rhinoceros eye genes ruled out pseudogenizations among these loci as possible molecular explanations for reputedly poor eyesight among members of this lineage, with selection pressure analyses instead revealing that while some loci (i.e. *GUCA1B* and *OPN4*) approach omega values indicative of neutral evolution in the white rhinoceros, most are evolving under purifying selection (Figure 2.10). Indeed, these findings fit with the recent characterization of a black rhinoceros retina suggesting that their vision may not be as poor as previously believed (Pettigrew and Manager 2008).

Nevertheless, the brief examination of these loci presented here provides a start for future studies aiming to uncover visual acuity of rhinoceroses and the molecular underpinnings of their eyesight.

4.3. Evolution of *UCP1* transcriptional regulatory elements across the mammalian phylogeny

My aim in Chapter 3 of this thesis was to trace the evolution of *UCP1* transcriptional regulatory elements in, not only rhinoceroses, but a total of 139 mammalian species as a continuation of my previous study that detailed the independent pseudogenization of thermogenic *UCP1* in several eutherian lineages (Gaudry et al. 2017). My comparative analyses reveal no evidence for a *UCP1* enhancer in marsupials and monotremes, bolstering the theory that this regulatory element arose with the advent UCP1-mediated NST in a stem eutherian mammal and may underlie the extremely high UCP1 concentrations in BAT compared to the relatively low expression levels of UCP2 and UCP3 in other tissues. Also, the CRE-4 and CCAAT promoter sites first described in mice and rats (Bouillaud et al. 1988; Kozak et al. 1994) are not conserved among other BAT expressing eutherians and thus, are not crucial for UCP1 transcription. Similarly, both the PRR and the CpG island are absent in some eutherians known to rely heavily upon BAT for thermoregulation (e.g. mice) and thus, are not essential for UCP1 expression. While the DNA methylation of the *UCP1* CpG island was once hypothesized to underlie tissue-specific expression of this protein, this seems unlikely given that this element is not conserved in all BAT expressing eutherians. Moreover, several *UCP1* enhancers transcription factor binding motifs (i.e. upTRE, dnTRE, CRE-2, RARE-2, NBRE, BRE-1, and BRE-2), differ in sequence identity relative to murid rodents (the

species in which they were first described) but remain conserved in other eutherian lineages, thus questioning their functionality. Indeed, these data may also reflect the evolution of differential mechanisms of transcriptional control among various eutherian lineages (i.e. rodents versus other euthterians). By contrast, the TATA box of the promoter as well as the CRE-3, PPRE, and RARE-3 motifs within the enhancer are highly conserved in nearly all eutherians and thus, likely to play a key role in UCP1 expression. Transcriptional regulatory regions are maintained in large-bodied species with presumably little need for BAT-mediated NST (e.g. rhinoceroses, camels, giraffe, and pinnipeds), suggesting that UCP1 may still play some functional role and emphasizing the need for UCP1 expression to be further researched in these lineages. While some species with *UCP1* inactivations retain conserved transcriptional regulatory regions, likely due to a slow molecular rate of neutral evolution linked with large body size, the TATA box of the promoter is highly degraded in cetaceans as well as the hyrax and the enhancer has been deleted in the bottlenose dolphin, killer whale, sperm whale, hyrax and pangolins. This latter finding suggests that the *UCP1* enhancer is not pleiotropic, as it would have been expected to be retained in the genome had it been involved in the transcriptional control of multiple gene products. With both BAT and UCP1 under extensive medical research as a promising avenue in the fight against human diabetes and obesity, studies such as that provide a deeper understanding of the evolution molecular mechanisms that may control *UCP1* transcription could prove to be extremely valuable in addition to broadening our understanding of the evolution of thermoregulation in eutherian mammals.

LITERATURE CITED

Alexander, G., Bennett, J.W., and Gemmell, R.T. (1975). Brown adipose tissue in the new-born calf (*Bos taurus*). *J. Physiol.* 244, 223-234.

Allentoft, M.E., Collins, M., Harker, D., Haile, J., Oskam, C.L., Hale, M.L., Campos, P.F., Samaniego, J.A., Gilbert, M.T.P., Willerslev, E., and Zhang, G. (2012). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. Lond. B Biol. Sci.* p.rspb20121745. DOI: 10.1098/rspb.2012.1745

Bailey, W.J., Fitch, D.H., Tagle, D.A., Czelusniak, J., Slightom, J.L., and Goodman, M. (1991). Molecular evolution of the psi eta-globin gene locus: gibbon phylogeny and the hominoid slowdown. *Mol. Biol. Evol*. 8, 155-184.

Berg, F., Gustafson, U., and Andersson, L. (2006). The uncoupling protein 1 gene (UCP1) is disrupted in the pig lineage: a genetic explanation for poor thermoregulation in piglets. *PLoS Genet*. 2, e129.

Bianco, A.C., and Silva, J.E. (1987). Optimal response of key enzymes and uncoupling protein to cold inbrown adipose tissue depends on local T_3 generation. *Am. J. Physiol.* 253, E255-E263

Binladen, J., Wiuf, C., Gilbert, M.T.P., Bunce, M., Barnett, R., Larson, G., Greenwood, A.D., Haile, J., Ho, S.Y., Hansen, A.J., and Willerslev, E. (2006). Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. *Genetics*. 172, 733-741.

Bird, A.P., and Wolffe, A.P. (1999). Methylation-induced repression—belts, braces, and chromatin. *Cell*. 99, 451-454.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.* 16*,* 6-21.

Boeskorov, G. (2001). Woolly rhino (*Coelodonta antiquitatis*) distribution in Northeast Asia. *Deinsea*. 8*,* 15-20.

Boeskorov, G.G., Lazarev, P.A., Sher, A.V., Davydov, S.P., Bakulina, N.T., Shchelchkova, M.V., Binladen, J., Willerslev, E., Buigues, B., and Tikhonov, A.N. (2011). Woolly rhino discovery in the lower Kolyma River. *Quat. Sci. Rev.* 30, 2262-2272.

Bromham, L. (2009). Why do species vary in their rate of molecular evolution?. *Biol. Lett*. rsbl-2009. DOI: 10.1098/rsbl2009.0136.

Bokar, J.A., Roesler, W.J., Vandenbark, G.R., Kaetzel, D.M., Hanson, R.W., and Nilson, J.H. (1988). Characterization of the cAMP responsive elements from the genes for the alpha-subunit of glycoprotein hormones and phosphoenolpyruvate

carboxykinase (GTP). Conserved features of nuclear protein binding between tissues and species. *J. Biol. Chem.* 263, 19740-19747.

Borges, R., Johnson, W.E., O'Brien, S.J., Vasconcelos, V., and Antunes, A. (2012). The role of gene duplication and unconstrained selective pressures in the melanopsin gene family evolution and vertebrate circadian rhythm regulation. *PloS one*, 7*,* e52413.

Brand, M.D., and Esteves, T.C. (2005). Physiological functions of the mitochondrial uncoupling proteins UCP2 and UCP3. *Cell metab.* 2, 85-93.

Brent, G.A., Moore, D.D., and Larsen, R.P. (1991). Thyroid hormone regulation of gene expression. *Annu. Rev. Physiol.* 53*,* 17-35.

Brook, S.M., Dudley, N., Mahood, S.P., Polet, G., Williams, A.C., Duckworth, J.W., Van Ngoc, T., and Long, B. (2014). Lessons learned from the loss of a flagship: The extinction of the Javan rhinoceros *Rhinoceros sondaicus annamiticus* from Vietnam. *Biol. Cons.* 174, 21-29.

Brook, S. van Coeverden de Groot, P., Mahood, S., and Long, B. (2011). Extinction of the Javan rhinoceros (*Rhinoceros sondaicus*) from Vietnam. WWF Vietnam Programme, Hanoi.

Brotherton, P., Endicott, P., Sanchez, J.J., Beaumont, M., Barnett, R., Austin, J., and Cooper, A. (2007). Novel high-resolution characterization of ancient DNA reveals C> U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res*. 35, 5717-5728.

Burset, M., Seledtsov, I.A., and Solovyev, V.V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 28, 4364-4375.

Bouillaud, F., Raimbault, S., and Ricquier, D. (1988). The gene for rat uncoupling protein: complete sequence, structure of primary transcript and evolutionary relationship between exons. *Biochem. Biophys. Res. Commun.* 157, 783-792.

Campbell, K.L., Storz, J.F., Signore, A.V., Moriyama, H., Catania, K.C., Payson, A., Bonaventura, J., Stetefeld, J., and Weber R.E. (2010). Molecular basis of a novel adaptation to hypoxic-hypercapnia in a strictly fossorial mole. *BMC Evol. Biol.* 10, 214. DOI: 10.1186/1471-2148-10-214

Cannon, B., and Nedergaard, J. (2004). Brown adipose tissue: function and physiological significance. *Physiol. Rev.* 84, 277-359.

Cédric, G., Neha, P., Roshan, P., Uttam, S., and Rajendra, G. (2016). Assessing and managing the rising rhino population in Kaziranga (India). *Ecol. Indic.* 66, 55-64.

Cerdeño, E. (1995). Cladistic analysis of the family Rhinocerotidae (Perissodactyla). *Am. Mus. Novit*. 3145, 1-25.

Cerdeño, E. (1998). Diversity and evolutionary trends of the Family Rhinocerotidae (Perissodactyla). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 141, 13-34.

Chifman, J., and Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*. 30, 3317-3324. DOI: 10.1093/bioinformatics/btu530.

Chou, J., Gupta, A., Yaduvanshi, S., Davidson, R., Nute, M., Mirarab, S., and Warnow, T. (2015). A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*. 16, 1. DOI: 10.1186/1471-2164-16-S10-S2

Cope, E.D. (1873). On the osteology of the extinct tapiroid Hyrachyus. *Proc. Am. Philos. Soc.* 13, 212-224.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.L., and Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15758-15763.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., and Driscoll, M. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5261-5266.

Degnan, J.H., and Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evolut*. 24, 332-340.

Deng, T., Hanta, R., and Jintasakul, P. (2013). A new species of Aceratherium (Rhinocerotidae, Perissodactyla) from the late Miocene of Nakhon Ratchasima, northeastern Thailand. *J. Vert. Paleontol.* 33, 977-985.

Deng, T., Wang, X., Fortelius, M., Li, Q., Wang, Y., Tseng, Z.J., Takeuchi, G.T., Saylor, J.E., Säilä, L.K., and Xie, G. (2011). Out of Tibet: Pliocene woolly rhino suggests high-plateau origin of Ice Age megaherbivores. *Science*, 333, 1285-1288.

Di Palma, F., Alfoldi, J., Johnson, J., Berlin, A., Gnerre, S., Jaffe, D., MacCallum, I., Young, S., Walker, B.J., and Lindblad-Toh, K. (2012). Draft genome of *Ceratotherium simum*. Unpublished - Direct submission.

Domning, D.P., Emry, R.J., Portell, R.W., Donovan, S.K., and Schindler, K.S. (1997). Oldest West Indian land mammal: rhinocerotoid ungulate from the Eocene of Jamaica. *J. Vert. Paleontol.* 17, 638-641.

Dong, C., Zhang, J., Qiao, J., and He, G. (2012). Positive selection and functional divergence after melanopsin gene duplication. *Biochem. Genet.* 50*,* 235-248.

Echtay, K. (2007). Mitochondrial uncoupling proteins-What is their physiological role? *Free. Radic. Biol. Med.* 43, 1351-1371.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 19, 1792-1797.

Endo, H., Kobayashi, H., Koyabu, D., Hayashida, A., Jogahara, T., Taru, H., Oishi, M., Itou, T., Koie, H., and Sakai, T. (2009). The morphological basis of the armor-like folded skin of the greater Indian rhinoceros as a thermoregulator. *Mamm. Study.* 35, 195-200.

Emerling, C.A., and Springer, M.S. (2014). Eyes underground: regression of visual protein networks in subterranean mammals. *Mol. Phylogenet. Evol. 78*, 260-270.

Feldmann, H.M., Golozoubova, V., Cannon, B., and Nedergaard, J. (2009). UCP1 ablation induces obesity and abolishes diet-induced thermogenesis in mice exempt from thermal stress by living at thermoneutrality. *Cell Metab.* 9*,* 203-209.

Fernando, P., Polet, G., Foead, N., Ng, L., Pastorini, J., and Melnick, D. (2006). Genetic diversity, phylogeny and conservation of the Javan rhinoceros (*Rhinoceros sondaicus*). *Conserv. Genet*. 7, 439-448.

Ferreira, S.M., Greaver, C., Knight, G.A., Knight, M.H., Smit, I.P.J., and Pienaar, D. (2015). Disruption of rhino demography by poachers may lead to population declines in Kruger National Park, South Africa. *PloS one.* 10, e0127783.

Foose, T.J., and van Strien, N.J. (1997). Asian rhinos: status survey and conservation action plan. Cambridge, UK: World Conserv. Monit. Cent. pp. 1-113.

Fortelius, M. (1983). The morphology and paleobiological significance of the horns of *Coelodonta antiquitatis* (Mammalia: Rhinocerotidae). *J. Vert. Paleontol*. 3, 125-135.

Galtier, N., and Daubin, V. (2008). Dealing with incongruence in phylogenomic analyses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 363*,* 4023-4029.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261-282.

Gatesy, J. Meredith, R.W., Janecka, J.E., Simmongs, M.P., Murphy, W.J., and Springer, M.S. (2016), Resolution of a concatenation/coalescence kerfuffle: partitioned coalescence support and a robust family-level tree for Mammalia. *Cladistics*. DOI: 10.1111/cla.12170

Gatesy, J., and Springer, M.S. (2014). Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol. Phylogenet. Evo.* 80:231-266.

Gaubert, P., Wozencraft, W.C., Cordeiro-Estrela, P., Veron, G., and Bininda-Emonds, O. (2005). Mosaics of convergences and noise in morphological phylogenies: What's in a viverrid-like carnivoran? *Syst. Biol*. 54, 865-894. DOI: 10.1080/10635150500232769

Gaudry, M.J., Jastroch, M., Treberg, J.R., Hofreiter, M., Paijmans, J.L.A., Starret, J., Wales, N., Signore, A.V., Springer, M.S., and Campbell, K.L. (2017). Inactivation of thermogenic UCP1 as a historical contingency in multiple placental mammal clades. *Sci. Adv.* 3:e1602878. DOI:10.1126/sciadv.1602878

Gojobori, T., Li, W.H., and Graur, D. (1982). Patterns of nucleotide substitutions in pseudogenes and functional genes. *J. Mol. Evol*. 18, 360-369.

Guérin, C. (1980). A propos des rhinocéros (Mammalia, Perissodactyla) néogènes et quaternaires d'Afrique: essai de synthèse sur les espèces et sur les gisements. In Proceedings 8th Panafrican congress prehistory and Quaternary studies*, Nairobi.* pp. 58-63.

Guérin, C. (1982). Les Rhinocerotidae (Mammalia, Perissodactyla) du Miocène terminal au Pleistoène supérieur d'Europe Occidentale comparés aux espèces actuelles: Tendances évolutives et relations phylogénétiques. *Geobios*, 15, 599-605.

Guérin, C. (1989). La famille des Rhinocerotidae (Mammalia, Perissodactyla): systématique, histoire, evolution, paleoecology. *Cranium*. 2, 3-14.

Groves, C.P. (1983). Phylogeny of the living species of rhinoceros. *Z. Zool. Syst. Evol.* 21, 293-313.

Hagelberg, E., and Clegg, J.B. (1991). Isolation and characterization of DNA from archaeological bone. *Proc. R. Sco. Lond. B Biol. Sci.* 244, 45-50.

Hariyadi, A.R.S., Sajuthi, D., Astuti, D.A., Alikodra, H.S., and Maheshwari, H. (2016). Analysis of nutrition quality and food digestibility in male Javan rhinoceros (*Rhinoceros sondaicus*) in Ujung Kulon National Park. *Pachyderm.* 57, 86-96.

Harms, M., and Seale, P. (2013). Brown and beige fat: development, function and therapeutic potential. *Nat. Med.* 19, 1252-1263.

Havmøller, R.G., Payne, J., Ramono, W., Ellis, S., Yoganand, K., Long, B., Dinerstein, E., Williams, A.C., Putra, R.H., Gawi, J., and Talukdar, B.K. (2016). Will current conservation responses save the critically endangered Sumatran rhinoceros *Dicerorhinus sumatrensis*? *Oryx*. 50, 355-359.

He, K., Shinohara, A., Jiang, X.L., and Campbell, K.L. (2014). Multilocus phylogeny of talpine moles (Talpini, Talpidae, Eulipotyphla) and its implications for systematics. *Mol. Phylogenet. Evol.* 70, 513-521.

He, X., and Zhang, J. (2006). Toward a molecular understanding of pleiotropy. *Genetics*. 173, 1885-1891.

Heaton, J.M. (1972). The distribution of brown adipose tissue in the human. *J. Anat.* 112, 35-39.

Heldmaier, G. (1971). Nonshivering thermogenesis and body size in mammals. *J. Comp. Physiol*. 73, 222-248.

Herpin, P., Damon, M., and Le Dividich, J. (2002). Development of thermoregulation and neonatal survival in pigs. *Livest. Prod. Sci.* 78, 25-45.

Hiley, P.G., (1977). The thermoregulatory response of the rhinoceros (*Diceros bicornis* and *Ceratotherium simum*) and the zebra (*Equus burchelli*) to diurnal temperature change. *E. Afr. Wildl. J.*, 15, 377.

Hiller, M., Schaar, B.T., and Bejerano, G. (2012). Hundreds of conserved non-coding genomic regions are independently lost in mammals. *Nucleic. Acids. Res.* 40, 11463-11476.

Hillis, D.M. (1987). Molecular versus morphological approaches to systematics. *Annu. Rev. Ecol. Evol. Syst.* 18, 23-42.

Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, A.V., and Pääbo, S., (2001b). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* 29, 4793-4799.

Hofreiter, M., Serre, D., Poinar, H.N., Kuch, M., and Pääbo, S., (2001a). Ancient DNA. *Nat. Rev. Genet.* 2, 353-359.

Hooijer, D.A. (1968). A rhinoceros from the late Miocene of Fort Ternan, Kenya. *Zool. Med. Leiden.* 43, 77-92.

Hooijer, D.A. (1978). "Rhinocerotidae," in *Evolution of African Mammals*, eds. V.J. Maglio, and H.B.S. Cooke (Cambridge, MA: Harvard University Press), 371-378.

Huelsenbeck, J.P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17, 754-755.

Hughes, D.A. Jastroch, M., Stoneking, M., and Klingenspor, M., (2009). Molecular evolution of UCP1 and the evolutionary history of mammalian non-shivering thermogenesis. *BMC. Evol. Biol.* 9, 4. doi: 10.1186/147-2148-9-4

Hutchins, M., and Kreger, M.D. (2006). Rhinoceros behaviour: implications for captive management and conservation. *Int. Zoo. Yb.* 40, 150-173.

Ishigaki, Y., Katagiri, H., Yamada, T., Ogihara, T., Imai, J., Uno, K., Hasegawa, Y., Gao, J., Ishihara, H., Shimosegawa, T., Sakoda, H., Asano, T., and Oka, Y. (2005). Dissipating excess energy stored in the liver is a potential treatment strategy for diabetes associated with obesity. *Diabetes*. 54, 322-332.

Jastroch, M., Withers, K.W., Taudien, S., Frappell, P.B., Helwig, M., Fromme, T., Hirschberg, V., Heldmaier, G., McAllan, B.M., Firth, B.T., Brumester, T., Platzer, M., and Klingenspor, M. (2008). Marsupial uncoupling protein 1 sheds light on the evolution of mammalian nonshivering thermogenesis. *Physiol. Genomics.* 32*,* 161-169.

Jastroch, M., Buckingham, J.A., Helwig, M., Klingenspor, M., and Brand, M.D. (2007). Functional characterization of UCP1 in the common carp: uncoupling activity in liver mitochondria and cold-induced expression in the brain. *J. Comp. Physiol. B.* 177, 743-752.

Kahlke, R.D., and Lacombat, F. (2008). The earliest immigration of woolly rhinoceros (*Coelodonta tologoijensis*, Rhinocerotidae, Mammalia) into Europe and its adaptive evolution in Palaearctic cold stage mammal faunas. *Quat. Sci. Rev.* 27, 1951-1961.

Kapur, V.V., and Bajpai, S., (2015). Oldest South Asian tapiromorph (Perissodactyla, Mammalia) from the Cambay Shale Formation, western India, with comments on its phylogenetic position and biogeographic implications. *The Palaeobotanist*. 64, 95-103.

Klingenspor, M., Fromme, T., Hughes, D.A., Manzke, L., Polymeropoulos, E., Riemann, T., and Trzcionkam, M. (2008). An ancient look at UCP1. *Biochim. Biophys. Acta, Bioenergetics*. 1777, 637-641.

Klingenspor, M., and Fromme, T. (2012). "Brown adipose tissue," in *Adipose tissue Biology*. ed. M.E. Symonds (New York, NY: Springer), 39-79.

Kiskinis, E., Hallberg, M., Christian, M., Olofsson, M., Dilworth, S.M., White, R., and Parker, M.G. (2007). RIP140 directs histone and DNA methylation to silence Ucp1 expression in white adipocytes. *EMBO J.* 26, 4831-4840.

Klug, J., Knapp, S., Castro, I., and Beato, M. (1994). Two distinct factors bind to the rabbit uteroglobin TATA-box region and are required for efficient transcription. *Mol. Cell. Biol.* 14, 6208-6218.

Kozak, U.C., Kopecky, J., Teisinger, J., Enerback, S., Boyer, B., and Kozak, L.P. (1994) An upstream enhancer regulating brown-fat-specific expression of the mitochondrial uncoupling protein gene. *Mol. Cell. Biol.* 14, 59-67.

Kumar, N., Liu, D., Wang, H., Robidoux, J., and Collins, S. (2008). Orphan nuclear receptor NOR-1 enhances 3′,5′-cyclic adenosine 5′-monophosphate-dependent uncoupling protein-1 gene transcription. *Mol. Endocrinol.* 22, 1057-1064.

Laursen, W.J., Mastrotto, M., Pesta, D., Funk, O.H., Goodman, J.B., Merriman, D.K., Ingolia, N., Shulman, G.I., Bagriantsev, S.N., and Gracheva, E.O. (2015). Neuronal UCP1 expression suggests a mechanism for local thermogenesis during hibernation. *Proc. Natl. Acad. Sci. U.S.A*. 112, 1607-1612.

