UNIVERSITY OF LEEDS

Ecological genomics of lactation strategies in pinnipeds

David Thomas Orr

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Faculty of Biological Sciences

School of Biology

July 2022

The candidate confirms that the work submitted is [his/her/their] own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

The world was a very different place when I embarked on this PhD and completing this piece of work would not have been possible without the support of a whole cast of friends, family, and colleagues who have helped me along the way.

Firstly, I would like to thank my partner and best friend, Jessie. She has been my rock through the entirety of my academic career. She has been my cheerleader, always pushing me to be the best version of myself. She has helped me always see the bigger picture and been there to help pick me up when I am down. To my parents, Dot and Andy, you have always supported and encouraged me, pushing me to follow my passions. You have both always been the best role models through your actions and thank you for making me the person I am today.

My academic supervisors, Dr Simon Goodman and Dr Mary O'Connell. Thank you for being superb examples of everything a supervisor should be. You have each played such a pivotal role in my development as a scientist and person over the past few years. Simon has introduced me to the wonderful world of pinnipeds – of which will stay with me forever. His infectious passion for pinnipeds has been a constant source of inspiration for me. Mary has been a fantastic mentor offering constant guidance and motivation, even during the dark days of the pandemic. She has been a great teacher during these years constantly seeking to improve, both my knowledge and critical thinking. Both of you have pushed me to reach my full potential and I have enjoyed working with you both so much. I would also like to thank Ian Carr, without Ian's optimistic and encouraging words my project would have been much more difficult.

I would also like to thank my friends and colleagues in both Leeds and Nottingham. I was very lucky to have such an unbelievably supportive group surrounding me. From words of encouragement, analysis support, coffee breaks, beers, and chats these people gave me everything I needed to succeed. I would like to say a special thank you to Michaela Agapiou, Peter Mulhair, and Fiona Whelan. These people are some of the most supportive people I have ever met and have been fantastic friends throughout my PhD.

I would also like to thank NERC Spheres for sponsoring this project and Leeds NGS facility and Nottingham DeepSeq facility in helping with the generation of data.

Abstract

Long standing evolutionary questions such as genetic basis of adaptations can now be investigated with genome-scale empirical data. In this thesis, pinnipeds (seals, sea lions, fur seals and walrus) are employed as a model species to investigate the evolution of novel phenotypes in Mammalia. In the 40 million years since pinniped divergence, they have evolved extensive variation in life-history strategies, with unique morphological, physiological, and behavioural adaptations. Much of this variation is related to the constraints imposed from spatial and temporal separation of terrestrial parental care with marine foraging. Whilst many of the key differences in ecology between pinniped species are understood to be underpinned by lactation related traits, the molecular basis of these trait differences is poorly understood. In Chapter 2, I describe high-quality *de novo* genomes assemblies for Caspian (*Pusa caspica*) and Hooded (*Cystophora cristata*) seals employing long and short read sequencing data. I developed and applied a novel annotation pipeline to identify protein coding regions of the generated assemblies. In Chapter 3, I resolve the phylogeny for pinnipeds within the Carnivora by combining the gene models generated in Chapter 2 with 15 additional publicly available pinniped genomes. I employed multiple phylogeny reconstruction methods identifying and excluding influences from genes outlying phylogenetic signals and resolving a congruent phylogeny across multiple reconstruction methods. In Chapter 4, I combine the species phylogeny and gene models, with codon-based models of evolution to perform genome-wide scans of selective pressure variation, identifying candidate genes under positive selection. I find signals in many genes known to influence milk properties in cattle, or lipid associated metabolic disease phenotypes in humans – which suggests lactation traits in pinnipeds have evolved because of selection on metabolic pathways which are conserved across mammals.

Abbreviations

ABCA7	ATP binding cassette subfamily A member 7
ACSL5	Acyl-CoA synthetase long chain family member 5
Adh	Alcohol dehydrogenase
ADIPOR1	Adiponectin receptor 1
ALMS1	ALMS1 centrosome and basal body associated protein
ASVR	Associated site rate variation
AU	Approximately unbiased
BEB	Bayes empirical Bayes
BI	Bayesian inference
BP	Bootstrap probabilities
BUSCO	Benchmarking Universal Single-Copy Orthologs
CACNA1E	Calcium voltage-gated channel subunit alpha1 E
CAMKK1	Calcium/calmodulin dependent protein kinase kinase 1
CD36	CD36 molecule
CEP63	Centrosomal protein 63
CERK	Ceramide kinase
CGN	Cingulin
CRB1	Protein crumbs homolog 1
CYP8B1	Cytochrome P450 family 8 subfamily B member 1
D	interspecific divergence
DGKQ	Diacylglycerol kinase theta
DGKQ	Diacylglycerol kinase theta
DLC	Duplication-loss-coalescence
dN	Non-synonymous substitutions per nonsynonymous site
DNA	Di-oxyribonucleic acid
dS	Synonymous substitutions per synonymous site
E	Expected base frequencies
EEF2K	Eukaryotic elongation factor 2 kinase
ELOVL2	ELOVL fatty acid elongase 2
ENPEP	Glutamyl aminopeptidase
EPAS1	Endothelial PAS domain protein 1
EPRS1	Glutamyl-prolyl-tRNA synthetase 1
EST	Expressed sequence tag
F13A1	Coagulation factor XIII A chain
FAM83H	Family with sequence similarity 83 member H
FASN	Fatty acid synthase
FDR	False discovery rates
GBF1	Golgi brefeldin A resistant guanine nucleotide exchange factor 1

Gbp	Gigabase pairs
gCF	Gene concordance factor
gDFP	Discordance due to polyphyly
GHITM	Growth hormone inducible transmembrane protein
GLA	Galactosidase alpha
GLS	Gene-wise lieklihood scores
GO	Gene ontology
GTR	General time reversal
GWAS	Genome-wide association studies
HGNC	HUGO Gene Nomenclature Committee
HMM	Hidden markov models
ILS	Incomplete lineage sorting
INPPL1	Inositol polyphosphate phosphatase like 1
JAK1	Janus kinase 1
JC69	Jukes-Cantor model
JTT	Jones, Taylor and Thornton
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIF21A	Kinesin family member 21A
KLF11	Kruppel like factor 11
LALBA	Alpha-lactalbumin
LAPSG	Lactation associated positively selected genes
LCA	Last common ancestor
LEP	Leptin
LIPC	Lipase C
LIPE	Hormone sensitive lipase E
LPCAT3	Lysophosphatidylcholine acyltransferase 3
LPL	Lipoprotein lipase
LPP	Local posterior probability
LRP5	LDL receptor related protein 5
LRP5	LDL receptor related protein 5
LRT	Likelihood ratio test
LS	Largest subtree
LTBP3	Latent transforming growth factor beta binding protein 3
MCL	Markov Cluster Algorithm
MCL1	MCL1 apoptosis regulator
MCMC	Markov chain Monte Carlo
MGF	Multi-gene orthologous family
MGI	Mouse genome informatics
MGST1	Microsomal glutathione S-transferase 1
MI	Maximum inclusion

MK	McDonald-Krietman
ML	Maximum likelihood
MO	Monophyletic outgroup
MOGAT1	Monoacylglycerol O-acyltransferase 1
MRCA	Most recent common ancestor
MROH1	Maestro heat like repeat family member 1
MSA	multiple sequence alignment
MSC	Multispecies coalescent
MUFA	Monounsaturated fatty acids
My	Million years
Mya	Million years ago
NADK	NAD kinase
NCF1	Neutrophil cytosolic factor 1
Ne	Effective population sizes
NEB	Naïve empirical Bayes
NEFA	non-esterified fatty acids
NEK11	NIMA related kinase 11
NGS	Next generation sequencing
NME8	Thioredoxin domain-containing protein 3
NOS3	Nitric oxide synthase 3
O	Observed base frequencies
OC	Orthologous cluster
OLC	Overlap-Layout-Consensus
ONT	Oxford Nanopore Technologies
P	Intra-specific polymorphisms
PDV	Phocidae Distemper Virus
PEX12	Peroxisomal Biogenesis Factor 12
PLCD3	Phospholipase C delta 3
PLIN1	Perlipin
PLIN2	Perilipin 2
PNPLA2	Adipose triglyceride lipase
PNPLA3	Patatin like phospholipase domain containing 3
PP	posterior probability
PPA	Posterior predictive analysis
PPA	Posterior predictive analyses
PPA-DIV	Site-specific base diversity
PSG	Positively selected gene
PSMC	Pairwise sequential Markovian coalescent
PUFA	polyunsaturated fatty acids
REG	Rapidly evolving genes

RF	Robinson-Foulds
RIN	RNA integrity number
RNA	Ribonucleic acid
RPAP3	RNA polymerase II associated protein 3
SACS	Sacsin molecular chaperone
SBS	Sequencing-by-synthesis
SEMA4G	Semaphorin 4G
SERPINE1	Serpin family E member 1
SFA	Short fatty acid
SGO	Single gene ortholog
SIGLEC1	Sialic acid binding Ig like lectin 1
SLC37A1	Solute carrier family 37 member 1
SLC51B	Solute carrier family 51 subunit beta
SNP	Single nucleotide polymorphism
SORCS1	Sortilin related VPS10 domain containing receptor 1
SPHK2	Sphingosine kinase 2
STAT5B	Signal transducer and activator of transcription 5B
STAT6	Signal transducer and activator of transcription 6
TAG	Triacylglycerol
TFRC	Transferrin receptor
THBS1	Thrombospondin 1
TMIGD3	Transmembrane and immunoglobulin domain containing 3
ToL	Tree of life
VAV3	Vav guanine nucleotide exchange factor 3
WNK3	WNK lysine deficient protein kinase 3
XDH	Xanthine dehydrogenase
ZC3H3	Zinc finger CCCH-type containing 3

List of Figures

List of Tables

Chapter 1: Introduction

A key challenge in evolutionary genomics is to understand the genetic basis of differences in adaptive traits among species, and how these relate to the selective pressures arising from environmental and ecological factors. Suites of novel adaptations often arise as species diversify to exploit new environments or ecological niches. Genomic analyses of taxa arising during such radiations have significantly advanced my understanding of the nature and genetic architecture of novel adaptations. Within Mammalia, marine mammals have developed some of most extreme adaptations in morphological, physiological, and life-history traits as they shifted from terrestrial to marine habitats. Pinnipeds are the group of marine mammals containing seals, sea lions, fur seals, and walrus and are unique in that their behaviour is constrained to breed on solid substrates (land and ice), but the rest of their life-history is dependent on the marine environment. This has led to the evolution of traits unique to Pinnipedia, with a suite of adaptations related to how they provision their young.

Some of the traits seen in Pinnipedia are extreme when compared to other mammals, such as the ability to temporarily pause lactation, in sea lions and fur seals, or the Hooded seal possessing the highest fat content of milk of any extant mammal. Such diversity within an evolutionary recent clade gives a unique opportunity to examine how environmental and ecological factors interact with molecular processes to drive the evolution of novel adaptations. In this thesis, I generate *de novo* assemblies for the previously unsequenced Caspian seal and Hooded seal genomes. I then use these genome assemblies and advanced phylogenetic tools to resolve some of the phylogenetic uncertainty in the current pinniped phylogeny. The phylogeny is then used to support analyses identifying genes showing species and lineage specific signatures of selection, and the unique adaptions of pinnipeds they might be associated with, focusing on lactation strategies.

1.1 Evolutionary history of pinnipeds

To fully understand the phenotypic adaptations in pinnipeds it is important to fully appreciate their evolutionary history and the ecological drivers that have led to the diverse group seen today. Pinnipeds are mammalian family which diverged from an arctoid carnivore common ancestor, a group which includes procyonids, mustelids and ursids, around 30-40 million years ago (Mya) (Berta et al., 2018). They have since diversified into 38 extant species, representing over 25% of current marine mammal species. They are comprised of three families - Phocidae (true seals), Otariidae (sea lions and fur seals), and Odobenidae (Walrus) (Berta et al. 2005). The relationships between the three families, with reference to the sister taxon of pinnipeds, has been contentious (Uhen, 2007). Two competing hypotheses have been

proposed – a monophyletic origin (Figure 1.1A), where Phocidae, Otariidae and Odobenidae all diverged from a common ancestor; and a diphyletic origin of pinnipeds (Figure 1.1B), in which Phocidae derived from basal Mustelidae, whilst Otarioids (Otariidae and Odobenidae) derived from basal Ursidae ancestor. Early paleontological and morphological evidence supported the latter hypothesis (Tedford, 1976; Repenning et al., 1979; de Muizon, 1982; Barnes, 1989; Wozencraft, 1989; Nojima, 1990; Kuhn and Frey, 2012; Koretsky et al., 2016). However, more recent reanalysis of morphological evidence, and overwhelming support from DNA sequence data has now led to acceptance of a monophyletic origin (Weber, 1904; Gregory, 1910; Davies, 1958; Wyss, 1987; Berta and Wyss, 1994; Flynn et al., 2005; Fulton and Strobeck, 2006; Kohno, 2006; Sato et al., 2006; Higdon et al., 2007; Yonezawa et al., 2009; Nyakatura and Binida-Emonds, 2012; Furbish 2015; Hassanian et al., 2021). Despite near-consensus support for a monophyletic origin hypothesis, recent studies have disputed the arctoid group as most closely related to pinnipeds. There is evidence to suggest Mustelidae, Ursidae, or an Ursidae-Mustelidae ancestry (Flynn and Nedbal, 1998; Delisle and Strobeck, 2005; Feijoo and Parada, 2017).

The earliest diverging lineages are known as the Pinnipedimorpha clade, within which fossils with characters suggestive of an Ursidae-sister hypothesis and Mustelidae-sister hypothesis are both present (Orlov, 1933; Savage, 1957; Tedford et al., 1994; Jacobs et al., 2009). When a fossil of a primitive Pinniped, subsequently named *Enaliarctos* was discovered, it was believed to be the earliest Pinniped, originating in the eastern North Pacific between 30.6 and 28 Mya (Berta et al., 2018). Fossils for 5 species of *Enaliarctos spp.* have been since been found (Berta et al., 2018). These fossils possess characters shared with archaic bears, such as similar cranial features and heterodont dentition (differentiated teeth such as incisors and molars), supporting a Ursidae-sister hypothesis (Berta, 1991). *Enaliarctos spp.* had key aquatic adaptations ubiquitous across modern day pinnipeds, including large eyes, specialised inner ear, and sensitive whiskers. Although, *Enaliarctos spp.* used both forelimb and hindlimbs for aquatic propulsion, unlike modern day pinnipeds, Phocidae use hind limbs and Otarioid predominantly use forelimbs (Mitchell and Tedford 1973, Flynn et al., 1988; Hunt and Barnes, 1994). More recently, fossils of Pinnipedimorpha with otter-like characters have been found, including *Puijila darwinii* which occupied the Arctic 24-20 Mya (Rybczynski et al., 2009). Their webbed feet appear to bridge the gap between the fully-flippered species and terrestrial carnivores. Through cladistic analyses of morphological traits, *Puijila* and other Pinnipedimorpha fossils cluster in early diverging lineages at the base of Pinnipedia. Thus, despite *P. darwinii* appearing to display transitional characters between terrestrial and aquatic lifestyles, this suggests that both *Enaliarctos spp.* and *Puijila spp.* are sister lineages to modern day pinnipeds rather than

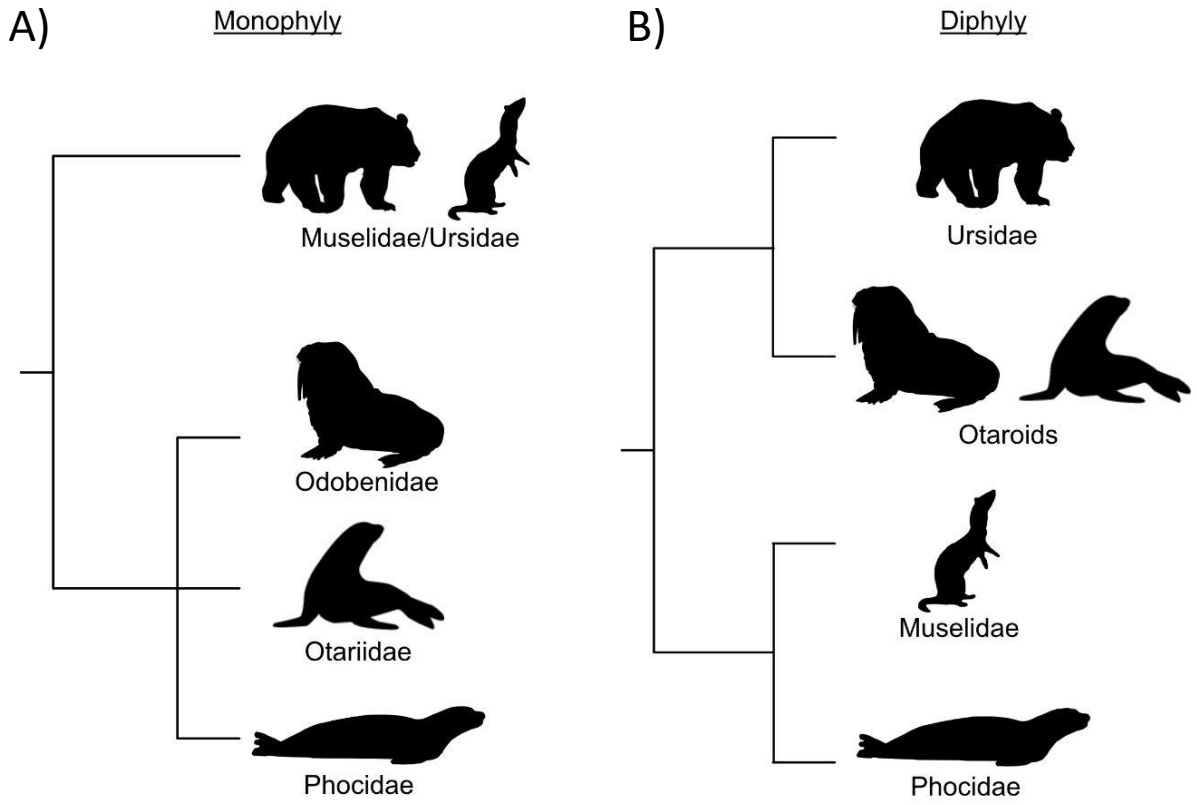


Figure 1.1. Phylogenetic origins of Pinnipedia. The alternative hypotheses of the evolutionary relationships between pinnipeds (A) monophyly origin hypothesis, in which a single ancestor of pinnipeds and (B) diphyly origin hypothesis in which Phocidae and mustelids exist as a sister taxon and Ursidae and Otarioids exist as a sister taxon. Adapted from Berta et al. (2015).

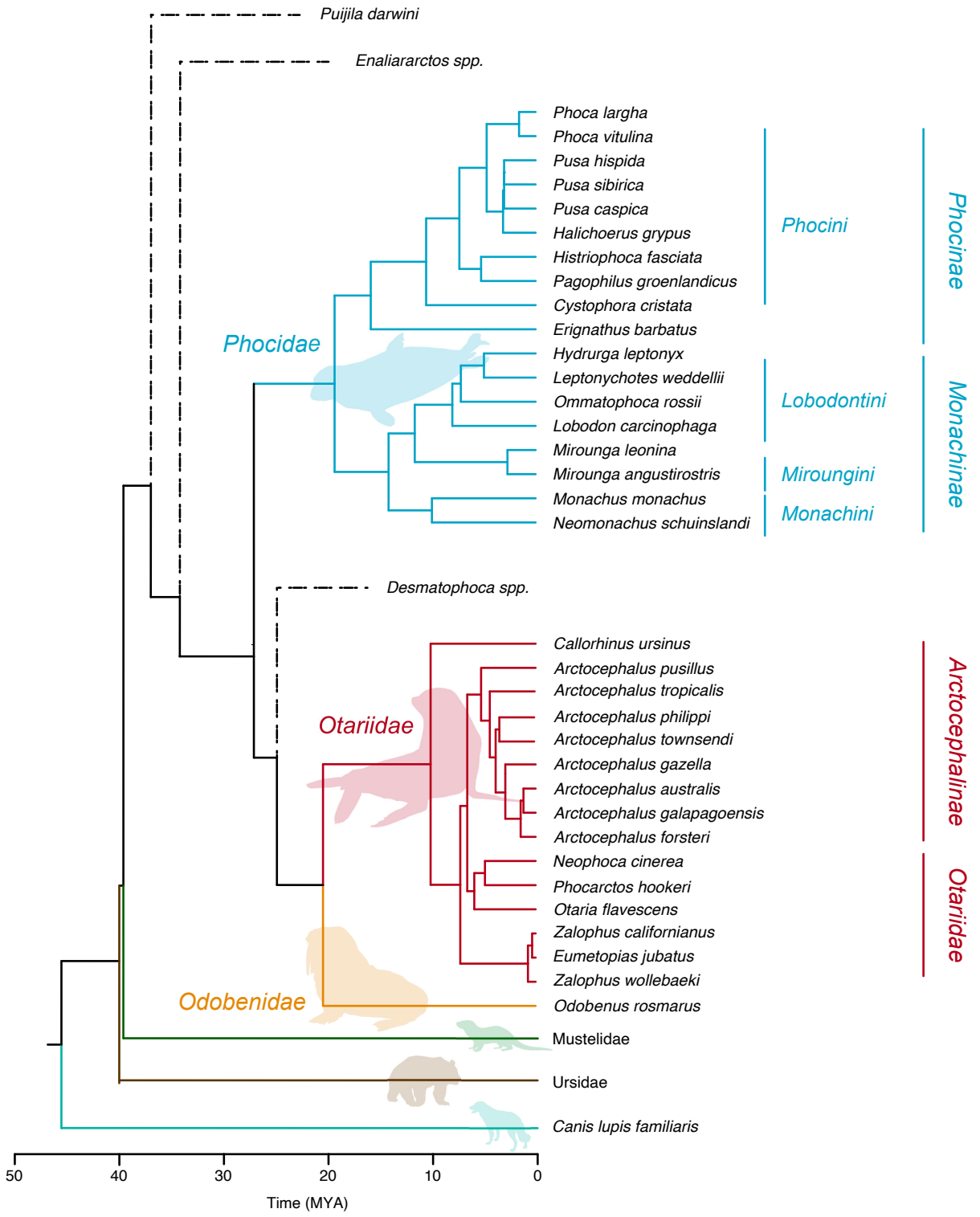


Figure 1.2. Time calibrated phylogeny of pinnipeds. Phylogeny of pinnipeds and their relationship within Carnivora. Adapted from Nyakatura and Bininda-Emonds, (2012); Lopes et al., (2020) and Paterson et al., (2020).

direct ancestors (Paterson et al., 2020). *P.darwinii* lived in a cool temperate environment with seasonally frozen lakes, suggesting the earliest pinnipeds were exclusively freshwater (Rybczynski et al., 2009). With many early pinnipeds being primarily freshwater or near-shore, the hypothesis of transitional phases leading to the extreme level of aquatic adaptations seen in extant species appears likely (Paterson et al., 2020).

A major diversification in pinniped evolution is that between Otarioid and Phocidae, which is thought to have occurred during episodic increases in Arctic Sea ice cover which began around 23 Mya (DeConto et al., 2008). The changes in ice cover would have increased waters habitable by krill, changing arctic food webs and creating new niches, which were exploitable by marine mammals. Phocidae have a rich fossil record when compared to most marine mammals, based on fossil evidence it is thought that early Phocidae evolved in the North Atlantic before dispersing into the whole of the Atlantic. Phocidae is comprised of two major clades – Phocinae and Monachinae, which split around 14.7 Mya. Phocinae then spread throughout the north Atlantic and Monachinae across the south (Figure 1.3) (Davis et al. 2004, Higdon et al. 2007, Fulton & Strobeck 2010). Shifting Arctic oceanic events appear to have coincided with speciation events in Phocidae. For instance, isotope records suggest the Arctic Ocean may have started fully circulating with global oceans at the time of the earliest diversification in Phocinae, the *Erignathini* (Haley et al., 2008). The divergence of the *Cystophorini* from the rest of Phocini is thought to have occurred at a similar time to the opening of the Bering Strait (Fulton and Strobeck, 2010). Oceanic events go some way to explain the diversification of Phocidae, but glaciation may have resulted in allopatric speciation, for instance between the ribbon and harp seal (Deméré et al., 2003). The *Pusa* clade has not yet been fully resolved, which makes determining the ecological factors responsible for the diversification of this clade more challenging. Nevertheless, there is a consensus that landlocked seals, such as the Caspian seal and Baikal seal *(Pusa sibirica*) appear to be a result of colonisation and subsequent isolation of inland water systems (Chapskii, 1955; McClaren, 1960).

Otariidae and Odobenidae split from the Phocidae lineage approximately 20 Mya (Figure 1.2) (Nyakatura and Bininda-Emonds, 2012; Yonezawa et al., 2009). Historically, Otariidae was thought to comprise of two monophyletic groups Arctocephalinae (fur seals) and Otariinae (sea lions), based on the layer of specialised underfur exhibited by the former. Molecular studies have since determined these subfamilies to be paraphyletic (Wynen et al., 2001; Yonezawa et al., 2009). Northern fur seal (*Callorhinus ursinus*) is in a sister clade to the rest of Otariidae, being the earliest diverging Otariid (Berta and Wyss, 1994; Berta

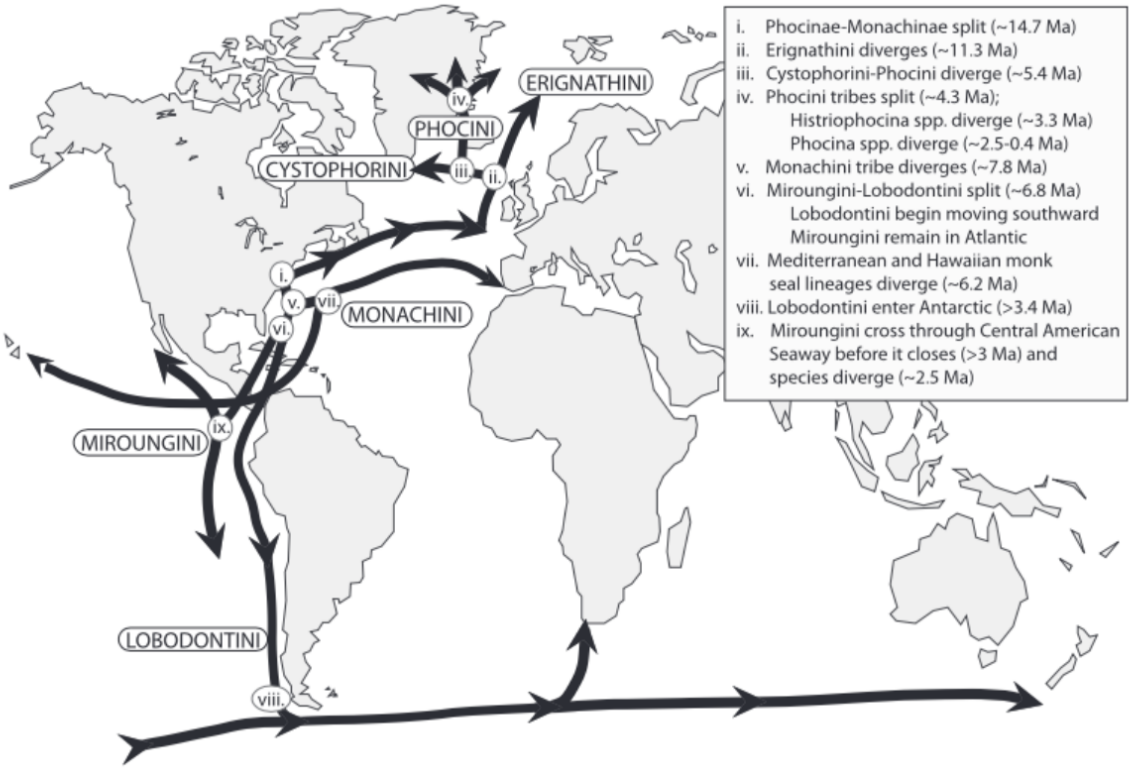


Figure 1.3. Hypothesised movements of Phocidae. The expected movement of Phocidae species during their evolution, with the time of diversification in the grey box. Adapted from Fulton and Strobeck, (2010).

et al, 2018). It is argued that the Otariidae achieved dispersal in the southern hemisphere by a single or several trans-equatorial crossings (Deméré et al., 2003; Koretsky & Barnes, 2006; Yonezawa et al., 2009) (Figure 1.4). The most recent molecular analysis supports diverging Otariid (Berta and Wyss, 1994; Berta et al, 2018). It is argued that the Otariidae achieved dispersal in the southern hemisphere by a single or several trans-equatorial crossings (Deméré et al., 2003; Koretsky & Barnes, 2006; Yonezawa et al., 2009) (Figure 1.4). The most recent molecular analysis supports multiple crossings. An initial influx into the southern hemisphere being aided through temperate conditions and high productivity approximately 3-5 Mya (Lopes et al., 2020). Fluctuating climates enabled an explosive radiation event which has resulted in recent diversity of southern hemisphere Otariidae, *Arctocephalus*, approximately 3 Mya (Nyakatura and Bininda-Emonds 2012; Lopes et al., 2020). Recent and rapid diversification of *Arctocephalus*, in addition to possible introgression, makes resolving a simple phylogenetic relationship for these taxa difficult (Churchill et al., 2014; Lopes et al., 2020).

Despite Walrus (*Odobenus rosmarus*) being the only surviving species of Odobenidae, rich fossil evidence has shown that this was a once diverse family, containing at least two other clades (Boessenecker and Churchill, 2013). Fossils of *Odobenus* have been recorded from as far south as southern USA and France, from the middle and late Pleistocene suggesting that the current range restriction to the Arctic is a recent phenomenon (Berta et al, 2018).

1.2 Evolution of Lactation Strategies in pinnipeds

Extant pinnipeds occupy a wide range of habitats, from tropical waters, e.g., Hawaiian monk seal, to the polar regions. Pinnipeds have adapted to various foraging ecologies, ranging from shallow coastal waters to deep, oceanic habitats. Different foraging ecologies have also emerged with a range of prey specialisations from the filter feeding of krill to predation of large prey such as Penguins and other pinnipeds (Kienle and Berta, 2018). The evolution of different lactation strategies is likely to be an important factor in facilitating the exploitation of different niches, whilst also coping with the spatial-temporal partitioning of marine foraging and terrestrial breeding. There are three lactation strategies present in extant pinnipeds: capital breeding strategy, income breeding strategy, and nursing strategy (Schultz and Bowen, 2005) (Figure 1.5). The capital breeding strategy is observed solely in Phocidae and involves a long seasonal build-up of energy reserves before hauling out on land for parturition. Young are provisioned for a short time frame during which the mother metabolises accumulated reserves whilst

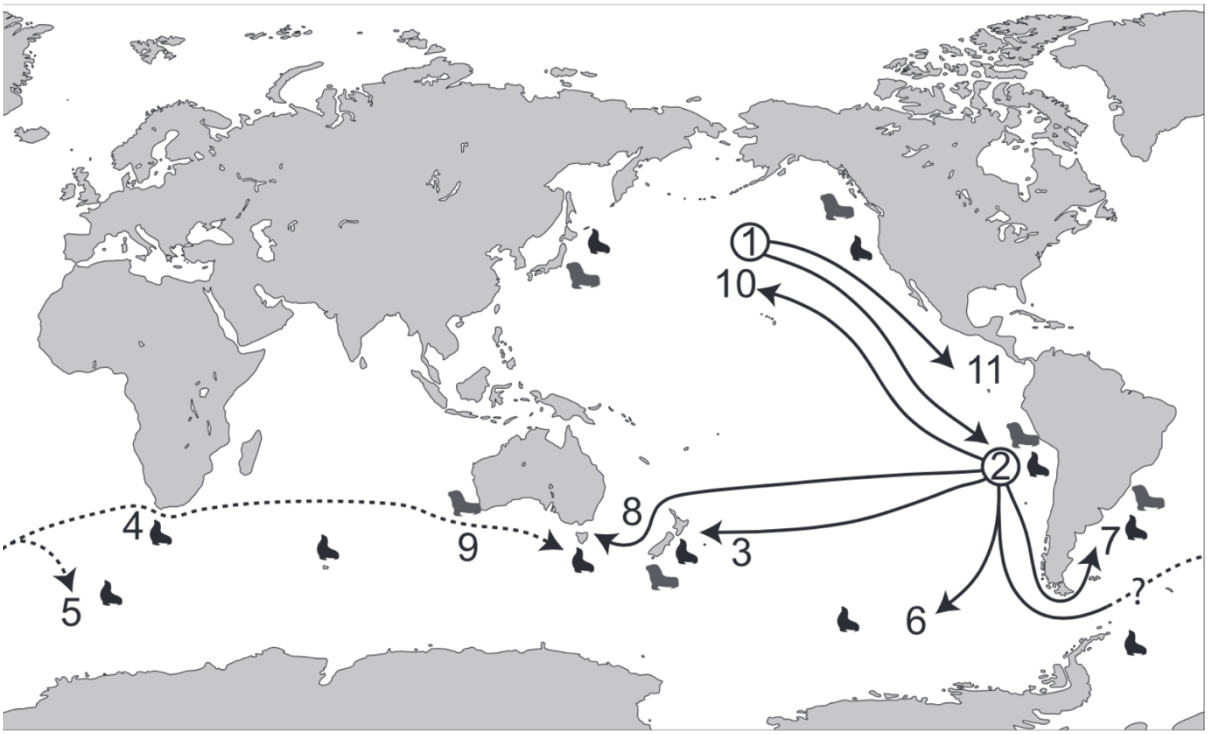


Figure 1.4. Hypothesised movement of the Otariidae. The expected movement of the Otariidae during their evolution. 1 represents the emergence of the archaic Otariidae species, with 2 representing their initial transequatorial dispersal and origin of southern hemisphere clade of Otariidae. 2-9 represent the rapid dispersal of the Arctocephalinae species along the southern hemisphere. 10 represents the transequatorial movement of San Fernandez fur seal (*Arctocephalus philippii),*, with 11 representing the dispersal of the Zalophus clade. Adapted from Churchill et al. (2014).

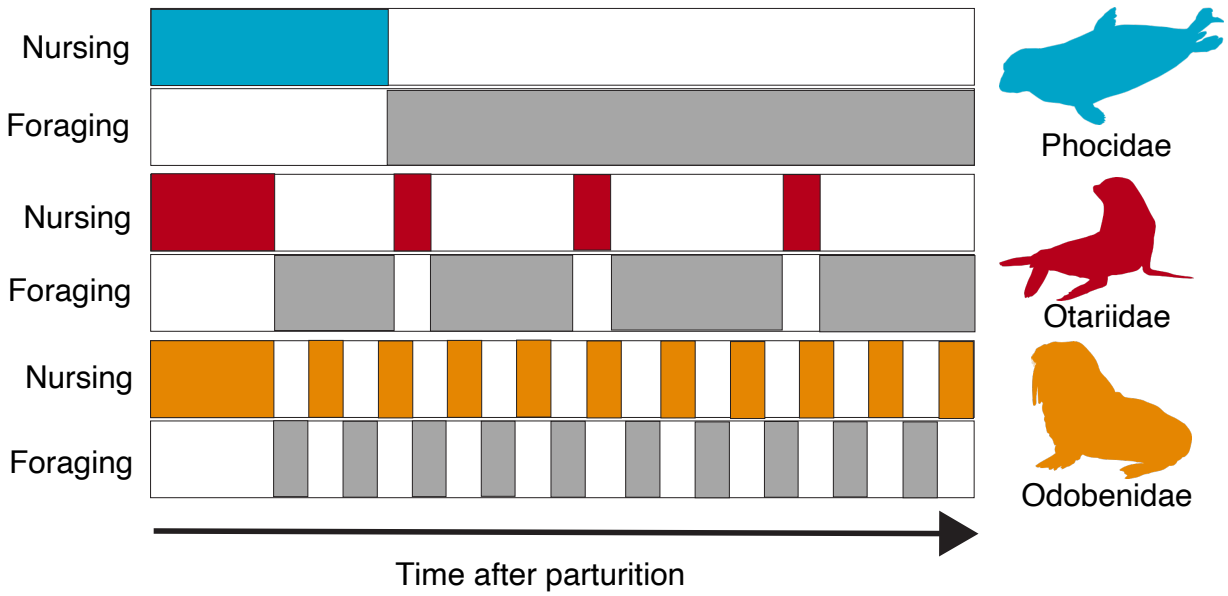


Figure 1.5. Lactation strategies as delineated by time allocation after parturition. A representative of the various lactation strategies seen in pinnipeds, with time scales representing the proportion of time spent on land nursing the young or in a marine environment foraging. The time scale is representative as a proportion of lactation period.

nursing. The income breeding strategy is ubiquitous in Otariidae, where in contrast to building up reserves throughout the year, females provision their young over months to years. During provisioning, mothers will make long intermittent foraging trips, ranging from days to multiple weeks, before returning to their young to continue nursing. Elements of income breeding are also present to some degree in some smaller bodied Phocidae species. In these species weaning of pups is still completed within a matter of weeks, but females make short foraging trips towards the end of the lactation period to supplement resources. Lastly, the nursing strategy is unique to the walrus. Walruses have the longest lactation length of any Pinniped, ranging from 1-5 years, during this time they are constantly foraging and provisioning the young (Clark et al., 2020).

A thorough understanding of pinniped evolutionary history is important to explain the phenotypic and behavioural diversity present in relation to the lactation strategies. Stem pinnipeds, including some *Enaliarctos spp.,* had small bodies and heterodont dentition, indicating they still chewed prey and frequently returned to land to consume prey (Repenning, 1976; Churchill and Clementz et al., 2015). This type of foraging indicates stem pinnipeds would perform a lactation strategy similar to that of their terrestrial ancestors, exploiting upwelling sites along coastal margins whilst provisioning young (Berta, 2018).

A thick layer of subcutaneous adipose tissue, blubber, is ubiquitous across pinnipeds. . Pinniped ecology and life-history suggests two primary drivers for the evolution of blubber: thermal insulation, and energy storage. Increased body size was prevalent in early Phocidae (Wyss, 1994), to decrease predation and increase thermal retention. Increasing body mass also increased the ability to store energy as blubber and facilitated the exploitation of colder environments that possessed seasonally highly abundant prey. Polar water increases in productivity in summer months due to extended sunlight hours, and ice could be used for parturition. The fundamental characteristics of capital breeding strategy appeared early in Phocidae. The unstable nature and predation risk in sea ice in turn presented a selective pressure for a decrease in lactation length with an increase in milk lipid levels and pup growth rates. Early members of the Phocidae lineage, i.e., ancient Monachinae, had large body sizes, which could have facilitated the initial separation of maternal foraging and lactation (Churchill et al., 2014). This separation would enable Phocidae to move to exploit habitats with less productive local waters. Extant Monachinae species exhibit a capital breeding strategy, with Elephant seals being an extreme example. Elephant seals consist of two species: the Northern elephant seal (*Mirounga angrustirostris*) and Southern elephant seal (*Mirounga leonina*), with

the former breeding on beaches or islands in the eastern Pacific Ocean, whereas the latter utilise remote islands across the Southern Ocean. Despite breeding at remote locations adult female Elephant seals spend up to 10 months of the year conducting extensive migrations to reach foraging grounds, sometimes several thousand kilometres from breeding grounds (Hindell et al., 2016). At these foraging sites Elephant seals perform continuous dives exceeding 100 minutes, reaching depths of more than 1,500m where they feed on mesopelagic fish (Robinson et al., 2012; Adachi et al., 2021). This extreme foraging pattern is facilitated by the large mass of Elephant seals, with female Southern elephant seals reaching up to 900kg and males up to 4000kg. During lactation periods, females use their reserves to produce energy rich milk whilst meeting maternal metabolic requirements. During this process adult females can lose up to 42% of their initial body mass (Costa et al., 1985).

The Phocinae seals evolved in the Northern arctic water, where Miocene cooling led to ice sheet formation with breeding on unstable pack ice (Fulton and Strobeck, 2010). The risk of ice break-up before pups are weaned and increased to exposure to predation, has resulted in the evolution of traits supporting reduced lactation periods. For example, the Hooded seal has a lactation length of 4-7 days, during which time the mothers produce the highest percentage fat milk of all Mammalia. During this time the young accumulate around 7kg of mass per day (Bowen et al., 1995). Even though some populations of northern Atlantic seals, such as the Grey seal (*Halichoerus grypus*) and Harbour seal (*Pusa vitulina*), now breed terrestrially, this is most likely a recent phenomenon exploiting land exposed after the glacial retreat from the end of the Last Glacial Period. These species have an abbreviated ancestral lactation strategy, and other remnants of an ice breeding past, such as a white lanugo coat in pups, although this is shed *in utero* in Harbour seals (Shaughnessy and Fay, 1977). Several north hemisphere Phocidae, including landlocked seals such as the Caspian seal, Ringed seal (*Pusa hispida*) and Baikal seal are all small bodied. This permits or even necessitates, a hybridisation between income-breeding and capital breeding strategies. Antarctic Phocidae species, such as the Weddell seal (*Leptonychotes weddellii*) have a more prolonged lactation length, with weaning occurring at 6-7 weeks post-partum. The physical development of pups possibly needed to be extended to reduce the risk of marine predation, with an absence of terrestrial predation (Reijnders et al., 1990).

Early Otariidae inhabited isolated terrestrial rookeries to avoid land-based predation and gain proximity to productive local prey areas, whilst retaining a relatively small body size (Churchill et al., 2014). Early Otariidae had evolved homodont dentition, suggesting an adaptation to foraging and consummation of

whole prey at sea (Costa, 1993). Climate cooling during the Miocene selected for increased body sizes in Otariidae, which enabled the exploitation of larger prey and distant offshore foraging trips (Churchill et al., 2014). Increased foraging distances would have selected for faster surface travel, and the ability for mammary glands to temporarily cease lactation whilst remaining functional, which is unique to Otariidae amongst mammals. The income breeding strategy is ubiquitous in Otariidae although the length of the lactation is variable between species, ranging from 3 – 36 months (Trillmich and Lechner, 1986; Donohue et al., 2002; Sapriza, 2018). Milk fat composition is also variable between species, with the fat content increasing with trip duration rather than the length of lactation (Boness and Don Bowen, 1996; Sapriza, 2018). It is hypothesised that the reduced frequency of offspring provisioning in combination with reduced or less predictable prey availability, selected for slower pup development (up to 2-3 years). This is observed in extant tropical Otariids – Galapagos sea lion (*Zalophus wollebaeki*) and California sea lion (*Zalophus californianus*). Otariidae at higher latitudes have access to more predictable marine environments, which has allowed them to retain a longer lactation period, whereas species at lower latitudes have developed shorter lactation lengths due to the short seasonal productivity periods (Sapriza, 2018).

The nursing strategy is only 28exhibited in one species and the only extant member of the Odobenidae family. Walrus are the most social pinniped species, regularly found in groups on ice floes or land, where they rest, moult, and provision young (Fay, 1981). Female walruses give birth on pack ice and after partition the mother will return to the sea for a few days. Unlike other pinnipeds the pup will join the mother in these foraging trips, with the pup remaining at the surface during dives (Kovacs and Lavigne, 1992). Walruses have a unique feeding mode, in which bivalves, which reside on the seafloor, are consumed by sucking the organism out its shell. It is thought that during the early stages of lactation the pup must learn this feeding mode, and created a selective pressure for this unique lactation strategy (Fisher and Stewart, 1997). The pups start to consume food at around 5 months, but it has been seen that they will consume milk and solid foods for up to 5 years (Clark et al., 2020). As a consequence of a lack of spatial-temporal separation between feeding and lactation in walrus, lactation pressures that have affected lactation length and milk energy content appear to have been relaxed. Which may possibly explain the low-fat content and long lactation period in this unique species.

1.3 Signatures of positive selection on protein coding elements in the evolution of novel traits in marine mammals

Marine mammals offer key macroevolutionary transitions, from terrestrial to land environments (McGowern et al., 2014). A recolonisation of the marine environment has occurred at least three times in mammals and led to several behavioural and physiological adaptations convergently evolving (Berta et al., 2005). This convergence creates a model system to investigate the genomic basis of ecological adaptations. To date convergence has been observed the genomic level, with marine mammal groups possessing amino acid substitutions in the same coding regions or even at identical residues in genes possibly responsible for these adaptations (Mirceta et al., 2013; Foote et al., 2015; Zhou et al., 2015; Lui et al., 2019).

In marine mammals, sensory systems have been tailored to suit an underwater system (Hank and Dehnhardt, 2013); bodies have become streamlined, resulting in appendage or limb loss in some species (Wang et al., 2013); and a whole remodelling of the respiratory system has been observed which enable marine mammals to occupy deep depths within the water column (Kooyman, 2006). Marine mammals can tolerate depths that would be detrimental to terrestrial mammals. Cardiovascular control and cell-level hypoxia tolerance adaptations have attracted great interest for their medical relevance (Hooker et al., 2012; Hopkins et al., 2001; Williams et al., 2021). With the emergence of genomics, understanding of the molecular underpinnings of specialised traits, including increased oxygen storage and hypoxia prevention, has become more widely known. Marine mammals have much higher levels of oxygen-storing myoglobin in their muscles than terrestrial mammals (Kooyman and Poganis, 1998; Poganis et al., 2011; Hooker et al., 2012; Párraga et al., 2018). Myoglobin structure is highly conserved in both terrestrial and marine mammals, although is found to be more stable in marine mammals when aggregating at high concentrations (Antonini et al., 1971). Mirceta et al. (2013) found that net surface charge of myoglobin was highly correlated with the maximal concentration of myoglobin. By investigating the sequences of myoglobin from different species amino acid substitutions were observed in lineages of mammals that endure aquatic or semi-aquatic lifestyles. It was then found that these substitutions increased the net surface charge of myoglobin, thus allowing marine mammals to have higher myoglobin concentrations, and thus higher oxygen concentrates, accumulate in muscle tissue.

Heat transfers around 25 times more quickly in water than it does air, giving significant implications for endothermic marine mammals. The ability to store energy in dense vascular layers of fat is a ubiquitous feature across marine mammals. Blubber serves as both an energy source and insulator, with a low conductivity around 1/10th of water, blubber can limit heat transfer (Favilla and Costa, 2020). Due to the prominent role of adipose tissue in obesity, the evolution of genes related to blubber composition are of medical significance (Zhou and Rui, 2013). Studies have centred around the leptin (*LEP*) gene, a crucial gene in controlling body weight (Zhang et al., 1994). Leptin was shown to be under positive selective pressure in species of cetacean and pinnipeds (Ortiz et al., 2011; Yu et al., 2011). Another common phenotypic adaptation in response to thermal loss, is the increase of a volume to surface area ratio, this makes marine mammals an excellent model to study the evolution of large body size. Large body size, and therefore large cell number, has been used to investigate Peto's paradox, the lack of correlation between cancer risk and body size (Peto et al., 1975). Signatures of positive selection in known cancer related genes were first seen in Bowhead whale (Keane et al., 2015), expanding this to more large, bodied cetacean lineages found, many cancer-linked pathways (including 33 genes that are mutated in human cancers) identified as under positive selection (Tollis et al., 2019).

The molecular underpinnings of lactation traits within marine mammals, including pinnipeds, is still poorly explored. So far studies have identified possible positively selected genes that could play a role in lipid transport (Noh et al., 2022), or could possibly prevent mammary gland involution (Reich and Arnould, 2007). To date, the majority of studies have investigated the molecular basis of lactation related traits through expression analyses, covered in more detail in Chapter 4. Within other mammals lineages morphological adaptations often require modifications in the protein coding regions of genomes. Thus, I hypothesise that positive selection is a significant factor in the diverse nature of traits driven by lactation strategy variation in pinnipeds.

1.4 Selective Pressure and Molecular Evolution

Darwin's seminal work observed that species change over time and that the process of natural selection, in which the frequency of traits within a population, are directly related to their impact on reproductive success, and subsequent retention (Darwin, 1859). In the field of molecular evolution, the modifications that occur in the sequences of cellular molecules such as DNA, RNA and proteins are studied to uncover patterns and processes of "descent with modification". Until the 1960s and the availability of empirical sequence data, natural selection was thought to play the predominant role in determining the fate of

alleles in a population (Duret, 2008). Under this hypothesis, positive selection drove advantageous alleles to fixation, and these fixed alleles contributed to the adaptations of a species to its environment. This 'selectionist' view suggested that nonadaptive processes only played minor roles in genome evolution, with most polymorphisms being assumed to be a product of balancing selection. Protein surveys challenged this view, with empirical data to examine theories suggesting that the majority of variation in molecular changes are caused by neutral mutations and not positive or balancing selection (Lynch, 2007).

1.4.1 Neutral theory of evolution

The neutral model of evolution suggests that the majority of variance across populations is caused by stochastic factors (Kimura, 1968). The theory states that random mutations create variation and alleles consequently move to fixation or extinction within a population. For neutral alleles, the probability of these alleles becoming fixed within a population (P_x) exists as inversely proportional to the effective population size (N_e) (Equation 1.2). The rate of mutations per individual per generation (μ) has an impact on substitution rate. In a diploid population, the number of novel mutations per generation is $2N_e\mu$. Thus, in the case of neutral mutations the rate of substitution (k) is equal to μ (Equation 1.3) and is independent of N_e. This is explained through mutation rates being higher in a population with large N_e but with reduced chances of fixation, whereas mutation rates will be lower in a population with small N_e but will have an increased probability of fixation.

Equation 1.2 Probability of fixation of neutral mutation within a diploid population

$$P_x = \frac{1}{2N_e}$$

Equation 1.3 The overall rate of substitution is equal to the rate of mutation for neutral mutations

$$k = 2N_e\mu\left(\frac{1}{2N_e}\right) = \mu$$

In populations that have experienced bottleneck events, the low population sizes have increased the probability of neutral mutations becoming fixed in the population. For beneficial or deleterious mutations, the probability of fixation within the population changes with strength of selection (i.e., selection coefficient "s") having an impact (Kimura, 1957). For a neutral mutation, the average time for a mutation to become fixed is $4N_e$ generations (Kimura, 1980), but for a non-neutral mutation "s" is a governing factor (Equation 1.4). Therefore, smaller N_e have increased levels of genetic drift and consequently have lower

genetic variability and rates of adaptive potential (Willi et al., 2006). Founder effects and bottlenecks cause N_e to occur at very low levels, these dynamics give an increased chance of losing genetic diversity.

Equation 1.4 The average time taken for fixation in non-neutral alleles

$$k = 4N_e s\mu$$

It was observed that rates of heterozygosity in populations were much more reduced than would be estimated from a neutral mode of evolution (Ayala et al., 1972). From these observations the 'nearly neutral model of evolution' was developed, accounting for substitutions that are slightly deleterious being removed in larger populations (Ohta, 1973). Under this theory, functionally important sites will be highly conserved, although substitutions can occur at functionally relaxed sites, which can be randomly driven to fixation through genetic drift (Nei et al., 2010).

1.4.2 Natural selection

As random mutations enter a population through genetic drift their impact on fitness (reproductive success) is assessed by natural selection. Natural selection acts on these mutations to either increase "fitness" (i.e., reproductive success), or prevent a decrease in "fitness" of the population. Positive selection occurs when mutations that increase the reproductive success of that individual, increasing the frequency of this allele in the population. Purifying, or negative selection, acts on mutations in the population that result in a decrease in fitness, and therefore decrease in frequency within the population.

Positive selection acts to only increase the reproductive success and this can act on sexual traits that offer no survival advantage. Sexual related traits can either be honest indictors of survival advantage or can impede the survival of individuals. For example, the bright blue feet of the blue footed booby (*Sula nebouxii*) are honest indicators, i.e., the shade of blue becomes duller as the condition of the individual decreases, therefore females are attracted to survival indictors (Velando et al., 2006). In contrast, female peacocks are attracted to the male's elaborate tail, showing they can thrive despite a handicap (Zhavi, 1999). Handicap features are often sexual dimorphic, and so the females will not carry the handicap but would inherit the successful alleles. Positive selection can also occur on a group level, kin selection has been used to explain altruism in communities of relatives (Foster et al., 2006). This is especially true for the haplodiploid Hymenoptera, who due to the unusual ploidity – haploid males and diploid females, sisters have a higher relatedness than they would to their own offspring (Hamilton, 1972).

The debate about what level positive selection contributes to molecular evolution is still ongoing (Chen and Zhang, 2020; Eyre-Walker, 2006; Lynch, 2007; Kern and Hahn, 2018; Nei et al., 2010). The ability to determine whether a mutation is neutral or not is complex. Neutral mutations are defined as having no positive or negative impact on the fitness of the organism. However, mutations proximal to positions that affect fitness can avoid recombination by hitchhiking with their non-neutral neighbours, presenting patterns akin to non-neutral sites (Barton et al., 2000; Eyre-Walker and Keightley, 2009; Pouyet et al., 2018). Therefore, neutral mutations can be hard to detect in comparison to non-neutral mutations.

1.4.3 Positive selection and functional shift in protein coding regions

Differentiating between a beneficial selected allele and a neutral allele driven to fixation through genetic drift can be difficult, especially considering that the adaptation can proceed from standing variation in a population and can target phenotypic properties that are affected by multiple sites across different genes (polygenic) (Buffalo and Coop, 2020). However, with the advent of sequencing technologies, it has become feasible to empirically assess models of evolution. Combined with long term experimental evolution studies we are now at a point where we can test core principles of my models of molecular evolution and natural selection. For example, Lenski's pioneering work on long term experimental evolution of *E.coli in vitro* (Lenski, 2017) and rapid next generation sequencing has led to the emergence of 'evolve and re-sequence' experiments (Tenaillon et al., 2017). With these powerful experimental systems, we can trace allele frequency changes through time directly from genomic data (Schlötterer et al., 2015). These approaches, whilst exciting for certain species, are not applicable for non-model species, due to large genome sizes requiring extensive sequencing, and difficulties in cultivation of large populations.

Taking an alternative approach to understanding the processes of mutation, we can assess selective pressure variation for proteins in organisms. This can be achieved using the ratio of non-synonymous substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous site (dS). Non-synonymous substitutions cause a substitution of the amino acid whereas synonymous substitutions cause a silent substitution (a change in the codon used but not in the amino acid encoded) (Figure 1.6). The assumption of this model is that dS is a proxy for neutrality, however synonymous substitutions have been shown to be subjected to selective pressures, with roles involved in protein stability and splicing (Hurst and Pal, 2001; Stoletzki, 2008). The ratio between dN and dS (ω) is used to detect the type of selective pressure acting at a particular region of a protein. The ratios can be classified into $\omega < 1$, $\omega = 1$

and ω > 1, which indicate purifying/negative selection, neutral evolution, and positive selection respectively.

Positive selection, or adaptive evolution, has been shown to be responsible for 41-45% of protein coding region substitutions in drosophila (Smith and Eyre-Walker, 2002; Welch, 2006) and approximately 30 to 60% within mammals (Kosiol et al., 2008), with common targets in immune system pathways (Schultz and Sackton, 2019). Despite this high level of adaptive evolution, it is regarded that most codons within protein coding genes are subjected to high levels of purifying selection (Hahn, 2018). The structure of proteins dictates that there are only small regions of proteins that can be altered without making the protein non-functional (Nielsen et al., 2005).

Positive selection leading to functional shift in the protein was theorised before it was recognised through empirical analysis (Yang, 1998; Hughes, 2007). Methods to detect positively selected sites within a gene are often detected using codon based *in silico* methods (Jeffares et al., 2015) covered in more detail in the next section. The functional shift of these proteins requires combining detection methods with *in vitro* analyses to empirically determine whether a change in function is observed (Hughes, 2008).

1.4.4 Methods for detecting selective pressure analyses

Various methods have been developed to detect signatures of selection variation within and between species. This thesis focuses on between species comparisons where a single representative of a given species is used to calculate variation in selective pressure over a set of genes. Population based methods of selection pressure variation detection, such as the McDonald-Krietman (McDonald and Krietman, 1991) and Tajima's D (Tajima, 1989), are not the focus of this thesis due to costs, time constraints and the endangered status of study species. These would be preferably complementary analysis for the future to determine levels of fixation or variation of the candidate positively selected sites within a species. Early techniques to assess selective pressure variation used codon degeneracy to inform nonsynonymous site and synonymous substitution rates (Li et al., 1985). Sites were categorised into three classes: (i) non-degenerate sites - where any substitution at any site were nonsynonymous, (ii) two-fold degenerate sites, where a transversion would result in a nonsynonymous substitution, and (iii) four-fold degenerate sites where any substitution at any site would be synonymous. This method, along with Nei and Gojobori (1986), assumed random nucleotide substitution among the four nucleotides, but the rate of transversion Is usually lower than that of transition. Thus, the number of synonymous sites is expected to be higher

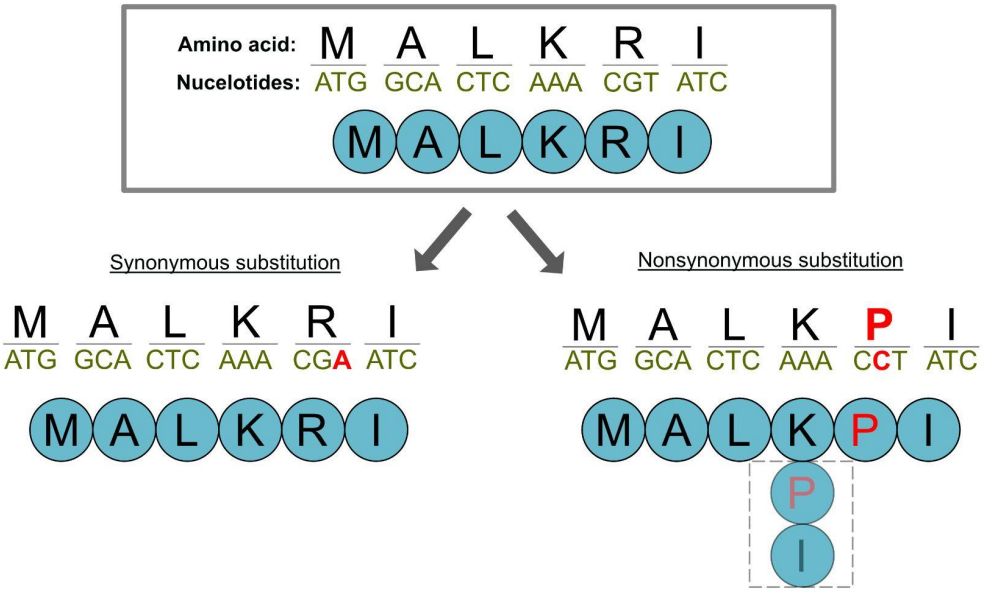


Figure 1.6. Representation of difference between a nonsynonymous substitution per synonymous site (dN) and synonymous substitution per synonymous site (dS). The synonymous nucleotide substitution (CGT to CGA) does not alter the amino acid and thus does not alter the sequence of the protein. The nonsynonymous nucleotide substitution (CGT to CCT) alters the composition of the protein by changing the amino acid (R to P), in some cases this may change the structure of the protein (dotted box).

than estimated and the resulting dS is overestimated (Li, 1993; Ina, 1995). Thus, the number of synonymous sites is expected to be higher than estimated and the resulting dS is overestimated (Li, 1993; Ina, 1995). Modified versions of these techniques were then introduced, using Kimura's two-parameter model (Kimura, 1980). Kimura's two-parameter model uses a weighted average of synonymous transitions at four-fold and two-fold degenerate sites and compares them to that of the weighted average of nonsynonymous transversions at nondegenerate and two-fold sites to calculate ω (Li, 1993; Ina, 1995). Using a weighted average can be consequential as it lacks the statistical power to detect positive selection when acting only a small number of sites within a sequence (Nei and Kumar, 2000). In addition, extreme codon-usage bias can affect estimations of dN and dS, and so *ad hoc* methods have been developed to account for both transitions and codon-usage biases (Yang and Nielsen, 2000). Distance based selective pressure methods make it difficult to extract information on the exact sites contributing to positive selection as a weighted average is used, also as no phylogenetic information is used, they are unable to look at lineage-specific selective pressures.

The McDonald-Krietman (MK) test, first applied on the alcohol dehydrogenase (*Adh*) gene in *Drosophila melanogaster*, is a phylogeny-based method developed as an extension of the Hudson-Kreitman-Aguadé test (Hudson et al., 1987; McDonald and Krietman, 1991). As with other models discussed in this section, the MK test uses a neutral model of evolution as a null (Kimura, 1968), which assumes most differences between sequences are accumulated by genetic drift. McDonald-Krietman compared interspecific (between species) divergence (D) and intra-specific (within species) polymorphisms (P). Following the assumption that most mutations are neutral or strongly deleterious and that synonymous mutations are neutral, it would be expected that the ratio of dN/dS at polymorphic sites (P_n/P_s) would be approximately equal to the ratio of dN/dS interspecific divergent sites. In contrast, if positive selection is acting on a region, sites are rapidly driven to fixation which would inflate D relative to P, thus $dN/dS > P_n/P_s$. A weakness of the MK test is the assumption that deleterious mutations are constantly strongly selected against is not always true. Aome slightly deleterious mutations will contribute to P_n without interspecific fixation, thus decreasing the power to detect positive selection in the region. A solution to this would be to remove low frequency mutations, which would remove slightly deleterious mutations, but this has been shown to be sensitive to the arbitrary cut off value (Fay et al., 2001; Charlesworth and Eyre-Walker, 2008). As this method is population based it is not suitable for my analysis but offers an avenue to validate the fixation of positively selected candidates within species.

Likelihood-based methods assess the probability of the data evolving under specified models of evolution. Likelihood-based methods use likelihood ratio tests (LRTs) to determine the model of evolution that best describes the observed data from a set of nested models (Anisimova et al., 2001). Assessing selective pressure variation though likelihood-based methods can be achieved using evaluation on a site-by-site basis, site-wise likelihood-ratio method (e.g., Massingham and Goldman, 2005), or on a whole sequence basis, using information from all sites within a sequence (Nielsen and Yang, 1998).

Codon based models of evolution consider the molecular evolution at a codon level, attempting to account for physiochemical properties of the encoded amino acids and thereby provide a more realistic interpretation of evolutionary patterns than amino acid models (Arenas, 2015). The Markov models of codon substitution, initially proposed by Goldman and Yang (1994), were developed to allow heterogeneous variation of ω values across sites (Nielsen and Yang, 1998; Yang et al., 2000) and branches in a phylogeny (Yang and Nielsen, 2002). CodeML within the PAML package (Yang, 2007) applies these models in a likelihood framework and combined with *post hoc* Bayesian analyses calculate probabilities of selective pressure variation across sites in a multiple sequence alignment (MSA). Defining branches (which can either be tips or monophyletic groups) as 'foreground', ω values in specified lineages of interest can be compared to that of background lineages.

The codon-based models of evolution are nested, with each model differing by a set number of parameters thus providing additional complexity. Heterogeneous codon models calculate ω across sites and are more sensitive than the distance models that incorporated weighted averages of synonymous rates (Nielsen and Yang, 1998). What follows is a brief description of the codon-based models implemented in Chapter 4 of this thesis and the LRTs that permit their comparison in a likelihood framework (Constantinides et al., *in prep*):

Site-specific models that available in CodeML (Yang, 2007), and described in Yang et al. (2000) are summarised in Table 1.1. The simplest model, model 0 (M0), assumes that all sites are evolving at an equal ratio ω, calculated as an average over the whole alignment. The model 3 (k=2) and model 3 (k=3), are discrete class models which are extensions of M0. Model 3 (k=2) allows for two unconstrained ratios, ω_0 and ω_1, over the sites, whereas Model 3 (k=3) accommodates an additional unconstrained ratio ω_3. Model 1a is a neutral model, where sites are partitioned into two classes, $\omega_0 = 0$ and $\omega_1 = 1$. Model 2a (Selection model) is an extension of model 1, adding one extra class, ω_2, which is unconstrained and may vary above

Table 1.1. Parameters of site codon models within CodeML (Yang, 2007). Models of site-specific variation in CodeML. ω is equal to dN/dS ratio over a MSA with ω_0 referring to purifying selection, ω_1 referring to neutral evolution and ω_2 referring to positive selection if > 1. The proportion of sites evolving under ω_i is denoted by p_i.