Lasken, R.S., and Stockwell, T.B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol*. 7, 19 DOI: 10.1186/1472-6750-7-19

Lee, T.I., and Young, R.A. (2000). Transcription of eukaryotic protein-coding genes. *Annu. Rev. Genet.* 34, 77-137.

Li, W.H., Wu, C.I., and Luo, C.C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* 21, 58-71.

Lin, J., Cao, C., Tao, C., Ye, R., Dong, M., Zheng, Q., Wang, C., Jiang, X., Yan, C., Li, K., Speakman, J.R., Wang, Y., Jin, W., and Zhao, J. (2017). Cold adaptation in pigs depends on UCP3 in beige adipocytes. *J. Mol. Cell. Biol.* 1-12. DOI: 10.1093/jmcb/mjx018

Loose, H. (1975). Pleistocene Rhinocerotidae of W. Europe with reference to the recent two-horned species of Africa and S. E. Asia. *Scripta Geol.* 33, 1-59.

Loots, G., and Ovcharenko, I. (2004). rVista 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.* 32, W217-W221.

Lorenzen, E.D., Nogués-Bravo, D., Orlando, L., Weinstock, J., Binladen, J., Marske, K.A., Ugan, A., Borregaard, M.K., Gilbert, M.T.P., Nielsen, R., and Ho, S.Y. (2011). Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*. 479, 359-364.

Mailloux, R.J., and Harper, M.E. (2011). Uncoupling proteins and the control of mitochondrial reactive oxygen species production. *Free. Radic. Biol. Med*. 51, 1106-1115.

Martin, A.P., and Palumbi, S.R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. U.S.A*. 90, 4087-4091.

Mantovani, R. (1999). The molecular biology of the CCAAT-binding factor NF-Y. *Gene*. 239*,* 15-27.

McGaugh, S., and Schwartz, T.S. (2017). Here and there, but not everywhere: repeated loss of uncoupling protein 1 in amniotes. *Biol. Lett.* 13, 20160749.

McNab, B.K. (1983). Energetics, body size, and the limits to endothermy. *J. Zool.* 199, 1-29.

Meredith, R.W., Gatesy, J., Cheng, J., and Springer, M.S. (2011). Pseudogenization of the tooth gene enamelysin (*MMP20*) in the common ancestor of baleen whales. *Proc. R. Soc. B*. 278, 993-1002. doi: 10.1098/rspb.2010.1280

Meredith, R.W., Janečka, J.E., Gatesy, J., Ryder, O.A., Fisher, C.A., Teeling, E.C., Goodbla, A., Eizirik, E., Simão, T.L., Stadler, T., and Rabosky, D.L. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science*, 334, 521-524.

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* No. 6. Pbdprot5448.

Milliken, T., Emslie, R.H., and Talukdar, B. (2009). "African and Asian rhinoceroses–status, conservation and trade," in A report from the IUCN Species Survival Commission (IUCN/SSC) African and Asian rhino specialist groups and TRAFFIC to the CITES secretariat pursuant to resolution conf. 9.14 (Rev. CoP14) and decision 14.89. Report to CITES 15th meeting (Doha, March 2010), CoP 15 Doc.45.1A annex: 1-18.

Mirarab, S., and Warnow T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*. 31, i44-i52. DOI: 10.1093/bioinformatics/btv234

Moore, D.I. (1978). Post-glacial vegetation in the South Patagonian territory of the giant ground sloth, Mylodon. *Bot. J. Linean. Soc.* 77, 177-202.

Moore, W.S. (1995). Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution.* 49, 718-726.

Morales, J.C., and Melnick, D.J. (1994). Molecular systematics of the living rhinoceros. *Mol. Phylogenet. Evol.* 3, 128-134.

Mzilikazi, N., Jastroch, M., Meyer, C.W., and Klingenspor, M. (2007). The molecular and biochemical basis of nonshivering thermogenesis in an African endemic mammal, *Elephantulus myurus*. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 293, R2120-R2127.

Nakajima, N., Horikoshi, M., and Roeder, R.G. (1988). Factors involved in specific transcription by mammalian RNA polymerase II: purification, genetic specificity, and TATA box-promoter interactions of TFIID. *Mol. Cell. Biol.* 8, 4028-4040.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol*. 32, 268-274.

Nowak, R.M. (1999). *Walker's Mammals of the World, 6th edition, Volume II*. Baltimore, MD: The Johns Hopkins University Press.

Oelkrug, R., Goetze, N., Exner, C., Lee, Y., Ganjam, G.K., Kutschke, M., Müller, S., Stöhr, S., Tschöp, M.H., Crichton, P.G., Heldmaier, G., Jastroch, M., and Meyer, C.W. (2013). Brown fat in a protoendothermic mammal fuels eutherian evolution. *Nat. Commun*. 4, 2140. doi:10.1038/ncomms3140.

Oelkrug, R., Polymeropoulos, E.T., and Jastroch, M., (2015). Brown adipose tissue: physiological function and evolutionary significance. *J. Comp. Physiol. B.* 185, 587-606.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., and Johnson, P.L. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*. 499, 74-78.

Orlando, L., Leonard, J.A., Thenot, A., Laudet, V., Guérin, C., and Hänni, C. (2003). Ancient DNA analysis reveals woolly rhino evolutionary relationships. *Mol. Phylogent. Evol.* 28, 485-499.

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., and Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annu. Rev. Genet. 38*, 645-679.

Patikoglou, G.A., Kim, J.L., Sun, L., Yang, S.H., Kodadek, T., and Burley, S.K. (1999). TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes. Dev.* 13, 3217-3230.

Payne, A.M., Downes, S.M., Bessant, D.A., Plant, C., Moore, T., Bird, A.C., and Bhattacharya, S.S., (1999). Genetic analysis of the guanylate cyclase activator 1B (GUCA1B) gene in patients with autosomal dominant retinal dystrophies. *J. Med. Genet.* 36, 691-693.

Pearson, L. E., Liwanag, H. E., Hammill, M. O., and Burns, J. M. (2014). To each its own: Thermoregulatory strategy varies among neonatal polar phocids. *Comp. Biochem. Physiol. A*. 178, 59-67.

Pedersen, S.B., Bruun, J.M., Kristensen, K., and Richelsen, B. (2001). Regulation of UCP1, UCP2, and UCP3 mRNA expression in brown adipose tissue, white adipose tissue, and skeletal muscle in rats by estrogen. *Biochem. Biophys. Res. Commun.* 288, 191-197.

Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., and Bejerano, G. (2013). Enhancers: five essential questions. *Nature Rev. Genet.* 14, 288-295.

Peterson, M.G., Tanese, N., Pugh, B.F., and Tjian, R. (1990). Functional domains and upstream activation properties of cloned human TATA binding protein. *Science*. 248, 1625.

Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., Egholm, M., Rothberg, J.M., and Leamon, J.H. (2006). Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*. *7*, 216.

Pocock. R.I. (1945). Some cranial and dental characters of the existing species of Asiatic rhinoceroses. *Proc. Zool. Soc. Lond*. 114, 437-450.

Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R.D., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., and Rampp, M. (2006). Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*. 311, 392-394.

Polymeropoulos, E.T. Jastroch, M., and Frappell, P.B. (2012). Absence of adaptive nonshivering thermogenesis in a marsupial, the fat-tailed dunnart (*Sminthopsis crassicaudata*). *J. Comp. Physiol. B.* 182, 393-401. DOI: 10.1007/s00360-011-0623-x.

Price, S.A., and Bininda-Emonds, O.R.P. (2009). A comprehensive phylogeny of extant horses, rhinos and tapirs (Perissodactyla) through data combination. *Zoosyst. Evol*. 85, 277-292.

Prothero, D.R., Guérin, C., and Manning, E. (1989). "The History of the Rhinocerotoidea," in *The evolution of Perissodactyls*, eds. D.R. Prothero, and R.M. Schoch, (New York, NY: Oxford University Press.), 321-340.

Prothero, D.R. 1992. “Fifty million years of rhinoceros evolution,” in *Scientific American book of dinosaurs*, ed. O.A. Reyder (San Diego, CA: Zoological Scoiety), 82-91.

Prothero, D.R. and Schoch, R.M. 1989. “Classification of the Perissodactyla,” in *The Evolution of Perissodactyls*, eds. D.R. Prothero and R.M. Schoch (New York, NY: Oxford University Press), 530-537.

Prothero, D. R., Manning, E., and Hnason, C.B. (1986). The phylogeny of the Rhinocerotoidea (Mammalia, Perissodactyla). *Zool. J. Linnean Soc*. 87, 341-366.

Przełecka, A. (1981). Seasonal changes in ultrastructure of brown adipose tissue in the common shrew (*Sorex araneus L*.). *Cell. Tissue. Res.* 214*,* 623-632.

Rabelo, R., Reyes, C., Schifman, A., and Silva, J.E. (1996). A complex retinoic acid response element in the uncoupling protein gene defines a novel role for retinoids in thermogenesis. *Endocrinology*. 137, 3488-3496.

Rabelo, R., Schifman, A., Rubio, A., Sheng, X., and Silva, J.E. (1995). Delineation of thyroid hormone-responsive sequences within a critical enhancer in the rat uncoupling protein gene. *Endocrinology*. 136, 1003-1013.

Radinsky, L.B. (1966). The families of the Rhinocerotoidea (Mammalia, Perissodactyla). *J. Mammal.* 47, 631-639.

Radinsky, L.B. (1969). The early evolution of the Perissodactyla. *Evolution.* 23, 308-328.

Reiss, K.Z. (2001). Using phylogenies to study convergence: The case of the ant-eating mammals. *Am. Zool.* 41, 507-525.

Ripple, W.J., Newsome, T.M., Wolf, C., Dirzo, R., Everatt, K.T., Galetti, M., Hayward, M.W., Kerley, G.I., Levi, T., Lindsey, P.A., and Macdonald, D.W. (2015). Collapse of the world’s largest herbivores. *Sci. Adv. 1,* e1400103.

Robinson, D.F., and Foulds, L.R. (1981). Comparison of phylogenetic trees. *Math. Biosci*. 53, 131-147.

Rohland, N., and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* 22, 939-946.

Rosenberg, M.S., and Kumar, S. (2001). Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc. Natl. Acad. Sci. U.S.A*. 98, 10751-10756.

Rowlatt, U., Mrosovsky, N., and English, A. (1971). A comparative survey of brown fat in the neck and axilla of mammals at birth. *Biol. Neonate.* 17, 53-83.

Sabina, J., and Leamon, J.H. (2015). Bias in whole genome amplification: Causes and considerations. *Methods Mol. Biol.* 1347, 15-41.

Saito, S., Saito, C.T., and Shingai, R. (2008). Adaptive evolution of the uncoupling protein 1 gene contributed to the acquisition of novel nonshivering thermogenesis in ancestral eutherian mammals. *Gene.* 408, 37-44.

Sato, M., Nakazawa, M., Usui, T., Tanimoto, N., Abe, H., and Ohguro, H. (2005). Mutations in the gene coding for guanylate cyclase-activating protein 2 (GUCA1B gene) in patients with autosomal dominant retinal dystrophies. *Graefe's Arch. Clin. Exp. Ophthalmol.* 243, 235-242.

Scornavacca, C., and Galtier, N. (2017). Incomplete lineage sorting in mammalian phylogenomics. *Syst. Biol.* 66, 112-120.

Sears, I.B., MacGinnitie, M.A., Kovacs, L.G., and Graves, R.A. (1996). Differentiation-dependent expression of the brown adipocyte uncoupling protein gene: regulation by peroxisome proliferator-activated receptor gamma. *Mol. Cell. Biol.* 16, 3410-3419.

Shore, A., Emes, R.D., Wessely, F., Kemp, P., Cillo, C., D'Armiento, M., Hoggard, N., and Lomax, M.A. (2012). A comparative approach to understanding tissue-specific expression of uncoupling protein 1 expression in adipose tissue. *Front. Genet.* 3, DIO: 10.3389/fgene.2012.00304.

Shore, A., Karamitri, A., Kemp, P., Speakman, J.R., and Lomax, M.A. (2010). Role of UCP1 enhancer methylation and chromatin remodeling in the control of UCP1 expression in murine adipose tissue. *Diabetologia*. 51, 1164-1173. DIO: 310.1007-s00125-010-1701-4.

Silva, M., and Downing, J.A. (1995). *CRC handbook of mammalian body masses*. Boca Raton, FL: CRC press.

Simpson, G.G. (1945). The principles of classification and a classification of mammals. *Bull. Am. Mus. Nat. Hist.* 86, 1-350.

Skinner, J.D., and Chimimba, C.T. (2005). *The Mammals of the Southern African Subregion, 3rd Edition.* Cambridge, UK: Cambridge University Press.

Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015 <http://www.repeatmasker.org>.

Song, S., Liu, L., Edwards, S.V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. U.S.A*. 109, 14942-14947.

Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. *Annu. Rev. Biochem.* 72, 449-479.

Springer, M.S., DeBry, R.W., Douady, C., Amrine, H.M., Madsen, O., de Jong, W.W., and Stanhope, M.J. (2001). Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. *Mol. Biol. Evol.* 18, 132-143.

Springer, M.S., Signore, A.V., Paijmans, J.L., Vélez-Juarbe, J., Domning, D.P., Bauer, C.E., He, K., Crerar, L., Campos, P.F., Murphy, W.J., Meredith, R.M., Gatesy, J., Willerslev, E., MacPhee, R.D.E., Hofreiter, M., and Campbell, K.L. (2015). Interordinal gene capture, the phylogenetic position of Steller's sea cow based on molecular and morphological data, and the macroevolutionary history of Sirenia. *Mol. Phylogenet. Evol.* 91, 178-193.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 22, 2688-2690.

Steiner C.C., and Ryder, O.A. (2011). Molecular phylogeny and evolution of the Perrissodactyla. *Zool. J. Linnean Soc.* 163, 1289-1303.

Stuart, A.J., and Lister, A. (2012). Extinction chronology of the woolly rhinoceros Coelodonta antiquitatis in the context of late Quaternary megafaunal extinctions in northern Eurasia. *Quat. Sci. Rev*. 51, 1-17.

Suh, A., Smeds, L., Ellegren, H. (2015). The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biology*. 13, e1002224. doi: 10.1371/journal.pbio.1002224

Sullivan, M.J., Petty, N.K., and Beatson, S.A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics*. 27, 1009-1010.

Swofford, D.L. (2002). *PAUP* version 4.0. Phylogenetic analysis using parsimony (and other methods)*. Sunerland, MA: Sinauer Associates.

Talukdar, B.K. (2009). Asian rhino specialist group report. *Pachyderm*. 46, 14-17.

Thomas, D.W., Dorais, M., and Bergeron, J.M. (1990). Winter energy budgets and cost of arousals for hibernating little brown bats, *Myotis lucifugus*. *J. Mamm.* 71, 475-479.

Tougard, C., Delefosse, T., Hänni, C., and Montgelard, C. (2001). Phylogenetic relationships of the five extant rhinoceros species (Rhinocerotidae, Perissodactyla) based on mitochondrial cytochrome b and 12S rRNA genes. *Mol. Phylogenet. Evol.* 19, 34-44.

Umesono, K., Murakami, K.K., Thompson, C.C., and Evans, R.M. (1991). Direct repeats as selective response elements for the thyroid hormone, retinoic acid, and vitamin D3 receptors. *Cell*. 65, 1255-1266.

Vawter, L., and Brown, W.M. (1986). Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock. *Science*. 234*,* 194-196.

Villarroya, F., Peyrou, M., and Giralt, M. (2017). Transcriptional regulation of the uncoupling protein-1 gene. *Biochimie*. 134, 86-92.

Wall, W.P. (1980). Cranial evidence for a proboscis in Cadurcodon and a review of snout structure in the family Amynodontidae (Perissodactyla, Rhinocerotoidea). *J. Paleo.* 54, 968-977.

Welker, F., Smith, G.M., Hutson, J.M., Kindler, L., Garcia-Moreno, A., Villaluenga, A., Turner, E., and Gaudzinski-Windheuser, S. (2017). Middle Pleistocene protein sequences from the rhinoceros genus Stephanorhinus and the phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae. *PeerJ*. 5, e3033.

Willerslev, E., Gilbert, M.T.P., Binladen, J., Ho, S.Y., Campos, P.F., Ratan, A., Tomsho, L.P., da Fonseca, R.R., Sher, A., Kuznetsova, T.V., and Nowak-Kemp, M. (2009). Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. *BMC Evol. Biol*. 9, 95.

Wu, J., Boström, P., Sparks, L.M., Ye, L., Choi, J.H., Giang, A.H., Khandekar, M., Virtanen, K.A., Nuutila, P., Schaart, G., Huang, K., Tu, H., van Marken Lichtenbelt, W.D., Hoeks, J., Enerbäck, S., Schrauwen, P., and Spiegelman, B.M. (2012). Beige adipocytes are a distinct type of thermogenic fat cell in mouse and human. *Cell*. 150*,* 366-376.

Xu, L., Thali, M., and Schaffner, W. (1991). Upstream box/TATA box order is the major determinant of the direction of transcription. *Nucleic. Acids. Res.* 19, 6699-6704.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586-1591.

Yubero, P., Vinas, O., Iglesias, R., Mampel T., Villarroya, F., and Giralt, M. (1994). Identification of tissue-specific protein binding domains in the 5'-proximal regulatory region of the rat mitochondrial brown fat uncoupling protein gene. *Biochem. Biophys. Res. Commun.* 204, 867-873.

Zafir, A.W.A., Payne, J., Mohamed, A., Lau, C.F., Sharma, D.S.K., Alfred, R., Williams, A.C., Nathan, S., Ramono, W.S., and Clements, G.R. (2011). Now or never: What will it take to save the Sumatran rhinoceros *Dicerorhinus sumatrensis* from extinction? *Oryx*. 45, 225-233.

Zin-Maung-Maung-Thein, T., Masanaru, T., Tsubamoto, T., Thaung-Htike, Egi, N., and Maung-Maung. (2008). A new species of Dicerorhinus (Rhinocerotidae) from the Plio-Pleistocene of Myanmar. *Palaeontology*. 51, 1419–1433.

Zwickl, D.J. (2006). GARLI: genetic algorithm for rapid likelihood inference. Version 2.0. (http://www.bio.utexas.edu/faculty/antisense/garli.Garli.html).

APPENDICES

Appendix 1. Genetic markers targeted using in-solution hybridization capture experiments. The 26 "Assembly of the Tree of Life" markers originate from the Meredith et al. (2011), 8 genetic markers originate from the Steiner and Ryder (2011) publication, while the 165 genetic markers are unique to this study stemming from 54 nuclear genes.

Gene abbreviation	Full gene name	Segment length (bp)
Assembly of the Tree of Life genes (Meredith et al. 2011)		
A2AB	Alpha-2B adrenergic receptor	1371
ADORA3	Adenosine A3 receptor	610
ADRB2	Adrenoceptor Beta 2	1178
APOB	Apolipoprotein B	2639
APP	Amyloid precursor protein	868
ATP7A	ATPase copper trasnporting Alpha polypeptide	726
BCHE	Butyrylcholinesterase	1225
BDNF	Brain Derived Neurotrophic Factor	866
BMI1	BMI1 Proto-Oncogene, Polycomb Ring Finger	686
BRCA1	Breast Cancer 1, Early Onset	3115
BRCA2	Breast Cancer 2, Early Onset	4472
CNR1	Cannabinoid Receptor 1	1233
CREM	CAMP Responsive Element Modulator	677
DMP1	Dentin Matrix Acidic Phosphoprotein 1	1609
EDG1	Endothelial Differentiation G-Protein Coupled Receptor 1	1146
ENAM	Enamelin	3189
FBN1	Fibrillin 1	1021
GHR	Growth Hormone Receptor	1180
IRBP	Interphotoreceptor retinoid binding protein	1486
PLCB4	Phospholipase C Beta 4	514
PNOC	Prepronociceptin	548
RAG1	Recombination Activating 1	2602
RAG2	Recombination Activating 2	806
TTN	Titin	1934
TYR	Tyrosinase	820
vWF	von Willebrand Factor	1361
Steiner and Ryder 2011 genes		
EDNRB	Endothelin Receptor Type B	873
KIT CDS	KIT Proto-Oncogene Receptor Tyrosine Kinase	644
KIT Intron	KIT Proto-Oncogene Receptor Tyrosine Kinase	1173
MC1R	Melanocortin 1 Receptor	1242
MITF	Melanogenesis Associated Transcription Factor	583
SNAI2	Snail Family Transcriptional Repressor 2	934
SOX10	SRY-Box 10	603

TBX15	T-Box 15	948
Genes unique to this study		
AMBN exon 1	Ameloblastin	412
AMBN exon 2	Ameloblastin	488
AMBN exon 3	Ameloblastin	362
AMBN exon 4	Ameloblastin	476
AMBN exon 5	Ameloblastin	361
AMBN exon 6	Ameloblastin	564
AMBN exon 7	Ameloblastin	500
AMBN exons 8 and 9	Ameloblastin	425
AMBN exon 10	Ameloblastin	476
AMBN exon 11	Ameloblastin	1508
ARR exon 4	Arrestin 3, retinal (x-arrestin)	437
ARR exon 6	Arrestin 3, retinal (x-arrestin)	490
ARR exon 7	Arrestin 3, retinal (x-arrestin)	464
ARR exon 9	Arrestin 3, retinal (x-arrestin)	489
ARR exon 10	Arrestin 3, retinal (x-arrestin)	453
ARR exon 11	Arrestin 3, retinal (x-arrestin)	452
ARR exon 12	Arrestin 3, retinal (x-arrestin)	534
ARR exon 13	Arrestin 3, retinal (x-arrestin)	507
CNGB3 exon 3	Cyclic Nucleotide Gated Channel Beta 3	554
CNGB3 exon 4	Cyclic Nucleotide Gated Channel Beta 3	478
CNGB3 exon 5	Cyclic Nucleotide Gated Channel Beta 3	484
CNGB3 exon 6	Cyclic Nucleotide Gated Channel Beta 3	582
CNGB3 exon 7	Cyclic Nucleotide Gated Channel Beta 3	338
CNGB3 exon 8	Cyclic Nucleotide Gated Channel Beta 3	446
CNGB3 exon 9	Cyclic Nucleotide Gated Channel Beta 3	364
CNGB3 exon 10	Cyclic Nucleotide Gated Channel Beta 3	471
CNGB3 exon 11	Cyclic Nucleotide Gated Channel Beta 3	477
CNGB3 exon 12	Cyclic Nucleotide Gated Channel Beta 3	529
CNGB3 exon 13	Cyclic Nucleotide Gated Channel Beta 3	430
CNGB3 exon 14	Cyclic Nucleotide Gated Channel Beta 3	446
CNGB3 exon 15	Cyclic Nucleotide Gated Channel Beta 3	547
CNGB3 exon 16	Cyclic Nucleotide Gated Channel Beta 3	450
CNGB3 exon 17	Cyclic Nucleotide Gated Channel Beta 3	546
CNGB3 exon 18	Cyclic Nucleotide Gated Channel Beta 3	559
CSN2 exon 1	Casein Beta	472
CSN2 exons 2 and 3	Casein Beta	574
CSN2 exon 4	Casein Beta	630
CSN2 exon 5	Casein Beta	909
CYGB exon 1	Cytoglobin	454
CYGB exon 2	Cytoglobin	607
CYGB exon 3	Cytoglobin	720