Model	Free parameters	Fixed parameters	Reference
Model 0	ω	n/a	(Goldman and Yang, 1994; Yang and Nielsen, 1998)
Model 1 (Neutral)	p_0, p_1 ($p_1 = 1 - p_0$)	$\omega_0 < 1$, $\omega_1 = 1$	(Nielsen and Yang, 1998; Yang et al. 2005)
Model 2 (Selection)	p_0, p_1, p_2($p_2 = 1 - p_0 - p_1$)	$\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$	(Nielsen and Yang, 1998; Yang et al. 2005)
Model 3	p_0, p_1, p_2($p_2 = 1 - p_0 - p_1$)	ω_0, ω_1, ω_2	(Yang et al., 2000)
Model 7 (beta)	n/a	p, q	(Yang et al., 2000)
Model 8 (beta & ω)	p_0, p_1 ($p_2 = 1 - p_0 - p_1$)	$\omega_s > 1$	(Yang et al., 2000)
Model 8a (beta & ω)	p_0, p_1 ($p_2 = 1 - p_0 - p_1$)	$\omega_s = 1$	(Yang et al., 2000)

1. Model 7 beta models allow ω to be beta distributed β(p,q), where p = 0 and q = 1. Model 8 (beta + ω > 1) is an extension of Model 7 and has a user-defined number of categories to approximate beta but is extended to include an additional class where ω > 1. Model 8a (beta + ω = 1) is the null hypothesis of model 8 where the additional category is set to be neutral, ω = 1.

The site-specific codon models in Table 1.1 were developed to test selective pressure variation across sites in an MSA (Goldman and Yang, 1994; Yang and Nielsen, 1998; Yang et al., 2000; Yang et al. 2005). Branch-site (Table 1.2) models allow ω to vary across sites and across branches in a phylogeny, identifying sites under positive selection lineages of interest (foreground). Model A, which is an extension of Model 1a, assumes two site classes are the same in both background and foreground lineages, with site class 0 containing $\omega_0 < 1$ and site class 2 containing $\omega_1 = 1$, but allows for an additional class, 2a and 2b. For 2a the foreground $\omega_2 > 1$ and the background $\omega_0 < 1$ and foreground, whereas 2b has a foreground $\omega_2 > 1$ and background $\omega_0 = 1$, and so allow for positive selection in the foreground and neutral or negative selection in the background. Model B is an extension of model 3 (K=3), where site classes were estimated from the data rather than set to $\omega_0 < 1$ and $\omega_0 = 1$. Model B did not test well during simulations with a much greater proportion of false positives (Zhang, 2004). Model A null was developed as the null hypothesis of model A, which differs from model A as the additional class is set as $\omega_2 = 1$, assuming neutral evolution in the foreground branch. The model A and model A null models have been shown to perform better than the previous model A and model B tests (Zhang et al., 2005) and will be applied in Chapter 4.

To reduce the risk of reporting results from a local, rather than global maxima multiple starting omega values are used from 0-10 as in previous publications (Yang et al. 1998, Yang 1997, Webb et al. 2017, Hyland et al., 2021). To assess the goodness-of-fit of each model to the data an LRT is performed between nested models (Nielsen and Yang 1998, Yang et al. 2000). Models are compared between complex parameter rich models and their simpler parameter sparse counterpart, with differences between the models being only the number of parameters. LRTs are calculated by twice the difference of the log likelihoods of both models (2Δl), followed by a χ^2 distribution. Degrees of freedom are derived from the number of additional parameters in the more complex model, if the 2Δl is greater than the critical value the more complex model is determined as significant (Table 1.3).

Table 1.2 Lineage specific models used within the CodeML analysis. Adapted from Yang, (2007).

Model	Parameters	Foreground	Background
M1Neutral	$p_0 : \omega_0 < 1$	n/a	n/a
modelA	$p_0 : \omega_0$, $p_1 : \omega_1$, $p_2 : \omega_0, \omega_2$, $p_3 : \omega_1, \omega_2$	$\omega_2 > 1$	$0 < \omega_0 < 1$, $\omega_1 = 1$
modelAnull	$p_0 : \omega_0$, $p_1 : \omega_1$, $p_2 : \omega_0, \omega_1$	$\omega_1 = 1$	$0 < \omega_0 < 1$, $\omega_1 = 1$

Table 1.3 Likelihood ratio test summary for models implemented in CodeML. *Comparison of nested models, degrees of freedom (df), multiplication of amount of difference in the lnL scores between models (Δl), and critical χ^2 values for each comparison. Adapted from Morgan et al. (2012).*

Comparison	df	Δl	Critical χ2 values
Model 0 vs Model 3 (k=2)	2	X2	≥ 5.99
Model 3 (k=2) vs Model 3 (k=3)	-	X1	≥ 1.00
Model 1a vs Model 2a	2	X2	≥ 5.99
Model 7 (beta) vs Model 8a (beta & ω)	2	X2	≥ 5.99
Model 8 (beta & ω) vs Model 8a (beta & ω)	1	X2	≥ 2.71 (0.05 significance)
Model 1a vs Model A	2	X2	≥ 5.99
Model A vs Model A null	1	X2	≥ 3.84 (0.05 significance)

The posterior probability (PP) of each specific site within the positively selected category is calculated using empirical Bayes estimates (EB) (Yang et al., 2005). Naïve empirical Bayes (NEB) estimates (Yang et al. 1998) and Bayes empirical Bayes (BEB) estimates (Yang et al. 2005) are both calculated. NEB estimates are susceptible to Type I errors in small datasets (Anisimova et al., 2002), and so BEB estimates are an improvement. BEB estimates assign a prior probability of observing positive selection to the model parameters, and integrates over their uncertainties (Yang et al., 2005).

Past studies using branch-site models have used tests to correct for multiple testing calculated in tests such as Benjamin-Holberg false discovery rates (FDR). It has since been argued that these tests are inappropriate for branch-site models, as these approaches assume the null model has uniformly distributed P values when true (Strimmer, 2008; Noble, 2009), and the distribution for the P values of the favoured null genes are rarely investigated. In the Model A model, it is assumed that a proportion of sites will be under positive selection in the foreground branch (p2: $\omega > 1$), whereas in the null model the same proportion of sites will evolve neutrally ($\omega = 1$). In reality, some sequences will be under strict purifying selection and have no proportion of sites neutrally evolving, leading to misspecification of the model. In a study from Potter et al. (2021), simulations of specified numbers of false positives were used to determine that a blind application of FDR tests is not appropriate for these tests that have already been deemed very conservative.

In Chapter 4 I use branch-site models of across protein coding regions of 5 lineages of pinniped to attempt to identify sites under positive selection, the information in this section provides the background to the theoretical basis of that analysis and its limitations.

1.5 Building Phylogenies from Genome Scale Datasets

Analyses, such as selective pressure variation discussed in section 1.2, are dependent on accurate and detailed relationships being known between the species in question. Phylogenetic trees are mathematical representations of evolutionary relationships and are named as such due to the tree-like structure that evolutionary relationships theoretically follow. Etymological features of phylogenetic trees continue along same theme with features of a phylogenetic tree named: root, branches, and leaves (Figure 1.7). The base of a phylogenetic tree, the root, being the most distant evolutionary ancestor of all the taxa in a tree. Each taxon is named as a leaf, or operational taxonomic unit (OTU), and relationships to other leaves in a tree are dependent on the evolutionary branch on which they evolve with the resulting formation referred to

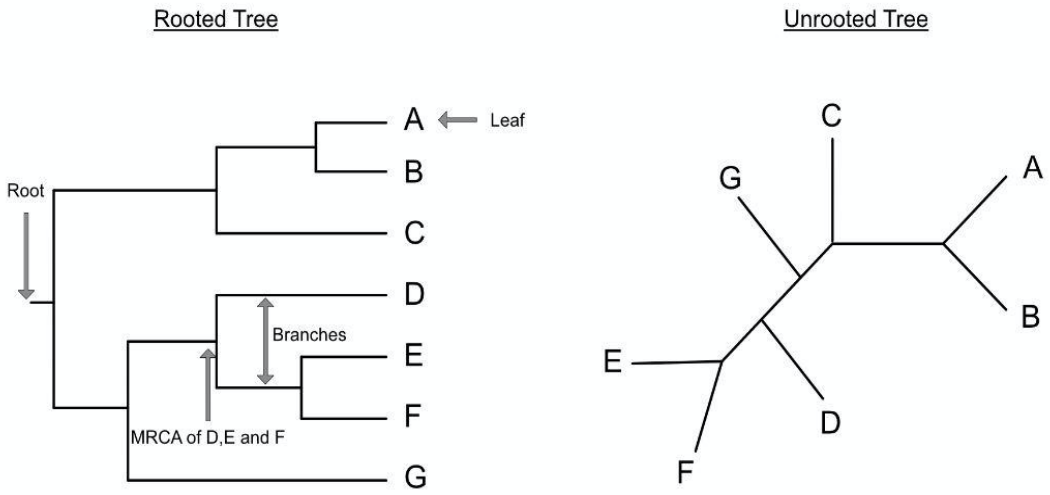


Figure 1.7. A comparison of unrooted and rooted tree. The figures show the same topology but in a rooted and unrooted tree. In this diagram D, E and F are an example of a monophyletic group/ clade.

as the tree's topology. The points at which branches split are referred to as nodes, taxa all resulting from a single node are named monophyletic groups in rooted trees and clans in unrooted tree (Wilkinson et al., 2007). The node for a monophyletic group represents the most recent common ancestor (MRCA) of the group (Figure 1.7), unrooted trees have no such ancestral node. Rooted trees can be very useful in determining directionality of patterns and processes across the tree, with branch lengths from the root either proportional to evolutionary rates or time, calibrated using fossil data, these rooted trees are called phylograms. Phylogenetic trees have been used since the origin of the evolutionary biology field and remain an integral tool.

1.5.1 Phylogenomic principles and practices

1.5.1.1 Data types in molecular reconstruction

Prior to the availability of molecular data such as DNA and protein sequencing, phylogenetic inference was dependent on the analysis of morphological comparisons. Using discrete characters to draw relationships between different species is an important resource in assessing phylogenetic relationships alongside molecular data (Li and Palci, 2015). Phylogenetic estimates though morphological traits can offer advantages of molecular data, especially in a deep-time levels (Keating et al., 2020). Their ability to integrate fossils, which can help calibrate molecular clocks and break long branches (Weins, 2001). However, molecular analyses offer an advantage as they can analyse many more discrete characters and incorporate empirically derived evolutionary models. It is widely regarded that no single data type contains all the information needed to comprehensively reconstruct phylogenetic trees or phylogenetic networks and the use of morphological and molecular datatypes are essential (Zhou and Zhang, 2016).

Molecular phylogenies, for a long time, were dependent on data from small fragments of DNA in the form of expressed sequence tags (ESTs) or partial sequences of mitochondrial genes. As methods to generate molecular data increased, mitochondrial genomes and complete genes been used in phylogenetic analyses (Springer et al., 2004; Chan et al., 2013; Foley et al., 2016; Prum et al., 2015). The advent of rapid and large-scale sequencing has significantly facilitated the growth of genomic data available to phylogenetic analyses, even for non-model organisms bringing us now firmly within the "phylo-genomic era". Although, the wealth of data has increased, complications including long branch attraction and incorrect orthologous assignments are still apparent (covered in more detail later in this section). The use of coding regions reduces the complexity in inferring homology and constructing alignments across

species, but the format of these coding regions can be dependent on the timescales being analysed. Nucleotide sequences are often used as they provide more phylogenetic information than amino acid sequences, with non-synonymous changes being apparent between samples. When dealing with deeper timescales amino acid data is commonly used, which removes noise produced by rapidly evolving synonymous sites (Zwick et al., 2012).

1.5.1.2 Taxon sampling

Limiting the number of taxa reduces the computational requirements of the analysis, whilst also reducing occurrences of common characters being shared across species by chance, rather than common descent (i.e., homoplasy) (Brandley et al., 2009). A denser sampling of taxa is well known to increase support for branch topology. Thus, increasing the accuracy of phylogenetic inferences, with support for monophyletic clades being drastically reduced or non-existent when sampling is reduced (Zwickl and Hillis, 2002; Philippe et al., 2009). Phylogenetic inference methods produce unrooted trees, and the inclusion of additional taxa with a known relationship to the clade of interest is then necessary to add directionality to the phylogeny. These 'outgroup' taxa can be an important addition to the analysis as they provide the power to define the root of the phylogeny. Reduced sampling can have downstream effects on orthologous sequence calling and alignments, but even deep taxonomic sampling may not be able to resolve short internal branches with low phylogenetic signal (Phillipe et al., 1994). To summarise, taxa sampling phylogenetic analysis requires considerable attention, this is especially true in genome scale analyses where computational resource and time is critical and systematic errors committed can accumulate and deliver incongruent results (Jeffroy et al., 2008).

1.5.1.3 Orthologous Groupings

A crucial assumption when building species phylogenies is an understanding of homologous relationships between the sequence data. Homologous sequences between species are derived either as a consequence of speciation (i.e., orthologs), or a duplication event (i.e., paralogs) (Fitch, 1970). The identification of orthologs between species and the distinguishing of orthologs from paralogs is imperative in phylogenetic comparative analyses (Koonin et al., 2005; Kristensen et al., 2011) as, unlike orthologs, paralogs do not reflect the evolutionary relationships across taxa (Figure 1.7).

Homology prediction tools can be either tree based or graph-based tools. Graph based orthology identification tools are the most popular used in phylogenetic inference analyses and are based on an

underlying assumption that orthologues across species have a higher similarity than any other gene in either species (Overbeek et al., 1999). Graph based methods use all-vs-all pairwise sequence comparisons, using tools such as BLAST (Altschul et al., 1990) or DIAMOND (Buchfink et al., 2015 which perform sequence similarity searches across all sequences and cluster sequences accordingly. Tree based methods initially use graph-based methods, with relaxed constraints, to infer relationships between species. Then assigning orthologs and paralogs through comparisons to the species topology (Trachana et al., 2011). The main limitation to tree-based search methods is that they require a high confidence species phylogeny, which is not always available (Gabaldón, 2008).

Increasing gene number has been argued to be key in resolving some long-standing phylogenetic issues. Although, complex issues can arise when gene duplications and losses occur in orthologous groupings, leading to ostensibly paralogous relationships, even when only single species representatives of an orthologous group are present (single copy genes) (Rokas and Carroll, 2005). Paralogs, homologous genes that have diverged as a consequence of duplication and speciation, can be misinterpreted as orthologs in phylogenetic analyses, confounding the phylogenetic signal and leading to erroneous conclusions. This 'hidden paralogy' can drive alternative topologies within phylogenies (Dolittle, 1999). To address the issue of hidden paralogy in this thesis, 'clan-check' was used to filter large scale gene sets for species phylogeny reconstruction (Siu-Ting et al., 2019). Clancheck is built on the understanding that groups of taxa should always group together in true orthologous gene trees due to their evolutionary history, and these 'clans' (Wilkinson et al., 2007), should group away from other species. 'clan' based paralog filtering approach has been used to attempt to resolve some of the conflicts within the Lissamphibia (a group of jawed invertebrates comprising three orders of amphibia) (Siu-Ting et al., 2019). Ancient amphibia have been known to be susceptible to whole genome duplication events, which exacerbates the detection of truly orthologous relationships between genes from different species (Mable et al., 2011). In the analysis, any gene trees that had species which violated 'incontestable' clans, such as mammals, were removed. Using this filtering method, Siu-Ting et al., (2019) were able to find a consensus tree and showed that support for alternative topologies was being driven by paralog inclusion.

1.5.2 Phylogenetic Reconstruction Using Molecular Datasets

1.5.2.1 Substitution Models

Substitution models of evolution describe how the composition of the sequences changes and its rate of change over time. Substitution models can be applied to nucleotide bases or amino acid residues and consist of exchange rate frequencies and the base frequencies in a composition vector. The first substitution models were nucleotide based with the simplest being the Jukes-Cantor model (JC69) (Jules and Cantor, 1969) which considered all exchange rates and all base frequencies to be equal. This model was extended by Kimura's K2P model (Kimura, 1980), which accounted for bases with equal chemical structures (transitions) having different rates than that of changes of bases with alternative chemical structures (transversions). Felsenstein (1981) extended the JC69 model to include unequal base frequencies.

Since then, models have developed through extensive parameterisation, and thereby increased complexity. The general time reversal (GTR) model (Tavare, 1986) for example, estimates base frequencies from the data and assigns rates to each possible nucleotide substitution, this model assumes that these base frequencies and rates of change are consistent across the dataset. Models have also been developed to model the evolution of amino acids, with the simplest example being the Poisson model (Bishop and Friday, 1985). This is an extension of the JC69 model but designed for the 20 different amino acid states rather than four nucleotides. Dayhoff (1968) developed an amino acid model that could account for the probability of changes of amino acids, this led to many empirical amino acid substitution models. Currently, the most frequently used amino acid substitution models are empirical. An example of a widely used empirical amino acid substitution method is the Jones, Taylor and Thornton model (JTT) (Jones et al., 1992), which is based on transmembrane proteins. Frequently employed nucleotide and amino acid models are listed in Table 1.4, along with the data from which they were calculated. Empirical analyses showed that substitutions per site do not follow a Poisson distribution, which would be expected if all sites evolved at the same rate across a sequence (Fitch and Margoliash, 1967). To account for associated site rate variation (ASRV) the +I parameter was introduced to account for invariable sites (Reeves, 1992) and gamma distributed rate variation (Γ), was more representative of biological data (Yang, 1994; Yang, 1996).

Table 1.4. Nucleotide and amino acid substitution. Adapted from IQ Tree (Minh et al., 2020)

Model	Data derivation	Reference
Nucleotide models		
JC/JC69	Equal substitution rates and equal base frequencies	*Jukes and Cantor, 1969*
F81	Equal rates but unequal base frequencies	*Felsenstein, 1981*
K80/K2P	Unequal transition/transversion rates and equal base frequencies	*Kimura, 1980*
HKY/HKY85	Unequal transition/transversion rates and unequal base frequencies	*Hasegawa, Kishino and Yano, 1985*
TN/TN93	Like HKY but unequal purine/pyrimidine rates	*Tamura and Nei, 1993*
K81/K3P	Three substitution type model and equal base frequency frequencies	*Kimura, 1981*
SYM	Symmetric model with unequal rates but equal base frequencies	*Zharkikh, 1994*
GTR	General time reversible model with unequal rates and unequal base frequencies	*Tavare, 1986*
Amino acid models		
Blosum62	BLOcks SUbstitution Matrix	*Henikoff and Henikoff, 1992*
cpREV	chloroplast matrix	*Adachi et al., 2000*
Dayhoff	General matrix	*Dayhoff et al., 1978*
DCMut	Revised Dayhoff matrix	*Kosiol and Goldman, 2005*
FLU	Influenza virus	*Dang et al., 2010*
HIVb	HIV between-patient matrix HIV-Bm	*Nickle et al., 2007*
HIVw	HIV within-patient matrix HIV-Wm	*Nickle et al., 2007*
JTT	General matrix	*Jones et al., 1992*
JTTDCMut	Revised JTT matrix	*Kosiol and Goldman, 2005*
LG	General matrix	*Le and Gascuel, 2008*
Poisson	Equal amino-acid exchange rates and frequencies	*Bishop and Friday, 1985*
PMB	Probability Matrix from Blocks, revised BLOSUM matrix	*Veerassamy et al., 2004*
rtREV	Retrovirus	*Dimmic et al., 2002*
VT	General ‘Variable Time’ matrix	*Mueller and Vingron, 2000*
WAG	General matrix	*Whelan and Goldman, 2001*

Homogeneous models, like those described in Table 1.3, are useful due to their short running times or ability to make use of small sequences. However, heterogeneous models of increased complexity account for rate of change of sites between lineages (rate heterogeneity) and sites within a sequence or among lineages that have biases towards certain residues (compositional heterogeneity), and as such are more biologically realistic. The CAT model, implemented within Phylobayes (Lartillot and Philippe, 2004), is an infinite mixture model that accounts for heterogeneity within the sequence. Using the CAT model each site or column within an alignment possesses a profile, which is estimated from the data. This profile is a probability vector which is defined by a simple replacement process, so that for each substitution possible there is a corresponding probability factor. The likelihood at each site or column of the alignment is then averaged over all possible processes within a profile. Using mixture models, such as CAT, with other homogeneous models such as GTR can greatly increase performance of phylogenetic reconstructions, by reducing incongruence caused by long branch attraction and sequence saturation (Philippe et al. 2009, 2011; Roure et al. 2013).

1.5.2.2 Methods of Phylogenetic Inference

Resolving evolutionary relationships from aligned sequences can be performed using methods from two categories: distance-based methods or character-based methods. Distance based methods involve calculating distances across all sequence pairs in an alignment, before clustering them based on the distance matrix, forming a resolved phylogeny (Kapli et al., 2020). Character based methods, such as Maximum likelihood (ML) and Bayesian inference (BI), compare all sequences in the alignments, but on a site-by-site basis calculating the probability score of each possible tree. ML methods of inference are implemented in many current phylogeny tools (Yang, 2000; Stamatakas, 2014; Minh et al., 2020). The underlying methodology of Maximum Likelihood inference is to optimise the branch lengths when given a substitution model and alignment , to produce a topology that has the Maximum Likelihood value (Nei and Kumar, 2000). Bayesian inference (BI) is similar to ML methods in that it uses a likelihood function but uses prior information to quantify uncertainties in parameters. Each possible tree (referred to hereafter as hypotheses) is used with the alignment and substitution model as a prior probability of the hypothesis and is then tested against the probability of observing all other hypotheses. Both BI and ML have high computational requirements and the appropriate choice of whether to use a ML or BI approach to phylogeny reconstruction is dependent on the phylogenetic question at hand. Regardless of the approach used a statistical analysis should be employed to test the confidence of resulting phylogenies.

From an aligned set of sequences, it is possible to analyse the data using a supermatrix approach, which concatenates all the sequences into a single sequence, or one can treat each sequence or gene as independent alignments which are allowed to have their own set of parameters, i.e., a supertree approach (Yang, 2006). The supermatrix alignment has increased length in comparison to individual gene alignments, and these long alignments are ideal for estimating model parameters in heterogeneous models such as CAT mixture models in Phylobayes (Lartillot and Philippe, 2004). Heterogeneous models can reduce the effects of LBA (Lartillot et al., 2007) but lack the ability to model evolutionary dynamics amongst different individual genes, which can result in incorrect or conflicting topologies (Scornavacca and Galtier, 2017). The supertree approach performs phylogenetic reconstructions using multiple evolutionary models at the individual gene level, heuristic algorithms are then used to assemble these 'subtrees' into a consensus tree (Bininda-Emonds, 2004; Yang, 2007).

Supertree analyses are useful in investigating potential artefacts within the sequence data and trees, such as incomplete lineage sorting (ILS). Incomplete lineage sorting is a phenomenon that occurs when there is sufficient genetic variability in an ancestral species that some variability continues to persist beyond the speciation event and ancestral polymorphisms are not 'sorted' prior to additional speciation events (Figure 1.8) (Avise, 1994). ILS is more common in populations that have large ancestral population sizes and short internal branches, e.g., adaptive radiation events in short timeframes (Maddison, 1997). The multispecies coalescent (MSC) model can be implemented which allows coalescence events to occur after speciation events and builds a probability distribution within gene trees for a species tree (Rannala et al., 2020). The debate over implementing a supermatrix or supertree approaches has long been discussed (Bryant and Hahn, 2020), the consensus being that supertree approaches reduce ILS, but supermatrix approaches are favoured when investigating deeper phylogenetic timescales, where the power of ILS is reduced (Bryant and Hahn, 2020). As both approaches have strengths and weaknesses, studies often apply both supertree and supermatrix approaches and test for congruence between the resulting phylogenies (Fernández et al., 2018; Jebb et al., 2020; Tarver et al., 2016).

1.5.2.3 Measuring confidence intervals on a phylogeny

Statistical approaches have been well established to determine if the variables of the substitution model (including as rates of change, composition, and branch lengths) are of adequate fit to the data and the support for the resulting topology. The confidence intervals provide a measure of support for each split in the tree and include bootstrapping and jack-knifing and for BI, posterior probabilities. Bootstrapping is

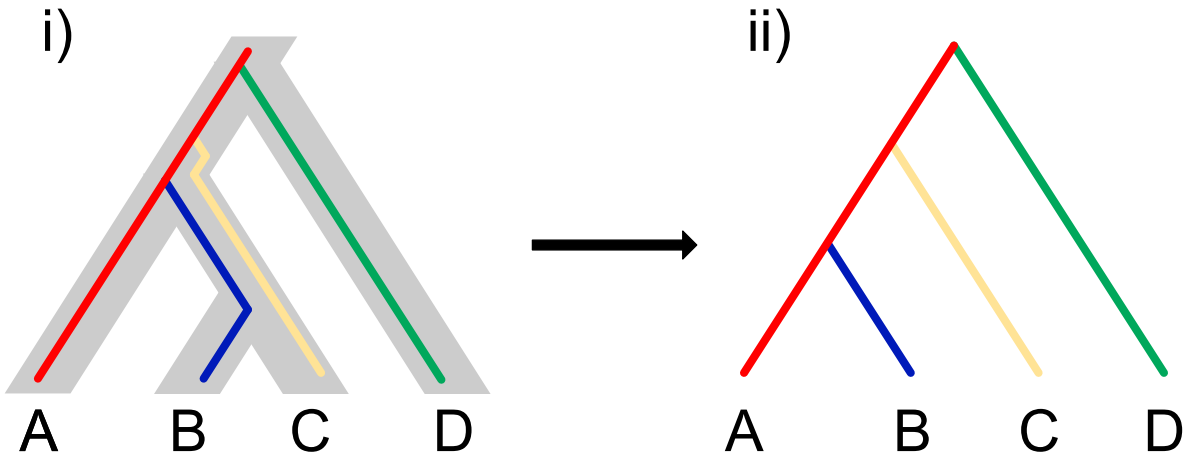


Figure 1.8. Incomplete lineage sorting representation. i) the grey bars show the true species phylogeny between the taxa and the coloured lines represent polymorphisms arising. Two polymorphisms are maintained in the population prior to the split between A, B and C. These two alleles are maintained in the lineage of B and C until their speciation event. ii) the resulting gene tree resolves A and B as sister species, although the B and C are true sister species. Adapted from Pollard et al. (2006).

a computational technique that estimates a statistic for which the underlying distribution is unknown or difficult to derive (Efron, 1982). It has been a common practice in evaluating the support of splits in the topology when using ML. Bootstrapping is a resampling technique in which the support of each split is assessed by resampling with data from the original dataset, with the support for a node derived from the percentage of bootstrap replicates that support the inferred split. Many phylogenetic studies use a cut-off of 70% as an indicator of sufficient support, derived from Hillis and Bull (1993), although bootstrap support can be difficult to interpret (Efron et al., 1996; Felsenstein and Kishino, 1993; Susko, 2009). Confidence levels of support for clades in BI are posterior probabilities calculated from the underlying Markov chain Monte Carlo (MCMC) simulations. Posterior probabilities thus represent the probability of a specific clade given the model and data; however posterior probabilities are sensitive to model misspecifications. Simple models can result in overinflated posterior probabilities, and incorrect topologies receiving support values near 100% (Huesenbeck and Rannala, 2004; Yang and Zhu, 2018).

As discussed, BI are very sensitive to the models on which they are based, thus tests to assess the confidence of underlying models against the data are crucial for infinite mixture models, like CAT, as the models are inferred directly from the data. Composition bias occurs due to evolutionary models assuming species with compositional homogeneity, similar frequencies of residues, are more closely related. For instance, taxa with adenine/thymine-rich regions would be incorrectly grouped (Kapli et al., 2020). Methods have been developed to account for compositional heterogeneity in the data, although these can be very parameter rich and computationally expensive. χ^2 tests are built into certain phylogenetic tools to test that the model is an adequate fit for the data, but these can suffer from high Type II error rates, an alternative method is posterior predictive analysis (PPA). PPA simulates data from a model using posterior predictive estimates from the MCMC, tests on the simulations are then used as a null against which the statistics of the empirical data are compared (Lartillot et al., 2007). An additional approach is to reduce the character set, i.e., reducing AT bias by recording A and T bases to purines and C and G to pyrimidines, although this approach can introduce additional biases through the removal of phylogenetic information (Susko and Roger, 2007).

If phylogenetic analysis has led to alternate phylogenies being produced it is possible to compare the topologies to determine whether a single topology is significantly better than the others. Initially, Kishino and Hasegawa (1989) devised a parametric test on the log likelihood scores of 2 trees using Z-scores (KH test). The KH test could only test two trees and did not account for corrections for multiple testing

(Goldman et al., 2000), therefore the SH test was introduced to allow the comparison of several trees (Shimodaira and Hasgwa, 1999). It was then noted by Strimmer and Rambault (2001) that the SH test was extremely conservative when testing large numbers of trees, and so in response the approximately unbiased (AU) test was derived (Shimodaira, 2002). The AU test compares bootstrap probabilities (BP) to calculate a P-value from the change in BP values across the topologies, which are then used to attempt to reject the alternate topology (Shimodaira, 2002).

In Chapter 3 I perform phylogenetic analyses using Maximum Likelihood and Bayesian inference methods to generate a single species phylogeny for Pinnipedia, the information presented above provides the background to my choice of methods and approach.

1.6 Genome Assembly and Annotation

All the methods highlighted so far in this section require DNA sequences from coding areas, following on from next generation sequencing technology revolution, many groups and consortia have emerged with the goal to sequence all of life (Hotaling et al., 2021). The catalogue of publicly available data is now vast, enhancing genomic reliant analyses. In this section, a brief history of genomic sequencing is covered along with the different methods that allow next generation sequence data to be transformed into usable gene models, thus providing the context for the sequencing, assembly and annotation approaches taken for the Caspian seal and Hooded seal in Chapter 2.

1.6.1 From Sample to Data

The first instances of determining the bases of a DNA molecule were shown in the publication of the Sanger sequence method (Sanger and Coulson, 1975) and the subsequent publication of the first complete genome of a bacteriophage φX174 (Sanger et al., 1977). Using chain terminating dideoxynucleotides, this method is the foundation for the developments that would enable future genome sequencing projects. Advancements in methods, instrumentation, and automation (Lander et al., 2001) significantly reduced the price of a genome assembly culminating in the initiation of the Human genome project in 1990 (Lander et al., 2001; IHGSC, 2004). The completion of the human genome project in 2003 forged advancements in the genomics field, with the early 2000s giving rise to a multitude of next generation sequencing (NGS)

technologies that were able to produce large quantities of data at significantly lowering costs (Goodwin et al., 2016).

Sequencing technologies applied different methods to produce the data but the most successful of these has been the sequencing-by-synthesis (SBS) approach deployed by Illumina sequencing platform. Despite the high accuracy and yield, one caveat of Illumina sequencing technologies is that of the read length obtained through sequencing. The read lengths are limited to 150 nucleotides per read, this can cause complications when performing *de novo* assemblies, especially in repeat-rich regions (Levy & Myers, 2016). Mate pair sequencing has been used in combination with paired end reads to counteract this caveat, using large insert sizes (the part of the DNA fragment between the sequence reads), in the magnitude of kilobases of base pairs it is possible to span repeat rich areas. Repeat rich areas can be larger than the maximum insert sizes and so can still result in highly fragmented regions (Wetzel et al., 2011).

In recent years, long read sequencing, or 3rd generation sequencing technologies, have become available. The prominent commercial entities behind this technology are Oxford Nanopore Technologies (ONT) and Pacific BioSciences (Pac-Bio), providing the opportunities to sequence genome fragments up to 2,300,000 base pairs (Payne et al., 2018). The methodologies behind ONT and PacBio systems differ slightly, although the error rate involved in base calling in these technologies has been their key drawback. In recent years, improvement of Nanopore base callers (Zhang et al., 2020) and the development of PacBio HiFi reads have meant that read lengths of tens of kilobases with error rates lower that 0.5% are now feasible (Hon et al., 2020). Long-range sequencing techniques have been developed to support both short and long-read sequencing projects. Long-range technologies use proximity of fragments within chromatin or barcoded large fragments of DNA to span megabases, becoming advantageous in the development of chromosomal level assemblies (Lieberman-Aiden et al., 2009).

The revolution in sequence technologies in combination with greatly reduced costs have allowed extensive applications of genomic sequences to be applied to non-model organisms, leading to ambitious whole genome consortiums, setting out to sequence, assemble and annotate the genomes of entire phyla (Koepfli et al., 2015; Lewin et al., 2018; Rhie et al., 2020). The storage and analysis of the data is now becoming the limiting factor in the generation of biological data (Papageorgiou et al., 2018; Hodcroft et al., 2021). Producing high quality genome assemblies from non-laboratory grown organisms is more routine, but complications concerning heterozygosity and attaining quality tissue samples in large enough

quantities have led to innovative advancements in genome assemblies (Koren et al., 2018; Oppert et al., 2019).

1.6.2 Generating Gene Models

1.6.2.1 De novo genome assembly

De novo genome assembly is a challenging task which involves assembling short fragments of DNA sequences (reads) into long contiguous sequences (contigs), which at the optimal size will be equivalent to chromosomes. Generating chromosome length contigs has recently become achievable due the decreasing costs of long-read sequencing and long-range sequencing, even for eukaryotic species with complex karyotypes (Du and Liang, 2019; Dudchenko et al., 2020). The complexity of non-model organisms has further exacerbated the problem, with polyploidy, repetitive rich regions and complex rearrangements generating uncertainties, which have led to specially devised algorithms with increased computational demands (Ott et al., 2018; Nowoshilow et al., 2018). Most eukaryotic species have genomes with large repetitive regions, mammals have repetitive regions spanning 25-50% of the genome, and it is these regions that are difficult to assemble when the repetitive region spans longer than the sequence read length, this is especially prominent at centromeres and telomeres (Miga et al., 2020). Graph based approaches (De Bruijn graph and Overlap-Layout-Consensus) use overlapping reads to assemble contigs and most assembly tools employ these strategies as they have been shown to be the most accurate and efficient (Miller et al., 2010).

In early genome builds, such as the human genome project, Overlap-Layout-Consensus (OLC) method was applied. OLC consists of three stages: overlaps between reads are identified, layout of all the reads and overlaps are placed into an overlap graph, and finally a consensus sequence is determined from the MSA of overlapping sequences (Idury and Waterman, 1995). An overlap graph (Figure 1.9) is produced from all-vs-all pairwise comparisons, in which the nodes represent reads, and edges represent overlaps that share K-mers between reads. The optimal path through the graph is derived using a Hamiltonian path, this is a NP-hard problem, and heuristic or greedy algorithms are used which are intensive. MSAs are then formed to collapse the overlapping reads into a consensus read. The OLC approach was initially developed to deal with long Sanger reads, as short, accurate Illumina based reads became popular, the low overlapping k-mer lengths and high computational requirements of OLC methods gave preference to De Bruijn graph approaches.

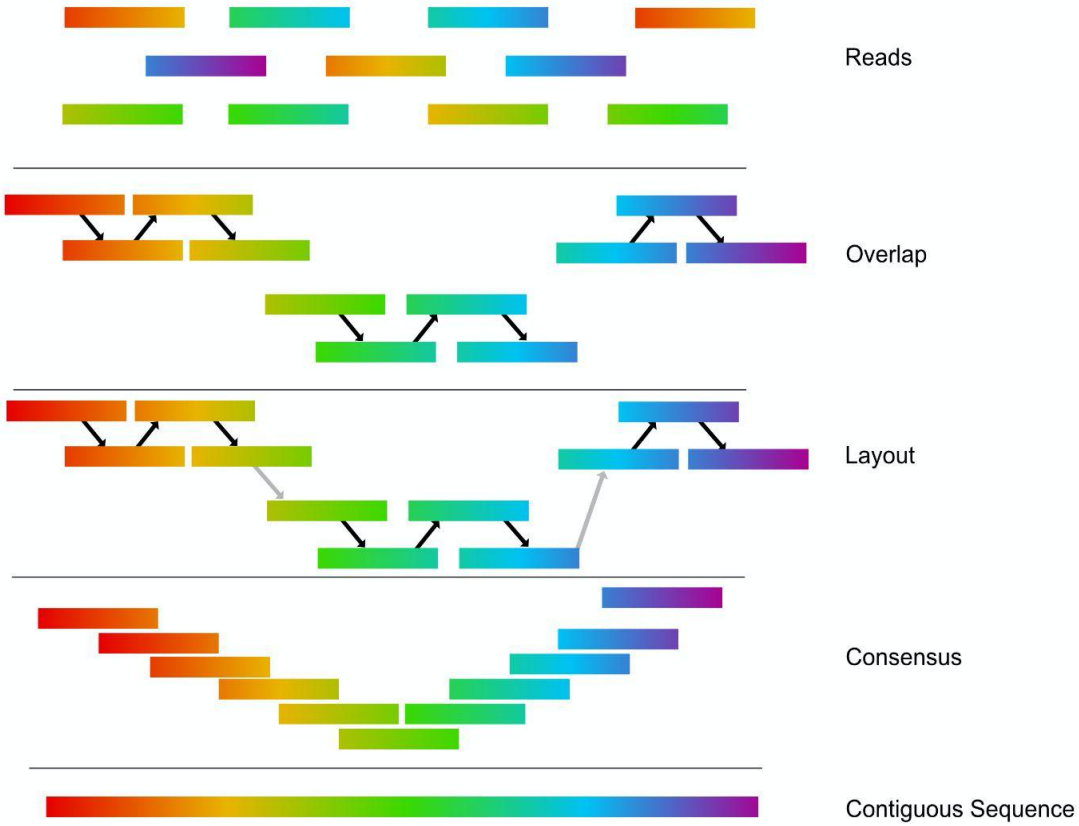


Figure 1.9. Overlap Layout Consensus method. Sequencing reads are clustered together using similarity measures, all-vs-all alignments are performed to find overlaps between the different reads within clusters converting into a graph (Overlap). A Hamiltonian path is used to find the most efficient pathway through the graph (Layout) before multiple sequence alignments are produced to resolve a consensus sequence.

The De Bruijn graph method uses kmer (short read substrings of length k) graphs (Figure 1.10), this avoids the all-vs-all pairwise comparisons necessary in OLC reducing computational load. Within a De Bruijn graph, nodes are represented by fractions of reads of length k, and each node is connected if they contain an identical sequence of k-1, avoiding the need for each read to be saved into memory. The value of k is directly consequential on the contigs outputted by this approach, if k is too low then small repetitive regions will not be able to be passed, if k is too large the graph will be disconnected resulting in many fragmented subgraphs. Tools such as KmerGenie (Chikhi and Medvedev, 2013) have been developed to assess the optimal kmer size for a given genome. Once the graph is produced Hamiltonian or Eulerian path methods are used to traverse the graph, generating contigs. Bubbles in the graph can be a major issue in De Bruijn graphs, these can be caused by heterozygous alleles in the data, causing alternative paths to be found. The sequence type and genome size are important factors to consider when choosing an assembly algorithm. OLC is more suitable for low coverage long reads and can consume large computational resources for larger genomes, whereas de Bruijn methods are preferable for high-coverage short reads and reduce computational resources (Li et al., 2011).

1.6.2.2 Gene annotation

To fully extract the information in a genome and compare across species, genes and other features of the genome need to be annotated. Many tools have been developed to annotate the different features of a genome, in this section the two facets of genome annotation most relevant to this thesis are covered: annotation of structures using *ab initio* trained and homology-based gene model prediction tools, and functional annotation of the features.

Structural annotation involves the identification of features of DNA from sequences, e.g., introns, exons, and promoters etc. Structural annotations were originally viewed as the annotation of protein-coding genes (Ejigu and Jung, 2020), but as the definition of “a gene” has evolved so to have the structural features. The prediction of these additional features such as noncoding genes (genes that code for RNA molecules) and pseudogenes, have been critical in expanding my understanding of genomes (Alexander et al., 2010). Repetitive regions are particularly challenging for annotation and assembly of genomes (Dominguez Del Angel et al., 2018), on average mammal genomes are composed of approximately 50% repetitive regions (Nishibuchi et al., 2017). To overcome these issues, genomes are often subjected to repeat masking, where repetitive elements are identified within the assembled genome and edited out of the genome. Genome assemblies undergo hard masking (which nucleotides being replaced with an ‘X’ or

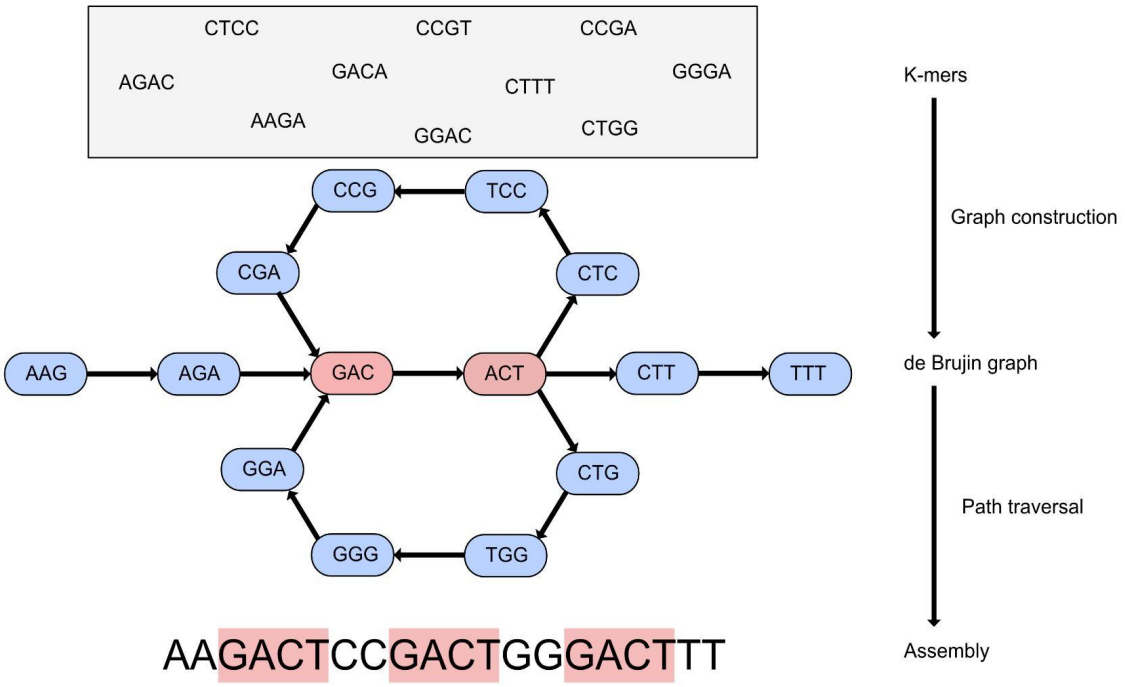


Figure 1.10. De Bruijn assembly workflow. Sequencing reads are broken down into k-mers (4-mers in this example), which are then used to produce a de Bruijn graph in which nodes are k-1mers and edges are formed from overlapping kmers. Eulerian or Hamiltonian paths are then used to find the most efficient path through the graph, combining this information results in the consensus sequence. Adapted from Berger et al. (2013).

'N') - or are softmasked (where nucleotides in repetitive regions are changed to a lowercase). Repeat identification tools generally rely on comprehensive databases of species-specific repeat sequences, DFam (Storer et al., 2021) and Repbase (Bao et al., 2015). However, repetitive elements lack conservation across species due to a lack of selective pressure (Hancock et al., 2001), and therefore there has been a need for tools that automate *de novo* repetitive element discovery (Flynn et al., 2020).

Pipelines have been developed for the automation of gene annotation with MAKER2 (Holt and Yandell, 2011) and BRAKER2 (Brůna et al., 2021) being popular tools. Both these tools use *ab initio*, RNA-seq, and homology-based steps that are used to predict gene models. *Ab initio*-based methods rely on gene predictions which previously acquired trained data, from closely related species, to construct models that can be then applied to the novel genome. Hidden markov models (HMM) are often used in these approaches to identify promoters, coding/ non-coding regions, and intron-exon junctions in the data. AUGUSTUS (Stanke et al., 2006) uses probability distributions from training data sets that are acquired either through conserved protein coding elements or a closely related species (Hoff et al., 2019). AUGUSTUS can be used with ESTs or RNA-seq, which can infer intron sizes with the sequences and assist with alternative splice sites (Ejigu and Jung, 2020). *Ab initio* tools have an advantage over other methods as they may identify models that possibly have low RNA-seq coverage or novel genes. In comparison calling gene models from closely related species which have existing annotations, will miss novel genes, although *ab initio* methods are susceptible to produce false positives. Homology-based tools are based on the underlying principle that protein coding sequences have retained conservation in positional homology between species. This allows a set of proteins from a closely related species to be used to identify homologs in the newly sequenced genomes. The reliance on the input data creates challenges when performing annotations with no closely related annotated species with no sequence alignments between the input data and novel genome. Thus, it is recommended to use various pieces of evidence to perform structural annotation, tools such as EVidenceModeler (EVM) has been developed to use multiple sources of genome annotation data with confidence weights to produce a consensus set of gene models (Haas et al., 2008). All automated techniques of annotation are susceptible to errors, and therefore many large genome consortiums are using manual annotation, in which all gene models are manually reviewed (Harrow et al., 2006).

Once gene models have been established an important process is to assess the function or biological association to that model. Comparison of sequence homology to heavily studied organisms, or databases,

can be performed using BLAST (Altschul et al., 1990) or DIAMOND (Buchfink et al., 2015), although alignments errors and misalignments, to unrelated regions of similarity, can result in erroneous classifications (Sasson et al., 2006). Novel protein sequences can be assessed using domains and functional sites, InterProScan (Jones et al., 2014) uses the InterPro database (Blum et al., 2021) to find overlapping information from protein families, domains, and functional sites.

It is important to assess genomes and their annotations to quantify quality, this is often done using BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão et al., 2015). This uses a deeply studied catalogue of single copy orthologs. The orthologs are then aligned against the gene models produced in the alignment and extrapolated to show completeness of the novel gene models produced. Annotation is an important step and essential in downstream phylogenetic analyses, although current pipelines for automation are still lacking. The advancement in long-read RNA-sequences (IsoSeq) is making this challenge easier and will improve the accuracy of isoform identification. It is important to realize that the annotation at a particular time will most likely not stand the test of time and organisms will need to undergo reannotation when new tools and methods are developed.

1.7 Aims of the thesis

1) **Assemble and annotate the genome of two species of Pinniped,** Caspian seal **and Hooded seal – Chapter 2**

In Chapter 2, I will use cutting edge next generation technology along with the most recent assembly tools to assemble the genomes of two previously unsequenced pinniped species, comparing the quality of the assemblies with other recently produced mammalian assemblies. These two assemblies will add to the catalogue of pinniped gnomes already in the public domain, whilst allowing further analyses into some of the properties of the genome.

2) **Investigate the evolutionary relationships of pinniped species and their position in the Carnivora tree – Chapter 3**

In Chapter 3 I use the protein coding regions established from Chapter 2, along with protein coding regions of other publicly available pinniped and Carnivora assemblies, to attempt to resolve the contentious relationships within Pinnipedia. I use a stringent filtering process to gather high confidence single gene orthologs, attempting to reach a consensus for the relationships surrounding the *Pusa* clade within Phocidae and test the Mustelidae sister hypothesis of Pinnipedia.

3) **Identify protein coding regions that contribute to the diverse lactation strategies that exist across Pinnipedia – Chapter 4**

In Chapter 4 I use the protein coding regions established from Chapter 2 along with the resolved relationships from Chapter 3 to identify protein coding regions under selection pressure variation across 5 clades of pinnipeds. Using *in silico* methods of selection for the 5 clades (*Pinnipedia, Phocidae, Otariidae,* Caspian seal*, Hooded seal*), I identify genes under positive selection. Publicly available collections of gene ontology and genotype-phenotype associations are then used to identify genes with strong links to lactation related functions. These are then described in the context of the phenotypic specialisations of the lineages to infer possible relation to functions.

Chapter 2: Genome Assembly and annotation of the Caspian seal and Hooded seal

2.1 Introduction

Pinnipedia is one of the most striking and diverse groups within the mammalian lineage. Since they split from their mostly terrestrial Mustelidae lineage approximately 45 Mya (Nyatakura and Bininda-Edmonds, 2012), pinnipeds now occupy waters and shores on every continent. There are three families within the pinniped group - Phocidae (true seals), Odobenidae (Walrus) and Otariidae (sea lions), all of which have extreme diversity in physiology and behaviours across, and even within, families. The application of whole genome sequencing to non-model organisms, is providing unique insights into the genome evolution and molecular underpinnings of a diversity of traits across the tree of life. With their unique phenotypic attributes and rapid adaptation pinnipeds make superb candidates for detailed genome evolution studies (Park et al., 2018).

The family Phocidae are generally referred to as the “true seals”, or the “earless seals” due to their lack of external ear flap (Berta, 2015). There are 19 species within the Phocidae family differing drastically in morphology, with large bodied Phocidae such as the Southern elephant seal weighing upto 4000kg, in comparison to the smaller bodied *Pusa spp.* weighing 80-120kg (Berta et al., 2018). Caspian seals are small bodied Phocidae endemic to the landlocked Caspian Sea, breeding on the ice sheet that forms in the northern Caspian Sea each winter (Wilson et al., 2017). The phylogenetic relationship of the Caspian seal is not fully resolved but current genetic data suggests a divergence from a common ancestor with the larger bodied Grey seal around 1-2 Mya. One of the most striking attributes of Caspian seals is its ability to tolerate a wide range of temperatures arising from the extreme continental climate of its habitat. Air temperatures range from -35°C in the winter months to +40°C in the summer. Thus, the Caspian seal is at the southern limit for sea ice formation in the northern hemisphere winter, but experiences sub-tropical conditions for much of the rest of the year. Caspian seals have undergone considerable decline in recent years, with population estimates of around 1,000,000 breeding females in the late 19th century to approximately 68,000 at present (Härkönen et al. 2012, Dmitrieva et al. 2015; Goodman & Dmitrieva 2016). The historical decline through the 20th century was driven by unsustainable commercial hunting (Härkönen et al. 2012). Primarily, fisheries bycatch but also habitat loss contributes to present day decline (Dmitrieva et al. 2013; Goodman and Dmitrieva 2016). In addition to human driven mortality, Caspian seals are also at threat from outbreaks of zoonotic Phocidae Distemper Virus (PDV) (Wilson et al., 2014), pollution, and climate change (Kajiwara et al., 2008; Kovacs et al., 2012). The overall population decline

has led to the IUCN Red List for Threatened Species to class Caspian seals as “Endangered” in 2008 (Goodman and Dmitrieva 2016).

Hooded seals are sexually dimorphic Phocidae, with males reaching over 400kg (Kovacs, 2002). These Phocidae follow an annual migratory movement cycle across a widespread range of central and western North Atlantic waters. Their name derives from the distinctive nasal lobe of males, which inflates into a prominent “hood” during courtship and dominance displays (Kovacs, 2009). Hooded seals have been known to feed at depths of up to 1000m for as long as one hour (Folkow and Blix, 1999). One of the most striking behaviours seen in Hooded seals is their abbreviated lactation period. The seals haul out on land to breed, after birth the mothers remain on land and fast during the lactation period. This lactation period ceases after just four days. In this short period the young can gain around 7kgs of mass per day whilst the mothers can lose up to 17% of their body mass during lactation (Kovacs and Lavigne, 1992; Lydersen et al., 1997). To facilitate this extraordinary mass transfer in such a short period of time Hooded seals have evolved the highest percentage of fat contained in their milk, compared to all other mammals (Debier et al., 1999). Past population declines have been attributed to commercial hunting which has been present for centuries, but in recent years the practice has been subjected to quotas and restrictions. Currently, the main threat to the Hooded seal is their dependency on Arctic Sea ice for breeding and lactation. With increased vulnerability from global warming, they are now considered amongst the most at-risk marine mammals, being listed as “vulnerable” by the IUCN (Laidre et al., 2008; Albouy et al., 2020).

Here I present the whole genome sequences of two previously unsequenced Phocidae species, Caspian seal and Hooded seal. I *de novo* assembled from third generation Pacific Bioscience and Oxford Nanopore technologies, derived reads, and error corrected using short read Illumina sequencing. I annotated these genome assemblies using a comprehensive approach using homology searches, RNASeq and *ab initio* methods which were assessed using gene presence analysis. I also calculate genome wide average heterozygosity and relate estimates to other mammal species with varying demographic histories and exposure to anthropogenic impacts. The genomic resources for this diverse group will enhance my understanding of adaptations observed across Mammalia and will also provide a valuable resource for investigations into genome evolution and species’ demographic history.

2.2 Methods and Materials

2.2.1 Sample collection

For DNA sequencing, muscle tissue sample was harvested from a fresh, dead male Caspian seal pup in the Kazakh region of the northeast Caspian Sea, the specimen was freshly stranded dead as a result of fishing by-catch, and 65g of muscle tissue of an adult male Hooded seal was obtained from the Norwegian Polar Institute, Tromsø, Norway. For RNA sequencing, tissue samples were obtained from heart, kidney, liver, lung, muscle, spleen and thymus were collected from the same deceased individual and placed in RNAlater. All samples were stored at -20°C immediately after collection for long term storage.

2.2.2 Library construction

For the Caspian seal DNA sequencing, I extracted from a small portion (approximately 5g) of muscle sample using a modified phenol-chloroform extraction method developed by Pacific Biosciences (PacBio, 2015). I performed four extractions (15S, 16S, 17S and 18S) and quantified for DNA yield on a Qubit fluorometer (Thermo Fisher Scientific) giving yield estimates between 72.7 and 222.8 ng/µl. DNA fragment sizes were analysed using TapeStation (Agilent) with at least 80% of fragments in all samples being over 6,000 bp, this was increased to 84% of fragments being over 12,000 bp after RNA removal and short fragment removal. Samples 15S and 16S were used to prepare Illumina sequence libraries and PacBio sequence libraries at the Next Generation Sequencing Facility, University of Leeds, UK. Illumina libraries were prepared using NEBNext Ultra DNA Library preparation protocol (New England Biolabs, Ipswich MA, USA), undergoing several rounds of shearing to allow a maximum fragment size of 500bp, resulting in an average DNA concentration of 13.6 ng/µl. Samples underwent two lanes of sequencing on 150bp paired end mode on an Illumina HiSeq 3000. The resulting raw output files were converted to fastq files using *BCL2FASTQ*. Samples 17S and 18S underwent RNA and short fragment removal to remove all fragments under 25,000 bp using Circulomics Short Read Eliminator and Circulomics Short Read Eliminator XS kits. Two libraries were prepared using the Genomic DNA by Ligation PromethION Kit (Oxford Nanopore Technologies) and run over two PromethION flow cells with Guppy version 3.0.5 basecaller (Oxford Nanopore Technologies, 2019). Libraries were generated for the PacBio Sequel II using the SMRTBELL template prep kit. After sequencing subreads were extracted from the PacBio SMRT Link software (Pacific Biosciences, 2020).

For RNA sequencing, a PureLink RNA Mini Kit (Thermofisher Scientific, MA, USA) was used on six tissues - heart, kidney, liver, lung, spleen and thymus for RNA extraction. RNA samples were analysed on a Bioanalyser (Aligent, Santa Clara, USA). RNA yields varied across the tissue samples from 21ng/µl to 207ng/µl. RNA integrity number (RIN) scores also varied with some tissues (Heart, Lung and Kidney) scoring below a RIN score of 6. To compensate for the lower RIN scores samples with lower RIN scores were fragmented for a shorter time, in comparison to the samples with a higher RIN score, and a NEB total RNA library prep (New England Biolabs, Ipswich MA, USA) was used to prevent any 3' end biases that would occur using a PolyA method. RNA libraries were then sequenced on an Illumina HiSeq 3000.

DNA for the Hooded seal was extracted using the same phenol-chloroform DNA extraction method that was used for Caspian seal was then run on two small pieces of the muscle tissue (approximately 5g each), resulting in two samples, 1H and 2H. Illumina libraries were performed using the NEBNext Ultra DNA Library preparation protocol (New England Biolabs, Ipswich MA, USA) on the 1H sample, after shearing a DNA concentration of 6.3 ng/µl was recorded. This sample underwent one lane of sequencing on 150bp paired end mode on an Illumina HiSeq 3000. The resulting raw output files were converted to fastq files using *BCL2FASTQ*. Libraries were prepared using the Genomic DNA by Ligation sequencing Kit (Oxford Nanopore Technologies) and run over two minION flow cells with Guppy version 3.0.5 basecaller (Oxford Nanopore Technologies, 2019). RNA sequence for the Hooded seal was already publicly available from a previous study (Hoff et al., 2017).

2.2.3 Genome Assembly

Quality control checks on the sequence reads (Table 2.1, 2.2) from each sequence lane of Illumina HiSeq were run through *FASTQC* (Andrews, 2010), to assess sequencing quality (Electronic appendix 2.1). The reads produced from long read sequencing methods, Pacific Biosciences Sequel, Nanopore minION and Nanopore promethION were analysed through *Nanoplot* (De Coster et al., 2018). Reads were trimmed to remove low quality scored bases and adaptor artifacts from the sequencing runs using *Trimmomatic* (Bolger et al., 2014) for the Illumina HiSeq reads and *PoreChop* (Wick, 2018) for the Nanopore reads. Sequence reads from PacBio runs already had the adaptors removed prior to downloading using the SMRT Link software (Pacific Biosciences, 2020). *Trimmomatic* was used with the commands "-phred 33", "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10", "SLIDINGWINDOW:4:30", "LEADING:30", "TRAILING:30" and "MINLEN:80". This removed Illumina TruSeq adaptors and regions of the reads below a quality score of 30 Phred as well as removing all resulting trimmed sequences less than 80 base pairs in length. *Porechop*

Table 2.1. Sequence statistics of Caspian seal. †Coverage based on California sea lion genome size genome size of 2.38 Gbp (Peart et al,. 2021). *Illumina HiSeq has uniform read length of 151bp.

Sequence technology	Illumina HiSeq 2000	ONT PromethION	PacBio Sequel II
Approximate coverage†	40x	20x	10x
Reads	642,524,234	6,917,022	3,469,106
Read N50 (bp)	151*	24,409	4,230
Maximum length read (bp)	151*	281,030	113,257

Table 2.2. Sequence statistics of Hooded seal. †Coverage based on California sea lion genome size genome size of 2.38 Gbp (Peart et al,. 2021). *Illumina HiSeq has uniform read length of 151bp.

Sequence technology	Illumina HiSeq 2000	ONT MinION
Approximate coverage†	19x	10x
Reads	313,470,818	2,737,899
Read N50 (bp)	151*	11,733
Maximum length read (bp)	151*	122,358

(Wick et al., 2017) was used with the default settings to remove all adaptor sequences from the ends of reads, in addition to finding internal adaptor sequences and splitting those into multiple reads.

Before performing any genome assemblies on the Caspian or the Hooded seal, I estimated genome sizes and heterozygosity of the samples from the Illumina raw reads (Figure 2.1). *Jellyfish* (Marçais and Kingsford, 2011) was first used to create histograms of k-mers before the characteristics of the assembly were estimated using *GenomeScope* (Vurture et al., 2017). This would give an insight into the state of the data and would assist choosing the assemblers that would benefit from my data, some assemblers are ideally suited to work with genomes that exhibit high levels of heterozygosity or polyploidy (Zhang et al., 2020). The karyotype and organisation of chromosomes within pinnipeds, and even within Phocidae, is well resolved and so it was estimated in high confidence that the Caspian seal and Hooded seal would both be diploid with 32=2N chromosomes (Árnason, 1974; Beklemisheva et al., 2016).

Two different assemblers, *Wtdbg2 (Redbean)* version 2.5 (Ruan and Li, 2019) and *Flye Assembler* version 2.7.1 (Kolmogorov et al., 2019), were tested on both species. After trimming, the long-read sequencing depth varied between the different species, approximately 30x for the Caspian seal and approximately 7x for the Hooded seal. Multiple assemblers were used to evaluate if different assemblers could produce more contiguous assemblies at differing coverage depths*. Flye assembler* was run using the default settings with "--iteraions 0" and "--pacbio-raw" flag with the Caspian seal long reads, whilst the Hooded seal was run using the "nano-raw" flag. *Wtdbg2* assembler was run with the "nanopore/ont' preset for both Caspian seal and Hooded seal assemblies.

2.2.4 Polishing

The process of using more accurate short reads to correct errors within the long read produced assemblies is known as "polishing". Polishing was conducted with *Pilon* (Walker et al., 2014) and *Medaka* (Oxford Nanopore Technologies, 2018) using the filtered Oxford Nanopore reads, in addition to the filtered short reads. *Snakemake* (Köster and Rahmann, 2012) pipelines were produced to perform long read polishing and multiple iterations of short read polishing. The filtered Oxford Nanopore reads were first mapped to the raw assembly using *minimap2* version 2.17 (Li, 2018), with the flag "ava-ont". The resulting "paf" mapping file was passed to *Racon* version 1.4.3 (Vaser et al., 2017), with default settings with the output file then run in *Medaka-gpu* version 0.9.1 (Oxford Nanopore Technologies, 2018). The short Illumina reads were then mapped to the generated long-read polished assembly using *Minimap2* with the flag "ax -sr".

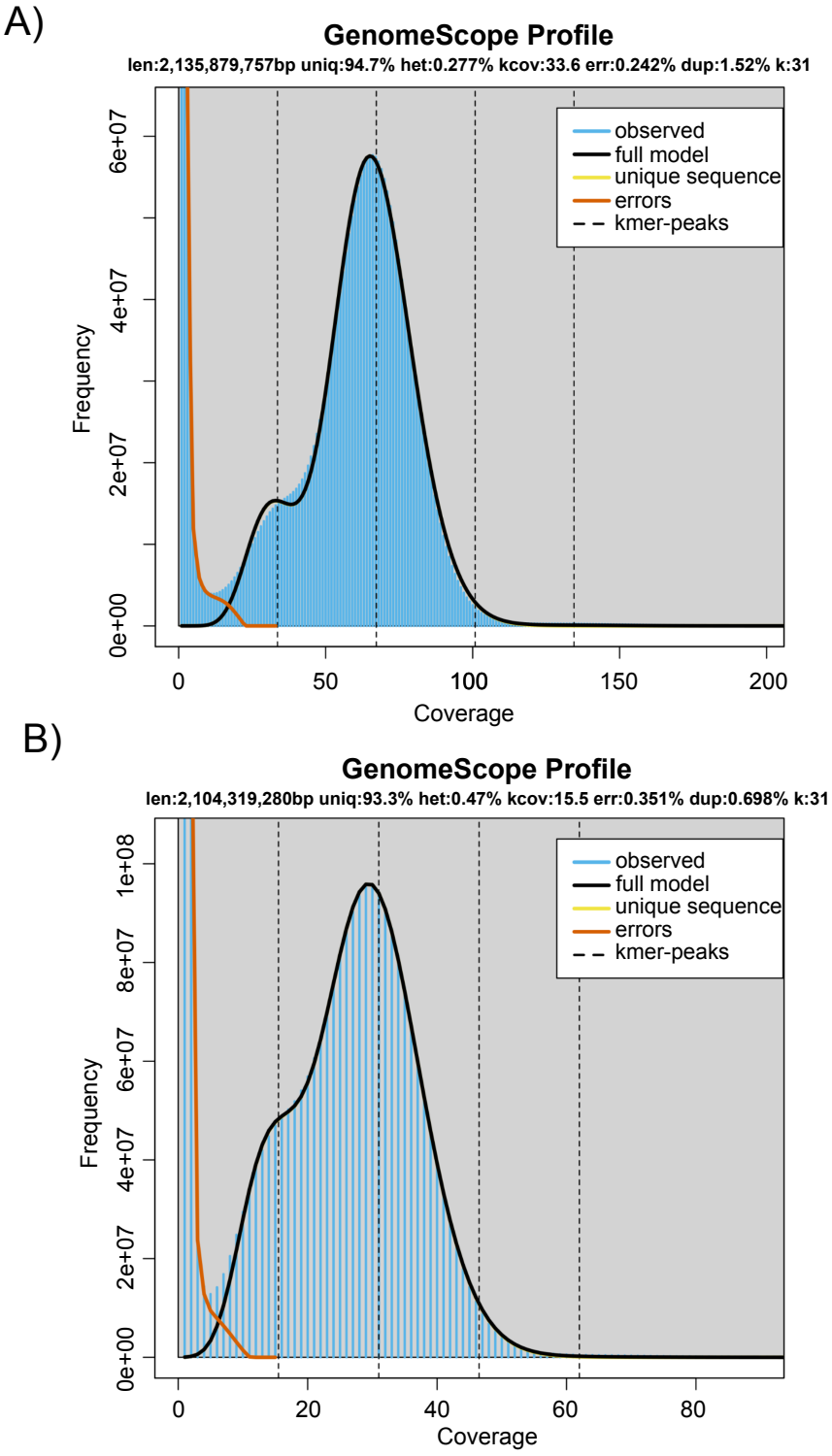


Figure 2.1. GenomeScope profiles for Caspian seal **and Hooded seal.** GenomeScope analysis estimating the heterozygosity and coverage of A) the Caspian seal and B) the Hooded seal, using kmer size of 31.