CYGB exon 4	Cytoglobin	418
GNAT2 exon 1	G Protein Subunit Alpha Transducin 2	480
GNAT2 exon 2	G Protein Subunit Alpha Transducin 2	486
GNAT2 exon 3	G Protein Subunit Alpha Transducin 2	492
GNAT2 exon 4	G Protein Subunit Alpha Transducin 2	525
GNAT2 exon 5	G Protein Subunit Alpha Transducin 2	526
GNAT2 exon 6	G Protein Subunit Alpha Transducin 2	619
GNAT2 exon 7	G Protein Subunit Alpha Transducin 2	613
GNAT2 exon 8	G Protein Subunit Alpha Transducin 2	527
GNGT2 exon 4	G Protein Subunit Gamma Transducin	434
GNGT2 exon 5	G Protein Subunit Gamma Transducin	448
GRK7 exon 1	G Protein-Coupled Receptor Kinase 7	899
GRK7 exon 2	G Protein-Coupled Receptor Kinase 7	1030
GRK7 exon 3	G Protein-Coupled Receptor Kinase 7	699
GRK7 exon 4	G Protein-Coupled Receptor Kinase 7	609
GUCA1B exon 1	Guanylate Cyclase Activator 1B	597
GUCA1B exon 2	Guanylate Cyclase Activator 1B	625
GUCA1B exon 3	Guanylate Cyclase Activator 1B	441
GUCA1B exon 4	Guanylate Cyclase Activator 1B	584
LEPR exon 1	Leptin Receptor	404
LEPR exon 2	Leptin Receptor	666
LEPR exon 3	Leptin Receptor	452
LEPR exon 4	Leptin Receptor	561
LEPR exon 5	Leptin Receptor	421
LEPR exon 6	Leptin Receptor	516
LEPR exon 7	Leptin Receptor	604
LEPR exon 8	Leptin Receptor	427
LEPR exon 9	Leptin Receptor	553
LEPR exon 10	Leptin Receptor	277
LEPR exon 11	Leptin Receptor	387
LEPR exon 12	Leptin Receptor	475
LEPR exon 13	Leptin Receptor	596
LEPR exon 14	Leptin Receptor	530
LEPR exon 15	Leptin Receptor	527
LEPR exon 16	Leptin Receptor	473
LEPR exon 17	Leptin Receptor	504
LEPR exon 18	Leptin Receptor	1209
NGB exon 1	Neuroglobin	128
NGB exon 2	Neuroglobin	460
NGB exon 3	Neuroglobin	393
NGB exon 4	Neuroglobin	504
OB exon 1	Leptin	718
OB exon 2	Leptin	756

OPN1LW exon 1	Opsin 1 (Cone Pigments), Long-Wave-Sensitive	440
OPN1LW exon 2	Opsin 1 (Cone Pigments), Long-Wave-Sensitive	632
OPN1LW exon 3	Opsin 1 (Cone Pigments), Long-Wave-Sensitive	548
OPN1LW exon 4	Opsin 1 (Cone Pigments), Long-Wave-Sensitive	489
OPN1LW exon 5	Opsin 1 (Cone Pigments), Long-Wave-Sensitive	403
OPN1LW exon 6	Opsin 1 (Cone Pigments), Long-Wave-Sensitive	590
OPN4 exon 1	Opsin 4	431
OPN4 exon 2	Opsin 4	562
OPN4 exon 3	Opsin 4	442
OPN4 exon 4	Opsin 4	510
OPN4 exon 5	Opsin 4	691
OPN4 exon 6	Opsin 4	375
OPN4 exon 7	Opsin 4	426
OPN4 exon 8	Opsin 4	518
OPN4 exon 9	Opsin 4	426
OPN4 exon 10	Opsin 4	599
OPSD exon 1	Rhodopsin	581
OPSD exon 2	Rhodopsin	547
OPSD exon 3	Rhodopsin	518
OPSD exon 4	Rhodopsin	588
OPSD exon 5	Rhodopsin	395
PDE6C exon 1	Phosphodiesterase 6C	899
PDE6C exons 2 and 3	Phosphodiesterase 6C	709
PDE6C exon 4	Phosphodiesterase 6C	472
PDE6C exon 5	Phosphodiesterase 6C	351
PDE6C exons 6 and 7	Phosphodiesterase 6C	580
PDE6C exon 8	Phosphodiesterase 6C	286
PDE6C exon 9	Phosphodiesterase 6C	576
PDE6C exon 10	Phosphodiesterase 6C	563
PDE6C exon 11	Phosphodiesterase 6C	397
PDE6C exon 12	Phosphodiesterase 6C	488
PDE6C exon 13	Phosphodiesterase 6C	417
PDE6C exon 14	Phosphodiesterase 6C	498
PDE6C exon 15	Phosphodiesterase 6C	363
PDE6C exon 16	Phosphodiesterase 6C	521
PDE6C exons 17 and 18	Phosphodiesterase 6C	628
PDE6C exons 19 and 20	Phosphodiesterase 6C	708
PDE6C exon 21	Phosphodiesterase 6C	499
PDE6C exon 22	Phosphodiesterase 6C	389
PDE6H exon 1	Phosphodiesterase 6H	526
PDE6H exon 2	Phosphodiesterase 6H	439
PDE6H exon 3	Phosphodiesterase 6H	504
Prestin exon 1	Solute Carrier Family 26 Member 5	577

Prestin exon 2	Solute Carrier Family 26 Member 5	498
Prestin exon 3	Solute Carrier Family 26 Member 5	514
Prestin exon 4	Solute Carrier Family 26 Member 5	612
Prestin exon 5	Solute Carrier Family 26 Member 5	551
Prestin exon 6	Solute Carrier Family 26 Member 5	546
Prestin exon 7	Solute Carrier Family 26 Member 5	413
Prestin exon 8	Solute Carrier Family 26 Member 5	536
Prestin exon 9	Solute Carrier Family 26 Member 5	546
Prestin exon 10	Solute Carrier Family 26 Member 5	518
Prestin exon 11	Solute Carrier Family 26 Member 5	407
Prestin exon 12	Solute Carrier Family 26 Member 5	444
Prestin exon 13	Solute Carrier Family 26 Member 5	511
Prestin exon 14	Solute Carrier Family 26 Member 5	524
Prestin exon 15	Solute Carrier Family 26 Member 5	560
Prestin exon 16	Solute Carrier Family 26 Member 5	660
Prestin exon 17	Solute Carrier Family 26 Member 5	406
Prestin exon 18	Solute Carrier Family 26 Member 5	521
SWS1 exon 1	Opsin 1 (Cone Pigments), Short-Wave-Sensitive	722
SWS1 exon 2	Opsin 1 (Cone Pigments), Short-Wave-Sensitive	465
SWS1 exon 3	Opsin 1 (Cone Pigments), Short-Wave-Sensitive	528
SWS1 exon 4	Opsin 1 (Cone Pigments), Short-Wave-Sensitive	635
SWS1 exon 5	Opsin 1 (Cone Pigments), Short-Wave-Sensitive	473
UCP1 exon 1	Uncoupling Protein 1	480
UCP1 exon 2	Uncoupling Protein 1	582
UCP1 exons 3 and 4	Uncoupling Protein 1	615
UCP1 exon 5	Uncoupling Protein 1	556
UCP1 exon 6	Uncoupling Protein 1	450
UCP2 exon 1	Uncoupling Protein 2	416
UCP2 exon 2	Uncoupling Protein 2	450
UCP2 exons 3 and 4	Uncoupling Protein 2	672
UCP2 exon 5	Uncoupling Protein 2	589
UCP2 exon 6	Uncoupling Protein 2	382
UCP3 exon 1	Uncoupling Protein 3	439
UCP3 exon 2	Uncoupling Protein 3	518
UCP3 exon 3	Uncoupling Protein 3	496
UCP3 exon 4	Uncoupling Protein 3	484
UCP3 exon 5	Uncoupling Protein 3	599
UCP3 exon 6	Uncoupling Protein 3	572

Appendix 2. Tissue types and quantities used to perform DNA extractions and create DNA libraries. For the black and Indian rhinoceroses, the DNA was previously extracted, thus the tissue type is not listed. Locations where the woolly rhinoceros fossils were acquired are listed as well as the barcodes for A and P adaptors. Barcoded P adaptors were only used for the ancient DNA libraries of the woolly rhinoceros to ensure the authenticity of both ends of the sequence reads. Asterisks represent phosphorothioate bonds that were used in adaptors for ancient DNA libraries to protect against endonuclease degradation.

Code	Species	DNA extraction material	A adaptor index	P adaptor index	Source
SR2	Sumatran rhinoceros	100 µL blood/leech	TAAGGAGAAC		R. Havmøller, T. Gilbert, and E. Willerslev, University of Copenhagen Leibniz-Institute for Zoo and Wildlife Research, Berlin, Germany
MT4	Malayan tapir	100 µL blood	TACCAAGATC		M. Berteslen, Copenhagen Zoo, Frederiksberg, Denmark T. Gilbert and E. Willerslev, University of Copenhagen
MT5	Malayan tapir	100 µL blood	CAGAAGGAAC		
IR	Indian rhinoceros	previously extracted	TTCTCATTGAAC		CRES
BR	Black rhinoceros	previously extracted	TAAGCCATTGTC		CRES
41JR446	Javan rhinoceros	100 mg bone	TCTAGCTCTTC		Peter van Coeverden de Groot, Queens University, Kingston, Ontario, Canada
43JR468	Javan rhinoceros	100 mg bone	TTCCTGCTTCAC		
WR1	Woolly rhinoceros	120 mg bone - China	TGAC*G*T*G*T	AGCT*G*C*G*T	T. Gilbert and E. Willerslev, University of Copenhagen
WR2	Woolly rhinoceros	113 mg tooth - Siberia	TCAC*T*A*G*T	AGA*T*A*T*C*T	
WR3	Woolly rhinoceros	176 mg bone - Indigirka	ATAG*A*G*C*T	TCTA*G*A*C*T	
WR4	Woolly rhinoceros	159 mg bone - Kolyma	ACAGCTGT	TCATGCGT	
WR5	Woolly rhinoceros	71 mg skin - Cherskii	TGAC*G*T*G*T	AGA*T*A*T*C*T	

Appendix 3. Accession numbers of all 199 genetic markers included in this study acquired from GenBank for eight species.

Gene abbreviasion	*Ceratotherium simum*	*Equus asinus*	*Equus caballus* (mongolian)	*Equus caballus* (throroughbred)	*Equus przewalskii*	*Sus scrofa*	*Bos taurus*	*Camelus ferus*
Assembly of the Tree of Life genes (Meredith et al. 2011)								
A2AB	AKZM01036161.1	JREZ01000530.1	ATDM01058753.1	AAWR02003437.1	ATBW01057699.1	AJKK01137175.1	DAAA02030114.1	AGVR01013673.1
ADORA3	AKZM01005752.1	JREZ01000141.1	ATDM01005511.1	AAWR02017081.1	ATBW01011722.1	AJKK01131255.1	DAAA02007523.1	AGVR01026070.1
ADRB2	AKZM01022125.1	JREZ01000329.1	ATDM01091287.1	AAWR02002274.1	ATBW01026832.1	AJKK01235047.1	DAAA02020602.1	AGVR01008536.1
APOB	AKZM01001011.1	JREZ01000059.1	ATDM01006810.1	AAWR02004366.1	ATBW01057833.1	AJKK01129282.1	DAAA02031729.1	AGVR01018504.1
APP	AKZM01023525.1	JREZ01000436.1	ATDM01022391.1	AAWR02022825.1	ATBW01010128.1	AJKK01215362.1	DAAA02000240.1	AGVR01033547.1
ATP7A	AKZM01049185.1	JREZ01000574.1	ATDM01030951.1	AAWR02034229.1	ATBW01089250.1	AJKK01149673.1	DAAA02072493.1	AGVR01063826.1
BCHE	AKZM01014441.1	JREZ01000432.1	ATDM01073484.1	AAWR02011449.1	ATBW01024171.1	AJKK01149474.1	DAAA02002286.1	AGVR01028033.1
BDNF	AKZM01001456.1	JREZ01000081.1	ATDM01078701.1	AAWR02018562.1	ATBW01002804.1	AJKK01204968.1	DAAA02041147.1	AGVR01060768.1
BMI1	AKZM01015508.1	JREZ01000553.1	ATDM01074630.1	AAWR02028728.1	ATBW01007532.1	AJKK01212992.1	DAAA02035442.1	AGVR01049498.1
BRCA1	AKZM01033323.1	JREZ01000009.1	ATDM01008130.1	AAWR02012618.1	ATBW01065891.1	AJKK01131722.1, AJKK01131723.1	DAAA02049201.1	AGVR01006136.1
BRCA2	AKZM01034892.1	JREZ01000110.1	ATDM01031919.1	AAWR02035369.1	ATBW01007591.1	AJKK01132312.1	DAAA02033078.1	AGVR01026502.1
CNR1	AKZM01008266.1	JREZ01000922.1	ATDM01014213.1	AAWR02014702.1	ATBW01016940.1	AJKK01239858.1	DAAA02026462.1	AGVR01016526.1
CREM	AKZM01015302.1	JREZ01000287.1	ATDM01064002.1	AAWR02028558.1	ATBW01069586.1	AJKK01227687.1	DAAA02035369.1	AGVR01042078.1
DMP1	AKZM01026006.1	JREZ01000199.1	ATDM01016919.1	AAWR02004809.1	ATBW01061324.1	AJKK01172986.1	DAAA02018360.1	AGVR01006876.1
EDG1	AKZM01005899.1	JREZ01000310.1	ATDM01018295.1	AAWR02017207.1	ATBW01018562.1	AJKK01214468.1	DAAA02007748.1	AGVR01044026.1
ENAM	AKZM01002861.1	JREZ01000259.1	ATDM01031704.1	AAWR02005019.1	ATBW01040478.1	AJKK01251004.1	DAAA02018021.1	AGVR01039102.1
FBN1	AKZM01007288.1	JREZ01000114.1	ATDM01032377.1	AAWR02001400.1	ATBW01047071.1	AJKK01265712.1	DAAA02029122.1	AGVR01047666.1
GHR	AKZM01010356.1	JREZ01000037.1	ATDM01004233.1	AAWR02016433.1	ATBW01029035.1	AJKK01139271.1	DAAA02050511.1	AGVR01049780.1
IRBP	AKZM01030114.1	JREZ01000188.1	ATDM01055010.1	AAWR02000275.1	ATBW01039778.1	AJKK01152680.1	DAAA02062187.1	AGVR01052084.1
PLCB4	AKZM01032150.1	JREZ01000071.1	ATDM01014420.1	AAWR02026310.1	ATBW01045679.1	AJKK01136240.1	DAAA02035027.1	AGVR01035622.1
PNOC	AKZM01031561.1	JREZ01000021.1	ATDM01018197.1	AAWR02032145.1	ATBW01064770.1	AJKK01178430.1, AJKK01178431.1	DAAA02021976.1	AGVR01022510.1
RAG1	AKZM01038803.1	JREZ01000302.1	ATDM01040481.1	AAWR02037054.1	ATBW01046869.1	AJKK01253333.1	DAAA02041329.1	AGVR01060669.1
RAG2	AKZM01038803.1	JREZ01000302.1	ATDM01064072.1	AAWR02037054.1	ATBW01046869.1	AJKK01253334.1	DAAA02041331.1	AGVR01060668.1
TTN	AKZM01018315.1	JREZ01000090.1	ATDM01004073.1	AAWR02006746.1	ATBW01012248.1	AJKK01247012.1	DAAA02004137.1	AGVR01055642.1
TYR	AKZM01032842.1	JREZ01000687.1	ATDM01005119.1	AAWR02017977.1	ATBW01011868.1	AJKK01258813.1	DAAA02062459.1	AGVR01047963.1
vWF	AKZM01041531.1	JREZ01000019.1	ATDM01031126.1	AAWR02029168.1	ATBW01009757.1	AJKK01155958.1	DAAA02014497.1	AGVR01007186.1
Steiner and Ryder 2011 genes								
EDNRB	AKZM01011074.1	JREZ01000046.1	ATDM01001915.1	AAWR02010931.1	ATBW01027999.1	AJKK01235865.1, AJKK01103932.1	DAAA02033626.1	AGVR01038532.1
KIT CDS	AKZM01002564.1	JREZ01000346.1	ATDM01000916.1	AAWR02005216.1	ATBW01032321.1	AJKK01266074.1	DAAA02017722.1	AGVR01013059.1
KIT Intron	AKZM01002564.1	JREZ01000346.1	ATDM01000916.1	AAWR02005216.1	ATBW01032321.1	AJKK01266074.1	DAAA02017722.1	AGVR01013059.1
MC1R	AKZM01039062.1	JREZ01000139.1	ATDM01036109.1	AAWR02027541.1	ATBW01074787.1	AJKK01260335.1	DAAA02046277.1	AGVR01003215.1
MITF	AKZM01004155.1	JREZ01000134.1	ATDM01037534.1	AAWR02032705.1	ATBW01002279.1	AJKK01156240.1	DAAA02054059.1	AGVR01034534.1
SNAI2	AKZM01026784.1	JREZ01000005.1	ATDM01020581.1	AAWR02040253.1	ATBW01077287.1	AJKK01128196.1	DAAA02038249.1	AGVR01010243.1
SOX10	AKZM01040169.1	JREZ01000305.1	ATDM01031165.1	AAWR02025014.1	ATBW01068581.1	NA	DAAA02014645.1	AGVR01005368.1
TBX15	AKZM01005633.1	JREZ01000011.1	ATDM01013936.1	AAWR02016963.1	ATBW01013404.1	AJKK01193204.1	DAAA02007365.1	AGVR01029492.1
Unique genes to this study								
AMBN exon 1	AKZM01002861.1	JREZ01000259.1	ATDM01085281.1	AAWR02005019.1	ATBW01040478.1	AJKK01251007.1	DAAA02018021.1	AGVR01039101.1
AMBN exon 2	AKZM01002861.1	JREZ01000259.1	ATDM01085281.1	AAWR02005019.1	ATBW01040478.1	AJKK01251007.1	DAAA02018021.1	AGVR01039101.1
AMBN exon 3	AKZM01002861.1	JREZ01000259.1	ATDM01089360.1	AAWR02005019.1	ATBW01040478.1	AJKK01251007.1	DAAA02018021.1	AGVR01039101.1
AMBN exon 4	AKZM01002861.1	JREZ01000259.1	ATDM01089360.1	AAWR02005019.1	ATBW01040478.1	AJKK01251007.1	DAAA02018021.1	AGVR01039101.1
AMBN exon 5	AKZM01002861.1	JREZ01000259.1	ATDM01089360.1	AAWR02005019.1	ATBW01040478.1	AJKK01251007.1	DAAA02018021.1	AGVR01039101.1
AMBN exon 6	AKZM01002861.1	JREZ01000259.1	ATDM01089360.1	AAWR02005019.1	ATBW01040478.1	NA	DAAA02018021.1	AGVR01039101.1
AMBN exon 7	AKZM01002861.1	JREZ01000259.1	ATDM01089360.1,ATDM01115842.1	AAWR02005019.1	ATBW01040478.1	AJKK01251006.1	DAAA02018021.1	AGVR01039101.1
AMBN exons 8 and 9	AKZM01002861.1	JREZ01000259.1	ATDM01115842.1	AAWR02005019.1	ATBW01040478.1	AJKK01251006.1	DAAA02018021.1	AGVR01039101.1
AMBN exon 10	AKZM01002861.1	JREZ01000259.1	ATDM01115842.1	AAWR02005019.1	ATBW01040478.1	AJKK01251006.1	DAAA02018021.1	AGVR01039101.1
AMBN exon 11	AKZM01002861.1	JREZ01000259.1	ATDM01109949.1	AAWR02005019.1	ATBW01040478.1	AJKK01251006.1	DAAA02018021.1	AGVR01039101.1
ARR exon 4	AKZM01043989.1	JREZ01000987.1	ATDM01007565.1	AAWR02039713.1	ATBW01072054.1	AJKK01201197.1	DAAA02072850.1	AGVR01052417.1
ARR exon 6	AKZM01043989.1	JREZ01000987.1	ATDM01007564.1	AAWR02039713.1	ATBW01072054.1	AJKK01201197.1	DAAA02072850.1	AGVR01052417.1
ARR exon 7	AKZM01043989.1	JREZ01000987.1	ATDM01007564.1	AAWR02039713.1	ATBW01072054.1	AJKK01201197.1	DAAA02072850.1	AGVR01052417.1
ARR exon 9	AKZM01043989.1	JREZ01000987.1	ATDM01007564.1	AAWR02039713.1	ATBW01072054.1	AJKK01201197.1	DAAA02072850.1	AGVR01052417.1
ARR exon 10	AKZM01043989.1	JREZ01000987.1	ATDM01007564.1	AAWR02039713.1	ATBW01072054.1	AJKK01201197.1	DAAA02072850.1	AGVR01052417.1
ARR exon 11	AKZM01043989.1	JREZ01000987.1	ATDM01007564.1	AAWR02039713.1	ATBW01072054.1	AJKK01201197.1	DAAA02072850.1	AGVR01052417.1
ARR exon 12	AKZM01043989.1	JREZ01000987.1	ATDM01007564.1	AAWR02039713.1	ATBW01072054.1	AJKK01201197.1	DAAA02072850.1	AGVR01052417.1