The resulting bam file was used with the long read polished assembly to further correct errors in *Pilon*, version 1.23. Multiple iterations of *Pilon* short read polishing can be effective. Iterations of Pilon were run until the assembly quality started to decrease due to over polishing (Electronic appendix 2.2). Each iteration was assessed by *BUSCO* (Benchmarking Universal Single Copy Orthologs) version 3.0.2 (Simão et al., 2015) mapping scores and short read mapping percentage.

2.2.5 Scaffolding using RNA-Seq Data

P_RNA_Scaffolder (CAFS-Bioinformatics, 2019), utilises information from paired end RNA-Seq to further scaffold contiguous sequences in the assembly. Mapping the RNA-Seq from the six different tissues (heart, kidney, liver, lung, spleen and thymus) to the polished assembly with *HiSat2* version 2.2 (Kim et al., 2019) and *BLAT* v.36 (Kent, 2002). *P_RNA_Scaffolder* then searches for pairs of reads that are mapped to different contigs, these reads are then orientated and scaffolded together. The resulting genomes were assessed for improvement though BUSCO score and short read mapping percentage.

2.2.6 Repeat Identification and Masking

Repeatmodeler version 2.0.1 (Flynn et al., 2020), was used to identify *de novo* repeats within the Caspian seal and Hooded seal assemblies. A creation of a Caspian seal custom repeat library and Hooded seal custom repeat library were each combined with mammalian sequences from Dfam_3.1 and RepBase-20181026. The generated repeat libraries were used in *RepeatMasker* version 4.0.9 (Smit et al., 2016) to soft mask repeat families identified in the sequences. *RepeatMasker* was used with the commands "-s" "-nolow" and "-engine crossmatch" to perform a highly sensitive run avoiding simple tandem or low complex sequences that may affect exon detection.

2.2.7 Gene Annotation

Three different methods were used to comprehensively capture and annotate gene structures within the assemblies, *ab initio* gene prediction, protein homology and RNA-Seq based prediction (Figure 3). First an *ab initio* method was performed using *SNAP* (Korf, 2004), *GlimmerHMM* version 3.0.4 (Majoros et al., 2004) and *Augustus* version 3.3.3 (Hoff and Stanke, 2018). To create a set of training genes of which to train the *ab initio* gene predictors, *BRAKER2* version 2.1.5 (Hoff et al, 2018) was used with the Caspian seal assembly. This training set of genes was filtered down with genes retained after aligning to proteins from the CanFam3 assembly (downloaded from RefSeq on 15/10/19) (filters of 80% subject coverage, 50% percent identity, and e-value of $e^{-0.05}$) using *DIAMOND* (Buchfink et al., 2017) and then processed through

the recommended Augustus training set generation protocol (Hoff et al., 2020), leaving a training set of 886 genes. This training set was used to train parameters for *SNAP* and *GlimmerHMM* whilst the "human" parameters from *Augustus* outperformed my Caspian seal parameters. Secondly, *GenomeThreader* version 1.7.3 (Gremme, 2013), was used to identify possible gene structures through sequence homology. Reviewed Uniprot (Bateman, 2021) protein sequences with experimental evidence at the protein or transcript level, were downloaded for the most well reviewed mammalian species (cat, dog, human and mouse) and aligned against the genome in using 'relaxed' filters. Harbour seal (*Phoca vitulia*) protein transcripts were downloaded from NCBI and aligned with 'stringent' filters. RNA-Seq paired reads were filtered using *Trimmomatic* (Bolger et al., 2014), with the same filters as the DNA although bases were passed to a threshold of Phred 20 quality score. Subsequently, these reads were *de novo* assembled into transcripts using *Trinity* version 2.9.1 (Grabherr et al., 2011). Then the generated transcripts were fed into the *Program to Align Spliced Assemblies (PASA)* version 2.4.1 (Haas et al., 2008), to refine and align the transcript models against the assembly. All the homology-based, RNASeq-based and *ab initio* predicted gene models were collated and provided as inputs for *EVidenceModeler* version 1.1.1 (Haas et al., 2008), this generated a consensus gene annotation. To assess putative gene names and confidence, the consensus gene models were annotated through alignment to the SwissProt database (downloaded 12/04/20) (UniProt, 2019) using *BLASTP*, version 2.2.31 (Altshul et al., 1990). Depending on the matching quality, genes were assigned as "COMPLETE", "PARTIAL" or "LOW QUALITY". Genes that did not return a hit to the SwissProt database were annotated using orthologous predictions through the *Emapper.py* script in *eggnog-mapper*, version 1.0.3 (Huerta-Cepas et al., 2017), and assigned as "PREDICTED". Unannotated genes were then scanned for domains using *InterProScan* and tagged with "HYPOTHETICAL" if any matches returned. Any gene models that did not receive annotation through the SwissProt database *BLAST*, *Emapper* or *InterProScan* (Jones et al., 2014) were removed from the final gene annotations.

2.2.8 Genomic Diversity

To compare genomic diversity of the Caspian seal in relation to that of the Hooded seal and other mammals the raw short Illumina reads were mapped against the assembly of the Californian sea lion. Short Illumina reads for the Caspian seal and the Hooded seal were aligned and filtered to the California sea lion reference assembly (NCBI:*GCA_009762305.2*) adhering to the GATK best practices pipeline (Van der Auwera et al. 2013). The raw reads were mapped to the reference assembly, variants were called using the *GATK HaplotypeCaller*, with site being called if the read depth was in the range of 50%-250% of the mean depth for the genome (67x and 30x coverage depth for the Caspian seal and Hooded seal

respectively). Variants were removed if they violated the guidelines of the GATK best practices and clustered SNPs (3 in a 10-base window) were removed. Genome-wide heterozygosity was calculated from the number of heterozygous genotypes divided by the total number of callable sites passing all filters.

2.3 Results

2.3.1 Genome Assembly and Quality Assessment

The assembled genome for the Caspian seal had a length of 2.34 Gbp (Gigabase pairs) consisting of 795 contigs with an N50 of 22.6 Mb, L90 of 108 and a maximum contig size of 103.76 Mb (Table 2.3) (Figure 2.2). The Hooded seal genome assembly had a more fragmented assembly (L90 = 2,353) and reduced genome size (2.29 Gbp) (Table 2.3) (Figure 2.2). Despite this, the contig N50 (1.06 Mb) of the assembly passes benchmarks of current expectations of modern reference assemblies (Rhie et al., 2020).

The quality of the assemblies was assessed through proportion of alignments of the short-read sequences and gene content. 99.82% of the short reads mapped to the Caspian assembly of which 97.88% mapped in proper pairs, i.e., forward and reverse reads corroborated configuration of the assembly. For the Hooded seal 98.8% of the short reads mapped to the Hooded seal assembly with 96.16% in proper pairs. The assemblies were also assessed through identification of single copy gene orthologs, using *BUSCO* v.3.0.2 (Simão et al., 2015). Single copy gene orthologs universal in the mammalian lineage (N=4,104) are aligned against the assembly with the proportion being found complete, in single copy and fragmented being used to assess the assembly quality. From the Caspian seal assembly 3,912 (95.3%) single copy gene orthologs were present, of which 3,874 (94.4%) were present in single copy with only 87 (2.1%) missing completely. The Hooded seal had 3,822 (92.3%) of single copy genes present, 3,789 (92.3%) of these possessed just one copy in the assembly and 130 (3.2%) were completely missing from the assembly. The gene presence scores of these assemblies are comparable with some of the highest quality non-model species currently available (Figure 2.3).

2.3.2 Genome Annotation and Quality Assessment

The annotation pipeline generated 23,144 and 23,297 gene models for the repeat masked Caspian seal and the Hooded seal assemblies, respectively. Filtering the gene models using *Emapper.py*, *InterproScan* and SwissProt reduced these numbers to 20,459 for the Caspian seal and 20,893 for the Hooded seal. The

Table 2.3. Summary statistics of generated assemblies and annotations.

	Caspian seal	Hooded seal
Total Length (Gbp)	2.38	2.29
Scaffolds	795	5,679
N50 (Mb)	22.62	1.5
L50	30	472
Max contig length (Mb)	103.76	22.62
GC %	41.38	41.39
Gene counts	20,459	20,893

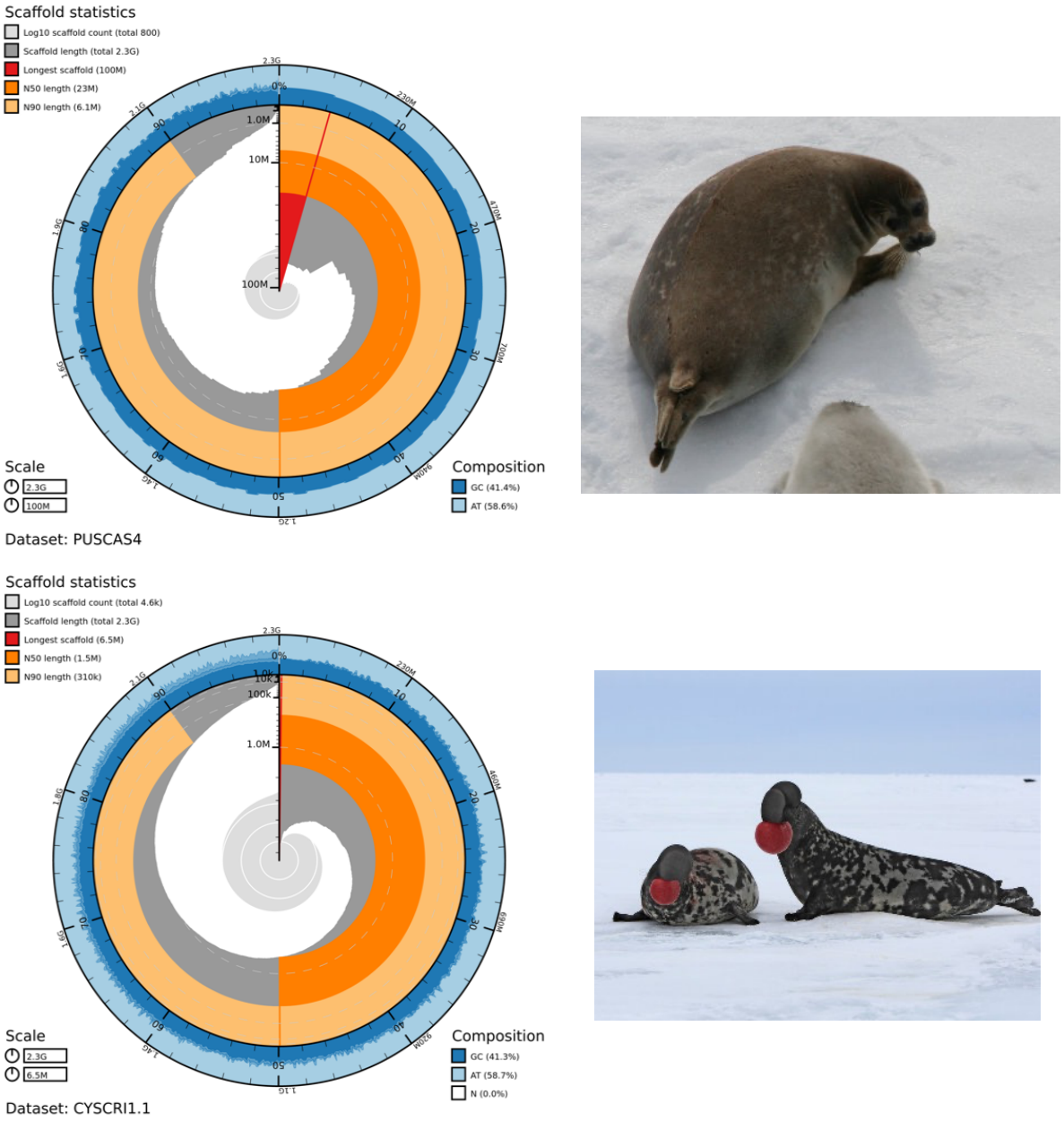


Figure 2.2 Snail plots of assembly statistics using Blobtools (Challis et al., 2020). Innermost dark grey arc displays the lengths of individual scaffolds in a decreasing manner, red segment represents the largest scaffold in the assembly, dark orange and light orange represent the N50 and N90 length respectively. Dark and light blue areas display the GC/AT content across the scaffolds. Top: snail plot of the Caspian seal , and female adult Caspian seal and her pup on the ice sheets in the Northeast Caspian Sea, Kazakh region. Bottom: snail plot of the Hooded seal , and two male Hooded seals fighting with the nasal “hood” displayed.

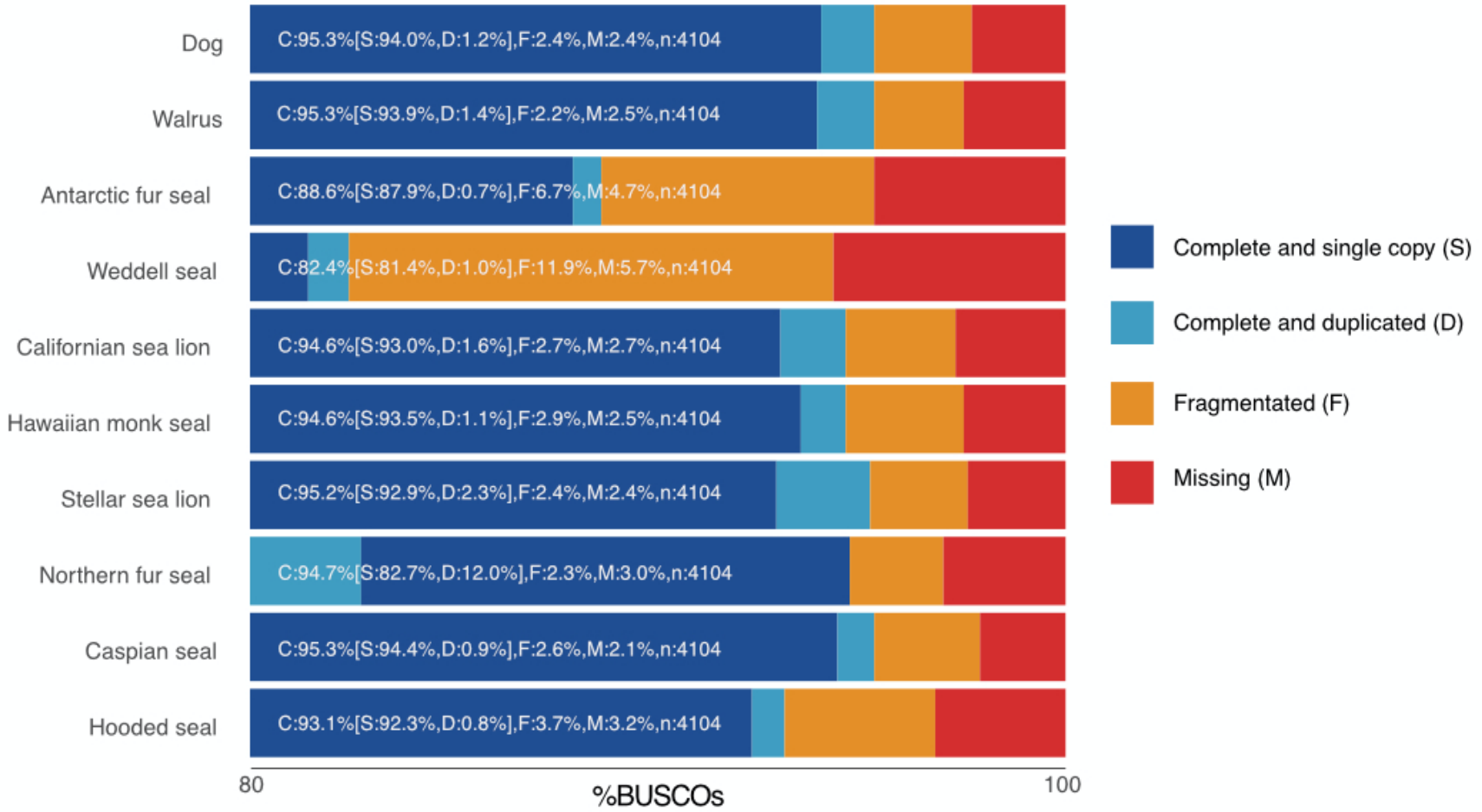


Figure 2.3. BUSCO summary of Carnivora assemblies. The percentage of complete (in single copy and duplicated copies), fragmented and missing single universal single copy gene orthologs from the mammalian lineage of the BUSCO database (N=4104), present in publicly available pinniped genomes. The dog genome (CanFam3.1) has been included for comparison, being the most sequenced Carnivora genome.

Hooded seal gave the higher gene model count. This could be a result of the more fragmented assembly having incomplete genes, the Caspian seal had more “Complete” matches and a lower number of “Low Quality” matches than the Hooded seal (Table 2.4). In addition, a BUSCO assessment revealed a higher percentage of single copy orthologous genes present in the Caspian seal annotation (95.1%) than the Hooded seal (90.3%). The final annotations were also assessed with DOGMA (Kemena et al,. 2019), which assesses a proteome set on the presence of a core number of well-known conserved domains and arrangements. Using the mammalian library of domains from -PfamScan, the Caspian seal returned a score of 95.79% and the Hooded seal returned a score of 90.54%. Using my annotation pipeline, I can be confident I have captured at least 90% of known mammal genes, with the Caspian seal expected to contain >95%. This compares favourably with the most complete pinniped assembly and annotation, California sea lion, which has approximately 96% of mammalian genes captured.

2.3.2.1 Genomic Diversity

I calculated genome wide average heterozygosity in the Caspian seal and Hooded seal to evaluate potential impacts on genetic diversity from recent demographic histories. To generate average genome wide heterozygosity, I used the California sea lion genome as an outgroup, this avoided reference genome bias. Mapping the short reads from both species to the California sea lion I found that both the Caspian seal and Hooded seal have average levels of heterozygosity, 0.00126 and 0.00174 respectively, when compared to a range of mammals from Robinson et al. (2016). The moderate level of heterozygosity suggests that the Caspian seal population has not lost significant amounts of genetic variation or shows signs of a bottleneck. Despite declining to around 10% of the estimated population size at the start of the 20th Century, the number of individuals remains large with around 68,000 breeding females. This provides some hope in the conservation of Caspian seal, demonstrating that sufficient genetic diversity is present to prevent the negative genetic consequences of bottleneck events. Although, this is only an estimate of average genome wide heterozygosity and further investigations into rates of heterozygosity and runs of heterozygosity are needed to demonstrate that this diversity is retained throughout the whole genome and not localized to specific regions.

Table 2.4. Genes models passed through each filtering step.

	Caspian seal	Hooded seal
SwissProt		
Complete	14,049	12,340
Partial	533	1,170
Low Quality	2,674	4,696
Emapper		
Predicted	173	199
InterproScan		
Hypothetical	3,030	2,488
Total	20,459	20,893

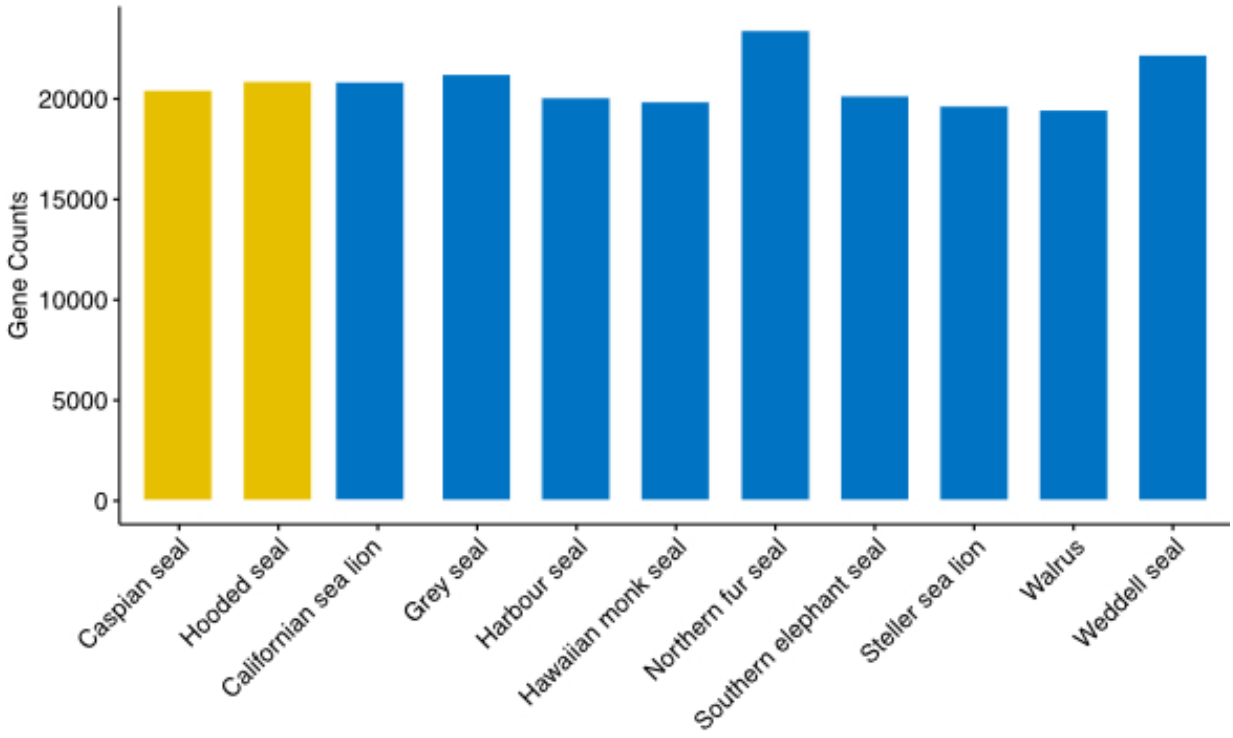


Figure 2.4. Gene count comparisons across pinniped assemblies. Gene counts of Caspian seal and Hooded seal in comparison to publicly available pinniped genomes.

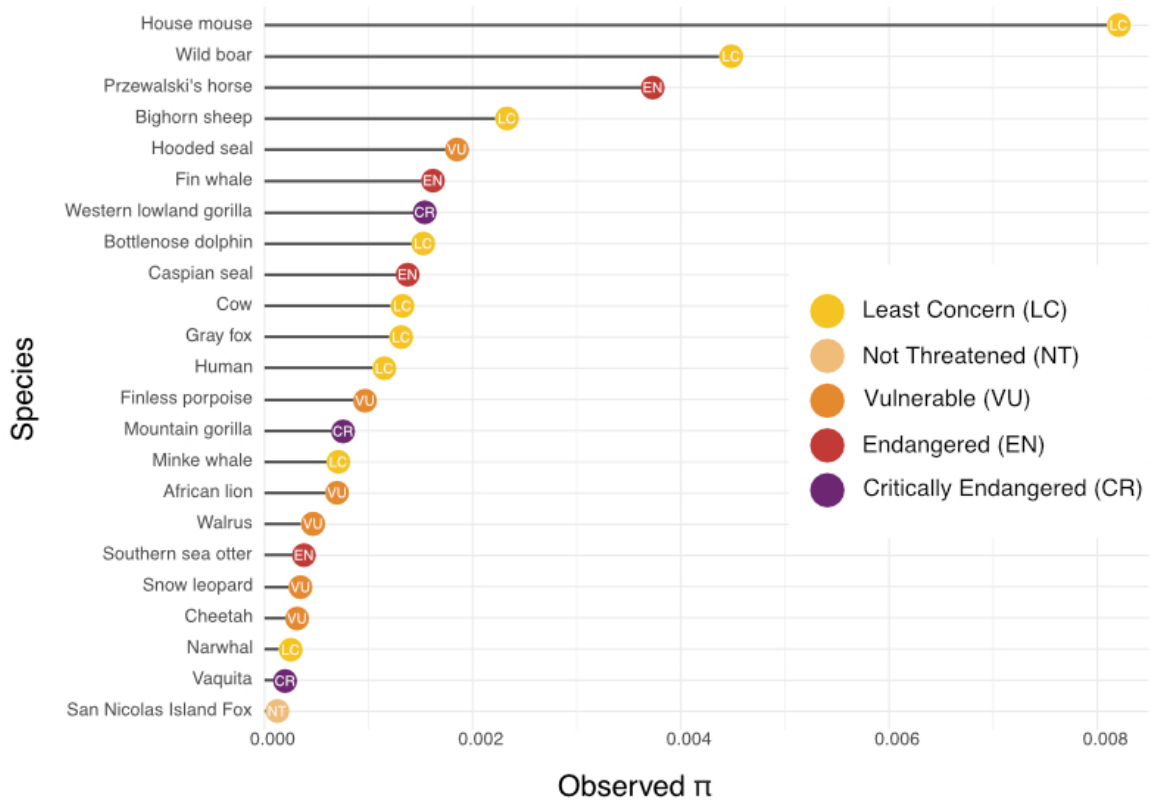


Figure 2.5. Observed heterozygosity comparison of Caspian seal **and Hooded seal**. Observed heterozygosity from a range of mammalian species, coloured by their IUCN red list rating, including the Caspian seal and Hooded seal.

2.4 Discussion

Pinnipedia is a major group of marine mammals, with extreme life histories and which has undergone recent evolutionary radiation. The genome information for pinnipeds is growing as they become more popular as a non-model species to investigate the genomic basis of physiological adaptations, some of which may be of interest for understanding non-infectious disease phenotypes in humans and agricultural animals. This analysis provides the first genome assembly and annotation of the Caspian seal and Hooded seal. These genomes and gene models will extend beyond this project and will provide a lasting resource for future pinniped genomic analyses.

Here we have used a single representative to produce a model of a genome that will be used to make inferences about a whole species. It is important to recognise the limitations of using extrapolating a single sample in this manner. In this analysis, one individual of Caspian and Hooded seal was used to generate a genome assembly and these assemblies were then used to produce gene sequences and assess levels of heterozygosity for the whole species. In the case of the Caspian seal, a panmictic species (Wilson et al, 2017), sampling issues may be reduced. For the Hooded seal it is important to be aware of stochastic or population level dynamics that may affect single populations and be cautious extrapolating these findings to the whole species. Using single genome can be successful in interesting patterns within the molecular data, but analyses should always be validated using many individuals across differing populations. Attempts were made to minimise possible errors that can be generated in the collection of the species, i.e., samples were harvested from fresh tissues of recently deceased species and stored in RNAlater immediately after harvesting. Despite this deliberately careful behaviour it is extremely difficult to mitigate possible issues such as, degradation and contamination of samples. Male individuals were used for this analysis, due to availability of resources. Being the homogametic sex allowed me to sample all possible chromosomes from a particular individual, although sex chromosomes would likely appear at a significantly lower coverage. In future analyses, it would be preferential to gather samples from male and female individuals to fully generate high coverage assemblies of autosomal and sex chromosomes, that can be used in analyses that maybe effected by sex biases.

This genome assembly was produced using long read ONT and PacBio data as a backbone for the genome assembly. The use of long reads allows the assembly of large repetitive sequences, short reads would not

be able to span over long repetitive regions and thus assembly of these regions would have been poor, if even achievable. The Caspian seal assembly was sequenced to a higher depth than that of the Hooded seal, 30x and 7x respectively, and thus the quality of the assembly and resulting annotation is impacted by these differences, the Caspian seal has a greater N50 and is less fragmented (able 1.3). The overall accuracy of long read data is lower than that of Illumina short reads (Amarasinghe et al., 2020), and so it was important for us to polish the long read only assembly with the short reads. I corrected for erroneous bases using this process and measured the accuracy of the mapping of the short reads as an indication of how accurate my assembly was. By performing multiple iterations of polishing the coding regions of my genome assemblies increased with the number of RNA reads perfectly mapping increasing by up to 10% (Appendix 2.1).

Many different genome assemblers have been produced since long reads began to become widely used in genome assemblies, all with differing results depending on the species and sequence depth (Murigneux et al., 2020; Wang et al., 2021). Fortunately, my species of interest were not susceptible to issues such as high heterozygosity, polyploidy or unusually high repetitive regions that could cause limitations in assembly. My only caveats were the different sequencer types and depths. To mitigate the impact of the sequence data I used two different tools with underlying assembly methods: *Flye assembler* (Kolmorgorov et al., 2019) and *Wtdbg2* (Ruan and Li, 2019). *Flye assembler* is based on generalised repeat graphs - these use the same fundamental principles as de Bruijn graphs, but use approximate sequence matches to call edges, this tolerates the higher noise in long reads. *Wtdbg2* (Ruan and Li, 2019) uses a combination of both de Bruijn and overlap-layout-consensus methods – breaking down sequences and merging into nodes based on similarity, building a “fuzzy de Bruijn” graph which allows for gaps and mismatches. Rather than using contiguity to assess my resulting alignments, which are liable to issues, such as small fragmentations, I used BUSCO scores as a proxy of genome completeness to choose superior assemblies.

Publicly available pinniped assemblies in the NCBI database such as the Grey seal, Northern elephant seal, Southern elephant seal, Hawaiian monk seal (*Neomonachus schauinslandi)*, Harbour seal and California sea lion, have assembly sizes ranging from 2.34 Gbp to 2.41 Gbp. The Weddell seal assembly is significantly larger than the other pinniped genomes. This was the first pinniped to be sequenced and used only short reads, the ungapped genome size is 2.22Gbp suggesting that the gaps in the gapped genome are overestimated in the assembly. The Caspian seal and Hooded seal assemblies produced in this chapter fall close to this range, with assembly sizes of 2.34 Gbp and 2.29 Gbp respectively. The Hooded seal assembly

is slightly smaller than expected, as *Wtdbg2* assembler produces smaller assemblies than expected when using nanopore reads. Despite this, the resulted assembly was higher quality in terms of contiguity and genome completeness than other methods. The use of long reads in genome assemblies allows us to accurately assemble large repetitive regions and structural variants that would not be possible with short reads alone. The use of other long-range sequencing technologies such as HiC and optical mapping (covered in Chapter 1.6.2) would allow the assembly of whole chromosome length scaffolds that would prevent mis-assembly. These technologies are widely used in current genome assembly analyses (Dudchenko et al., 2018). I was limited by financial resources and time, and thus the long-read sequencing was preferential as my main aims were to develop genomes with highly accurate and contiguous coding sequences, for which long reads were sufficient. In future analyses of chromosomal arrangements or synteny it would be beneficial to incorporate long-range sequencing to supplement these assemblies.

Using an automated annotation pipeline, I identified 20,459 and 20,893 gene models in the Caspian seal and Hooded seal respectively. BUSCO and DOGMA searches were used to assess the gene completeness, BUSCO assesses the proportion of single copy gene orthologs which have been found to be ubiquitous across Mammalia, whereas DOGMA uses conserved domains across mammalian. DOGMA is more likely to find fast evolving or diverged orthologs as present, as they are more likely to have retained their domain but could be diverged sufficiently to be lost by the BLAST search of BUSCO. It is expected that I would lose some of the gene models due to the fragmentated nature of my assembly, although my annotations had very high BUSCO and DOGMA scores comparable to annotations that have using long-range sequencing (Figure 1.3). The Hooded seal genome assembly was more fragmented than the Caspian seal genome assembly and this was reflected in the number of genes in the annotation, with the extra genes expected to be as a result of fragmented coding regions being called as separate genes. The BUSCO score reflects this also with Hooded seal have an extra 1.1% of fragmented genes compared to the Caspian seal. As there was little variation between the BUSCO scores of my genome assemblies and that of more contiguous genome assemblies (such as the Dog or California sea lion), I expect that much of the fragmentation comes from intergenic regions.

Automatic gene model calling, suffers from the caveat that homology with other species plays a vital part in the pipeline and so the quality of annotations is dependent on the quality of the annotations of closely related species. In this chapter, I attempted to overcome this by firstly only using the highest quality closely related species, in this case the Harbour seal, and also supplementing the homology search using

other evidence-based gene model callers. I used PASA genome annotation software for my annotation and gave weightings to the different sources of gene models. Gene models derived from more reputable sources, for instance the assembled transcriptome, were given the highest weighting, whereas homology-based gene models and *ab initio* gene models were given the lowest weighting. I would have benefited from the use of long read RNAseq (isoseq), as this would have generated full gene transcripts without assembly and could have decreased fragmentation in coding regions of my assembly. The cost and expertise needed for the library preparation and sequencing of "iso-seq" reads was not available for this anlayisis. Automatic gene model annotation can be improved by manual curation, this is further discussed in section 1.6.2.2 and all pinniped genomes would significantly benefit from manual curation of at least one species. It would be beneficial for a manual curation tool to be implemented in an online server for the community to participate in, which would greatly increase the quality of the available annotations.

2.5 Conclusion

In this chapter, I present the first genome assemblies for the Caspian seal and Hooded seal. These will serve as an important resource to advance the understanding of marine mammal evolution and conservation. I have assembled the genomes to a high degree of accuracy, with gene content levels comparable to that of some of the most researched genome assemblies such as the California sea lion.

Chapter 3: Phylogenetic Analysis of Pinnipedia

3.1 Introduction

3.1.1 Conflict in the pinniped phylogeny

With the development of next generation sequencing technologies (NGS), the rate at which high quality genomic data can be generated is increasing. With this increase in data, it was speculated that the issue of incongruence across phylogenies would be greatly reduced (McCormack et al., 2013; Rokas et al., 2013, Liu et al., 2019). However, even with the application of these larger datasets, contentious relationships across the tree of life (ToL) remain (Wicket et al., 2014; Xi et al., 2014; Arcila et al., 2017; Miyashita et al., 2019). Biological processes such as incomplete lineage sorting (ILS), gene duplications/losses and hybridisation events have led to genomes resembling a mosaic of evolutionary histories, presenting a significant challenge for unravelling the true phylogenetic history of species (Scornavacc and Galtier, 2017).

There have been disagreements between the evolutionary relationships of clades within the mammalian class. For instance, the resolution for the root of placental mammals could not be established until composition and rate heterogeneity was accounted for (Morgan et al., 2013; and Tarver et al., 2016). Despite modelling advancements and increased data sampling, contentious relationships still exist in the wider mammalian tree (McCormack et al., 2011; Springer, 2013; Foley et al., 2016). Within Mammalia, Carnivora is an order of nearly 300 species (Wozencraft, 1993). Carnivora are of prominent interest for their conservation status with some of the species, such as Giant Panda, Red panda, and Monk seals, falling within the top 150 evolutionary distinct and globally endangered mammals (EDGE, 2021). Carnivora includes terrestrial and aquatic species and together with their global distribution makes this order a fascinating subject of evolutionary history. The phylogeny of the Carnivora is well studied but even the most recent phylogenies, with near complete species coverage, primarily rely on mitochondrial genes or a combination of mitochondrial genes and a small subset of nuclear genes (Nyakatura and Binida-Emonds, 2012; Hassanian et al., 2021), rather than genome scale data. Thus, some recently diverged taxa still lack sufficient data to resolve phylogenetic relationships with high confidence.

Within the Carnivora, pinniped phylogeny and paleobiology have remained ambiguous, and morphological and molecular evidence are conflicting (Berta et al., 2018). Pinnipeds are members of the carnivoran clade Arctoidea (Figure 3.1), but their placement within this clade has been heavily debated. Indeed, early debates argued that pinnipeds are not a truly monophyletic group with morphological data

supporting the diphyly of pinnipeds (de Muizon 1982, Barnes 1989). The diphyletic hypothesis suggests that Odobenidae and Otariidae shared a common ancestor with Ursidae, however Phocidae share a more recent relationship to the Mustelidae clade (Tedford, 1976; Repenning et al., 1979; de Muizon, 1982; Barnes, 1989; Wozencraft, 1989; Nojima, 1990) (covered in further detail in Chapter 1.1). Evidence from the auditory morphology (Wyss, 1987) significantly shifted the narrative, finding overwhelming support for an alternative, single monophyletic relationship between Odobenidae, Otariidae and Phocidae, although the sister group was still unclear (Weber, 1904; Gregory, 1910; Davies, 1958). In recent analyses a small number of studies show continued support of a diphyletic origin (Kuhn and Frey, 2012; Koretsky et al., 2016) although this relationship was only achieved using exclusively morphological evidence. Incorporation of molecular evidence presents an unambiguous monophyletic relationship, with support for Mustelidae sister group hypothesis (Berta and Wyss, 1994; Flynn et al., 2005; Fulton and Strobeck, 2006; Higdon et al., 2007; Sato et al., 2006; Kohno, 2006; Yonezawa et al., 2009; Nyakatura and Binida-Emonds, 2012; Furbish 2015; Hassanian et al., 2021). To date only two studies have challenged the Mustelidae sister group hypothesis using genetic data. Firstly, Delisle and Strobeck (2005) used 12 mitochondrial genes to produce a phylogeny with posterior probabilities and bootstrap support values of 0.44 and 42% respectively, with the authors concluding the clade formed a polytomy. Feijoo and Parada (2017) performed Maximum Likelihood and quartet-based analyses with concatenated data for 29 nuclear genes (23,495 bp) to determine relationships across *Arctoidea*. Finding support that pinnipeds diverged from a common ancestor to both Mustelidae and Ursidae, suggesting a Ursidae/Mustelidae sister hypothesis. The methodology of this analysis was criticised, and alignments were found to be erroneous, with mismatched introns/exons, and incorrectly assigned homology, with paralogs incorrectly identified as orthologs, in approximately 38% of the 29 genes. Reanalysis after accounting for these errors concluded in support for a Mustelidae sister group hypothesis (Gately and Springer, 2018).

As with other mammalian orders with rapidly diversifying lineages (McGowen et al., 2020; Vanderpool et al., 2020), areas of contention persist within two of the three families of pinnipeds. Although the monophyly of the Otariidae family has consistently been recovered (Árnason et al. 2006, Higdon et al. 2007, Yonezawa et al. 2009, Churchill et al. 2014, Boessenecker & Churchill 2015), intra-family relationships have been disputed (Berta et al., 2018). For Otariids, cladistic analyses using the dense fur pelage, mandible and dental characteristics support the presence of two sub-families, Arctocephalinae (fur seals) and Otariinae (sea lions) (Berta & Deméré 1986, Barnes et al. 2006). Arctocephalinae are distributed primarily across the southern hemisphere, with the Northern fur seal as the sole occupier of

the northern hemisphere. Whereas Otariinae are widely distributed across northern and southern hemispheres, except the north Atlantic. The monophyletic relationship between Arctocephalinae has become contentious, with recent morphological and molecular evidence supporting a paraphyletic relationship between these subfamilies. These analyses find that the southern hemisphere Arctocephalinae and southern hemisphere Otariinae form a clade which is sister to that of the northern hemisphere Otariinae. Thus, suggesting the dense fur phenotype, that is only present in Arctocephalinae has been independently lost twice during Otariidae evolution. Conflicting relationships have been observed within sub-families, contradicting previous taxonomic naming. Southern hemisphere fur seal (*Arctocephalus spp.*) relationships remain contentious, most likely due to their recent, rapid diversification approximately 2.5 – 3 Mya (Lopes et al., 2020). Arctocephalus has been described as a paraphyletic (Berta and Churchill, 2012), however Nyakatura and Binida-Emonds (2012) argued that this status was premature. The taxonomical nomenclature of the genera has persisted even with a whole genome analysis of Otariidae, which described a monophyletic relationship between all southern hemisphere *Arctocephalinae* (Lopes et al., 2020). Contention of a monophyletic relationship between species of the *Zalophus* genera has also been proposed using whole genome data. California sea lion and Steller sea lion formed a clade, with Galapagos sea lion (*Zalophus wollebaeki*) described as a sister clade (Lopes et al., 2020).

Currently the remaining ambiguities for relationships within pinnipeds are between the Grey seal, Pusa spp. (Caspian seal, Ringed seal, and Baikal seal) and Phoca spp., Largha seal (*Phoca largha*) and Harbour seal (*Phoca vitulina*) (Table 3.1). Several alternative topologies for this group have been suggested by different studies. Initial morphological evidence and mitochondrial DNA fragments resolved Phoca spp. and Pusa spp. as monophyletic groups with Grey seal as an outgroup to the Phoca-Pusa clade (Figure 3.2A) (Burns and Fay, 1970; de Muizon, 1982; Perry et al., 1995; Binida-Emonds et al., 1999). Analyses that incorporated mitochondrial genes found Grey seal to be clustered with the Pusa clade, although from these analyses Ringed seal was the lone of Pusa representative (Figure 3.2B) (Davis et al., 2004; Delisle and Strobeck, 2005). Once additional Pusa spp. were included in the analyses conflicting relationships arose. Caspian seal and Grey seal were found to form a monophyletic clade, with Ringed seal placed as outgroup to both Phoca and other Pusa species (Figure 3.2C), although these relationships were weakly supported (Fulton and Strobeck, 2006; Palo and Väinölä, 2006; Árnason et al., 2006; Higdon et al., 2007). Complete mitochondrial genomes and nuclear genes yielded an alternative position for Grey seal, suggesting Grey seal as an outgroup to the Pusa and Phoca species, similar to the initial morphological

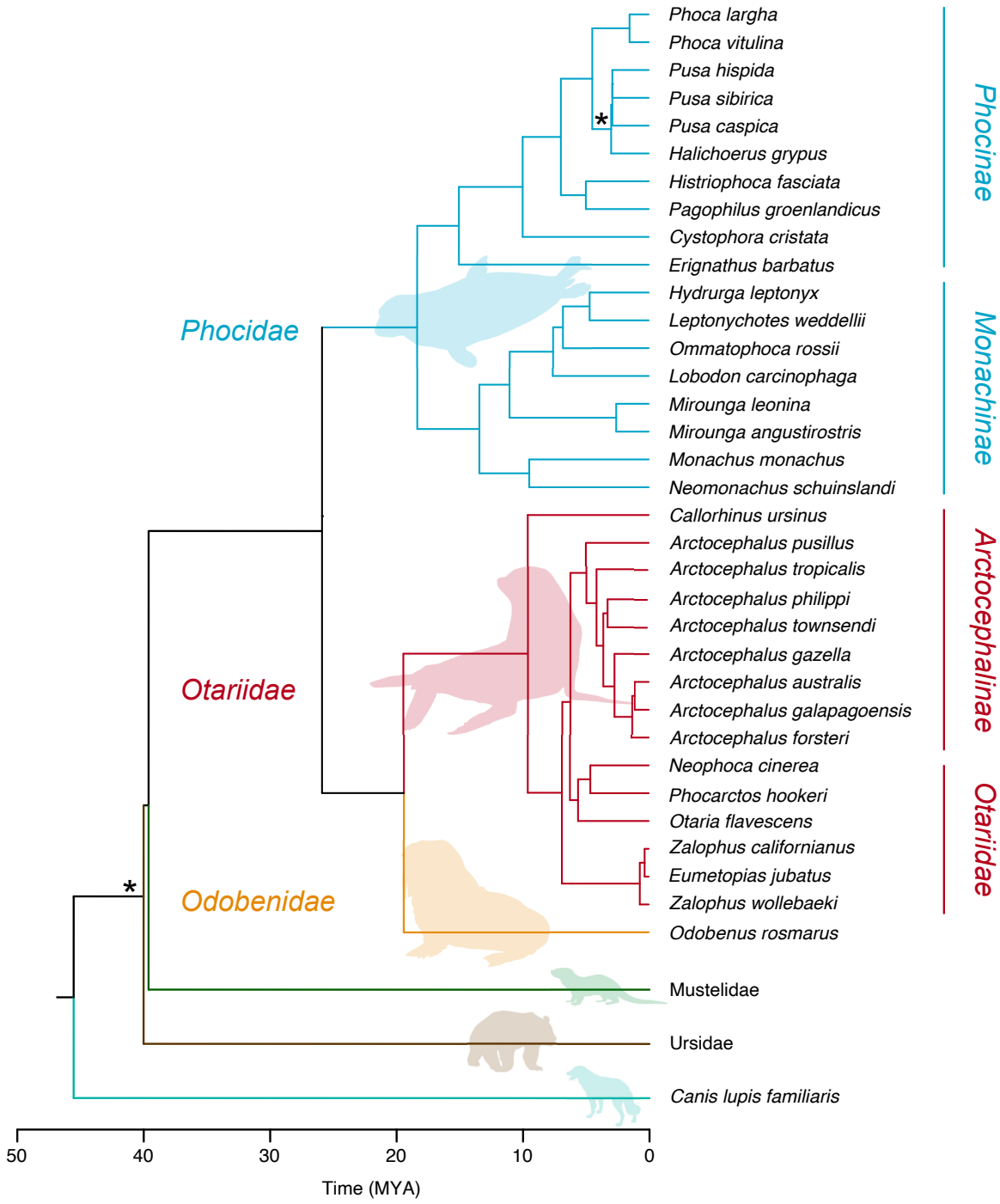


Figure 3.1 Current placement of pinniped families and subfamilies within Carnivora phylogeny. Divergence dating taken from TimeTree.org, with Otariidae relationships inferred from Lopes et al. (2020). Areas of controversy addressed in this chapter are highlighted by asterisks.

to being an outgroup to *Pusa-Phoca* (Figure 3.2D). The most comprehensive phylogenetic analysis to date, Nyakatura and Binida-Emonds (2012) used 41 nuclear genes, 18 tRNAs and 15 mitochondrial genes in a

analyses (Fulton and Strobeck, 2009; Fulton and Strobeck, 2010). Fulton and Strobeck (2010) also found support for Grey seal being an outgroup to *Pusa*, although this had a lower support (approximately 70%) in comparison supertree analysis. This found a relationship grouping Caspian seal and Grey seal as sister species within Pusa clade, with Phoca spp. as an outgroup to a Pusa (Figure 3.2E).

To date many different datasets have been applied to try and resolve the discrepancies within the pinniped and Carnivora lineages (Table 3.1). However not all phylogenetic datasets are improved through increasing the volume of data (Philippe and Roure, 2011). None of these previous attempts to resolve the positions took a critical view of data quality and model fit. Incorrectly assigning orthology has been shown to cause incongruities in the inferred topologies (Springer and Gatesy, 2018). Statistical tests to infer phylogenetic confidence have progressed in recent years. Tests such as the SH test (Shimodiara and Hasegawa, 1999) and KH test (Kishino and Hasewaga, 1989) were shown to be too conservative and unable to reject incorrect relationships when testing across many trees (Shimodaira, 2002). The Approximately Unbiased (AU) test (Shimodaira, 2002) has been shown to overcome this issue (see Introduction section 1.3.2.3).

Mutation rate and compositional heterogeneity have been shown to be influenced by many factors including life history traits, ecological factors, and evolutionary history. Not taking these factors into account in modelling can increase susceptibilities to systemic errors, such as long branch attraction (LBA) (Chira and Thomas, 2016; Wang et al., 2018; Yang et al., 2018). Heterogeneity within the data can be modelled through two different classes: the rate of substitution rates across the sites, called the rate heterogeneity, and composition of bases across the data - compositional heterogeneity (Sheffield, 2012; Morgan et al., 2013; Moran et al., 2015). Rate heterogeneity can be modelled using advanced homogeneous models, defined by a set number of rates in a gamma distribution. Models such as the Phylobayes CAT model (Lartillot et al., 2009), which account for different probabilities of character evolution at different sites, are needed to account for compositional heterogeneity (Moran et al., 2015). To my knowledge, no study to date has investigated the evolution of pinnipeds using models that account for heterogeneity and tested these models to see if they accurately fit the data, despite evidence showing that not accounting for heterogeneity can have implications on the resulting topologies (Morgan et al., 2013). In this chapter, two aspects of the resolution of the Pinnipedia are investigated: (i) at the family level, attempting to determine the order of 3 major families within Carnivora – Mustelidae, Ursidae and

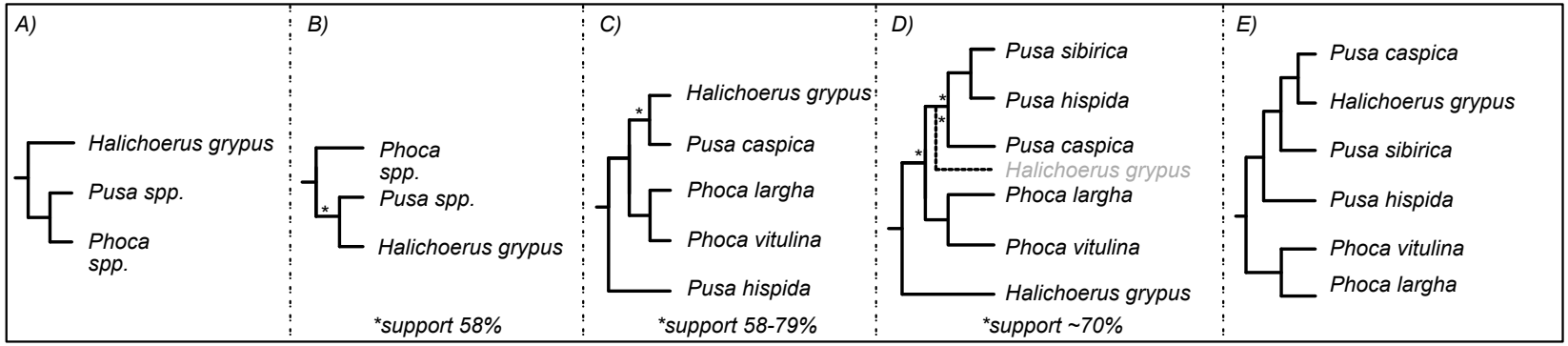


Figure 3.2 Changes in the predicted *Pusa-Phoca* phylogeny. As more publications have been released data types (i.e., mtDNA and nuclear DNA) and species sampling density has increased. A) shows the phylogeny inferred by morphological data, fragmented mtDNA and a small number of complete mt genes; B) complete mt genes; C) complete mt genes and a small number of nuclear genes; D) whole mt genomes and few nuclear genes; E) the *Pusa-Phoca* clade in the most comprehensive phylogeny published to date. All support values are 100% unless otherwise indicated. Publications corresponding to the phylogenies are detailed in Table 3.1.

Table 3.1. Previous publications of Phocidae phylogeny. The species refers to the number of Phocinae species in the analysis and the topology corresponds to which topology of the *Phocini* tribe (categorised in Figure 3.2) was resolved in the study.

Author, year	Species	Data Type	Data size	Topology
Burns and Fay, 1970	6	Morphological	-	A
de Muizon, 1982	6	Morphological	-	A
Perry et al., 1995	6	mtDNA	240 bp	A
Binida-Emonds et al., 1999	19	Morphological + mtDNA	-	A
Davis et al., 2004	15	mtDNA	10,842 bp	B
Delisle and Strobeck, 2005	15	mtDNA	30,421 bp	B
Fulton and Strobeck, 2006	18	mtDNA	3,601 aa	C
Palo and Väinöla, 2006	9	mtDNA	3,369 bp	C
Arnason et al., 2006	18	rRNA, mtDNA	14,336 bp	C
Higdon et al., 2007	19	Nuclear, mtDNA	26,818 bp	C
Fulton and Strobeck 2009	18	Nuclear, mtDNA	8,935 bp + mt genomes	D
Fulton and Strobeck 2010	18	Nuclear, mtDNA	8,935 bp + mt genomes	D
Nyakatura and Binida-Emonds, 2012	19	Nuclear, tRNA, mtDNA	43,834 bp	E

Pinnipedia, and (ii) provide consensus to the evolution of the Phocini tribe within Phocidae, providing a correct placement for the position of Grey seal within *Pusa spp.* A strict filtering approach was employed to ensure the orthologous groups used in the reconstruction are truly orthologous and provide adequate levels of phylogenetic signal, removing data with outlying levels of phylogenetic signal. Three independent approaches to reconstruction were taken: a Bayesian framework, a Maximum Likelihood framework and a coalescent supertree framework to reconstruct a phylogeny using single copy orthologs of protein coding DNA sequences. Homogeneous and heterogeneous models were evaluated for their fit to the data and their impact on the resulting topologies. I resolve Grey seal as a sister species to Caspian seal, further validating its position within the *Pusa* clade. I also confirm Mustelidae as an outgroup to the pinniped clade adding support for the Pinniped-Mustelidae hypothesis.

3.2 Materials and methods

3.2.1 Carnivora Dataset Assembly

3.2.1.1 Taxon sampling

Species were chosen from all three marine mammal orders (Carnivora, Cetartiodactyla, and Sirenia) in addition to four terrestrial mammal orders (Perissodactyla, Rodentia, Primates, and Proboscidea) included to reduce large phylogenetic distances within my dataset within the mammalian clade, with an additional non-placental mammal species, Platypus (*Ornithorhynchus anatinus*) (Table 3.2). All downloaded sequences were acquired in October 2020. Protein coding genes for non-pinniped species were downloaded from Ensembl v.101 (Yates et al., 2020) or Ensembl Rapid Release (Ensembl, 2020) using the FTP directories. Protein coding genes for nine of the pinnipeds were download from the Ensembl Rapid Release (Ensembl, 2020) or NCBI RefSeq database (O'Leary et al., 2016). Five additional pinniped species were acquired as unannotated genome assemblies. Of these, three species (Northern elephant seal, Bearded seal (*Erignathus barbatus*), Largha seal) were downloaded from the DNAZoo database (Dudchenko et al., 2016). One species, *Arctocephalus gazella*, was obtained from the NCBI RefSeq database (O'Leary et al., 2016); and one species, Ringed seal, was acquired through personal communication with the Baikal Ringed seal Genome Project group (P. Auvinen, University of Helsinki). These genomes were passed through the genome annotation pipeline described in detail in Chapter 2. The protein coding genes for two species, Caspian seal and Hooded seal, were annotated and extracted

Table 3.2. List of 36 species present in dataset. [1] species removed from 33taxa dataset to create 32taxa dataset, [2] species added to 33taxa dataset to create 32taxa dataset, and [3] species added to 32taxa dataset to create 34taxa dataset.

Common Name	***Species name***	**Order**	**Database**	**Genome version**
American beaver	*Castor canadensis*	Rodentia	Ensembl	C.can_genome_v1.0
Antarctic fur seal	*Arctocephalus gazella*	Carnivora	NCBI Refseq	arcGaz3
Bearded seal	*Erignathus barbatus*	Carnivora	DNA Zoo	Erignathus_barbatus_HiC
Blue whale[2]	*Balaenoptera musculus*	Cetartiodactyla	Ensembl	mBalMus1.v2
Californian sea lion	*Zalophus californianus*	Carnivora	Ensembl	mZalCal1.pri
Caspian seal	*Pusa caspica*	Carnivora	Orr et al., 2021	puscas4
Cat	*Felis catus*	Carnivora	Ensembl	Felis_catus_9.0
Cow	*Bos taurus*	Cetartiodactyla	Ensembl	ARS-UCD1.2
Dog	*Canis lupis familiaris*	Carnivora	Ensembl	CanFam3.1
Elephant	*Loxodonta africana*	Proboscidea	Ensembl	Loxafr3.0
Florida Manatee	*Trichechus manatus latirostris*	Sirenia	Ensembl Rapid Release	GCA_000243295.1
Giant panda	*Ailuropoda melanoleuca*	Carnivora	Ensembl Rapid Release	GCA_002007445.2
Grey seal	*Halichoerus grypus*	Carnivora	NCBI Refseq	Tufts_HGry_1.1
Harbour seal	*Phoca vitulina*	Carnivora	Ensembl Rapid Release	GCA_004348235.1
Hawaiian monk seal	*Neomonachus schauinslandi*	Carnivora	Ensembl Rapid Release	GCA_002201575.1
Humpback whale[1]	*Megaptera novaeangliae*	Cetartiodactyla	Tollis et al., 2019	GCA_004329385.1
Hooded seal	*Cystophora cristata*	Carnivora	Orr et al., 2021	cyscri1.1
Horse	*Equus caballus*	Perissodactyla	Ensembl	EquCab3.0
Human	*Homo sapiens*	Primates	Ensembl	GRCh38.p13
Largha seal[3]	*Phoca largha*	Carnivora	DNA Zoo	Phoca_largha_HiC
Mediterranean monk seal[1]	*Monachus monachus*	Carnivora	per comms	MMS_114

Mouse	*Mus musculus*	Rodentia	Ensembl	GRCm39
Mouse lemur	*Microcebus murinus*	Primates	Ensembl	Mmur_3.0
Northern elephant seal	*Mirounga angustirostris*	Carnivora	DNA Zoo	Mirounga_angustirostris_HiC
Northern fur seal	*Callorhinus ursinus*	Carnivora	Ensembl Rapid Release	GCA_003265705.1
Pig	*Sus scrofa*	Cetartiodactyla	Ensembl	Sscrofa11.1
Platypus	*Ornithorhynchus anatinus*	Monotremata	Ensembl Rapid Release	GCA_004115215.2
Polar bear	*Ursus maritimus*	Carnivora	Ensembl	UrsMar_1.0
Ringed seal[3]	*Pusa hispida*	Carnivora	Samaii Ringed seal Genome Project	Pushis
Sable	*Martes zibellina*	Carnivora	Ensembl Rapid Release	GCA_012583365.1
Sea otter	*Enhydra lutris kenyoni*	Carnivora	Ensembl Rapid Release	GCA_002288905.2
Southern elephant seal	*Mirounga leonina*	Carnivora	Ensembl Rapid Release	GCA_011800145.1
Sperm whale	*Physeter catodon*	Cetartiodactyla	Ensembl	ASM283717v2
Steller sea lion	*Eumetopias jubatus*	Carnivora	Ensembl Rapid Release	GCA_004028035.1
Walrus	*Odobenus rosmarus*	Carnivora	NCBI Refseq	Oros_1.0
Weddell seal	*Leptonychotes weddellii*	Carnivora	NCBI Refseq	LepWed1.0

as per Chapter 2. The annotation of Mediterran monk seal (*Monachus monachus*) was obtained from an external collaborator (Gaughren *pers comm.*, 2020). This genome was produced through the mapping of short read data to the Hawaiian monk seal genome. To generate the annotation data for this species, `getfasta` within the *bedtools* package was used with the GFF3 file from Hawaiian monk seal and Mediterranean monk seal genome fasta file. The annotation files for *Megaptera novaeangliae* were downloaded from the genome assembly additional files (Tollis et al., 2019).

3.2.1.2 Orthologous group assignment

Prior to clustering, protein coding sequences were cleaned, with sequences containing internal stop codes or incomplete sequences being removed, and nucleotides translated to amino acids. Ensembl or RefSeq headers were parsed to retain only the longest sequence for each gene identification sequence, this ensures no isoforms were present in the sequences and each gene has only a single representative sequence. Following this, sequences were cleaned and translated using the *'vespa_clean.py'* and *'vespa_translate.py'* commands from VESPA (Webb, Walsh and O'Connell, 2016). These scripts remove all sequences that possess an internal stop codon or length that is not a multiple of 3 before translating the nucleotide sequences into an amino acid format.

OrthoFinder2 (Emms & Kelly, 2019) was used with all protein coding sequences from the species in Table 3.2 to identify 'Orthogroups'. Orthofinder2 uses the clustering method originally developed for Orthofinder (Emms & Kelly, 2015) to determine 'Orthogroups'. Orthogroups are defined as a group of genes descended from a single gene in the last common ancestor (LCA) of a group of species. Due to gene duplication events these may contain 1-1, 1-to-many or many-to-many 'co-ortholog' relationships (Figure 3.3A). Initial relationships between sequences are first identified through an all-versus-all search from all species using *Diamond* v0.9.24 (Buchfink et al., 2015). For each species pair, Diamond bit scores are normalised by gene length and phylogenetic distance, to prevent the biased exclusion of small gene lengths and over-inclusion of large gene lengths into clusters during the Markov Cluster Algorithm (MCL) step (Figure 3.3C). Reciprocal best length-normalised hits are used to define the lower limit for the inclusion in an orthogroup, with only hits scoring higher than the lowest scoring reciprocal best length-normalised hit being included within an orthogroup graph (Figure 3.3D). Orthogroup graphs are then passed to MCL to convert into orthogroups by similarity scores, breaking apart clusters that have low similarity and retaining clusters of high similarity (Figure 3.3F).

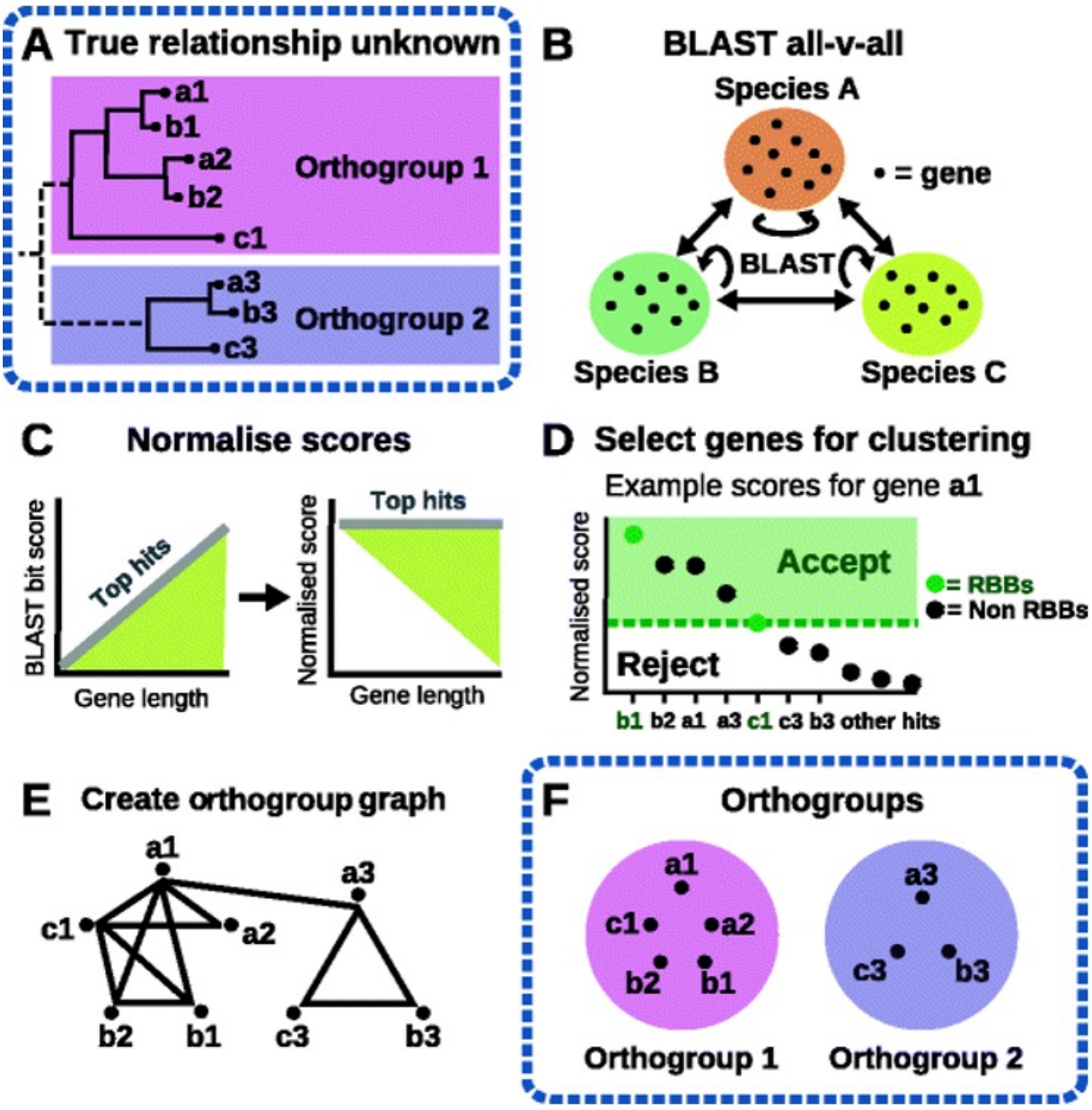


Figure 3.3 Steps of Orthogroup creation in Orthofinder. A) unknown orthogroups that are intended to be recovered; B) BLAST search of all genes vs all genes; C) Gene length and phylogenetic distance normalisation of BLAST bit scores to give the scores to be used for orthogroup inference; D) Selection of putative cognate gene-pairs from normalised BLAST scores; E) Construction of an orthogroup graph, genes are nodes and pairs of genes connected by edges, with edge weights equal to normalised bit scores; F) Clustering of genes into discrete orthogroups using MCL. Figure adapted from Emms & Kelley, (2015).