ARR exon 13	AKZM01043989.1	JREZ01000987.1	ATDM01007564.1	AAWR02039713.1	ATBW01072054.1	AJKK01201197.1	DAAA02072850.1	AGVR01052417.1
CNGB3 exon 3	AKZM01047478.1	JREZ01000162.1	ATDM01036422.1	AAWR02029565.1	ATBW01068458.1	AJKK01137143.1	DAAA02039567.1	AGVR01052577.1
CNGB3 exon 4	AKZM01047482.1	JREZ01000162.1	ATDM01096649.1	AAWR02029567.1	ATBW01068459.1	AJKK01247529.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 5	AKZM01047482.1	JREZ01000162.1	ATDM01096649.1	AAWR02029567.1	ATBW01068459.1	AJKK01247529.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 6	AKZM01047482.1	JREZ01000162.1	ATDM01096649.1	AAWR02029567.1	ATBW01068459.1	AJKK01247529.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 7	AKZM01047482.1	JREZ01000162.1	ATDM01052526.1	AAWR02029567.1	ATBW01128993.1	AJKK01247529.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 8	AKZM01047482.1	JREZ01000162.1	ATDM01052526.1	AAWR02029567.1	ATBW01068459.1	AJKK01247529.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 9	AKZM01047483.1	JREZ01000162.1	ATDM01052526.1	AAWR02029567.1	ATBW01068459.1	AJKK01247530.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 10	AKZM01047483.1	JREZ01000162.1	ATDM01052526.1	AAWR02029567.1	ATBW01068459.1	AJKK01247530.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 11	AKZM01047483.1	JREZ01000162.1	ATDM01052526.1	AAWR02029567.1	ATBW01068459.1	AJKK01247530.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 12	AKZM01047484.1	JREZ01000162.1	ATDM01087137.1	AAWR02029567.1	ATBW01068459.1	AJKK01247531.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 13	AKZM01047484.1	JREZ01000162.1	ATDM01087138.1	AAWR02029567.1	ATBW01068957.1	AJKK01247531.1	DAAA02039573.1	AGVR01052575.1
CNGB3 exon 14	AKZM01047485.1	JREZ01000162.1	ATDM01072602.1	AAWR02029567.1	ATBW01068956.1	AJKK01247531.1	DAAA02039573.1	AGVR01052574.1
CNGB3 exon 15	AKZM01047485.1	JREZ01000162.1	ATDM01024214.1	AAWR02029567.1	ATBW01068956.1	AJKK01247532.1	DAAA02039573.1	AGVR01052574.1
CNGB3 exon 16	AKZM01047486.1	JREZ01000162.1	ATDM01024214.1	AAWR02029567.1	ATBW01068956.1	AJKK01247533.1	DAAA02039573.1	AGVR01052573.1
CNGB3 exon 17	AKZM01047486.1	JREZ01000162.1	ATDM01024214.1	AAWR02029567.1	ATBW01068956.1	AJKK01247533.1	DAAA02039573.1	AGVR01052573.1
CNGB3 exon 18	AKZM01047486.1	JREZ01000162.1	ATDM01024214.1	AAWR02029567.1	ATBW01068956.1	AJKK01247533.1	DAAA02039573.1	AGVR01052573.1
CSN2 exon 1	AKZM01002847.1	JREZ01000339.1	ATDM01011211.1	AAWR02005034.1	ATBW01062094.1	AJKK01251026.1	DAAA02018006.1	AGVR01039100.1
CSN2 exons 2 and 3	AKZM01002847.1	JREZ01000339.1	ATDM01011211.1	AAWR02005034.1	ATBW01062094.1	AJKK01251026.1	DAAA02018006.1	AGVR01039100.1
CSN2 exon 4	AKZM01002847.1	JREZ01000339.1	ATDM01011211.1	AAWR02005034.1	ATBW01062094.1	AJKK01251026.1	DAAA02018006.1	AGVR01039100.1
CSN2 exon 5	AKZM01002847.1	JREZ01000339.1	ATDM01011211.1	AAWR02005034.1	ATBW01062094.1	AJKK01251026.1	DAAA02018006.1	AGVR01039100.1
CYGB exon 1	AKZM01030605.1	JREZ01000017.1	ATDM01064460.1	AAWR02012335.1	ATBW01061788.1	NA	DAAA02049501.1	NA
CYGB exon 2	AKZM01030605.1	JREZ01000017.1	ATDM01064460.1	AAWR02012335.1	ATBW01061788.1	AJKK01147258.1	DAAA02049501.1	AGVR01010096.1
CYGB exon 3	AKZM01030605.1	JREZ01000017.1	ATDM01064460.1	AAWR02012335.1	ATBW01061788.1	AJKK01147258.1	DAAA02049501.1	AGVR01010096.1
CYGB exon 4	AKZM01030605.1	JREZ01000017.1	ATDM01064460.1	AAWR02012335.1	ATBW01061788.1	AJKK01147257.1	DAAA02049501.1	AGVR01010096.1
GNAT2 exon 1	AKZM01005790.1	JREZ01000096.1	ATDM01012623.1	AAWR02017108.1	ATBW01026953.1	AJKK01131148.1	DAAA02007573.1	AGVR01019414.1
GNAT2 exon 2	AKZM01005790.1	JREZ01000096.1	ATDM01012623.1	AAWR02017108.1	ATBW01026953.1	AJKK01131148.1	DAAA02007573.1	AGVR01019414.1
GNAT2 exon 3	AKZM01005790.1	JREZ01000096.1	ATDM01012623.1	AAWR02017108.1	ATBW01026953.1	AJKK01131148.1	DAAA02007573.1	AGVR01019414.1
GNAT2 exon 4	AKZM01005790.1	JREZ01000096.1	ATDM01012623.1	AAWR02017108.1	ATBW01026953.1	AJKK01131148.1	DAAA02007573.1	AGVR01019414.1
GNAT2 exon 5	AKZM01005790.1	JREZ01000096.1	ATDM01012623.1	AAWR02017108.1	ATBW01026953.1	AJKK01131148.1	DAAA02007573.1	AGVR01019414.1
GNAT2 exon 6	AKZM01005790.1	JREZ01000096.1	ATDM01012623.1	AAWR02017108.1	ATBW01026953.1	AJKK01131148.1	DAAA02007573.1	AGVR01019414.1
GNAT2 exon 7	AKZM01005790.1	JREZ01000096.1	ATDM01012623.1	AAWR02017108.1	ATBW01026953.1	AJKK01131148.1	DAAA02007573.1	AGVR01019414.1
GNAT2 exon 8	AKZM01005790.1	JREZ01000096.1	ATDM01012623.1	AAWR02017108.1	ATBW01026953.1	AJKK01131148.1	DAAA02007573.1	AGVR01019414.1
GNGT2 exon 4	AKZM01033488.1	JREZ01000009.1	ATDM01051771.1	AAWR02012793.1	ATBW01021376.1	AJKK01248228.1	DAAA02049054.1	AGVR01052845.1
GNGT2 exon 5	AKZM01033488.1	JREZ01000009.1	ATDM01051771.1	AAWR02012793.1	ATBW01021376.1	AJKK01248228.1	DAAA02049054.1	AGVR01052845.1
GRK7 exon 1	AKZM01003148.1	JREZ01000032.1	ATDM01077466.1	AAWR02009142.1	ATBW01003742.1	AJKK01216535.1	DAAA02002888.1	AGVR01043136.1
GRK7 exon 2	AKZM01003148.1	JREZ01000032.1	ATDM01077466.1	AAWR02009142.1	ATBW01003742.1	AJKK01216535.1	DAAA02002888.1	AGVR01043136.1
GRK7 exon 3	AKZM01003148.1	JREZ01000032.1	ATDM01048952.1	AAWR02009142.1	ATBW01003742.1	AJKK01216534.1	DAAA02002888.1	AGVR01043136.1
GRK7 exon 4	AKZM01003148.1	JREZ01000032.1	ATDM01048951.1	AAWR02009142.1	ATBW01003743.1	AJKK01216534.1	DAAA02002888.1	AGVR01043136.1
GUCA1B exon 1	AKZM01013614.1	JREZ01000079.1	ATDM01015580.1	AAWR02010145.1	ATBW01044977.1	AJKK01242116.1	DAAA02055068.1	AGVR01043408.1
GUCA1B exon 2	AKZM01013615.1	JREZ01000079.1	ATDM01015580.1	AAWR02010145.1	ATBW01044977.1	AJKK01242117.1	DAAA02055068.1	AGVR01043408.1
GUCA1B exon 3	AKZM01013615.1	JREZ01000079.1	ATDM01015580.1	AAWR02010145.1	ATBW01044977.1	AJKK01242117.1	DAAA02055068.1	AGVR01043408.1
GUCA1B exon 4	AKZM01013615.1	JREZ01000079.1	ATDM01015580.1	AAWR02010145.1	ATBW01044977.1	AJKK01242117.1	DAAA02055068.1	AGVR01043408.1
LEPR exon 1	AKZM01027889.1	JREZ01000336.1	ATDM01032082.1	AAWR02017655.1	ATBW01063244.1	AJKK01143100.1	DAAA02008515.1	AGVR01022567.1
LEPR exon 2	AKZM01027889.1	JREZ01000336.1	ATDM01032082.1	AAWR02017655.1	ATBW01063244.1	AJKK01143100.1	DAAA02008515.1	AGVR01022567.1
LEPR exon 3	AKZM01027889.1	JREZ01000336.1	ATDM01032082.1	AAWR02017655.1	ATBW01063244.1	AJKK01143100.1	DAAA02008515.1	AGVR01022567.1
LEPR exon 4	AKZM01027889.1	JREZ01000336.1	ATDM01032082.1	AAWR02017655.1	ATBW01063244.1	AJKK01001444.1	DAAA02008515.1	AGVR01022567.1
LEPR exon 5	AKZM01027889.1	JREZ01000336.1	ATDM01032082.1	AAWR02017655.1	ATBW01063244.1	AJKK01110747.1	DAAA02008515.1	AGVR01022567.1
LEPR exon 6	AKZM01027889.1	JREZ01000336.1	ATDM01032082.1	AAWR02017655.1	ATBW01063244.1	AJKK01143103.1	DAAA02008515.1	AGVR01022567.1
LEPR exon 7	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143104.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 8	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143104.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 9	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 10	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 11	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 12	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 13	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 14	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 15	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1

LEPR exon 16	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 17	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1
LEPR exon 18	AKZM01027889.1	JREZ01000336.1	ATDM01019114.1	AAWR02017655.1	ATBW01063243.1	AJKK01143105.1	DAAA02008515.1	AGVR01053198.1
NGB exon 1	AKZM01017413.1	JREZ01000179.1	NA	AAWR02019722.1	ATBW01004906.1	NA	DAAA02029719.1	AGVR01029765.1
NGB exon 2	AKZM01017413.1	JREZ01000179.1	ATDM01062162.1	AAWR02019722.1	ATBW01004905.1, ATBW01004906.1	AJKK01151737.1	DAAA02029719.1	AGVR01029765.1
NGB exon 3	AKZM01017413.1	JREZ01000179.1	ATDM01062162.1	AAWR02019722.1	ATBW01004904.1, ATBW01004905.1	AJKK01151737.1	DAAA02029719.1	AGVR01029765.1
NGB exon 4	AKZM01017413.1	JREZ01000179.1	ATDM01062162.1	AAWR02019722.1	ATBW01004904.1	AJKK01151737.1	DAAA02029719.1	AGVR01029765.1
OB exon 1	AKZM01001590.1	JREZ01000358.1	ATDM01020475.1	AAWR02014458.1	ATBW01001166.1	AJKK01171588.1	DAAA02011496.1	AGVR01030184.1
OB exon 2	AKZM01001590.1	JREZ01000358.1	ATDM01020475.1	AAWR02014458.1	ATBW01001166.1	AJKK01171588.1	DAAA02011496.1	AGVR01030184.1
OPN1LW exon 1	AKZM01052144.1	JREZ01000855.1	ATDM01074989.1	AAWR02040653.1	ATBW01076034.1	NA	DAAA02070091.1	AGVR01039360.1
OPN1LW exon 2	AKZM01052144.1	JREZ01000855.1	ATDM01074988.1	AAWR02040653.1	ATBW01076034.1	AJKK01261022.1	DAAA02070091.1	AGVR01039362.1
OPN1LW exon 3	AKZM01052144.1	JREZ01000855.1	ATDM01074988.1	AAWR02040653.1	ATBW01076034.1	AJKK01261022.1	DAAA02070091.1	AGVR01039362.1
OPN1LW exon 4	AKZM01052144.1	JREZ01000855.1	NA	AAWR02040653.1	ATBW01076034.1	AJKK01261022.1	DAAA02070091.1	AGVR01039362.1
OPN1LW exon 5	AKZM01052144.1	JREZ01000855.1	ATDM01088214.1	AAWR02040653.1	ATBW01076035.1	NA	DAAA02070091.1	AGVR01039362.1
OPN1LW exon 6	AKZM01052144.1	JREZ01000855.1	ATDM01088214.1	AAWR02040653.1	ATBW01076035.1	AJKK01261025.1	DAAA02070091.1	AGVR01039362.1
OPN4 exon 1	AKZM01030156.1	JREZ01000188.1	ATDM01120149.1	AAWR02000300.1	ATBW01019328.1	AJKK01151600.1	DAAA02062163.1	AGVR01054137.1
OPN4 exon 2	AKZM01030156.1	JREZ01000188.1	ATDM01120148.1	AAWR02000300.1	ATBW01019328.1	AJKK01151600.1	DAAA02062163.1	AGVR01054137.1
OPN4 exon 3	AKZM01030156.1	JREZ01000188.1	NA	AAWR02000300.1	ATBW01019328.1	AJKK01151600.1	DAAA02062163.1	AGVR01054137.1
OPN4 exon 4	AKZM01030156.1	JREZ01000151.1	ATDM01020136.1	AAWR02032513.1	ATBW01041208.1	AJKK01151600.1	DAAA02062163.1	AGVR01054137.1
OPN4 exon 5	AKZM01030156.1	JREZ01000188.1	ATDM01085013.1, ATDM01085014.1	AAWR02000300.1	ATBW01019328.1	AJKK01151600.1	DAAA02062163.1	AGVR01054137.1
OPN4 exon 6	AKZM01030156.1	JREZ01000188.1	ATDM01085014.1	AAWR02000300.1	ATBW01019328.1	AJKK01151600.1	DAAA02062163.1	AGVR01054137.1
OPN4 exon 7	AKZM01030156.1	JREZ01000188.1	ATDM01085014.1	AAWR02000300.1	ATBW01019328.1	AJKK01151600.1	DAAA02062163.1	AGVR01054137.1
OPN4 exon 8	AKZM01030156.1	JREZ01000188.1	ATDM01085014.1	AAWR02000300.1	ATBW01019328.1	AJKK01151600.1	DAAA02062163.1	AGVR01054137.1
OPN4 exon 9	AKZM01030156.1	JREZ01000188.1	ATDM01085014.1	AAWR02000300.1	ATBW01019327.1	AJKK01151600.1	DAAA02062163.1	AGVR01054137.1
OPN4 exon 10	AKZM01030156.1	JREZ01000188.1	ATDM01084069.1	AAWR02000300.1	ATBW01019327.1	AJKK01151601.1	DAAA02062163.1	AGVR01054137.1
OPSD exon 1	AKZM01050235.1	JREZ01000151.1	ATDM01020136.1	AAWR02032513.1	NA	AJKK01234352.1	DAAA02054571.1	AGVR01019666.1
OPSD exon 2	AKZM01050235.1	JREZ01000151.1	ATDM01020136.1	AAWR02032513.1	ATBW01041208.1	AJKK01234352.1	DAAA02054571.1	AGVR01019666.1
OPSD exon 3	AKZM01050235.1	JREZ01000151.1	ATDM01020136.1	AAWR02032513.1	ATBW01041208.1	AJKK01234352.1	DAAA02054571.1	AGVR01019667.1
OPSD exon 4	AKZM01050235.1	JREZ01000151.1	ATDM01020136.1	AAWR02032513.1	ATBW01041208.1	AJKK01234352.1	DAAA02054571.1	AGVR01019667.1
OPSD exon 5	AKZM01050235.1	JREZ01000151.1	ATDM01020136.1	NA	ATBW01041208.1	AJKK01234352.1	DAAA02054571.1	AGVR01019667.1, AGVR01019668.1
PDE6C exon 1	AKZM01020618.1	JREZ01000510.1	ATDM01015083.1	AAWR02007805.1	ATBW01021885.1	AJKK01184645.1	DAAA02058833.1	AGVR01027645.1
PDE6C exons 2 and 3	AKZM01020619.1	JREZ01000510.1	ATDM01015083.1	AAWR02007805.1	ATBW01021885.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 4	AKZM01020619.1	JREZ01000510.1	ATDM01015083.1	AAWR02007805.1	ATBW01021885.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 5	AKZM01020619.1	JREZ01000510.1	ATDM01015083.1	AAWR02007805.1	ATBW01021885.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exons 6 and 7	AKZM01020619.1	JREZ01000510.1	ATDM01015083.1	AAWR02007805.1	ATBW01021885.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 8	AKZM01020619.1	JREZ01000510.1	ATDM01015083.1	AAWR02007805.1	ATBW01021886.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 9	AKZM01020619.1	JREZ01000510.1	ATDM01015082.1	AAWR02007805.1	ATBW01021886.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 10	AKZM01020619.1	JREZ01000510.1	ATDM01015082.1	AAWR02007805.1	ATBW01021886.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 11	AKZM01020619.1	JREZ01000510.1	ATDM01015082.1	AAWR02007805.1	ATBW01021886.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 12	AKZM01020619.1	JREZ01000510.1	ATDM01015081.1	AAWR02007805.1	ATBW01021886.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 13	AKZM01020619.1	JREZ01000510.1	ATDM01015081.1	AAWR02007805.1	ATBW01021887.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 14	AKZM01020619.1	JREZ01000510.1	ATDM01015081.1	AAWR02007805.1	ATBW01021887.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 15	AKZM01020619.1	JREZ01000510.1	ATDM01015081.1	AAWR02007805.1	ATBW01021887.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 16	AKZM01020619.1	JREZ01000510.1	ATDM01015081.1	AAWR02007804.1	ATBW01021887.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exons 17 and 18	AKZM01020619.1	JREZ01000510.1	ATDM01015081.1	AAWR02007804.1	ATBW01021887.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exons 19 and 20	AKZM01020619.1	JREZ01000510.1	ATDM01015081.1	AAWR02007804.1	ATBW01021887.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 21	AKZM01020619.1	JREZ01000510.1	ATDM01015081.1	AAWR02007804.1	ATBW01021887.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6C exon 22	AKZM01020619.1	JREZ01000510.1	ATDM01015081.1	AAWR02007804.1	ATBW01021887.1	AJKK01184646.1	DAAA02058833.1	AGVR01027644.1
PDE6H exon 1	AKZM01035109.1	JREZ01000091.1	ATDM01039860.1	AAWR02029326.1	ATBW01098005.1	AJKK01204475.1	DAAA02014228.1	AGVR01047241.1
PDE6H exon 2	AKZM01035109.1	JREZ01000091.1	ATDM01039860.1	AAWR02029326.1	ATBW01099410.1	AJKK01204475.1	DAAA02014228.1	AGVR01047241.1
PDE6H exon 3	AKZM01035109.1	JREZ01000091.1	ATDM01039859.1	AAWR02029326.1	ATBW01116333.1	AJKK01204475.1	DAAA02014228.1	AGVR01047241.1
Prestin exon 1	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031606.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 2	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031606.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 3	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 4	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 5	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1