Using the initial run containing 33 taxa (33Taxa_Dataset) (Table 3.2), 23,290 orthologous groups were identified using OrthoFinder2, of these 990 were SGOs that had representatives of all Carnivora species in my dataset. The number of orthologous groups containing all species is limited by the lowest quality coding sequences. The gene counts show that two species, Mediterranean monk seal and Humpback whale were of poor quality and as a result contained in only 8,687 (38.6%) and 10,259 (45.6%) orthologous groups respectively (Figure 3.4). In comparison, the other taxa were present in an average of 15,216 (68.7%) orthologous groups, with the Platypus being present in the least number of orthologous groups, 13,924 (61.9%), which was expected due to the phylogenetic distance of this outgroup species. *Orthofinder2* was run again with the Mediterranean monk seal removed and Humpback whale replaced by the Blue whale (32taxa dataset), this resulted in 23,761 orthologous groups being identified, with an increase of 1,781 single gene orthologs (SGOs) that had representatives of all Carnivora species in the dataset. Subsequent to the *Orthofinder2* run of the 32taxa dataset, two more pinniped genomes became available to use and so *Orthofinder2* was run an additional time using the previously computed *BLAST* searches and the 34taxa dataset. This resulted in 24,334 orthologous groups of which 2,279 had representatives of all Carnivora species in the dataset. Non-Carnivora species were removed from these SGOs as they were intended to be included in the selective pressure variation analyses (Chapter 4).

3.2.1.3 Multiple sequence alignments

Increases in divergence between sequences can cause errors in alignments, mis-aligned sequences across species have been shown to impact negatively on phylogenetic analyses (Ogden and Rosenberg, 2006; Fletcher and Yang, 2010). Due to the large quantity of orthologous groups present, alignments were generated and evaluated *in silico*. Alignments were generated using a 'progressive approach', *MAFFT* (Katoh and Standley, 2013), *MUSCLE* (Edgar, 2004) and *CLUSTAL Omega* (Sievers et al., 2011), and a 'constituency-based method', *T-COFFEE* (Notredame, Higgins and Heringa, 2000). The progressive approach builds an estimated tree using relationships between sequences, this tree is then used as a guide, aligning the most closely related sequences first, then progressively adding the next most related sequences. A caveat of the progressive approach is that errors made in the early stages of the alignment become irreversible, meaning alignments become stuck in local minima. Consistency-based methods estimate pairwise alignments for all pairs of sequences, then use the score of these pairwise alignments as constraints to maximise the agreement of the global alignment with the pairwise alignments (Chatzou et al., 2016). This process can achieve a more accurate, but more computationally demanding and slower alignment process. *AQUA* (Muller et al., 2010) was modified to incorporate *CLUSTAL Omega* and *T-*

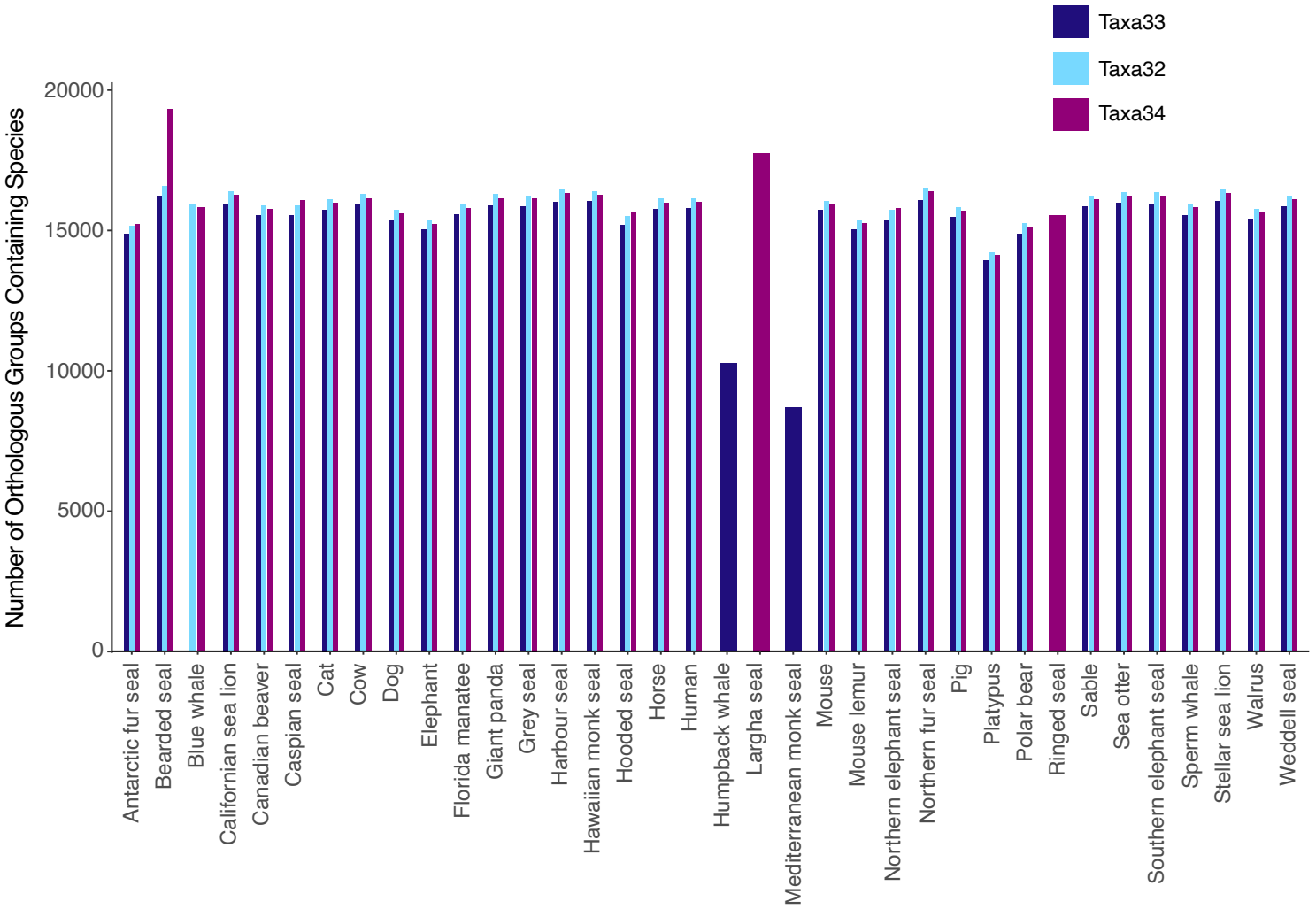


Figure 3.4 Number of orthologous groups per species with increasing taxa. Orthologous groups produced using Orthofinder using sampling from 33Taxa_Dataset (red), 32Taxa_Dataset (blue) - removed low quality species Mediterranean monk seal and humpback whale but added blue whale, and 34Taxa_Dataset (green) - adding Largha seal and Ringed seal.

COFFEE. Thus, each multiple sequence alignment (MSA) was then processed with an alignment refinement program, *RASCAL* (Thompson, Theirry and Poch, 2003), to attempt to enhance the quality of the MSA, before being evaluated using *norMD* (Thompson et al., 2001). From the six different MSAs constructed for each orthologous group, the MSA with the highest *norMD* score was used. From a visual check of a subset of the alignments it was seen that some alignments contain large amounts of gaps in the sequences, which can have major impacts on the resulting phylogeny (Darriba et al., 2016). The Antarctic fur seal was observed to have a substantial number of gaps in its sequence due to poor genome quality and so was removed from the MSAs. The consequences of trimming alignments to remove gaps is a debated process in phylogenetics, thus trimming was conducted in an automated and conservative manner using *TRIMAL* (Capella-Gutiérrez, Silla-Martínez and Gabaldón, 2009) with the parameter 'gappyout'. This removes all gaps based on a gap distribution obtained from the alignment and this method has been shown to reliably produce high quality alignments in mammalian datasets (Steenwyk et al., 2020).

3.2.1.4 Filtering of orthologous groups

The 2,279 orthologous group MSAs then underwent three filtration steps to ensure the dataset contained truly orthologous groupings and informative signal to infer bifurcations. The first step was to assess the sequences for "sequence saturation". Sequence saturation occurs as a result of multiple or identical substitutions at the same site in different sequences of an alignment. This can cause the divergent rate from distant lineages to be greatly underestimated, and phylogenetic inference has been shown to work best with only slightly saturated datasets (Philippe et al., 2011). To assess the levels of sequence saturation at the amino acid level, neighbour-joining phylogenetic trees were built in PAUP (Felsenstein, 1993) using a basic p-distance and JTT model of evolution. Sequences that have undergone sequence saturation would be expected to have much shorter branch lengths when trees are built using the p-distance matrices, which directly compares sequence similarity, compared to the JTT models, which accounts for multiple substitutions. This was run using the script *run_phylip.sh* (Electronic appendix, 3.1). The sum of the branch lengths for each p-distance calculated tree were plotted against the sum of the branch lengths using the JTT model, for each orthologous group on a single graph (Figure 3.5C). The graph was then assessed by eye, identifying any alignments that possessed substantially greater branch lengths calculated by the JTT model in comparison to those calculated by the p-distance model; four sequences were then removed from the dataset as putative sequences possessing sequence saturation.

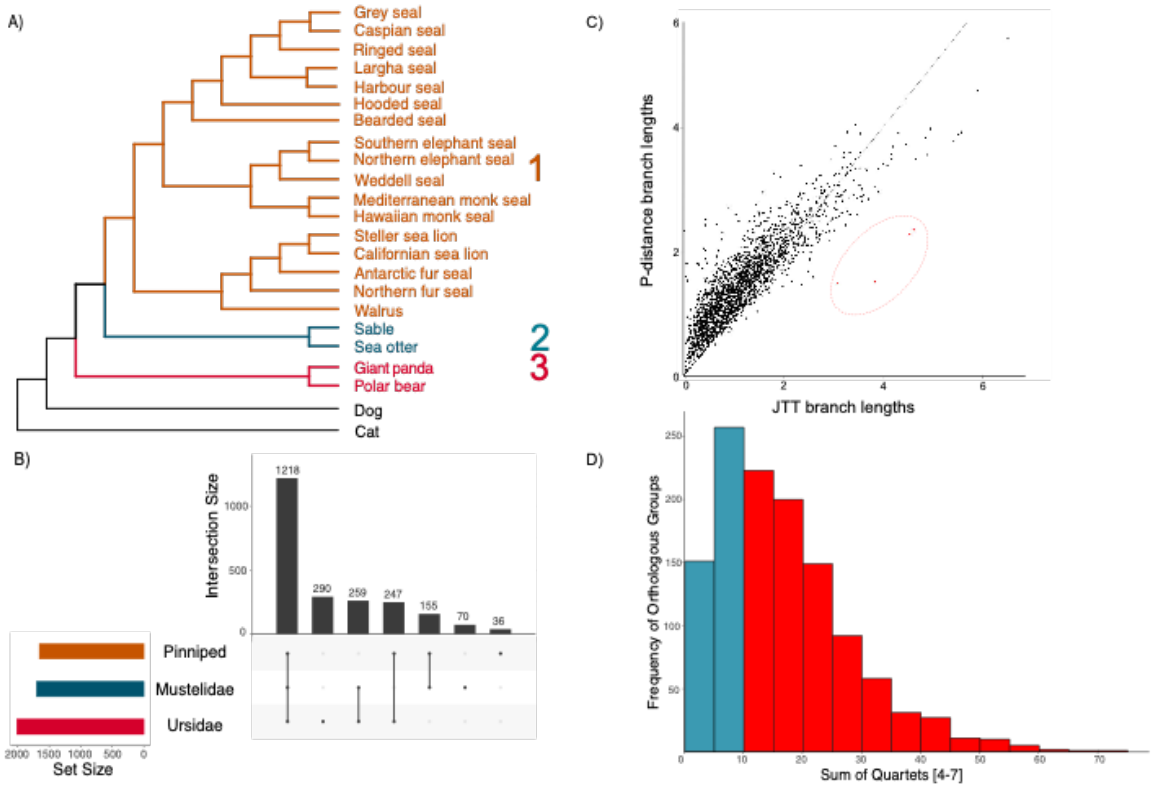


Figure 3.5 A representation of the orthologous groups lost during the filtering steps. A) Unrooted tree coloured by the different 'clans' used in the Clan_check step and B) upset plot of the number of orthologous groups that do not violate particular 'clans'. C) Sequence saturation tests showing the p-distance produced tree bench length of each orthologous group (N=2,275) against the JTT model branch lengths, with outliers represented with red dots in the dotted oval. D) Frequency of orthologous groups against the sum of quartets from segments (4-7) from their likelihood mapping analysis, discarded orthologous groups with a sum >10 shown in red.

3.2.1.5 Ortholog enrichment filter

To account for hidden paralogy within the dataset, *Clan_Check* (https://github.com/ChrisCreevey/clan_check) was employed. *Clan_check* uses unrooted single copy phylogenetic trees and determines whether the 'clans' (Wilkinson et al., 2007), unrooted analogues of a monophyletic group or clade, are violated. For each of the 2,275 MSAs, *IQ-TREE* (Nguyen et al., 2015) was run using automatic model selection (Kalyaanamoorthy et al., 2017) and 1000 ultrafast bootstrap replicates (Hoang et al., 2018). This produced a tree, which will be referred to as a 'gene tree'. Clans to be tested within *clan_check* were assigned at an order level, accounting for hybridisation that has been shown to occur within Carnivora (Kumar et al., 2017; Savriama et al., 2018). As *clan_check* requires at least 2 species per clan the following clans were assigned: Pinnipedia, Mustelidae, Ursidae (Figure 3.5A). The gene trees were concatenated into a single file and *clan_check* was run using the script *clancheck_prep.sh (*Electronic appendix, 3.2*)*. A maximum level strictness was applied from the output of *clan_check,* in which only trees which had no violations in any of the three clans were retained. This was carried out using *find_nonviolate_trees.py* (code supplied by Peter Mulhair) and the subsequent alignments that passed this step were identified using *get_OG_files.py* (code supplied by Peter Mulhair). 1,218 orthologous groups were found to have no violations of the specified 'clans' and retained as high confidence SGOs, whereas the 1,057 SGOs with clan violations were discarded.

3.2.1.6 Phylogenetic signal filter

To assess the level of phylogenetic signal present in each of the 1,218 SGOs, likelihood mapping was performed using *IQ-TREE* (Nguyen et al., 2015) using the command 'lmap'. Likelihood mapping works by breaking each gene tree down into quartets (Figure 3.6A), and the likelihood for each of the three possible configurations being calculated using Maximum Likelihood (Strimmer and von Haeseler, 1997). From the Maximum Likelihood score, the posterior probabilities are calculated. The three posterior probabilities, which will sum to one, are converted to coordinates and plotted onto a triangle diagram. If there is strong signal in the data then the probability of a particular configuration would be more like than the other two, thus the coordinates would be in one of three corners of the triangle (regions 1-3) (Figure 3.6B). Whereas, if there is conflict in the signal the coordinates would be in the centre or between the corners (regions 4-7) (Figure 3.6C). This is repeated for all possible quartets that can be obtained from each gene tree. If more than 10% of the data fall in regions 4-7 (Figure 3.6F) then the gene tree is discarded as having insufficient phylogenetic signal. From the 1,218 orthologous groups assessed, 406 were determined to have sufficient phylogenetic signal. The amino acid alignments (referred to as the "406AA" dataset) and

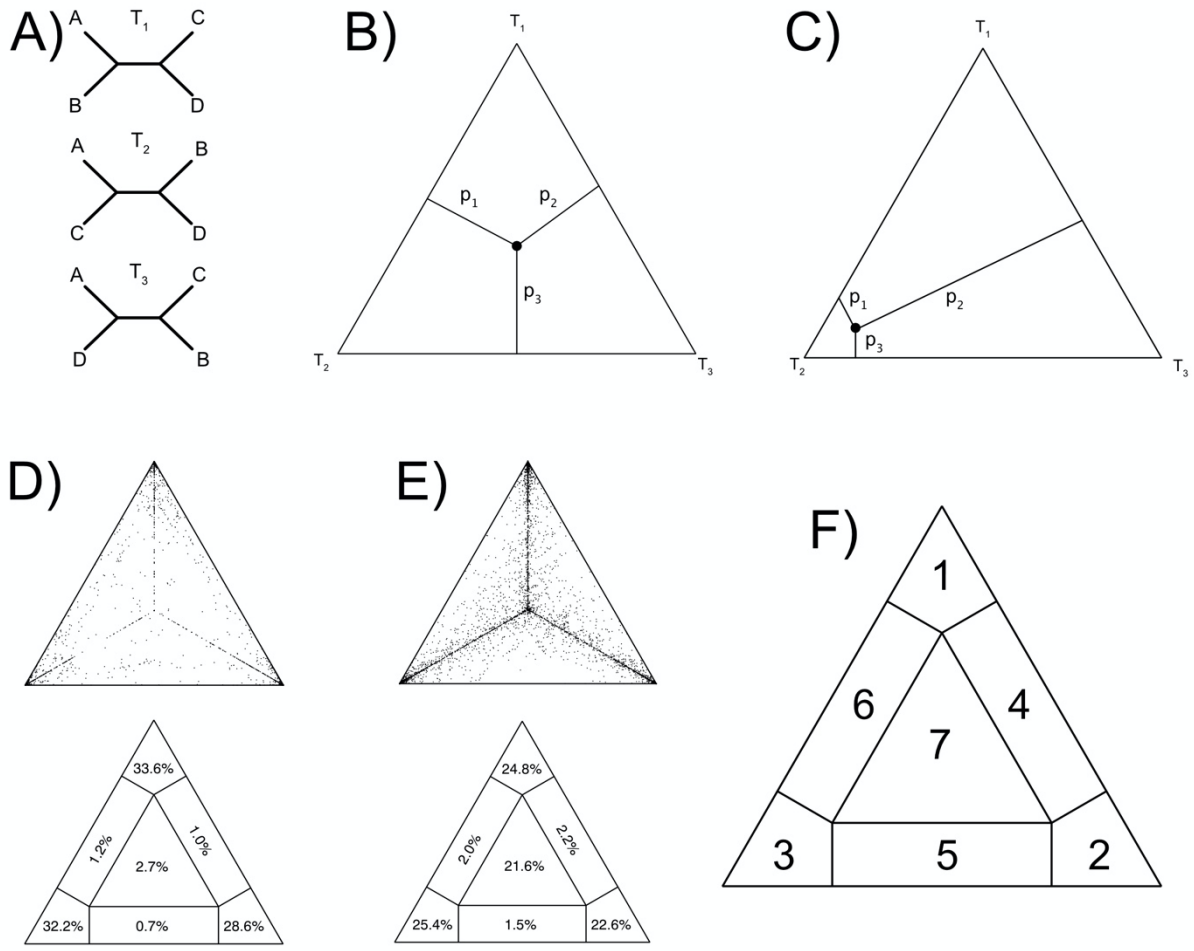


Figure 3.6. Steps of assessing phylogenetic signal of an SGO. A) A gene tree is broken down into all possible quartets; B) the likelihood of the quartet plotted on to a triangle, where all topologies have equal likelihood; C) the likelihood of the quartet plotted on to a triangle, where topology T_2 has the highest likelihood; D) Triangle plot that would result in SGO being retained as < 10% of plots falling into segments 4-7; E) Triangle plot that would result in SGO being filtered SGOs > 10% of plots falling into segments 4-7; F) Triangle plot boundary segments.

the corresponding original nucleotide files for these 406 SGOs were used to generate aligned nucleotide alignments (i.e., the "406NUC" dataset), using '*vespa.py create_database'* and '*vespa.py map_alignments*' in Vespa (Webb, Walsh and O'Connell, 2016).

3.2.1.7 Final alignment dataset

Some species used in the formation of the orthologous groups were removed from the phylogenetic analyses; these OGs were designed to be used in the positive selection analyses in Chapter 4. In total, I had 406 alignments that satisfied my filtering criteria. Two resulting datasets were used in this study: 406AA and 406NUC, which are detailed in Table 3.3. The nucleotide format of the 406 OGs (406NUC) has considerably more parsimony informative sites, this dataset was included to address whether the increased phylogenetic signal from the alignments could enhance the resolution of some the polytomies in the phylogeny.

3.2.2 Phylogenetic reconstruction

3.2.2.1 Phylogenetic analysis using a supermatrices approach

To generate supermatrices at both the amino acid and nucleotide levels, the orthologous groups of 406AA and 406NUC were each concatenated into a single alignment using *catsequences* (https://github.com/ChrisCreevey/catsequences). Phylogenetic reconstructions of both supermatrices (i.e., 406AA_cat and 406NUC_cat) were carried out using a Maximum Likelihood (ML) framework in *IQ-TREE* (Nguyen et al., 2015) and a Bayesian inference in *Phylobayes-MPI* v.4.1 (Lartillot, 2013). Automatic model selection (Kalyaanamoorthy et al., 2017) and 1000 ultrafast bootstrap replicates (Hoang et al., 2018) were performed in *IQTREE*. Two independent Monte Carlo Markov chains (MCMC) were run in *Phylobayes-MPI*, after constant sites were removed using *'-dc'* parameter, under the CAT+GTR model, with a gamma distribution consisting of four rate categories. The chains were run with sampling every 10 cycles until convergence was reached. Convergence between the chains was assessed using '*tracecomp*' and *'bpcomp' in Phylobayes-MPI*. Chains reached convergence when all maximum differences of the parameters within the tracefile produced by '*tracecomp*' function were < 0.3 with a minimum effective size > 100 and maximum difference if 0.1 produced by the '*bpcomp'* function. Burn in for the chains was estimated by visual assessment of the trace files in *Tracer* v1.7.1 (Rambaut et al., 2018).

Table 3.3. Summary statistics of datasets used in reconstruction

Data set	Taxa	Average OG Length	Concatenated Alignment length	Variable sites	Parsimony informative sites
406AA	21	1070.29 aa	434535 aa	103978	43261
406NUC	21	3210.85 bp	1303605 bp	298431	132452

3.2.2.2 Phylogenetic analysis using a supertree approach

ASTRAL-III (Zhang et al., 2018) can overcome issues of discordance present when attempting to resolve genes that have evolved under the multi-species coalescent process. This involves inferring gene trees for all the sequences in the alignment. The gene trees are then interrogated to resolve a species tree which shares a maximum number of quartets topologies with the gene trees, using a super tree process. By accounting for within branch coalescence for species, disconcordance factors such as incomplete lineage sorting can be overcome. *IQ-TREE* (Nguyen et al., 2015) was run for each orthologous group with an automatic model selection (Kalyaanamoorthy et al., 2017) and 1000 ultrafast bootstrap replicates (Hoang et al., 2018) on 406AA and 406NUC. The resulting gene trees were concatenated into one file and provided to ASTRAL-III (Zhang et al., 2018) to perform coalescent based phylogenetic reconstruction. The resulting trees were assessed for branch support using the proportion of concordant quartets around each node. ASTRAL-III does not add terminal branch lengths to the phylogeny, and so these were added to three phylogenies retrospectively using the python script *add-bl.py* (https://github.com/smirarab/global/blob/master/src/mirphyl/utils/add-bl.py). The full process is summarised in Figure 3.7.

3.2.2.3 Assessing support for contentious relationships

The resultant species topologies were assessed using the approximately unbiased (AU) tests (Shimodaira, 2002), assessing how many genes had the power to reject alternative topologies. For each of the SGOs in 406AA, 'idealised' trees were created, generating phylogenies for each of the topologies being assessed, using the script '*create_phtrees.py*' and the '*generatetrees*' function within *Clan* (Creevey & McInerney, 2005). All topologies for each SGO were assessed using AU tests performed in *IQ-TREE* (Nguyen et al., 2015) with automatic model selection (Kalyaanamoorthy et al., 2017) and 10,000 RELL replicates (Kishno et al., 1990). The resulting files generated by IQ-TREE were parsed and the number of gene trees that could reject all possible alternatives was recorded.

Concordance factors for each species tree topology were calculated in *IQ-TREE* using all 406AA gene trees. Due to the low support for some of the contentious clades, each orthologous group was assessed for significant outliers in phylogenetic signal. Parsimoniously informative sites were calculated using Phykit (Steenwyk et al., 2021) for each of the orthologous groups. The methodology in Shen (2017) was replicated to calculate the difference in gene-wise likelihood scores (ΔGLS) (Equation 3.1) between the orthologous groups using *RaxML-HPC* (Stamatakis, 2014) with the specified model set to

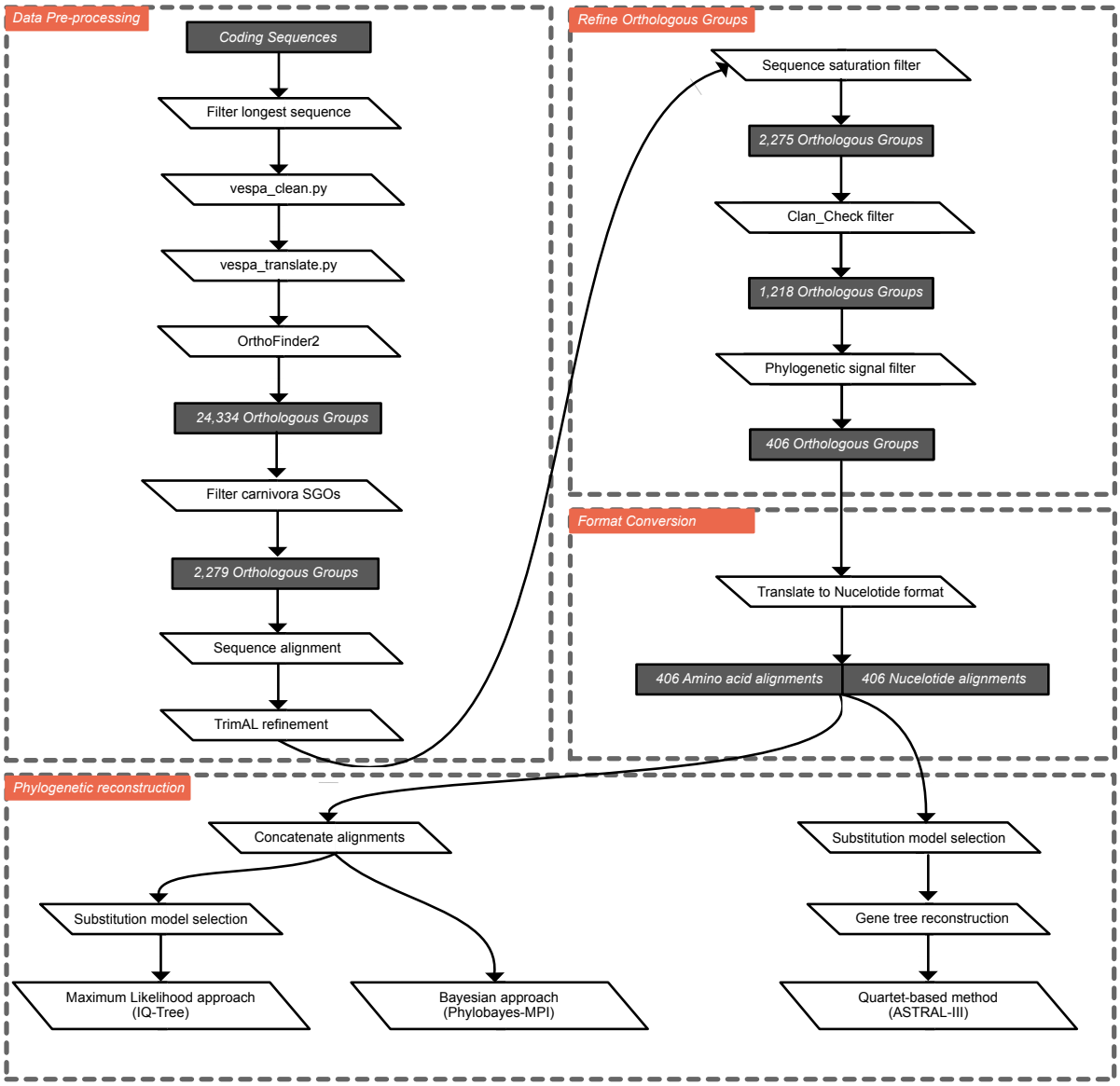


Figure 3.7 Schematic of phylogenetic reconstruction process. The diagram follows the process from raw coding sequences to resolved phylogenetic trees.

"PROTGAMMAJTT", as previously calculated in *ModelFinder* (Kalyaanamoorthy et al., 2017). The different topologies were contained within a tree file specified using the '*-z*' command to generate site likelihood scores. The site likelihood scores were calculated total over a "gene" to generate the gene likelihood score (Equation 3.1).

Equation 3.1. Calculating gene wise log likelihood score. *lnL* represents the log-likelihood score from ML analysis of a topology (T1, T2 and T3), for each 'gene' (G).

$$\Delta GLS_j = \frac{|lnL(G_j|T1) - lnL(G_j|T2)| + |lnL(G_j|T1) - lnL(G_j|T3)| + |lnL(G_j|T2) - lnL(G_j|T3)|}{3}$$

3.2.3 Testing for heterogeneity in the data sets

3.2.3.1 Test of composition heterogeneity

Prior to any phylogeny reconstruction, compositional heterogeneity was assessed in *IQ-TREE* using a chi-squared test (χ^2) test. The χ^2 test is used to assess the homogeneity of character composition across each sequence of an alignment, calculating whether the observed base frequencies (O) fall within the distribution of expected base frequencies (E) of data (Equation 3.2), where *k* is equal to the size of the alphabet (4 for nucleotides and 20 for amino acids). This test does not assume phylogenetic relatedness in the data and so can suffer from Type II error (Kumar and Gadagkar, 2001).

Equation 3.2. Chi-Squared Equation

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

3.2.3.2 Posterior predictive analyses

Phylogeny reconstruction was performed using the optimal model calculated in *ModelFinder* (Kalaanamoorthy et al., 2017) and the CAT model in Phylobayes. To assess whether the models adequately describe the patterns in the data, I carried out posterior predictive analyses (PPA) on each data set. PPA simulates data from a model using posterior parameter estimates from Bayesian phylogenetic analyses. The central concepts of PPAs are covered in Chapter 1.2.3.2.

Site-specific base diversity (PPA-DIV) measures the mean number of amino acids/nucleotides observed at each site. Using the model and parameters from the MCMC run for each data set, simulated data sets were created. PPA-DIV was calculated from the simulated datasets, if the real data fell within the distribution of the simulated datasets, then the composition of the data is well described by the employed model. The posterior predictive analyses were run using *Phylobayes* 4.1b (Lartillot et al., 2009) using the '*readmpi*' function with the '*-ppred*' option, using burn in values previously calculated in the phylogenetic analyses. For each PPA the deviation from the null expectation was given in the form of Z-scores, with a Z-score > 2 indicating the model does not fit the data.

3.2.4 Divergence time estimation

3.2.4.1 Divergence dating using time calibrations

To estimate the divergence times of speciation events across the phylogeny a white noise autocorrelated relaxed clock was employed in *Phylobayes* 4.1 (Lartillot et al., 2009). The optimal tree resulting from the reconstruction of the species phylogeny was supplied as input for the calculation. The outgroup chosen was "cat" as it represents the most basal split on the Carnivora tree given out sampling. Eight calibration points were taken from TimeTree (Hedges et al., 2015) (Table 3.4). Calibrations were chosen to account for each divergence event for pinnipeds within Carnivora. The Bearded seal was used as the representative of Phocidae, as the earliest divergent on the family and avoid any bias from using a Phocini species, as this was my clade of interest. The age of the root was specified as a gamma prior of mean 50 My (million years), with a deviation of 4 My, thse values being the date and confidence intervals from TimeTree (Hedges et al., 2015). 2 chains were concurrently run with convergence being accepted when differences > 0.3 and effective sizes > 100 for all parameters when using '*tracecomp'*, with appropriate burn in times visually assessed using Tracer v1.7.1 (Rambaut et al., 2018). Final trees were parsed using TreeAnnotator (Bouckaert et al., 2019).