Prestin exon 6	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 7	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 8	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 9	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 10	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 11	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 12	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271734.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 13	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271733.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 14	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271733.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 15	AKZM01042683.1	JREZ01000083.1	ATDM01025897.1	AAWR02031605.1	ATBW01080502.1	AJKK01271733.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 16	AKZM01042683.1	JREZ01000083.1	ATDM01025898.1	AAWR02031605.1	ATBW01080502.1	AJKK01271733.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 17	AKZM01042683.1	JREZ01000083.1	ATDM01025898.1	AAWR02031605.1	ATBW01080502.1	AJKK01271733.1	DAAA02010508.1	AGVR01021840.1
Prestin exon 18	AKZM01042683.1	JREZ01000083.1	ATDM01025898.1	AAWR02031605.1	ATBW01080502.1	AJKK01271733.1	DAAA02010508.1	AGVR01021840.1
SWS1 exon 1	AKZM01001577.1	JREZ01000358.1	ATDM01011751.1	AAWR02014467.1	ATBW01001178.1	AJKK01171610.1	DAAA02011497.1	AGVR01030176.1
SWS1 exon 2	AKZM01001577.1	JREZ01000358.1	ATDM01011751.1	AAWR02014467.1	ATBW01001178.1	AJKK01171610.1	DAAA02011497.1	AGVR01030176.1
SWS1 exon 3	AKZM01001577.1	JREZ01000358.1	ATDM01011751.1	AAWR02014467.1	ATBW01001178.1	AJKK01171610.1	DAAA02011497.1	AGVR01030176.1
SWS1 exon 4	AKZM01001577.1	JREZ01000358.1	ATDM01011751.1	AAWR02014467.1	ATBW01001178.1	AJKK01171610.1	DAAA02011497.1	AGVR01030176.1
SWS1 exon 5	AKZM01001577.1	JREZ01000358.1	ATDM01011751.1	AAWR02014467.1	ATBW01001178.1	AJKK01171609.1, AJKK01171610.1	DAAA02011497.1	AGVR01030176.1
UCP1 exon 1	AKZM01017598.1	JREZ01000001.1	ATDM01029350.1	AAWR02018851.1	ATBW01036321.1, ATBW0.036322.1	AJKK01245297.1	DAAA02044420.1	AGVR01051296.1
UCP1 exon 2	AKZM01017598.1	JREZ01000001.1	ATDM01029350.1	AAWR02018851.1	ATBW0.036321.1	AJKK01245297.1	DAAA02044421.1	AGVR01051296.1
UCP1 exons 3 and 4	AKZM01017598.1	JREZ01000001.1	ATDM01029350.1	AAWR02018851.1	ATBW0.036321.1	NA	DAAA02044421.1	AGVR01051296.1
UCP1 exon 5	AKZM01017598.1	JREZ01000001.1	ATDM01029350.1	AAWR02018851.1	ATBW0.036321.1	NA	DAAA02044421.1	AGVR01051296.1
UCP1 exon 6	AKZM01017598.1	JREZ01000001.1	ATDM01029350.1	AAWR02018851.1	ATBW0.036320.1	AJKK01245296.1	DAAA02044421.1	AGVR01051296.1
UCP2 exon 1	AKZM01032604.1	JREZ01000192.1	ATDM01012085.1	AAWR02018167.1	ATBW01058700.1	AJKK01176960.1	DAAA02041041.1	AGVR01023127.1
UCP2 exon 2	AKZM01032604.1	JREZ01000192.1	ATDM01012085.1	AAWR02018167.1	ATBW01058700.1, ATBW01058701.1	AJKK01176960.1	DAAA02041041.1	AGVR01023127.1
UCP2 exons 3 and 4	AKZM01032604.1	JREZ01000192.1	ATDM01012085.1	AAWR02018167.1	ATBW01058701.1	AJKK01176960.1	DAAA02041041.1	AGVR01023127.1
UCP2 exon 5	AKZM01032604.1	JREZ01000192.1	ATDM01012085.1	AAWR02018167.1	ATBW01058701.1	AJKK01176960.1	DAAA02041041.1	AGVR01023127.1
UCP2 exon 6	AKZM01032604.1	JREZ01000192.1	ATDM01012085.1	AAWR02018167.1	ATBW01058701.1	AJKK01176960.1	DAAA02041041.1	AGVR01023127.1
UCP3 exon 1	AKZM01032604.1	JREZ01000192.1	ATDM01086957.1	AAWR02018167.1	ATBW01058699.1	AJKK01176964.1	DAAA02041042.1	AGVR01023127.1
UCP3 exon 2	AKZM01032604.1	JREZ01000192.1	ATDM01086957.1	AAWR02018167.1	ATBW01058699.1	AJKK01176964.1	DAAA02041042.1	AGVR01023127.1
UCP3 exon 3	AKZM01032604.1	JREZ01000192.1	ATDM01086957.1	AAWR02018167.1	ATBW01058699.1	AJKK01176964.1	DAAA02041042.1	AGVR01023127.1
UCP3 exon 4	AKZM01032604.1	JREZ01000192.1	ATDM01086957.1	AAWR02018167.1	ATBW01058699.1	AJKK01176964.1	DAAA02041042.1	AGVR01023127.1
UCP3 exon 5	AKZM01032604.1	JREZ01000192.1	ATDM01086957.1	AAWR02018167.1	ATBW01058699.1	AJKK01176964.1	DAAA02041042.1	AGVR01023127.1
UCP3 exon 6	AKZM01032604.1	JREZ01000192.1	ATDM01079948.1	AAWR02018167.1	ATBW01058699.1	AJKK01176963.1	DAAA02041042.1	AGVR01023127.1

Appendix 4. Ion Torrent chip types and DNA libraries sequenced in each run. On three occasions multiple DNA libraries were pooled together and sequenced simultaneously. Sequencing runs using 318 v2 BC chips were performed to re-sequence several libraries in an attempt to improve the coverage.

Run	Chip type	Libraries sequenced
1	314 v2 BC	BR
2	314 v2 BC	IR
3	314 v2 BC	SR2
4	314 v2 BC	MT4, MT5
5	314 v2 BC	WR2
6	314 v2 BC	WR3
7	314 v2 BC	WR4
8	314 v2 BC	JR41
9	314 v2 BC	JR43
10	318 v2 BC	WR2, WR3, WR4
11	318 v2 BC	IR, BR, JR41, JR43

Appendix 5. Presence and absence of the *UCP1* enhancer, putative regulatory region (PPR), and CpG island in 139 mammalian species. Xs = absent, / = inconclusive due to insufficient data, * = sixteen species with recently published genome projects since the Gaudry et al. 2017 publication. Accession numbers are also provided for contigs and SRA projects.

Species name	Enhancer	PPR	CPG island	Accession number
Montremata				
Ornithorhynchus anatinus	X	X	X	NW_001794248.1
Marsupialia				
Monodelphis domestica	X	X	X	AAFR03015618.1
Macropus eugenii	/	/	X	ABQO020217652.1
Sarcophilus harrisii	/	/	X	AEFK01228715.1
Xenarthra				
Choloepus hoffmanni	X	/	/	
Dasypus novemcinctus	X	X	Yes	AAGV03181320.1
Mylodon darwinii	Yes	/	/	SRX327588
Afrotheria				
Chrysochloris asiatica	Yes	Yes	X	AMDV01244955.1
Dugong dugon	Yes	Yes	/	This study
Echinops telfairi	Yes	Yes	Yes	AAIY02209271.1
Elephantulus edwardii	Yes	Yes	X	AMGZ01097263.1
Elephas maximus	Yes	Yes	Yes	SRX1015608; SRX1015606; SRX1015604; SRX1015603
Hydrodamalis gigas	Yes	/	/	This study
Loxodonta africana	Yes	Yes	Yes	AAGU03034821.1
Mammuthus primigenius	Yes	Yes	/	SRX1015727; SRX1015732; SRX1015743; SRX1015748; SRX001906; ERP008929
Orycteropus afer	Yes	Yes	X	ALYB01104541.1
Procavia capensis	X	Yes	Yes	ABRQ02143236.1
Trichechus manatus latirostris	Yes	Yes	Yes	AHIN01109623.1
Laurasiatheria				
Acinonyx jubatus	Yes	Yes	Yes	LLWD01000416.1
Ailuropoda melanoleuca	Yes	Yes	Yes	LNAT01000144.1
Balaena mysticetus	Yes	Yes	Yes	SRX790318, SRX790317, SRX790316, SRX790303, SRX790319
Balaenoptera acutorostrata	Yes	Yes	Yes	ATDI01065547.1
Balaenoptera bonaerensis	Yes	Yes	Yes	BAUQ01197845.1
Balaenoptera physalus	Yes	Yes	Yes	SRX1571086, SRX323050

Bison bison	Yes	Yes	Yes	JPYT01100523.1
Bos grunniens	Yes	Yes	Yes	AGSK01075302.1
Bos indicus	/	Yes	/	AGFL01142554.1
Bos taurus	Yes	Yes	Yes	DAAA02044420.1
Bubalus bubalis	Yes	Yes	Yes	AWWX01630119.1
Camelus dromedarius *	Yes	Yes	Yes	LSZX01012659.1
Camelus ferus	Yes	Yes	Yes	AGVR01051296.1; AGVR01051297.1
Canis lupus familiaris	Yes	X	/	AAEX03011713.1
Capra aegagrus	Yes	Yes	Yes	CBYH010071014.1
Capra hircus	Yes	Yes	/	AJPT01162992.1; AJPT01162993.1
Capreolus capreolus	Yes	Yes	Yes	CCMK010092645.1; CCMK010104759.1
Ceratotherium simum	Yes	Yes	Yes	AKZM01017598.1
Coelodonta antiquitatis	Yes	Yes	Yes	This study
Condylura cristata	Yes	X	X	AJFV01047153.1
Dicerorhinus sumatrensis	Yes	Yes	Yes	This study
Diceros bicornis	Yes	Yes	Yes	This study
Eidolon helvum	Yes	Yes	/	AWHC01286101.1; AWHC01029981.1
Eptesicus fuscus	Yes	X	Yes	ALEH01005956.1
Equus asinus	Yes	Yes	Yes	JREZ01000001.1
Equus caballus	Yes	Yes	/	AAWR02018850.1; AAWR02018851.1
Equus przewalskii	Yes	Yes	Yes	ATBW01036321.1; ATBW01036322.1
Erinaceus europaeus	/	/	X	
Felis catus	Yes	Yes	Yes	AANG02062919.1
Giraffa camelopardalis *	Yes	Yes	X	LVKQ01071482.1
Hipposideros armiger *	Yes	Yes	Yes	NW_017731683.1
Leptonychotes weddellii	Yes	Yes	Yes	APMU01115165.1; APMU01141180.1
Lipotes vexillifer	Yes	Yes	Yes	AUPI01000024.1
Lycaon pictus *	Yes	X	/	LPRB01000019.1
Manis javanica *	X	X	X	NW_016530114.1
Manis pentadactyla	X	X	X	JPTV01131901.1
Megaderma lyra	Yes	/	/	AWHB01167753.1; AWHB01348443.1; AWHB01348444.1
Miniopterus natalensis *	Yes	X	Yes	NW_015504404.1
Mustela putorius furo	Yes	Yes	Yes	AGTQ01041845.1
Myotis brandtii	Yes	X	Yes	ANKR01273867.1; ANKR01273868.1
Myotis davidii	Yes	X	Yes	ALWT01125743.1
Myotis lucifugus	Yes	X	Yes	AAPE02001462.1
Odobenus rosmarus	Yes	Yes	Yes	ANOP01028105.1
Okapia johnstoni *	Yes	Yes	Yes	LVCL010093660.1; LVCL010093662.1
Orcinus orca	X	X	X	ANOL02004931.1
Ovis aries	Yes	Yes	Yes	AMGL01037664.1; JN604985.1
Panthera pardus *	Yes	Yes	Yes	NW_017619848.1
Panthera tigris altaica	Yes	Yes	Yes	ATCQ01112915.1

Panthera uncia	Yes	Yes	/	SRX273036
Pantholops hodgsonii	Yes	Yes	Yes	AGTT01188813.1
Physeter macrocephalus	X	Yes	Yes	AWZP01062081.1
Pteropus alecto	Yes	Yes	Yes	ALWS01011689.1
Pteropus vampyrus	Yes	Yes	Yes	ABRP02126915.1
Rhinoceros unicornis	Yes	Yes	Yes	This study
Rhinolophus ferrumequinum	Yes	Yes	Yes	AWHA01040305.1
Rhinolophus sinicus *	Yes	Yes	Yes	NW_017738992.1
Rousettus aegyptiacus *	Yes	Yes	Yes	NW_015494583.1
Sorex araneus	Yes	X	Yes	AALT02056093.1
Sus cebrifrons	Yes	/	/	ERX953604-ERX953626; ERX149172
Sus scrofa	Yes	X	X	LUXQ01106311.1
Sus verrucosus	Yes	/	/	ERX1054048-ERX1054067; ERX149174
Tapirus indicus	Yes	/	Yes	This study
Tursiops truncatus	X	X	X	ABRN02199412.1
Ursus maritimus	Yes	Yes	/	AVOR01014285.1; AVOR01014286.1
Vicugna pacos	Yes	Yes	Yes	ABRR02134987.1; ABRR02134989.1
Euarchontoglires				
Aotus nancymaae	Yes	Yes	Yes	JYKP01215429.1
Apodemus sylvaticus	Yes	/	X	LIPJ01452544.1; LIPJ01184746.1; LIPJ01447868.1; LIPJ01014497.1
Callithrix jacchus	Yes	Yes	Yes	ACFV01002817.1
Cavia aperea	/	/	/	AVPZ01000778.1
Cavia porcellus	Yes	Yes	Yes	AAKN02011801.1
Cebus capuchinis *	Yes	Yes	Yes	NW_016107319.1
Cercocebus atys	Yes	Yes	Yes	JZLG01060688.1
Chinchilla lanigera	Yes	Yes	Yes	AGCD01027651.1
Chlorocebus sabaeus	Yes	Yes	Yes	AQIB01017419.1
Colobus angolensis	Yes	Yes	Yes	JYKR01122839.1
Cricetulus griseus	Yes	Yes	X	AFTD01128393.1; AFTD01128394.1
Daubentonia madagascariensis	Yes	Yes	/	AGTM011584638.1; AGTM011584996.1; AGTM011708528.1; AGTM012010142.1; AGTM011594144.1
Dipodomys ordii	Yes	X	Yes	ABRO02057411.1
Ellobius lutescens *	Yes	X	Yes	LOEQ01000193.1
Ellobius talpinus *	Yes	X	Yes	LOJH01032235.1
Eulemur flavifrons	Yes	Yes	Yes	LGHW01000184.1
Eulemur macaco	Yes	Yes	Yes	LGHX01000184.1
Fukomys damarensis	Yes	Yes	Yes	AYUG01151056.1
Galeopterus variegatus	Yes	Yes	/	JMZW01045215.1; JMZW01045216.1
Gorilla gorilla gorilla	Yes	Yes	Yes	NW_004002547.1
Heterocephalus glaber	Yes	Yes	Yes	AFSB01162372.1; AFSB01162373.1

Homo sapiens	Yes	Yes	Yes	NG_012139.1
Jaculus jaculus	Yes	Yes	Yes	AKZC01091543.1
Macaca fascicularis	Yes	Yes	Yes	CAEC01514737.1
Macaca mulatta	Yes	Yes	Yes	AANU01271750.1
Macaca nemestrina	Yes	Yes	Yes	JZLF01028562.1
Mandrillus leucophaeus	Yes	Yes	Yes	JYKQ01107154.1; JYKQ01107155.1
Marmota marmota	Yes	Yes	Yes	CZRN01000015.1
Mesocricetus auratus	Yes	X	Yes	APMT01116524.1; NM_001281332.1
Microcebus murinus	Yes	Yes	Yes	ABDC01082367.1
Microtus agrestis	Yes	X	Yes	LIQJ01004042.1
Microtus ochrogaster	Yes	X	X	AHZW01157105.1; AHZW01157106.1
Mus musculus	Yes	X	X	CAAA01024310.1
Mus spretus *	Yes	X	X	LVXV01001867.1
Myodes glareolus	Yes	/	/	LIPI01003929.1
Nannospalax galili	Yes	Yes	X	AXCS01128925.1
Nasalis larvatus	Yes	Yes	Yes	JMHX01319533.1
Neotoma lepida *	/	/	X	LZPO01075894.1
Nomascus leucogenys	Yes	Yes	Yes	ADFV01177960.1
Ochotona princeps	Yes	X	X	ALIT01060999.1
Octodon degus	Yes	Yes	Yes	AJSA01193669.1; AJSA01193670.1; AJSA01193671.1
Oryctolagus cuniculus	Yes	Yes	Yes	AAGW02045633.1
Otolemur garnettii	Yes	Yes	X	AAQR03074138.1
Pan paniscus	Yes	Yes	Yes	AJFE01070904.1
Pan troglodytes	Yes	Yes	/	AACZ03032212.1; AACZ03032213.1
Papio anubis	Yes	Yes	Yes	AHZZ01043343.1
Peromyscus maniculatus	Yes	X	X	AYHN01134223.1
Pongo abelii	Yes	Yes	Yes	ABGA01062109.1
Propithecus coquereli	Yes	Yes	Yes	JZKE01017273.1
Rattus norvegicus	Yes	X	X	AAHX01097782.1
Rhinopithecus bieti *	Yes	Yes	Yes	NW_016805762.1
Rhinopithecus roxellana	Yes	Yes	Yes	JABR01098768.1
Saimiri boliviensis	Yes	Yes	Yes	AGCE01051213.1
Spermophilus tridecemlineatus	Yes	Yes	Yes	AGTP01049378.1
Tarsius syrichta	Yes	Yes	/	ABRT02355486.1
Tupaia belangeri chinensis	Yes	Yes	Yes	ALAR01031045.1

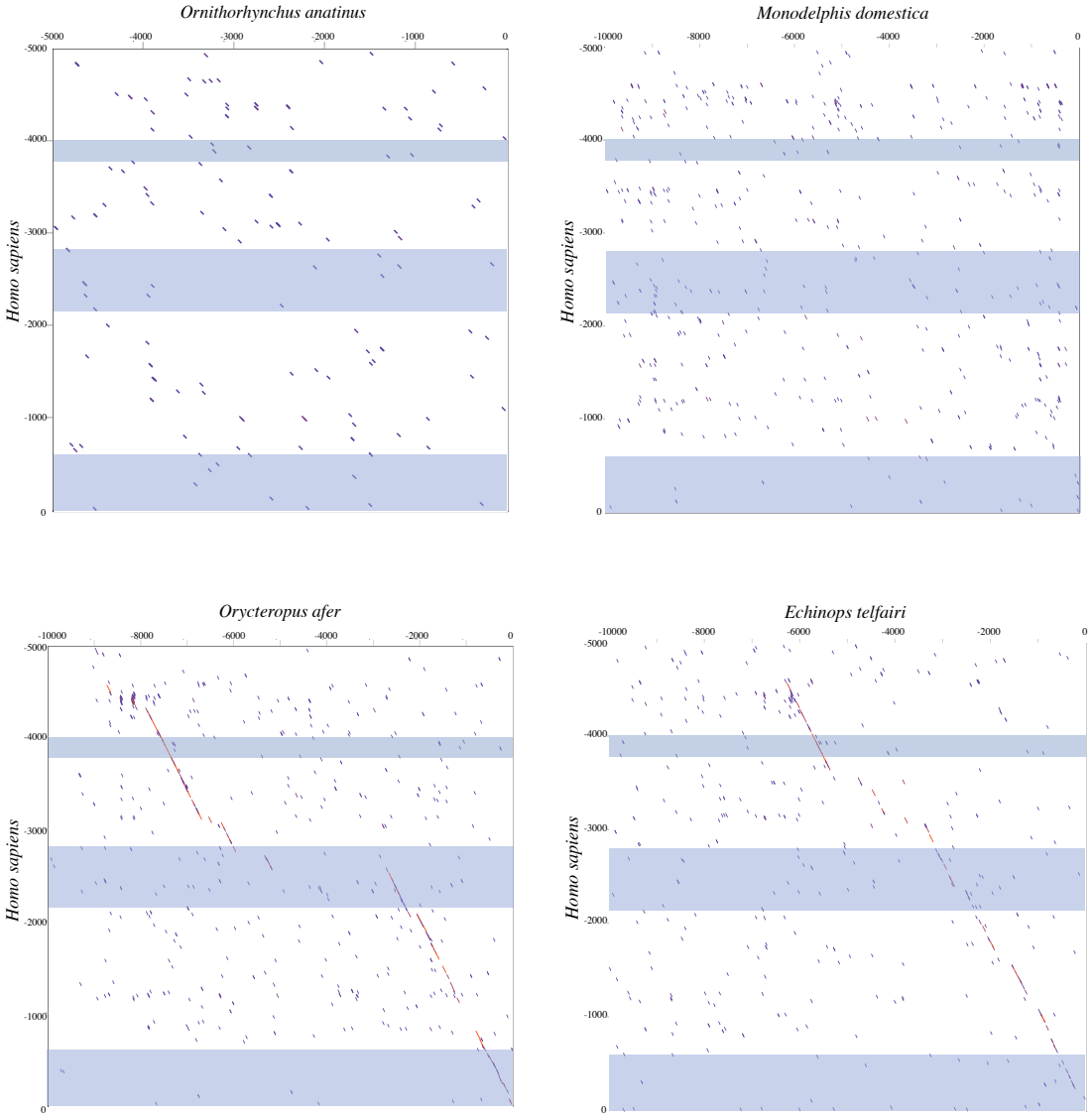
Appendix 6. Possible transcription factor binding motifs within the PRR of selected species screened using rVista 2.0. Duplicates sites were removed. Position is indicated relative to the start of the PRR sequence and the strand is indicated with + or - symbols.

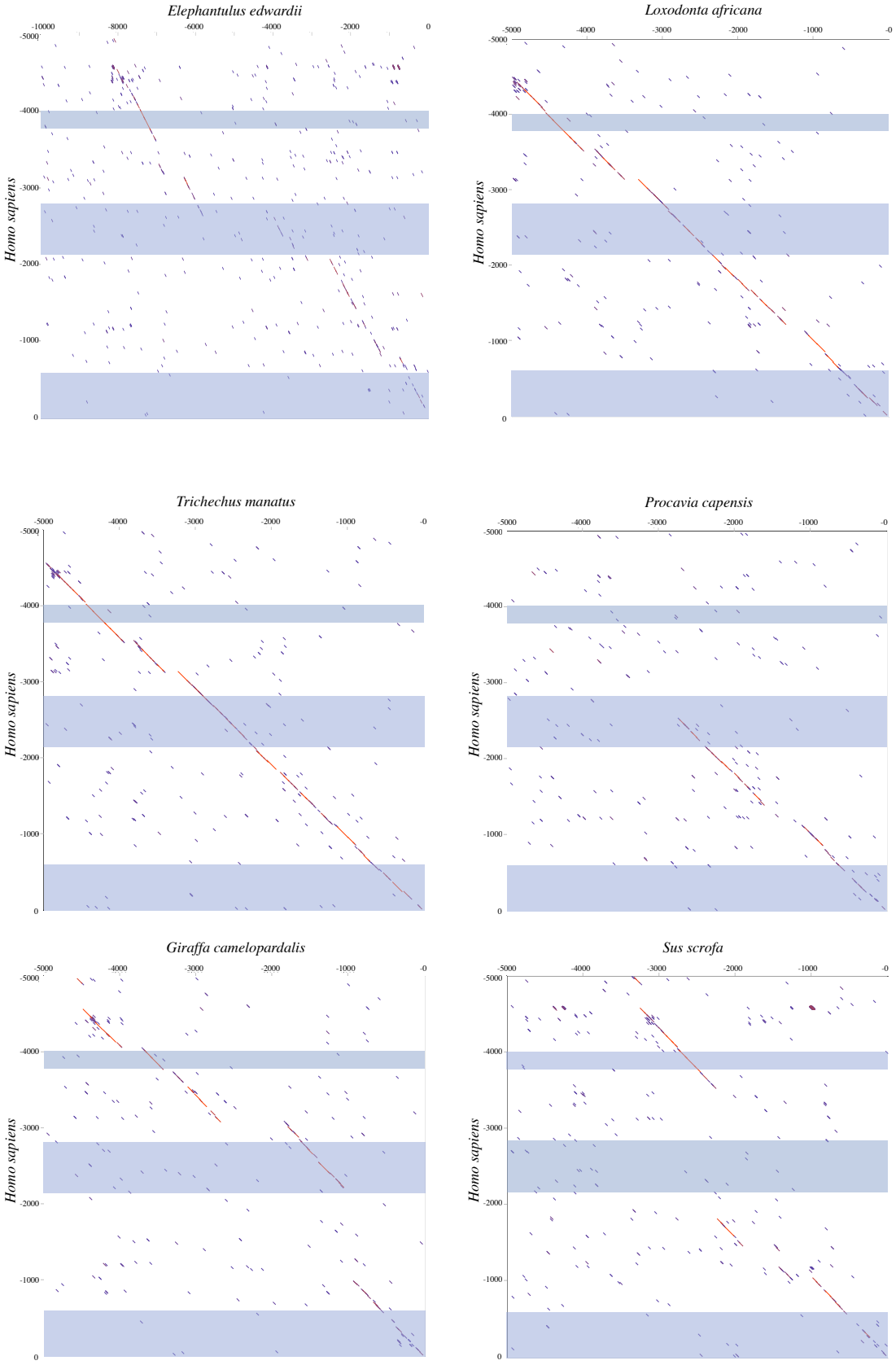
Species	Motif	Position	Sequence
Homo sapiens	CREB	24 +	catggCATCAgttc
	DR3	227 -	cagaGGTTCACTAGAGTCaac
	DR4	230 -	agGTTCACTAGAGTCAa
Marmota marmota	PPAR_DR1	50 -	tGGTCAAAGGACt
	DR4	326 -	tgGGTCCCTTAAGGTCa
	DR1	393 -	TGACACTTATCCc
Oryctolagus cuniculus	CREB	373 -	ccTAACATCAcc
	CEBP	519 -	gcTCCATTGCCTAACTCt
	PPAR_DR1	592 +	tGGCCCTTGGCCc
	PPAR_DR1	601 +	gCCCCTTTGTCCc
Camelus ferus	CEBP	271 -	taTACATTTGGGCATACt
	CEBP	503 -	tgTTCCTTTCCTAATTGt
	CREB	636 -	tgtCATCAcct
Bos taurus	CREB	149 +	CGTCAg
	CEBP	240 -	taTGCATTATAACAAACa
	CEBP	471 -	tgTTTCTTTCCTAATTTg
	PPAR_DR1	487 +	tGACCTTTGATAa
	PPAR_DR1	542 +	tGACCCTTGACCc
Giraffa camelopardalis	CREB	150 +	CGTCAg
	CREB	476 -	tgTTTCTTTCCTAATTTg
	PPAR_DR1	492 +	tGACCTTTGATAa
	PPAR_DR1	547 +	tGACCCTTGACCc
Balaenoptera acutorostrata	DR1	96 +	aGGGGAAGGGACA
	CEBP	518 -	taTTTCTTTCCTAACTTt
	PPAR_DR1	587 +	tGGCCCTTGACCc
	DR1	587 -	TGGCCCTTGACCc
	DR1	594 -	TGACCCCTTTCCc
Lipotes vexillifer	DR3	291 +	accGAACATTCTCAATCtgct
	CEBP	509 -	taTTTCTTTCCTAACTTt
	PPAR_DR1	580 +	tGGCCCTTGACCc
	DR1	587 -	TGACCCCTTTCCc

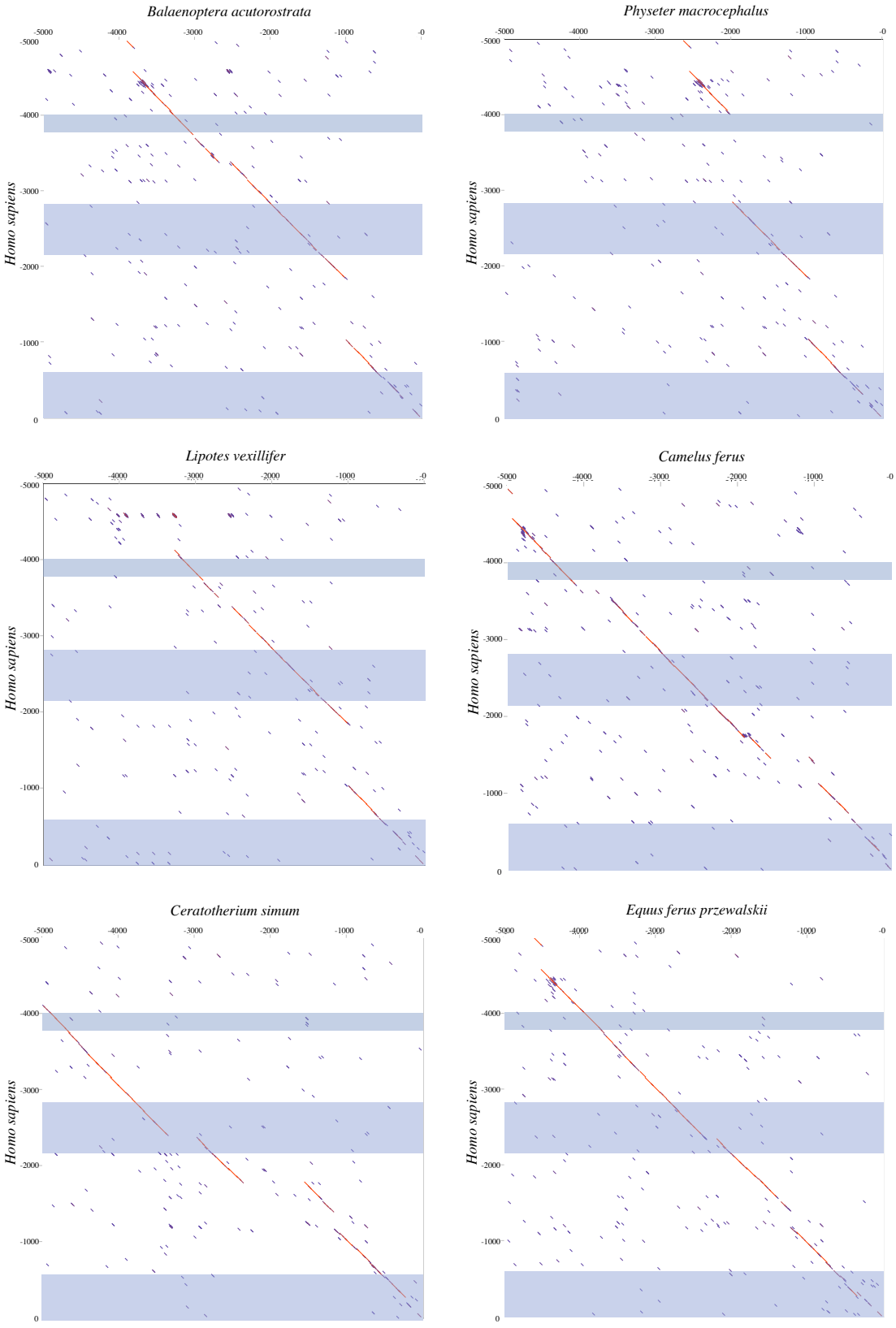
Ceratotherium simum	DR1	108 +	aGGGGAAGGGACA
	DR4	246 -	agGATCACTAGAGTTAg
	CEBP	284 -	taTACATTTAGTCATACt
	DR3	304 +	accGAACATTCTCAATCtctg
	DR4	425 +	tGTCCTCTTTTGACAtt
	PPAR_DR1	453 +	tCACACTTGACCc
Equus przewalskii	CEBP	9 +	cTTTCACAAtcc
	CREB	36 -	caTAGCGTCAgt
	CREB	41 +	CGTCAg
	DR4	234 -	agGTTCACTAGAGTTAg
	PPAR_DR1	537 +	tTACCTTTGACCa
	DR1	592 -	TGGTCCTTGACCc
	CREB	667 +	ttGCTGACTccc
Equus caballus	DR4	224 -	agGTTCACTAGAGTTAg
	PPAR_DR1	524 +	tTACCTTTGACCa
	DR1	579 -	TGGTCCTTGACCc
	CREB	654 +	ttGCTGACTccc
Pteroupus vampyrus	CREB	37 +	catagCATCAgctc
	DR4	408 +	tGTCCTCTTTTGACAtt
	PPAR_DR1	575 +	tGGCCCTTGACCc
	DR1	582 -	TGACCCCTTTCCt
Ailuropoda melanoleuca	DR1	85 +	aGGGGAAGGGACA
	CREB	505 +	ttGATGAGGccc
	DR1	554 -	TGGCCCATGACCc
	PPAR_DR1	561 +	tGACCCTTTGCCt
	CREB	628 +	ttGCTGACTccc
Odobenus rosmarus	DR1	92 +	aGGGGAAGGGACA
	DR4	406 +	tGTCCTCTTTTGACAtt
	DR1	567 -	TGGCCCATGACCc
	PPAR_DR1	574 +	tGACCCTTTTCCt
	CREB	670 +	ttGCTGACTccc
Panthera pardus	DR4	240 +	tGTCCTCTTTTGACAca
Leptonychotes weddellii	DR1	90 +	aGGGGAAGGGACA
	DR4	403 +	tGTCCTCTTTTGACAtt
	DR1	564 -	TGGCCCATGACCc
	PPAR_DR1	571 +	tGACCCTTTTCCt

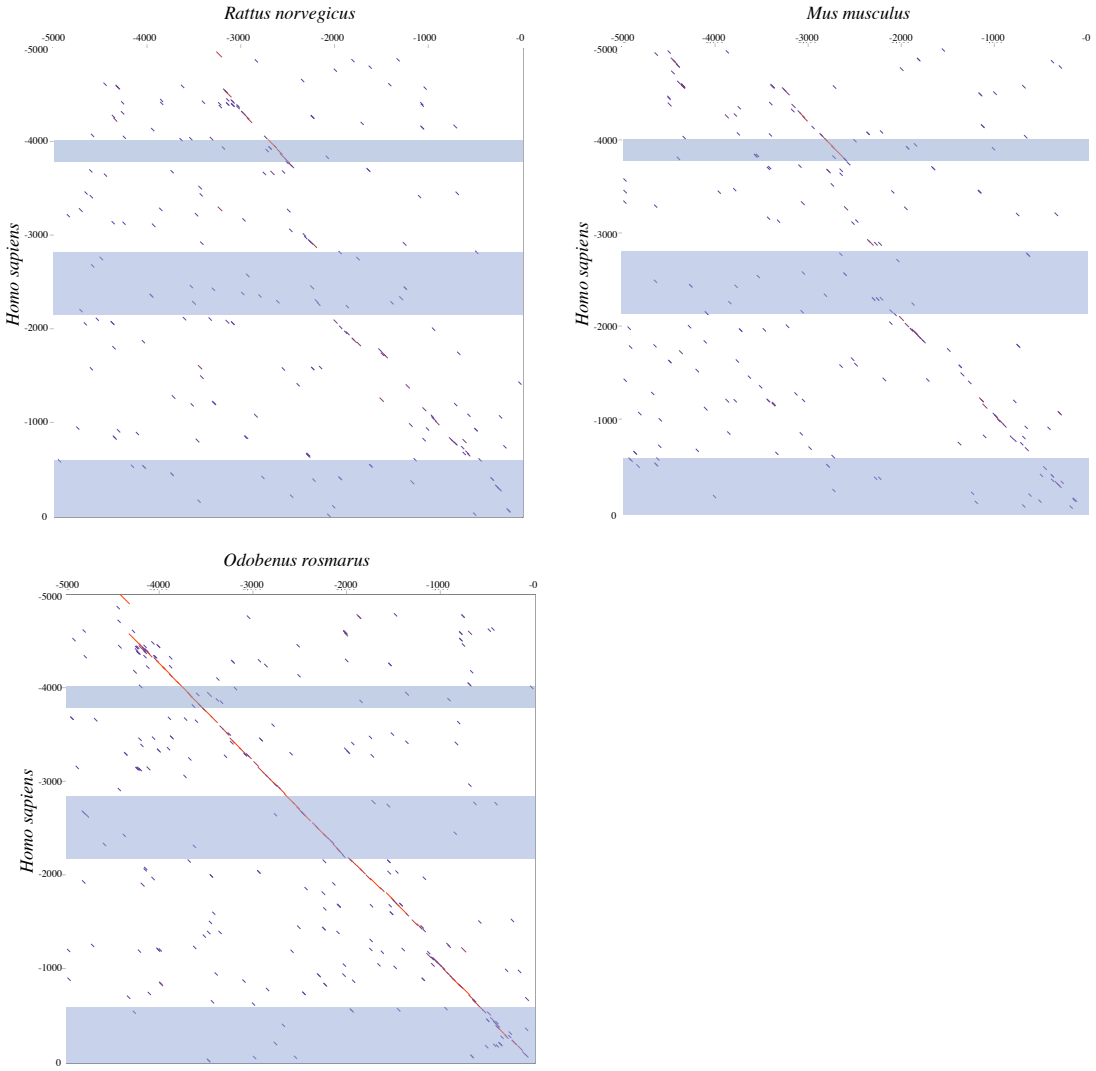
	CREB	671 +	ttGCTGACTccc
Procavia capensis	CREB	59 -	ccTAACATCAcc
	DR1	273 -	TGGTCCTTGACCt
	CREB	278 +	cttgaCCTCAttgc
	CREB	280 +	TGACCTca
Loxodonta africana	CREB	32 +	acataCATCAgctc
	CREB	347 -	caTAACATCAcc
	CREB	424 -	tTGACG
	PPAR_DR1	566 +	tGGCCCTTGACCc
Trichechus manatus	CREB	140 -	tgAGGTCA
	CREB	369 -	taaCATCACCaa
	PPAR_DR1	587 +	tGGCCCTTGACCc
Echinops telfairi	PPAR_DR1	189 -	gGGTCAAGGATCa
	CREB	326 -	ccTGACATCAct

Appendix 7. Dot plots of the 5,000 or 10,000 bp upstream of *UCP1* exon 1 of select mammalian species compared to the upstream sequence of humans. Blue shading represents the *UCP1* enhancer (~-4000 to -3800 in human), putative regulatory region (~-2700 to -2500 in human), and promoter/CpG island (-600 to 0 in human), in that order, from top to bottom.

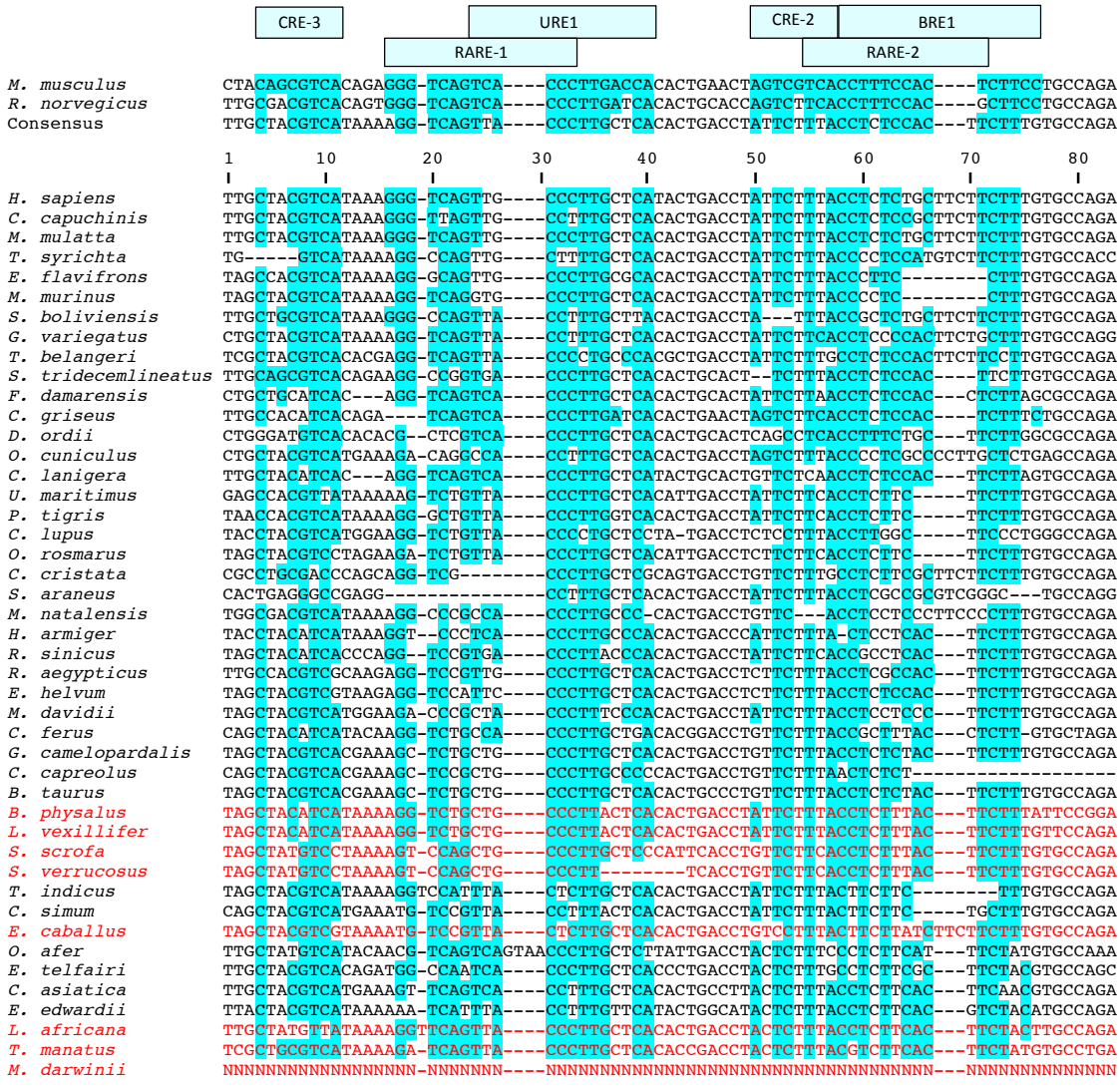


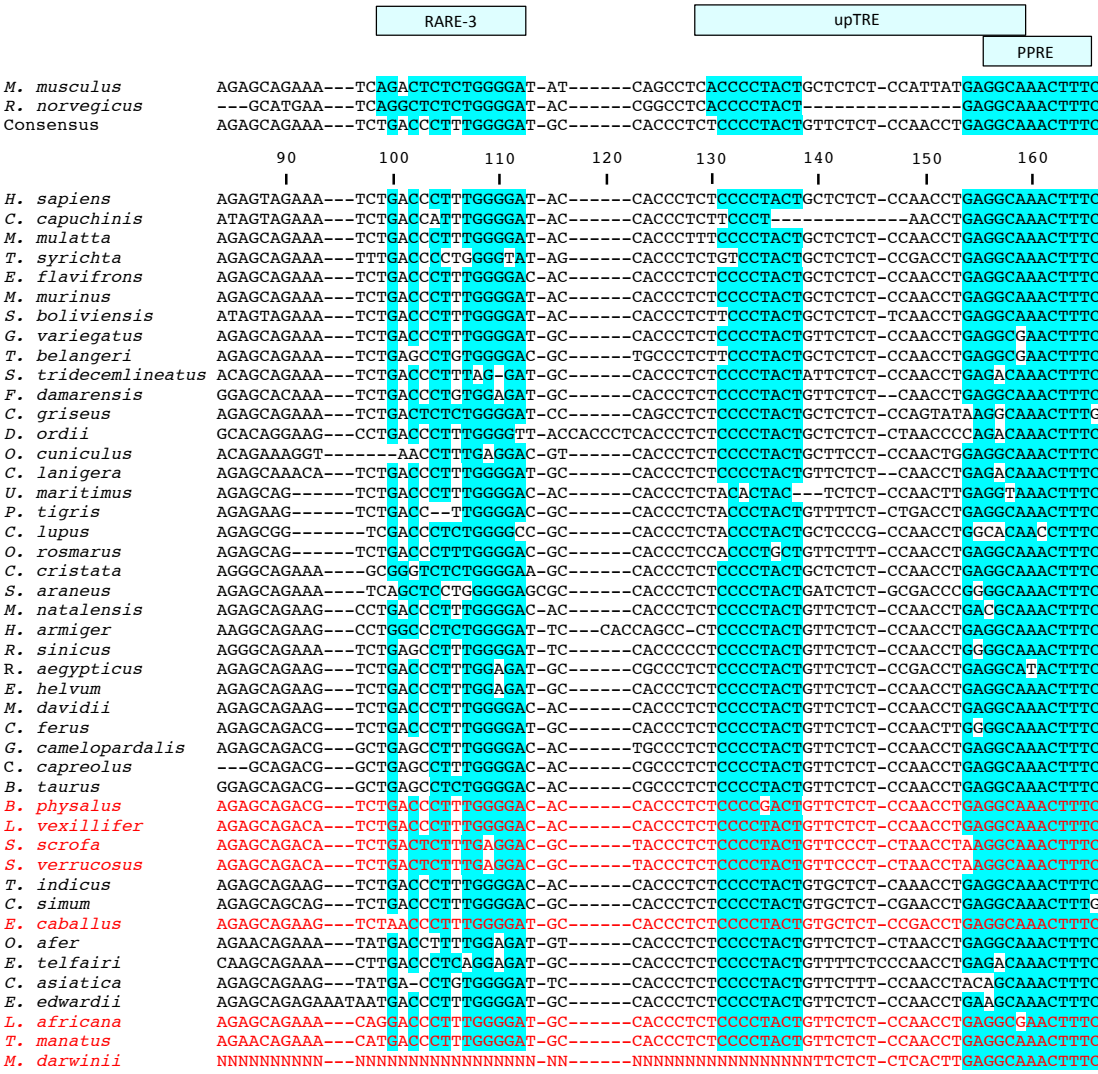

Elephantulus edwardii
Loxodonta africana
Trichechus manatus
Procavia capensis
Giraffa camelopardalis
Sus scrofa
Homo sapiens

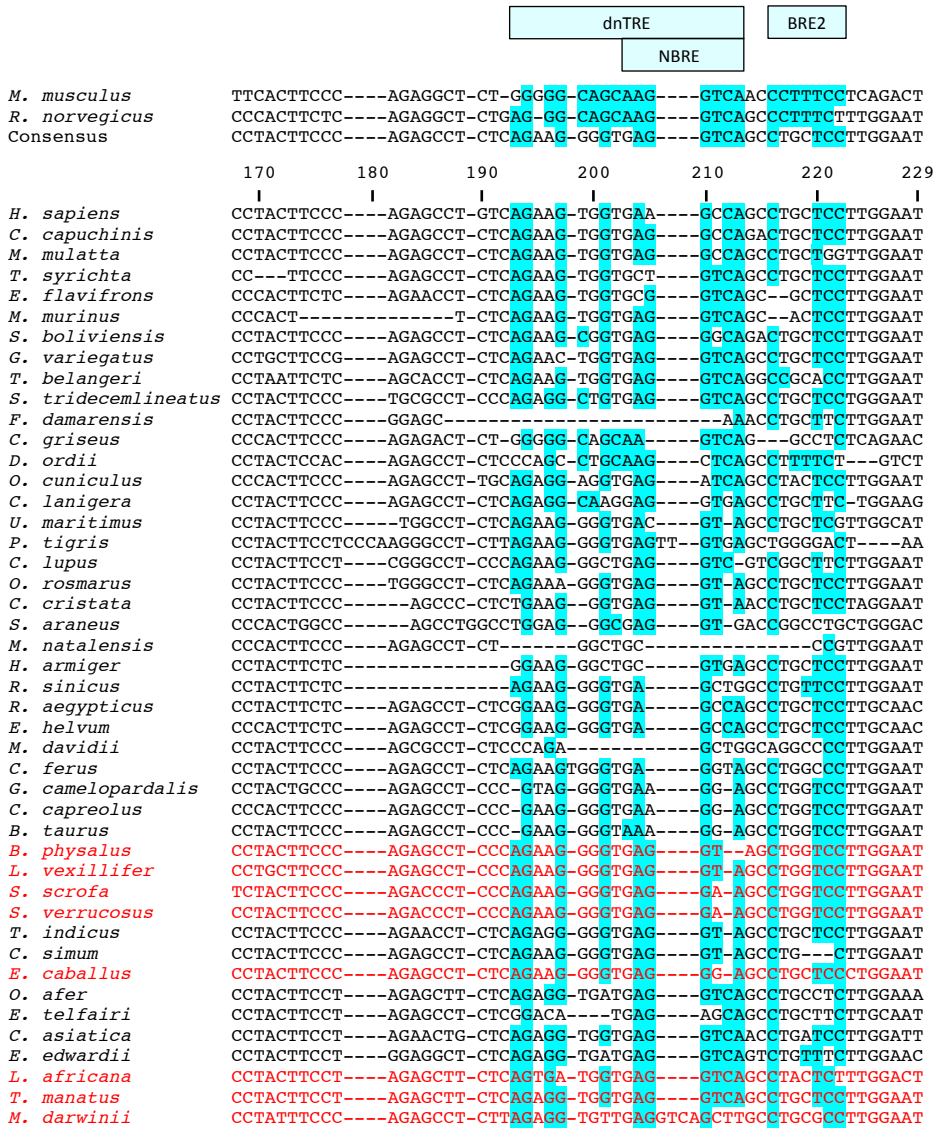

Balaenoptera acutorostrata
Physeter macrocephalus
Lipotes vexillifer
Camelus ferus
Ceratotherium simum
Equus ferus przewalskii
Homo sapiens
-5000
-4000
-3000
-2000
-1000
-0
0


Rattus norvegicus
Mus musculus
Odobenus rosmarus
Homo sapiens
-5000
-4000
-3000
-2000
-1000
-0
0

Appendix 8. *UCP1* enhancer alignment for select eutherian species. Sequences highlighted in blue denote the degree of conservation relative to transcription factor binding sites first described in mice or rats (see also Figure 3.2). The consensus sequence represents the simple majority based on species for which the *UCP1* gene is intact. Species with documented *UCP1* pseudogenes (Gaudry et al. 2017) are denoted in red font and were not included in the consensus calculations.




RARE-3
upTRE
PPRE
M. musculus AGAGCAGAAA---TCAGACTCTCTGGGGAT-AT------CAGCCTCACCCCTACTGCTCTCT-CCATTATGAGGCAAACTTTCT
R. norvegicus ---GCATGAA---TCAGGCTCTCTGGGGAT-AC------CGGCCTCACCCCTACT----------GAGGCAAACTTTCT
Consensus AGAGCAGAAA---TCTGACCCTTTGGGGAT-GC------CACCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCAAACTTTCT
90 100 110 120 130 140 150 160
H. sapiens AGAGTAGAAA---TCTGACCCTTTGGGGAT-AC------CACCCTCTCCCCTACTGCTCTCT-CCAACCTGAGGCAAACTTTCT
C. capuchinis ATAGTAGAAA---TCTGACCATTTGGGGAT-AC------CACCCTCTTCCCT----------AACCTGAGGCAAACTTTCT
M. mulatta AGAGCAGAAA---TCTGACCCTTTGGGGAT-AC------CACCCTTTCCCCTACTGCTCTCT-CCAACCTGAGGCAAACTTTCT
T. syrichta AGAGCAGAAA---TTTGACCCCTGGGGTAT-AG------CACCCTCTGTCCTACTGCTCTCT-CCGACCTGAGGCAAACTTTCT
E. flavifrons AGAGCAGAAA---TCTGACCCTTTGGGGAC-AC------CACCCTCTCCCCTACTGCTCTCT-CCAACCTGAGGCAAACTTTCC
M. murinus AGAGCAGAAA---TCTGACCCTTTGGGGAT-AC------CACCCTCTCCCCTACTGCTCTCT-CCAACCTGAGGCAAACTTTCC
S. boliviensis ATAGTAGAAA---TCTGACCCTTTGGGGAT-AC------CACCCTCTTCCCTACTGCTCTCT-TCAACCTGAGGCAAACTTTCT
G. variegatus AGAGCAGAAA---TCTGACCCTTTGGGGAT-GC------CACCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCGAACTTTCC
T. belangeri AGAGCAGAAA---TCTGAGCCTGTGGGGAC-GC------TGCCCTCTTCCCTACTGCTCTCT-CCAACCTGAGGCGAACTTTCT
S. tridecemlineatus ACAGCAGAAA---TCTGACCCTTTAG-GAT-GC------CACCCTCTCCCCTACTATTCTCT-CCAACCTGAGACAAACTTTCT
F. damarensis GGAGCACAAA---TCTGACCCTGTGGAGAT-GC------CACCCTCTCCCCTACTGTTCTCT--CAACCTGAGGCAAACTTTCT
C. griseus AGAGCAGAAA---TCTGACTCTCTGGGGAT-CC------CAGCCTCTCCCCTACTGCTCTCT-CCAGTATAAGGCAAACTTTGT
D. ordii GCACAGGAAG---CCTGACCCTTTGGGGTT-ACCACCCTCACCCTCTCCCCTACTGCTCTCT-CTAACCCCAGACAAACTTTCT
O. cuniculus ACAGAAAGGT-------AACCTTTGAGGAC-GT------CACCCTCTCCCCTACTGCTTCCT-CCAACTGGAGGCAAACTTTCT
C. lanigera AGAGCAAACA---TCTGACCCTTTGGGGAT-GC------CACCCTCTCCCCTACTGTTCTCT--CAACCTGAGACAAACTTTCT
U. maritimus AGAGCAG------TCTGACCCTTTGGGGAC-AC------CACCCTCTACACTAC---TCTCT-CCAACTTGAGGTAAACTTTCT
P. tigris AGAGAAG------TCTGACC--TTGGGGAC-GC------CACCCTCTACCCTACTGTTTTCT-CTGACCTGAGGCAAACTTTCT
C. lupus AGAGCGG-------TCGACCCTCTGGGGCC-GC------CACCCTCTACCCTACTGCTCCCG-CCAACCTGGCACAACCTTTCT
O. rosmarus AGAGCAG------TCTGACCCTTTGGGGAC-GC------CACCCTCCACCCTGCTGTTCTTT-CCAACCTGAGGCAAACTTTCT
C. cristata AGGGCAGAAA----GCGGGTCTCTGGGGAA-GC------CACCCTCTCCCCTACTGCTCTCT-CCAACCTGAGGCAAACTTTCC
S. araneus AGAGCAGAAA----TCAGCTCCTGGGGGAGCGC------CACCCTCTCCCCTACTGATCTCT-GCGACCCGGGGCAAACTTTCT
M. natalensis AGAGCAGAAG---CCTGACCCTTTGGGGAC-AC------CACCCTCTCCCCTACTGTTCTCT-CCAACCTGACGCAAACTTTCT
H. armiger AAGGCAGAAG---CCTGGCCCTCTGGGGAT-TC---CACCAGCC-CTCCCCTACTGTTCTCT-CCAACCTGAGGCAAACTTTCT
R. sinicus AGGGCAGAAA---TCTGAGCCTTTGGGGAT-TC------CACCCCCTCCCCTACTGTTCTCT-CCAACCTGGGGCAAACTTTCT
R. aegypticus AGAGCAGAAG---TCTGACCCTTTGGAGAT-GC------CGCCCTCTCCCCTACTGTTCTCT-CCGACCTGAGGCATACTTTCT
E. helvum AGAGCAGAAG---TCTGACCCTTTGGAGAT-GC------CACCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCAAACTTTCT
M. davidii AGAGCAGAAG---TCTGACCCTTTGGGGAC-AC------CACCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCAAACTTTCG
C. ferus AGAGCAGACG---TCTGACCCTTTGGGGAT-GC------CACCCTCTCCCCTACTGTTCTCT-CCAACTTGGGGCAAACTTTCT
G. camelopardalis AGAGCAGACG---GCTGAGCCTTTGGGGAC-AC------TGCCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCAAACTTTCT
C. capreolus ---GCAGACG---GCTGAGCCTTTGGGGAC-AC------CGCCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCAAACTTTCT
B. taurus GGAGCAGACG---GCTGAGCCTCTGGGGAC-AC------CGCCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCAAACTTTCC
B. physalus AGAGCAGACG---TCTGACCCTTTGGGGAC-AC------CACCCTCTCCCCGACTGTTCTCT-CCAACCTGAGGCAAACTTTCT
L. vexillifer AGAGCAGACA---TCTGACCCTTTGGGGAC-AC------CACCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCAAACTTTCT
S. scrofa AGAGCAGACA---TCTGACTCTTTGAGGAC-GC------TACCCTCTCCCCTACTGTTCCCT-CTAACCTAAGGCAAACTTTCT
S. verrucosus AGAGCAGACA---TCTGACTCTTTGAGGAC-GC------TACCCTCTCCCCTACTGTTCCCT-CTAACCTAAGGCAAACTTTCT
T. indicus AGAGCAGAAG---TCTGACCCTTTGGGGAC-AC------CACCCTCTCCCCTACTGTGCTCT-CAAACCTGAGGCAAACTTTCT
C. simum AGAGCAGCAG---TCTGACCCTTTGGGGAC-GC------CACCCTCTCCCCTACTGTGCTCT-CGAACCTGAGGCAAACTTTGT
E. caballus AGAGCAGAAG---TCTAACCCTTTGGGGAT-GC------CACCCTCTCCCCTACTGTGCTCT-CCGACCTGAGGCAAACTTTCT
O. afer AGAACAGAAA---TATGACCTTTTGGAGAT-GT------CACCCTCTCCCCTACTGTTCTCT-CTAACCTGAGGCAAACTTTCT
E. telfairi CAAGCAGAAA---CTTGACCCTCAGGAGAT-GC------CACCCTCTCCCCTACTGTTTTCTCCCAACCTGAGACAAACTTTCT
C. asiatica AGAGCAGAAG---TATGA-CCTGTGGGGAT-TC------CACCCTCTCCCCTACTGTTCTTT-CCAACCTACAGCAAACTTTCT
E. edwardii AGAGCAGAGAAATAATGACCCTTTGGGGAT-GC------CACCCTCTCCCCTACTGTTCTCT-CCAACCTGAAGCAAACTTTCT
L. africana AGAGCAGAAA---CAGGACCCTTTGGGGAT-GC------CACCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCGAACTTTCT
T. manatus AGAACAGAAA---CATGACCCTTTGGGGAT-GC------CACCCTCTCCCCTACTGTTCTCT-CCAACCTGAGGCAAACTTTCT
M. darwinii NNNNNNNNNN---NNNNNNNNNNNNNNNNN-NN------NNNNNNNNNNNNNNNNNNTTCTCT-CTCACTTGAGGCAAACTTTCT



Appendix 9. rVista enhancer predicted transcription factor binding motifs for selected species and the consensus sequence from Appendix 8.

Consensus	
Motif name	**location/strand/sequence**
CREB_Q2	2 (-) | tgcTACGTCAta
CREB_Q4	2 (-) | tgCTACGTCAta
E4F1_Q6	3 (+) | gcTACGTCAT
ATF6_01	4 (-) | CTACGTCA
CREBATF_Q6	4 (-) | ctaCGTCAt
CREB_01	4 (-) | ctACGTCA
CREB_02	4 (-) | ctaCGTCATAaa
CREB_Q4_01	4 (-) | ctaCGTCAtaa
DR4_Q2	5 (-) | taCGTCATAAAAGGTCa
CREB_Q3	7 (+) | CGTCAt
SRF_C	8 (+) | gTCATAAAAGGTCAg
SRF_Q4	8 (+) | gTCATAAAAGGtcagtta
ERR1_Q2	10 (+) | cataaaAGGTCAgt
RORA1_01	10 (+) | cataaaAGGTCAg
HNF4ALPHA_Q6	10 (-) | catAAAAGGTCAg
RUSH1A_02	11 (-) | atAAAAGgtc
FXR_Q3	14 (+) | aAAGGTCAGTTACC
ZBRK1_01	14 (-) | AAAGGTCAGTTACCC
MYB_Q3	15 (+) | aaggtCAGTTa
T3R_Q6	15 (-) | aAGGTCagt
FXR_IR1_Q6	16 (+) | AGGTCAGTTACCc
FXR_IR1_Q6	16 (-) | aGGTCAGTTACCC
MYB_Q5_01	17 (-) | ggtcaGTTa
SZF11_01	17 (-) | GGTCAGTTACCCTTG
VMYB_02	19 (-) | tCAGTTacc
ZTA_Q2	23 (+) | ttacccTTGCTCA
PPARG_02	26 (+) | CCCTTGCTCACACTGACCTATTC
PPARG_02	26 (-) | CCCTTGCTCACACTGACCTATTC
AR_Q6	27 (-) | cctTGCTCa
ER_Q6	28 (+) | cttgctcacacTGACCtat
ER_Q6	28 (-) | cttGCTCAcactgacctat
ER_Q6_02	35 (-) | acacTGACCTa
RORA1_01	38 (-) | cTGACCTattctt
RORA2_01	38 (-) | cTGACCTATTCtt
PPARA_01	39 (-) | TGACCTATTCTTTACCTCTC
CAAT_C	41 (+) | aCCTATTCTTTACCTCTCCACTTCT
GABP_B	70 (+) | gCCAGAAGAGca
PPAR_DR1_Q2	81 (+) | aGAAATCTGACCc
MEF3_B	81 (-) | aGAAATCTGACCc
IK1_01	93 (+) | ctttGGGGAtgcc
NFKAPPAB50_01	97 (+) | GGGGATGCCA
AHRARNT_02	98 (-) | GGGATGCCACCCTCTCCCC
HOX13_01	102 (+) | tGCCACCCTCTCCCCTACTGTTCTCTCCAa
KROX_Q6	103 (+) | gcCACCCTCtcccc
MAZ_Q6	109 (-) | CTCTCCCc
GR_Q6_01	118 (+) | acTGTTCt
PR_Q2	119 (-) | cTGTTCtctc
MEF3_B	127 (-) | tCCAACCTGAGGc
MAF_Q6	137 (-) | ggcaaACTTTCTCCTA
PPARG_03	139 (-) | CAAACTTTCTCCtactt
LYF1_01	153 (-) | ctTCCCAGA
STAT_Q6	165 (-) | TCTCAGAAggggt
TAXCREB_01	171 (+) | aaGGGGTGAGGtcag
TBX5_Q5	171 (-) | aAGGGGTGAg
CREB_02	173 (+) | ggGGTGAGGtca
VJUN_01	173 (+) | gggGTGAGGTCAGcct
CREB_Q2_01	173 (-) | ggggTGAGGtcagc
VJUN_01	173 (-) | gggGTGAGGTCAGcct
ATF3_Q6	174 (+) | gggTGAGGTCAGCc
ATF_01	174 (+) | gggTGAGGTcagcc
CREB_Q4_01	174 (+) | gggTGAGGtca
T3R_01	174 (+) | gggTGAGGTCAGcctg
ATF3_Q6	174 (-) | gGGTGAGGTCAgcc
ATF_01	174 (-) | gggtgAGGTCAgcc
ATF_B	174 (-) | gggtgAGGTCAg
CREBP1_Q2	175 (+) | ggTGAGGTCAgc
CREB_Q2	175 (+) | ggTGAGGTCagc
CREB_Q4	175 (+) | ggTGAGGTCAgc
CREBP1_Q2	175 (-) | ggTGAGGTCAgc
CREB_Q2	175 (-) | ggtGAGGTCAgc
CREB_Q4	175 (-) | ggTGAGGTCAgc
ATF_B	176 (+) | gTGAGGTcagcc
CREBP1CJUN_01	177 (-) | TGAGGTCA
CREB_01	177 (-) | tgAGGTCA
ER_Q6_02	178 (+) | gAGGTCAgcct

Homo sapiens	
Motif name	**location/strand/sequence**
CREB_Q2	2 (-) | tgcTACGTCAta
CREB_Q4	2 (-) | tgCTACGTCAta
E4F1_Q6	3 (+) | gcTACGTCAT
ATF6_01	4 (-) | CTACGTCA
CREBATF_Q6	4 (-) | ctaCGTCAt
CREB_01	4 (-) | ctACGTCA
CREB_02	4 (-) | ctaCGTCATAaa
CREB_Q4_01	4 (-) | ctaCGTCAtaa
CREB_Q3	7 (+) | CGTCAt
ZBRK1_01	14 (-) | AAGGGTCAGTTGCCC
MYB_Q3	15 (+) | agggtCAGTTg
FXR_IR1_Q6	16 (-) | gGGTCAGTTGCCC
MYB_Q5_01	17 (-) | ggtcaGTTg
MYB_Q6	18 (-) | gtcaGTTgcc
VMYB_02	19 (-) | tCAGTTgcc
RFX_Q6	20 (+) | caGTTGCcc
ZTA_Q2	23 (+) | ttgcccTTGCTCA
SF1_Q6	24 (+) | TGCCCTTG
AR_Q6	27 (-) | cctTGCTCa
ER_Q6	28 (-) | cttGCTCAtactgacctat
XBP1_01	35 (+) | atacTGACCTAttcttt
ER_Q6_02	35 (-) | atacTGACCTa
RORA1_01	38 (-) | cTGACCTattctt
RORA2_01	38 (-) | cTGACCTATTCtt
PPARA_01	39 (-) | TGACCTATTCTTTACCTCTC
CAAT_C	41 (+) | aCCTATTCTTTACCTCTCTGCTTCT
MYOGNF1_01	51 (+) | tacCTCTCTGCTTCTTCTTTGTGCCAgaa
STAT4_01	83 (-) | taGAAatc
PPAR_DR1_Q2	84 (+) | aGAAATCTGACCc
MEF3_B	84 (-) | aGAAATCTGACCc
IK1_01	96 (+) | ctttGGGGAtacc
AHRARNT_02	101 (-) | GGGATACCACCCTCTCCCC
ZIC2_01	105 (-) | tACCACCct
MAZ_Q6	112 (-) | CTCTCCCc
MEF3_B	130 (-) | tCCAACCTGAGGc
MAF_Q6	140 (-) | ggcaaACTTTCTCCTA
PPARG_03	142 (-) | CAAACTTTCTCCtactt
STAT1_01	151 (-) | tcctacTTCCCAGAGcctgtc
LYF1_01	156 (-) | ctTCCCAGA
NRSF_01	163 (-) | GAGCCTGTCAGAAGTGGTGAA
MEIS1_01	164 (-) | agcCTGTCAgaa
PAX5_01	173 (-) | gaagtgGTGAAGCCAGCCTGCTCCTTGg
CREB_02	176 (+) | gtGGTGAAGcca
SMAD4_Q6	176 (+) | GTGGTGAAGCCAGCc
CREBP1_Q2	178 (+) | ggTGAAGCCAgc
CREB_Q2	178 (+) | ggTGAAGCCagc
CREB_Q4	178 (+) | ggTGAAGCCAgc

Bos taurus

Motif name	location/strand/sequence
PAX3_B	1 (+) | tagctacGTCACGAAagctct
ATF_01	1 (-) | tagctACGTCAcga
CREB_Q2_01	2 (+) | agctaCGTCAcgaa
ATF4_Q2	2 (-) | agCTACGTCAcg
CREBP1_Q2	2 (-) | agCTACGTCAcg
CREB_Q2	2 (-) | agcTACGTCAcg
CREB_Q4	2 (-) | agCTACGTCAcg
E4F1_Q6	3 (+) | gcTACGTCAC
ATF6_01	4 (-) | CTACGTCA
CREBATF_Q6	4 (-) | ctaCGTCAc
CREB_01	4 (-) | ctACGTCA
CREB_02	4 (-) | ctaCGTCACGaa
CREB_Q4_01	4 (-) | ctaCGTCAcga
CREB_Q3	7 (+) | CGTCAc
AHR_Q5	7 (-) | cgtCACGAaag
E4F1_Q6	8 (-) | GTCACGAAag
ZBRK1_01	14 (-) | AAAGCTCTGCTGCCC
SF1_Q6	24 (+) | TGCCCTTG
AR_Q6	27 (-) | cctTGCTCa
ER_Q6	28 (-) | cttGCTCAcactgccctgt
FOXO4_02	43 (+) | ctgTTCTTTACctc
PR_Q2	43 (-) | cTGTTCttta
ZTA_Q2	67 (-) | TGTGCCAgaggag
SZF11_01	71 (+) | CCAGAGGAGCAGACG
BEL1_B	81 (+) | aGACGGCTGAGCCTCTGGGGACACCGcc
SZF11_01	84 (-) | CGGCTGAGCCTCTGG
EBF_Q6	92 (-) | cCTCTGGGGAc
ZBRK1_01	98 (+) | GGGACACCGCCCTCT
SP1_Q4_01	100 (-) | gacaCCGCCCtct
SP1_Q6	100 (-) | gacACCGCCCtct
SP1_01	101 (-) | acACCGCCct
KROX_Q6	103 (+) | acCGCCCTCtcccc
NFMUE1_Q6	104 (+) | CCGCCCTCT
MAZ_Q6	109 (-) | CTCTCCCc
GR_Q6_01	118 (+) | acTGTTCt
PR_Q2	119 (-) | cTGTTCtctc
MEF3_B	127 (-) | tCCAACCTGAGGc
MAF_Q6	137 (-) | ggcaaACTTTCCCCTA
PPARG_03	139 (-) | CAAACTTTCCCCtactt
STAF_02	143 (+) | cttTCCCCTACTTCCCAGAgc
MZF1_02	144 (-) | tttCCCCTacttc
STAT1_01	148 (-) | ccctacTTCCCAGAGcctccc
DEAF1_01	152 (-) | acttccCAGAGCCTCCCGAAGGGGt
DEAF1_02	152 (-) | acttcCCAGAGCCTCCCGAAggggt
LYF1_01	153 (-) | ctTCCCAGA
LYF1_01	163 (-) | ccTCCCGAA
AP2_Q6	164 (+) | ctCCCGAAGggg
ZID_01	174 (-) | GGTAAAGGAGCCt

Echinops telfairi

Motif name	location/strand/sequence
ATF_01	1 (-) | ttgctACGTCAcag
CREBP1_Q2	2 (-) | tgCTACGTCAca
CREB_Q2	2 (-) | tgcTACGTCAca
CREB_Q4	2 (-) | tgCTACGTCAca
E4F1_Q6	3 (+) | gcTACGTCAC
ATF6_01	4 (-) | CTACGTCA
CREBATF_Q6	4 (-) | ctaCGTCAc
CREB_01	4 (-) | ctACGTCA
CREB_Q4_01	4 (-) | ctaCGTCAcag
CREB_Q3	7 (+) | CGTCAc
E12_Q6	10 (+) | caCAGATGgcc
YY1_Q6	11 (-) | acAGATGGC
PPARG_01	12 (-) | cagaTGGCCAATCACCCTtgc
NFY_Q6_01	13 (+) | agatggCCAATca
NFMUE1_Q6	13 (-) | AGATGGCCA
CAAT_01	14 (+) | gatggCCAATca
NFY_01	14 (+) | gatggCCAATcaccct
NFY_C	14 (-) | gatGGCCAATCAcc
NFY_Q6	16 (+) | tggCCAATcac
GATA1_03	16 (-) | tGGCCAATCaccct
FXR_Q3	17 (-) | GGCCAATCACCCTt
CAAT_C	18 (+) | gCCAATCACCCTTGCTCACCCTGAC
ZEC_01	19 (-) | CCAATCACCCTTG
ZTA_Q2	23 (+) | tcacccTTGCTCA
LXR_Q3	23 (-) | TCACCCTTGCTCACCCtg
PPARG_02	26 (+) | CCCTTGCTCACCCTGACCTACTC
PPARG_02	26 (-) | CCCTTGCTCACCCTGACCTACTC
AR_Q6	27 (-) | cctTGCTCa
ER_Q6	28 (+) | cttgctcacccTGACCtac
ER_Q6	28 (-) | cttGCTCAccctgacctac
ER_Q6_02	35 (-) | acccTGACCTa
RORA2_01	38 (-) | cTGACCTACTCtt
DR3_Q4	40 (+) | gacCTACTCTTTGCCTCttcg
CAAT_C	41 (+) | aCCTACTCTTTGCCTCTTCGCTTCT
HIF1_Q3	62 (+) | ttctACGTGCcagc
NMYC_01	62 (+) | ttcTACGTGcca
HIF1_Q5	63 (+) | tctACGTGccag
NF1_Q6	63 (-) | tctacgtgccaGCCAAgc
AHRHIF_Q6	65 (+) | taCGTGcca
E2F_Q6_01	69 (+) | tgccAGCCAAgc
NFY_01	70 (+) | gccagCCAAGcagaaa
NFY_C	70 (-) | gccAGCCAAGCAga
ALPHACP1_01	72 (+) | CAGCCAAGCAG
GCNF_01	80 (-) | caGAAACTTGACCCTCag
DR1_Q3	81 (-) | AGAAACTTGACCc
MEF3_B	81 (-) | aGAAACTTGACCc
TTF1_Q6	91 (+) | ccctCAGGagat
RFX_Q6	98 (+) | gaGATGCca
AHRARNT_02	98 (-) | GAGATGCCACCCTCTCCCC
CHOP_01	99 (+) | agaTGCCACcctc
HOX13_01	102 (+) | tGCCACCCTCTCCCCTACTGTTTTCTCCCa
KROX_Q6	103 (+) | gcCACCCTCtcccc
MAZ_Q6	109 (-) | CTCTCCCc
LYF1_01	125 (-) | tcTCCCAAC
FOXP3_Q4	132 (-) | aCCTGAGACAAActttc
PPARG_03	140 (-) | CAAACTTTCTCCtactt
TEL2_Q6	151 (+) | cTACTTCCTA
PEA3_Q6	153 (+) | ACTTCCt
LUN1_01	156 (-) | TCCTAGAGCCTCTCGGA
TFIIA_Q6	173 (+) | CATGAGAGCagc
NERF_Q2	180 (-) | gcagcctgCTTCTTGcaa

Condylura cristata

Motif name	location/strand/sequence
HIF1_Q3	10 (-) | cccaGCAGGTcgcc
E47_01	11 (+) | ccaGCAGGTCgccct
HEB_Q6	12 (-) | CAGCAGGt
AHRARNT_02	21 (-) | GCCCTTGCTCGCAGTGACC
ER_Q6	24 (+) | cttgctcgcagTGACCtgt
ER_Q6	24 (-) | cttGCTCGcagtgacctgt
PPARA_02	24 (-) | cttgCTCGCAGTGACCtgt
XBP1_01	31 (+) | gcagTGACCTGttcttt
ER_Q6_02	31 (-) | gcagTGACCTg
HNF4_DR1_Q3	40 (+) | TGTTCTTTGCCTc
MYOGNF1_01	47 (+) | tgcCTCTTCGCTTCTTCTTTGTGCCAgaa
E2F1_Q3_01	64 (+) | tttGTGCCAGAAGGgc
ZTA_Q2	66 (-) | TGTGCCAgaaggg
GABP_B	69 (+) | gCCAGAAGGGca
DR3_Q4	71 (-) | cagaAGGGCAGAAAGCGGgtc
NERF_Q2	76 (+) | gggCAGAAAGcgggtctc
ZF5_B	83 (-) | aaGCGGGtctctg
OLF1_01	84 (-) | agcgggtCTCTGGGGAAGccac
EBF_Q6	90 (-) | tCTCTGGGGAa
XPF1_Q6	92 (+) | TCTGGGGAAG
SMAD4_Q6	93 (+) | CTGGGGAAGCCACCc
AHRARNT_02	96 (-) | GGGAAGCCACCCTCTCCCC
STAT3_02	96 (-) | gGGAAgcc
KROX_Q6	101 (+) | gcCACCCTCtcccc
MAZ_Q6	107 (-) | CTCTCCCc
MEF3_B	125 (-) | tCCAACCTGAGGc
MAF_Q6	135 (-) | ggcaaACTTTCCCCTA
PPARG_03	137 (-) | CAAACTTTCCCCtactt
STAF_02	141 (+) | cttTCCCCTACTTCCCAGCcc
MZF1_02	142 (-) | tttCCCCTacttc
SP1_01	153 (-) | tcCCAGCCcc
SP1_Q2_01	154 (+) | cCCAGCCCct
DEC_Q1	158 (+) | gccCCTCTGAagg
SP3_Q3	159 (-) | CCCCTCTGAAGGGT
CREB_02	168 (+) | agGGTGAGGtaa
CREB_Q2_01	168 (-) | agggTGAGGtaacc
MEF3_B	169 (+) | gGGTGAGGTAACc
ATF3_Q6	169 (-) | gGGTGAGGTAAcct
CREBP1_Q2	170 (+) | ggTGAGGTAAcc
CREB_Q2	170 (+) | ggTGAGGTAacc
CREB_Q4	170 (+) | ggTGAGGTAAcc
E4F1_Q6	171 (-) | GTGAGGTAac
MEF3_B	175 (-) | gGTAACCTGCTCc

Canis lupus

Motif name	location/strand/sequence
PAX3_B	1 (+) | tacctacGTCATGGAaggtct
ATF_01	1 (-) | tacctACGTCAtgg
CREB_Q2_01	2 (+) | acctaCGTCAtgga
ATF4_Q2	2 (-) | acCTACGTCAtg
CREBP1_Q2	2 (-) | acCTACGTCAtg
CREB_Q2	2 (-) | accTACGTCAtg
CREB_Q4	2 (-) | acCTACGTCAtg
E4F1_Q6	3 (+) | ccTACGTCAT
TFIII_Q6	3 (-) | CCTACGTCa
ATF6_01	4 (-) | CTACGTCA
CREBATF_Q6	4 (-) | ctaCGTCAt
CREB_01	4 (-) | ctACGTCA
CREB_02	4 (-) | ctaCGTCATGga
CREB_Q4_01	4 (-) | ctaCGTCAtgg
CREB_Q3	7 (+) | CGTCAt
FXR_Q3	14 (+) | gAAGGTCTGTTACC
ZBRK1_01	14 (-) | GAAGGTCTGTTACCC
FXR_IR1_Q6	16 (-) | aGGTCTGTTACCC
SZF11_01	17 (-) | GGTCTGTTACCCCTG
RFX1_02	20 (-) | ctGTTACCCCTGCTCCta
ARP1_01	30 (+) | TGCTCCTATGACCTct
ATF_01	35 (+) | ctaTGACCTctcct
ERR1_Q2	36 (-) | taTGACCTctcctt
DR4_Q2	38 (+) | tGACCTCTCCTTTACct
T3R_Q6	41 (+) | cctCTCCTt
SF1_Q6	49 (+) | TTACCTTG
STAT3_02	56 (+) | ggcTTCCc
GABP_B	67 (+) | gCCAGAAGAGcg
SMAD4_Q6	88 (+) | CTGGGGCCGCCACCc
MUSCLE_INI_B	89 (+) | tggggccgcCAccctctaccc
PAX5_02	90 (-) | gggggccGCCACCCTCTACCCTACtgctc
AHRARNT_02	91 (-) | GGGCCGCCACCCTCTACCC
YY1_02	94 (+) | ccgccACCCTCTACCCTact
HOX13_01	95 (+) | cGCCACCCTCTACCCTACTGCTCCCGCCAa
E2F_03	114 (-) | gctcCCGCCAAC
E2F_Q6_01	115 (+) | ctccCGCCAAcc
E2F_Q4_01	116 (+) | tccCGCCAACc
E2F1_Q4_01	117 (-) | ccCGCCAAC
E2F1_Q6_01	117 (-) | ccCGCCAACc
E2F_Q2	117 (-) | ccCGCC
E2F_Q3_01	117 (-) | ccCGCCAAC
E2F1_Q3	118 (-) | cCGCCAac
E2F1_Q4	118 (-) | CCGCCAAc
E2F1_Q6	118 (-) | cCGCCAAc
E2F_Q6	118 (-) | cCGCCAAC
ZTA_Q2	121 (+) | ccaaccTGGCACA
FOXP3_Q4	124 (-) | aCCTGGCACAACctttc
PPARG_03	132 (-) | CAACCTTTCTCCtactt
ETS1_B	139 (-) | tcTCCTACTTCCTCg
ETS_Q4	141 (+) | tcctaCTTCCTc
TEL2_Q6	143 (+) | cTACTTCCTC
ETS_Q6	145 (+) | acTTCCtc
PEA3_Q6	145 (+) | ACTTCCt
CETS168_Q6	145 (-) | aCTTCCTC
GLI_Q2	153 (-) | gGGCCTCCCAga
ZIC3_01	154 (-) | ggcctCCCa
LYF1_01	156 (-) | ccTCCCAGA
STAT_Q6	158 (-) | TCCCAGAAgggct
VJUN_01	166 (-) | gggCTGAGGTCGTcgg
ATF3_Q6	167 (+) | ggcTGAGGTCGTCg
CREB_Q4	168 (+) | gcTGAGGTCGtc
STAT_01	183 (+) | TTCTTGGAA

Mus musculus	
Motif name	**location/strand/sequence**
CREB_Q4	2 (-) | taCAGCGTCAca
WHN_B	3 (-) | acaGCGTcaca
LXR_Q3	4 (+) | caGCGTCACAGAGGGTCA
CREBATF_Q6	4 (-) | cagCGTCAc
DR4_Q2	5 (-) | agCGTCACAGAGGGTCa
LXR_DR4_Q3	6 (-) | GCGTCACAGAGGGTCA
CREB_Q3	7 (+) | CGTCAc
FXR_Q3	14 (+) | gAGGGTCAGTCACC
SMAD4_Q6	14 (+) | GAGGGTCAGTCACCc
BACH1_01	15 (+) | aggGTCAGTCACcct
PXR_Q2	15 (+) | agGGTCAgtcac
AP1_01	16 (+) | gggTCAGTCAccc
FXR_IR1_Q6	16 (+) | GGGTCAGTCACCc
FXR_IR1_Q6	16 (-) | gGGTCAGTCACCC
BACH2_01	17 (+) | ggTCAGTCAcc
AP1_Q4	17 (-) | ggtcAGTCAcc
AP1_Q6	17 (-) | ggtcAGTCAcc
FXR_Q3	17 (-) | GGTCAGTCACCCTt
PAX4_01	17 (-) | ggtcagtCACCCTTGACCACa
PAX2_01	18 (+) | gtcaGTCACCcttgaccac
PAX2_01	19 (-) | tcagtcaccCTTGACcaca
PPARG_01	19 (-) | tcagTCACCCTTGACCACact
HNF4ALPHA_Q6	22 (+) | gTCACCCTTGacc
COUP_01	23 (+) | TCACCCTTGACCac
COUP_DR1_Q6	23 (+) | TCACCCTTGACca
HNF4_DR1_Q3	23 (+) | TCACCCTTGACCa
PPAR_DR1_Q2	23 (+) | tCACCCTTGACCa
DR1_Q3	23 (-) | TCACCCTTGACCa
PPARG_03	23 (-) | TCACCCTTGACCacact
T3R_01	25 (-) | acccTTGACCACActg
ATF1_Q6	27 (+) | CCTTGACCACA
CBF_02	27 (-) | ccttgaCCACactgaa
PPARA_02	28 (-) | cttgACCACACTGAACtag
AML1_01	32 (-) | ACCACA
AML1_Q6	32 (-) | ACCACA
HNF4_Q6_02	39 (-) | TGAACT
ATF_01	42 (-) | actagTCGTCAcct
CREB_Q2_01	43 (+) | ctagtCGTCAcctt
CREB_Q2	43 (-) | ctaGTCGTCAcc
COUPTF_Q6	45 (+) | agtcgTCACCTTTCCACtcttcc
CREBATF_Q6	45 (-) | agtCGTCAc
CREB_Q4_01	45 (-) | agtCGTCAcct
PAX3_01	47 (+) | TCGTCACCTttcc
CREB_Q3	48 (+) | CGTCAc
MAF_Q6	57 (-) | tccacTCTTCCTGCCA
GABP_B	59 (-) | caCTCTTCCTGc
NRF2_01	61 (-) | ctCTTCCTGC
KAISO_01	64 (+) | tTCCTGCCAg
GABP_B	69 (+) | gCCAGAAGAGca
OLF1_01	85 (-) | tcagactCTCTGGGGATAtcag
EBF_Q6	91 (-) | tCTCTGGGGAt
PAX4_03	104 (+) | cagcctCACCCc
TBX5_Q5	108 (+) | cTCACCCCTa
EFC_Q6	128 (+) | cattaTGAGGCAAa
MAF_Q6	136 (-) | ggcaaACTTTCTTTCA
IRF_Q6	142 (+) | ctttcTTTCACTTcc
BLIMP1_Q6	144 (-) | ttcTTTCACTTCCc
IRF_Q6_01	146 (-) | cTTTCACTTcc
NKX25_Q5	147 (+) | tttCACTTcc
IRF1_Q6	148 (+) | TTCACTT
STAT1_02	150 (+) | cacTTCCc
STAT3_02	150 (+) | cacTTCCc
LYF1_01	152 (-) | ctTCCCAGA
LUN1_01	154 (+) | TCCCAGAGGCTCTGGGG
LUN1_01	154 (-) | TCCCAGAGGCTCTGGGG
EGR_Q6	165 (+) | CTGGGGGcagc
HNF4_Q6_03	168 (+) | gGGGCA
VDR_Q3	168 (+) | GGGGCAGCAAGGtca
ERR1_Q2	171 (+) | gcagcaAGGTCAac
RORA1_01	171 (+) | gcagcaAGGTCAa
T3R_01	172 (+) | cagCAAGGTCAAccct
SF1_Q6	175 (-) | CAAGGTCA
ER_Q6_02	176 (+) | aAGGTCAaccc

Rattus norvegicus	
Motif name	**location/strand/sequence**
ATF_01	1 (-) | ttgcgACGTCAcag
CREB_Q4	2 (+) | tgCGACGTCAca
CREBP1_Q2	2 (-) | tgCGACGTCAca
CREB_Q2	2 (-) | tgcGACGTCAca
CREB_Q4	2 (-) | tgCGACGTCAca
ATF_B	3 (+) | gCGACGTcacag
E4F1_Q6	3 (+) | gcGACGTCAC
CREBP1CJUN_01	4 (+) | CGACGTCA
LXR_Q3	4 (+) | cgACGTCACAGTGGGTCA
ATF1_Q6	4 (-) | CGACGTCACAG
ATF6_01	4 (-) | CGACGTCA
CREBATF_Q6	4 (-) | cgaCGTCAc
CREB_01	4 (-) | cgACGTCA
CREB_Q4_01	4 (-) | cgaCGTCAcag
DR4_Q2	5 (-) | gaCGTCACAGTGGGTCa
LXR_DR4_Q3	6 (-) | ACGTCACAGTGGGTCA
CREB_Q3	7 (+) | CGTCAc
AP1_Q2_01	10 (-) | cacagtggGTCA
AP2REP_01	12 (+) | CAGTggg
SMAD4_Q6	14 (+) | GTGGGTCAGTCACCc
ER_Q6	14 (-) | gtgGGTCAgtcacccttga
PXR_Q2	15 (+) | tgGGTCAgtcac
AP1_01	16 (+) | gggTCAGTCAccc
FXR_IR1_Q6	16 (+) | GGGTCAGTCACCc
FXR_IR1_Q6	16 (-) | gGGTCAGTCACCC
BACH2_01	17 (+) | ggTCAGTCAcc
AP1_Q4	17 (-) | ggtcAGTCAcc
AP1_Q6	17 (-) | ggtcAGTCAcc
FXR_Q3	17 (-) | GGTCAGTCACCCTt
PAX4_01	17 (-) | ggtcagtCACCCTTGATCACa
PAX2_01	18 (+) | gtcaGTCACCcttgatcac
PAX2_01	19 (-) | tcagtcaccCTTGATcaca
DR1_Q3	23 (-) | TCACCCTTGATCa
VMAF_01	27 (+) | ccttGATCACACTGCacca
SP3_Q3	36 (-) | CACTGCACCAGTCT
AREB6_02	48 (+) | cttCACCTttcc
STAF_02	53 (+) | cctTTCCACGCTTCCTGCCag
NERF_Q2	55 (-) | tttccacgCTTCCTGcca
WHN_B	56 (+) | ttccACGCttc
AHR_01	56 (-) | ttcCACGCTTCCTGCcag
ETS1_B	56 (-) | ttCCACGCTTCCTGc
MAF_Q6	57 (-) | tccacGCTTCCTGCCA
ETS_Q4	58 (+) | ccacgCTTCCTg
ETS2_B	58 (-) | ccacgcTTCCTGcc
NF1_Q6	58 (-) | ccacgcttcctGCCAGag
GABP_B	59 (-) | caCGCTTCCTGc
ELK1_02	60 (-) | acgcTTCCTgccag
CETS1P54_01	61 (-) | cgctTCCTGc
NRF2_01	61 (-) | cgCTTCCTGC
ETS_Q6	62 (+) | gcTTCCtg
CETS168_Q6	62 (-) | gCTTCCTG
KAISO_01	64 (+) | tTCCTGCCAg
BACH1_01	74 (+) | agcATGAATCAGgct
BACH1_01	74 (-) | agcATGAATCAGgct
AP1_01	75 (+) | gcaTGAATCAggc
BACH2_01	76 (+) | caTGAATCAgg
NRF2_Q4	76 (-) | caTGAATCAGGCt
AP1_C	77 (+) | aTGAATCAg
MAF_Q6_01	77 (-) | aTGAATCaggc
OLF1_01	82 (-) | tcaggctCTCTGGGGATAccgg
EBF_Q6	88 (-) | tCTCTGGGGAt
GATA2_01	93 (+) | gggGATaccg
NFKAPPAB50_01	93 (+) | GGGGATACCG
TAXCREB_01	100 (-) | ccggCCTCACCCCta
PAX4_03	101 (+) | cggcctCACCCc
CREB_02	101 (-) | cggCCTCACCcc
SREBP_Q3	102 (+) | ggccTCACCCCT
TBX5_Q5	105 (+) | cTCACCCCTa
CREB_Q2_01	112 (-) | ctacTGAGGcaaac
MAF_Q6_01	113 (+) | tactGAGGCAa
NFE2_01	113 (+) | tACTGAGGCAa
DR3_Q4	119 (+) | ggcAAACTTTCTCCCACttct
MAF_Q6	119 (-) | ggcaaACTTTCTCCCA
PPARG_03	121 (-) | CAAACTTTCTCCcactt
AHR_01	128 (-) | tctCCCACTTCTCAGagg
NKX25_Q5	130 (+) | tccCACTTct
NRSE_B	138 (+) | cTCAGAGGCTCTGAGGGCAGC
ATF1_Q6	146 (+) | CTCTGAGGGCA
VDR_Q3	151 (+) | AGGGCAGCAAGGtca
ERR1_Q2	154 (+) | gcagcaAGGTCAgc
SF1_Q6	158 (-) | CAAGGTCA
ER_Q6_02	159 (+) | aAGGTCAgccc
SRF_Q5_01	164 (-) | cagccCTTTCTTTGG