3.2.4.2 Pairwise sequential Markovian coalescent tests

To estimate the divergence times of the species in the Phocini tribe I used pairwise sequential Markovian coalescent (PSMC) analyses as in (Li and Durbin, 2011). PSMC uses heterozygotic sites across regions of the genome, combined with information on the mutation rates to estimate effective population sizes (N_e) for a given species or population under a coalescent framework. PSMC plots of N_e over time can be produced for different species then overlaid to estimate species divergence times. PSMC was intended to

Table 3.4 Calibrations used for divergence time estimations. All dates are obtained from TimeTree (Hodges et al., 2015).

Species node	Relationship	Upper - lower soft age priors (Mya)
Cat – Bearded seal	Felidae - Phocidae	57 - 52
Dog – Bearded seal	Canidae - Phocidae	52 - 49
Giant panda – Bearded seal	Ursidae - Phocidae	43 - 37
Sable - Bearded seal	Mustelidae - Phocidae	42 - 37
Walrus - Bearded seal	Odobenidae - Phocidae	28.9 - 23.1
Walrus – California sea lion	Odobenidae - Otariidae	22.1 - 16.8
Southern elephant seal – Bearded seal	Phocidae – Monachinae	21.8 - 15
Hooded seal - Ringed seal	Cystophoca - Phocini	13.9 - 6.2

be performed on the 5 species of interest which had genome assemblies: Grey seal, Harbour seal, Caspian seal, Ringed seal, and Largha seal. PSMC analysis uses unphased, raw short read data and this was obtained through GenBank for one of the five species of interest: Grey seal, Harbour seal raw reads were not publicly available as the contained some human contamination. Other reads were either sequenced *de novo* through this thesis or by collaborators (Table 3.2).

Reads were mapped to their corresponding reference assembly *BWA-mem* v07.17 (Li, 2013), *samtools* (Li et al., 2009) and *bcftools* (Li, 2011) were used to calculate base coverage information and generate VCF files using commands 'mpileup' and 'call' respectively. VCF files were converted into variant informative FASTQ files using *VCFtools* (Danecek et al., 2011) and PSMC command `fq2psmcfa` was used to convert the FASTQ file to a 'psmc' file. N_e was inferred across 64 free atomic time intervals (4+25*2+4+6) set using the '-p' parameter with the options '-N 25 -t15 -r5'. A mutation rate of 7.0 x e^{-09} (Stoffel et al., 2018) and generation times extracted from the IUCN red list species profile (https://www.iucnredlist.org/) were specified using the '-u' and '-g' parameters respectively. Plots generated using 'psmc_plot.pl' in the *PSMC* package (https://github.com/lh3/psmc), an R script was used to overlay and annotate multiple plots.

3.3 Results

3.3.1 Resolving the pinniped phylogeny using high confidence 1:1 orthogroups

3.3.1.1 Supermatrix approach

Using the 406AA dataset, the model JTT+F+R6 was estimated as the best fitting model using ModelFinder in *IQ-Tree*. 'R' refers to the FreeRate model (Yang, 1995; Soubrier et al., 2012), which is a generalised Gamma model (Yang, 1994) with relaxed assumptions of gamma-distributed rates, and 'F' refers to empirical codon frequencies counted from the data. ModelFinder returned GTR+F+R6 as best fitting model for the 406NUC supermatrix, the LnL scores for all the tested models from both datasets are given in Electronic appendix 3.3.

Phylogeny reconstruction using a supermatrix was carried out using two methods: (i) ML method in *IQTree* (Nguyen et al., 2015), and (ii) BI in *Phylobayes* (Lartillot et al., 2009). The ML phylogeny reconstructed using 406AA_cat obtained 100% support for all clades (Figure 3.8). In this phylogeny, Mustelidae resolved as the sister clade to Pinnipedia, with the Ursidae clade being sister to the Mustelidae-Pinnipedia clade.

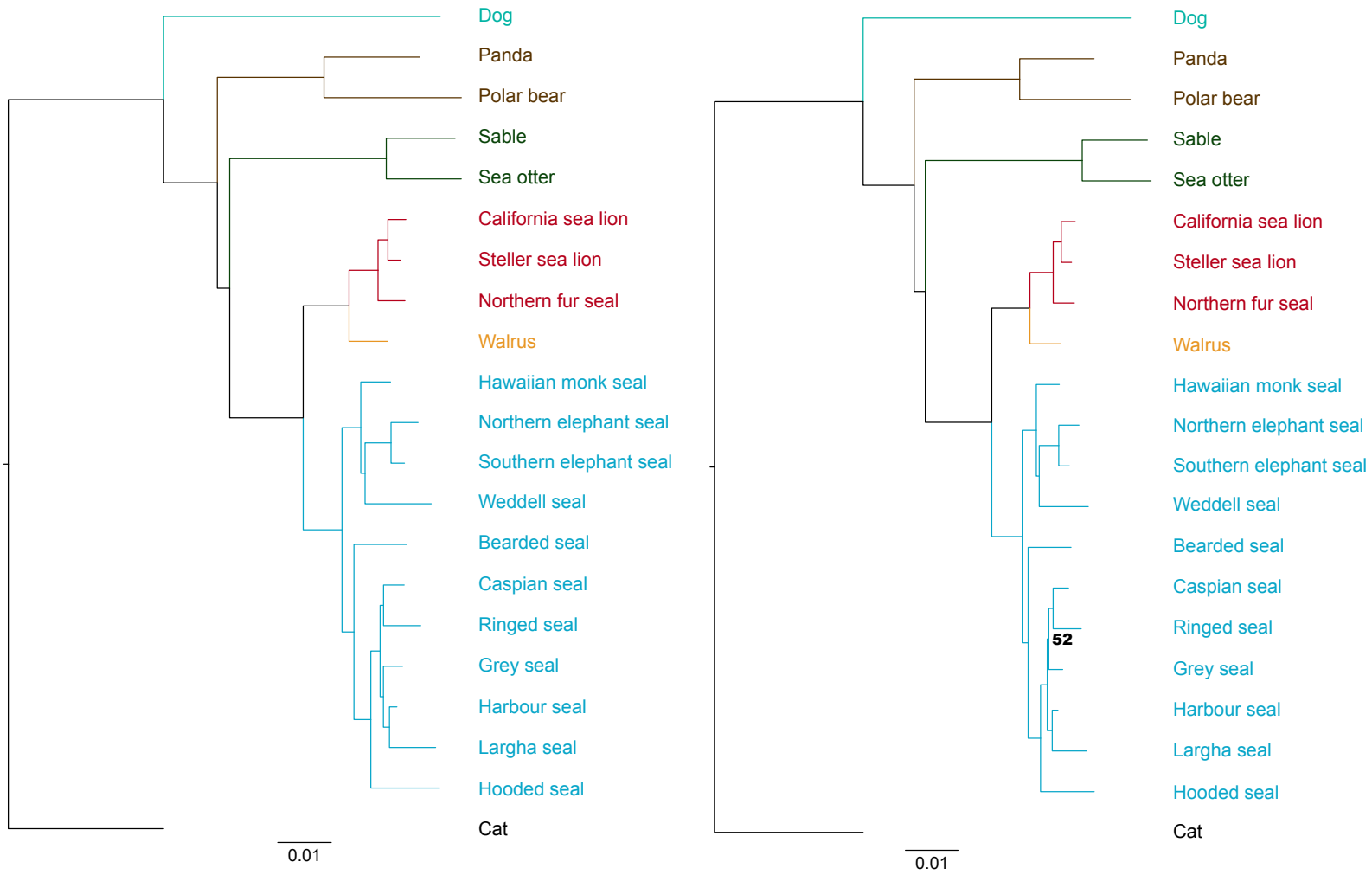


Figure 3.8. Phylogenetic reconstruction of 406 concatenated sequences through Maximum Likelihood method. Phylogenetic reconstruction of selected Carnivora species using the concatenated sequences in a Maximum Likelihood framework using IQ-Tree (Nguyen et al., 2015). Branch support is represented by ultra-fast bootstrap supports from 1000 replicates in, with supports less than 100 shown. Left) 406AA_cat sequences used with the optimal substitution model JTT+F+R6. Right) 406NUC_cat sequences used with the optimal substitution model GTR+F+R6.

The *Phocini* tribe differed from previously published pinniped phylogenies with Ringed seal forming a clade with Caspian seal*,* and *Harbour seal, Largha seal* and Grey seal forming a sister clade. This topology closely resembles the alternative topology produced by Fulton and Strobeck (2010) (Figure 3.2D), but with Grey seal existing as an outgroup to the *Phoca spp*.. The 406NUC_cat dataset produced a topology identical to the alternative topology by Fulton and Strobeck (2010) (Figure 3.2D), although the level of branch support was low for the basal *Pusa* node (bootstrap proportion: 52) (Figure 3.8).

PhyloBayes uses non-parametric methods to model among-site variations. Applying Dirichlet process mixtures to model site-specific profiles over 20 amino acid sites, or 4 nucleotide bases, which are then combined with globally defined exchange rates, which can be fixed to empirical estimates (i.e., JTT) or inferred from the data (CAT-GTR). The CAT-GTR method is expected to best fit real data, especially when using large datasets (Lartillot et al., 2015). The two chains reached convergence after 695 cycles with a burn in of 350 for the 406AA_cat dataset and after 4500 cycles for the 406NUC_cat dataset, with a burn in of 2500**.** The phylogeny run using the 406AA_cat dataset gave rise to an alternate topology within the Phocini tribe, suggesting that the Ringed seal and Caspian seal formed a *Pusa* clade, with Grey seal as sister to this clade. All nodes received posterior probabilities of 1, suggesting high support for all the configurations. The 406NUC_cat dataset generated alternative topologies, with extremely long branch lengths for *Canis familiaris* and *Felis catus*, the phylogeny also suggested Ursidae family as sister to Pinnipedia, not seen in any other phylogeny from this study (Figure 3.9). Within Phocidae, Grey seal was contained within the Pusa clade, being more closely related to Caspian seal than Ringed seal. Although the branch lengths for the species within the Phocini clan are extremely small in comparison to the outgroups.

3.3.1.2 Supertree approach

For each alignment in 406AA and 406NUC, gene trees were inferred by *IQ-Tree* (Nguyen et al., 2015) with the best fitting substitution model for each orthologous group. *ASTRAL-III* assigns node support values based on Bayesian posterior probabilities named “local posterior probabilities” (LPPs). LPPs are generated by calculating a “quartet support” of each of the three possible combinations of a quartet; these values are then used against the configuration in the final species topology to derive a posterior probability (PP) that the final topology is correct. Both 406AA and 406NUC resolved the same phylogenies through the supertree approach (Figure 3.10), although 406AA had PP of 0.95 at the Caspian seal*,* Ringed seal, Grey seal root, suggesting some slight uncertainty. The *Pusa* clade is resolved in an identical manner as 406NUC

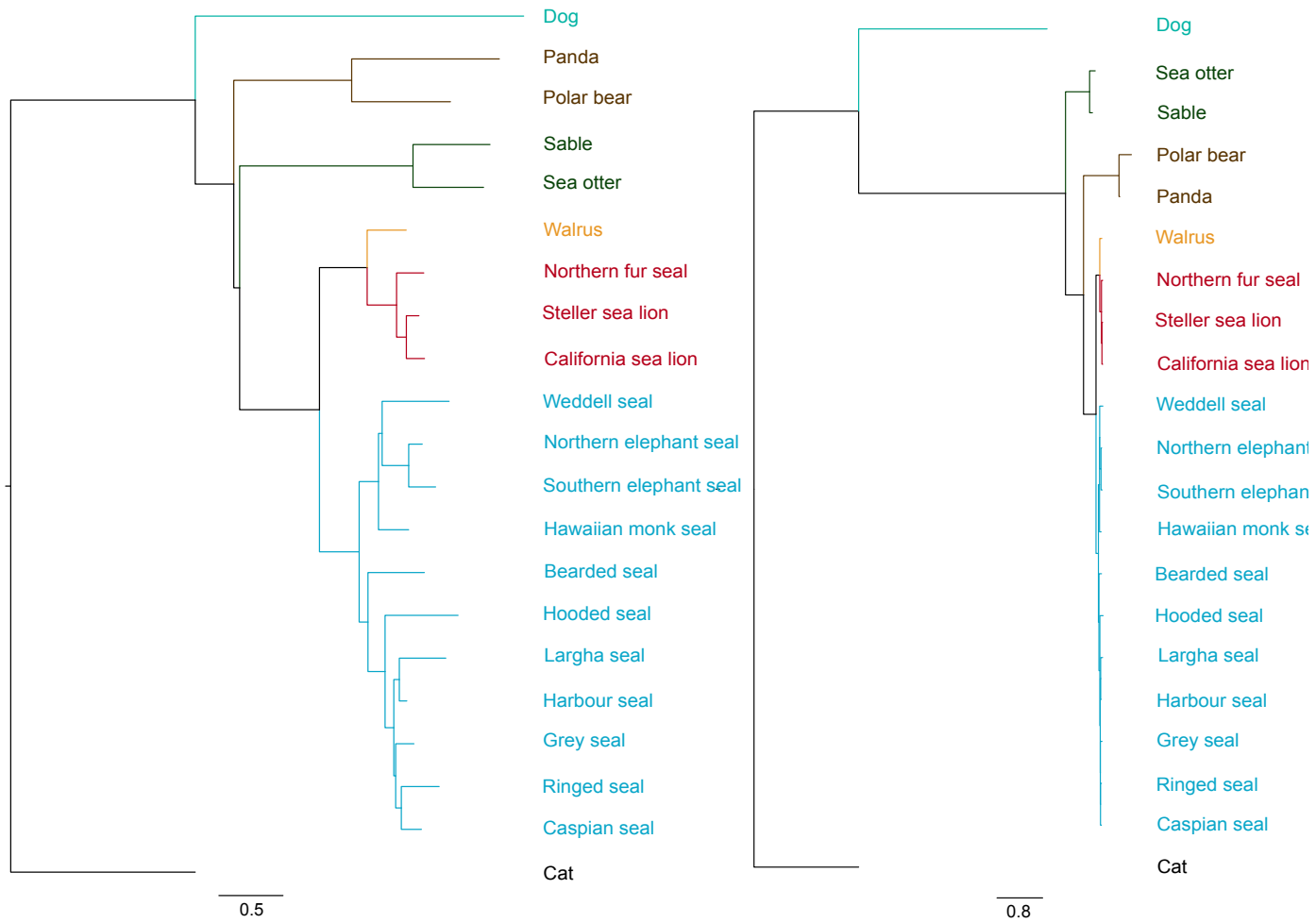


Figure 3.9. Phylogenetic reconstruction of 406 concatenated sequences through Bayesian inference method. Phylogenetic reconstruction of selected Carnivora species using the (left) concatenated amino acid sequences (406AA_cat) and (right) concatenated nucleotide sequences (406NUC_cat) in a Bayesian framework using PhyloBayes (Lartillot et al., 2009), using the substitution model CATGTR. Branch support is represented by posterior probabilities (PP), only PP < 1 are shown.

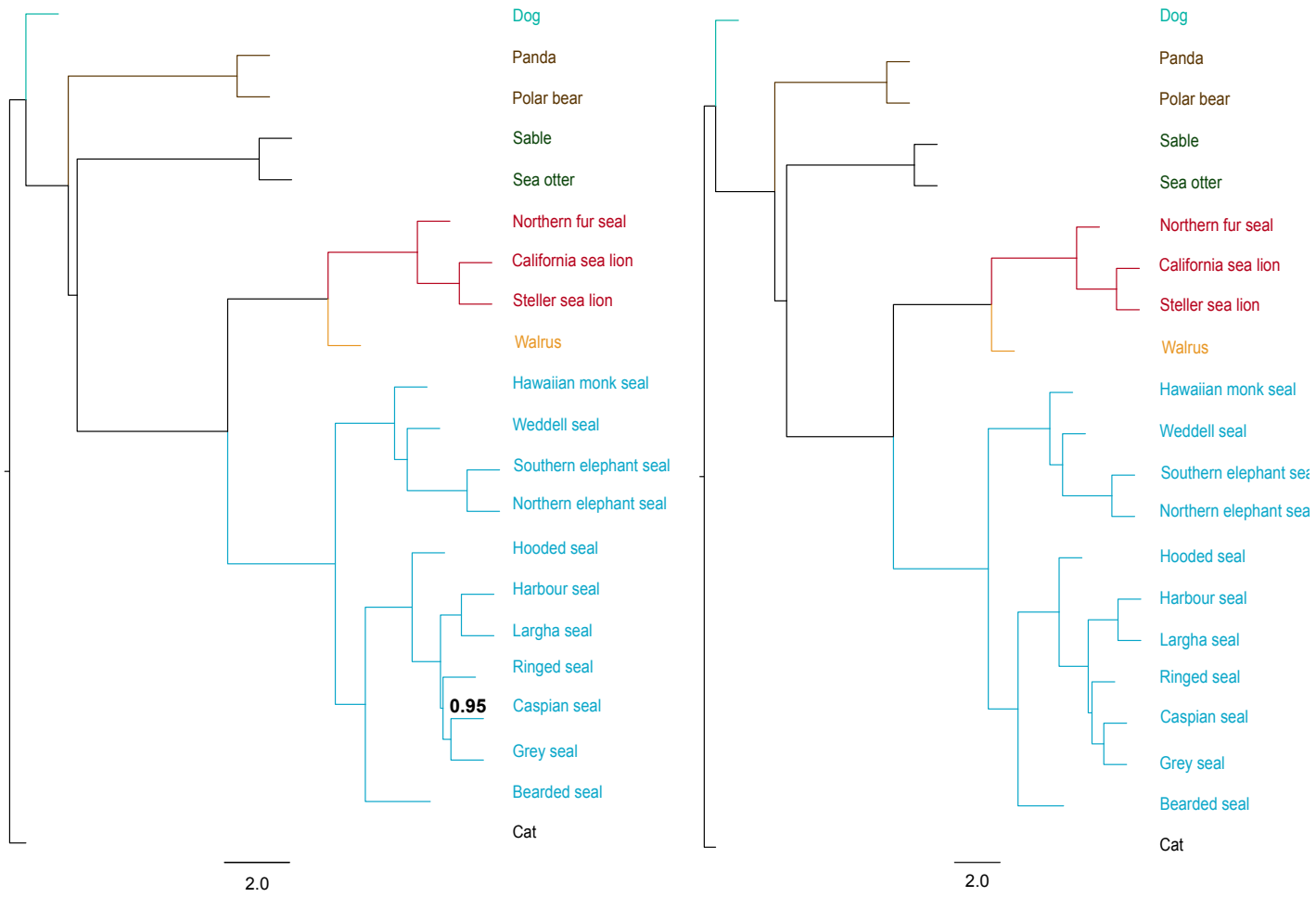


Figure 3.10. Phylogenetic reconstruction of 406 non concatenated sequences through supertree method. Phylogenetic reconstruction of selected Carnivora species using the (left) amino acid sequences (406AA) and (right) nucleotide sequences (406NUC) in a supertree framework using ASTRAL-III (Zhang et al., 2018), substitution model assessment and gene trees for each orthologous group were performed individually in IQ-Tree (Nguyen et al., 2015). Branch support is represented by local posterior probabilities (LPPs), only LPPs < 1 are displayed.

Bayesian approach, increasing the support for Caspian seal and Grey seal to form a clade in the true species topology.

3.3.2 Does the removal of phylogenetic signal outliers improve congruence?

The resolution of phylogenies across ToL has relied on a very small number of genes or sites (Shen et al., 2017). The aim of this section is to understand if some individual genes have a significant impact on a particular phylogeny, identifying the genes that heavily skew the inferred topology and contribute to the incongruence seen between approaches. It is hypothesised that removing genes that possess large differences of likelihood support between topologies will reduce bias between individual topologies. Although I am removing genes from my data, only a very small percentage of the overall data will be discarded, in respect to the entire dataset.

The initial gene support from each topology: 406AA supertree (T1), 406AA Maximum Likelihood (T2), and 406AA Bayesian (T3) were calculated using a gene concordance factor (gCF), quantifying the percentage of genes that agree with each topology (Figure 3.11). The gCF value for the nodes conveying Ursidae being sister to Mustelidae, and pinniped was only 46.06% (187 genes). The major clades within Pinnipedia all received high levels of support, with the Otariidae-Phocidae sister hypothesis resolved with 99.26% (403) of genes, and Walrus (Odobenidae)-Otariidae sister hypothesis receiving support from 97.04% (394) of genes. General trends of support were lower within the pinniped families. The support between tribes in Monachinae ranged from 81.03% (for the Monichini tribe as outgroup to the rest of Monchinae) to 50.49% of genes supporting Miroungini and Lobodontini being sister clades. Within the Phocini tribe, the level of discordance due to polyphyly (gDFP) is high in comparison to the rest of the phylogeny (33.9 – 61.82%). Support for the clustering of Grey seal with Largha seal and Harbour seal received very low support, with only 9.61% (39 genes) supporting this topology. More than double (20.69%, 84 genes) support Grey seal as a clustering with Caspian seal and Ringed seal within the Pusa clade. Only 16.5% (67) of genes support Ringed seal as sister to Caspian seal whereas 32.02% (130) genes are in agreement with Grey seal being sister to Caspian seal.

Identifying genes with the highest levels of phylogenetic signal would help determine whether a small number of genes were having a large impact on the incongruence between resulting phylogenies. For this analysis, phylogenetic signal is defined as the difference in log-likelihood scores between the three alternative resolutions, T1, T2 and T3 (Equation 2). The log-likelihood scores of a gene are calculated by

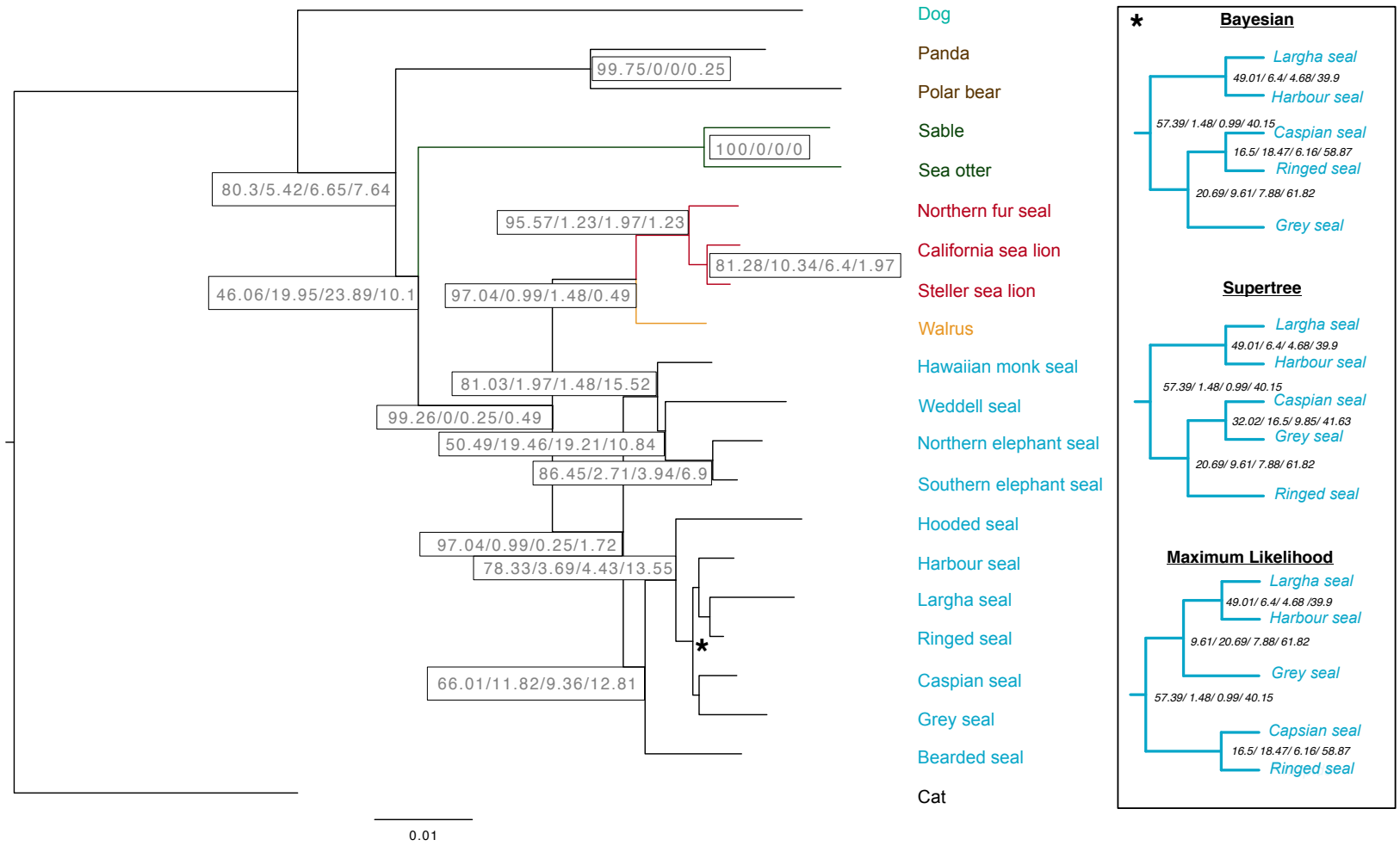


Figure 3.11 Levels of support counts of genes for each node in phylogeny. Values refer to: gCF (proportions of genes in concordance with the topology) / gDF1 (Gene discordance factor for nearest neighbour 1 branch) / gDF2 (Gene discordance factor for nearest neighbour 2 branch) / gDFP (Gene discordance factor due to polyphyly). Three different topologies for Pusa clade are detailed in the box.

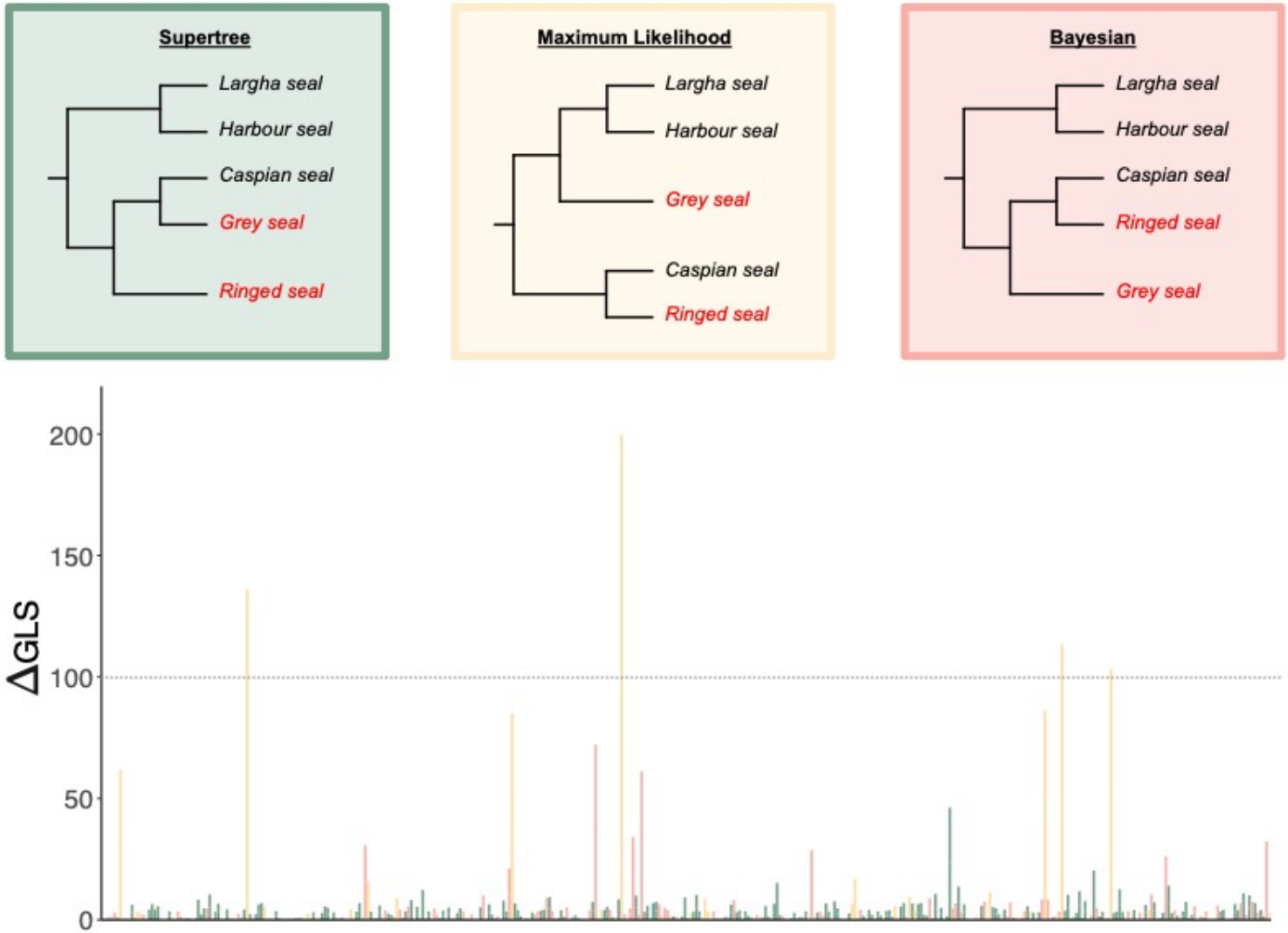


Figure 3.12 Phylogenetic signal differences (ΔGLS) between the best fitting topology and alternate topologies from the three different methods of phylogenetic reconstructions. Each bar represents a MSA with colours indicating the best fitting topology: yellow bars for the Maximum Likelihood derived topology, red bars are for the topology resolved using a Bayesian approach, and green bars indicate a supertree methodology.

extracting the total value of the log-likelihood scores of all the sites within a gene. From the 406 genes, 210 (51.72%) had a likelihood score that favoured the T1 topology, 91 (22.41%) favoured the T2 phylogeny and 105 (25.86%) favoured T3. The distribution of *Δ*GLS across the 406 genes shows that a very small proportion, 4 genes, had a *Δ*GLS over 100, whereas 397 of genes have a *Δ*GLS of less than 50 (Figure 3.12), all of which favoured the T2 topology. Given this large disparity in *Δ*GLS across the different genes suggests that a small number of genes are skewing the overall topology seen when concatenating the alignments. In Shen (2017), it was seen that removing just 5 genes substantially changed the topology of the tree, to retain as much data as possible the genes with the highest 1% (4 genes) of *Δ*GLS were removed from the dataset to create dataset 402AA. The four genes removed were OG008969/ ZN609 (*Δ*GLS = 199.65), OG0007790/ ERCC6 (*Δ*GLS = 136.05), OG0010807/ C2D1B (*Δ*GLS = 113.38), and OG0011205/ N4BP2 (*Δ*GLS = 103.10).

Phylogenetic reconstructions were performed using the 3 different approaches again using the new 402AA dataset. The supertree method using ASTRAL-III remained the same topology with the reduced 402AA dataset as it did with the 406AA dataset, with Caspian seal and Grey seal remaining in the same clade and Ringed seal as sister. The branch support for the Caspian seal – Grey seal clade was also increased, from 0.95 to 1 (Figure 3.13A). After using the 402AA in the Maximum Likelihood approach, the topology resolved to be identical to that seen in the supertree approach, but the branch support for the Caspian seal – Grey seal is decreased, with a support of 71% (Figure 3.13B). The only approach now displaying incongruence between the phylogenies is the Bayesian approach ((Figure 3C), which reached convergence after 535 cycles with an initial burn in of 150. The Bayesian approach resolved Ringed seal and Caspian seal in their own clade with Grey seal as sister to this clade, showing posterior probabilities of 1, replicating the relationships observed in the alternative topology of Futon and Strobeck (2010) (Figure 3.2D).

The removal of just 4 genes significantly improved congruence across the three approaches, the 1% removal methodology was run for another iteration, just comparing trees T1 and T3. From the 402 genes, 258 (63.55%) favoured T1 whereas 148 (36.45%) favoured T3 (Figure 3.14). The differences of GLS between the two topologies was significantly reduced when comparing the two resolutions, with only 12 genes having a *Δ*GLS greater than 10. Only the top 1% (4 genes) were removed from the subsequent analysis which resulted in a dataset of 398 genes. The four genes removed from the 402AA database were:

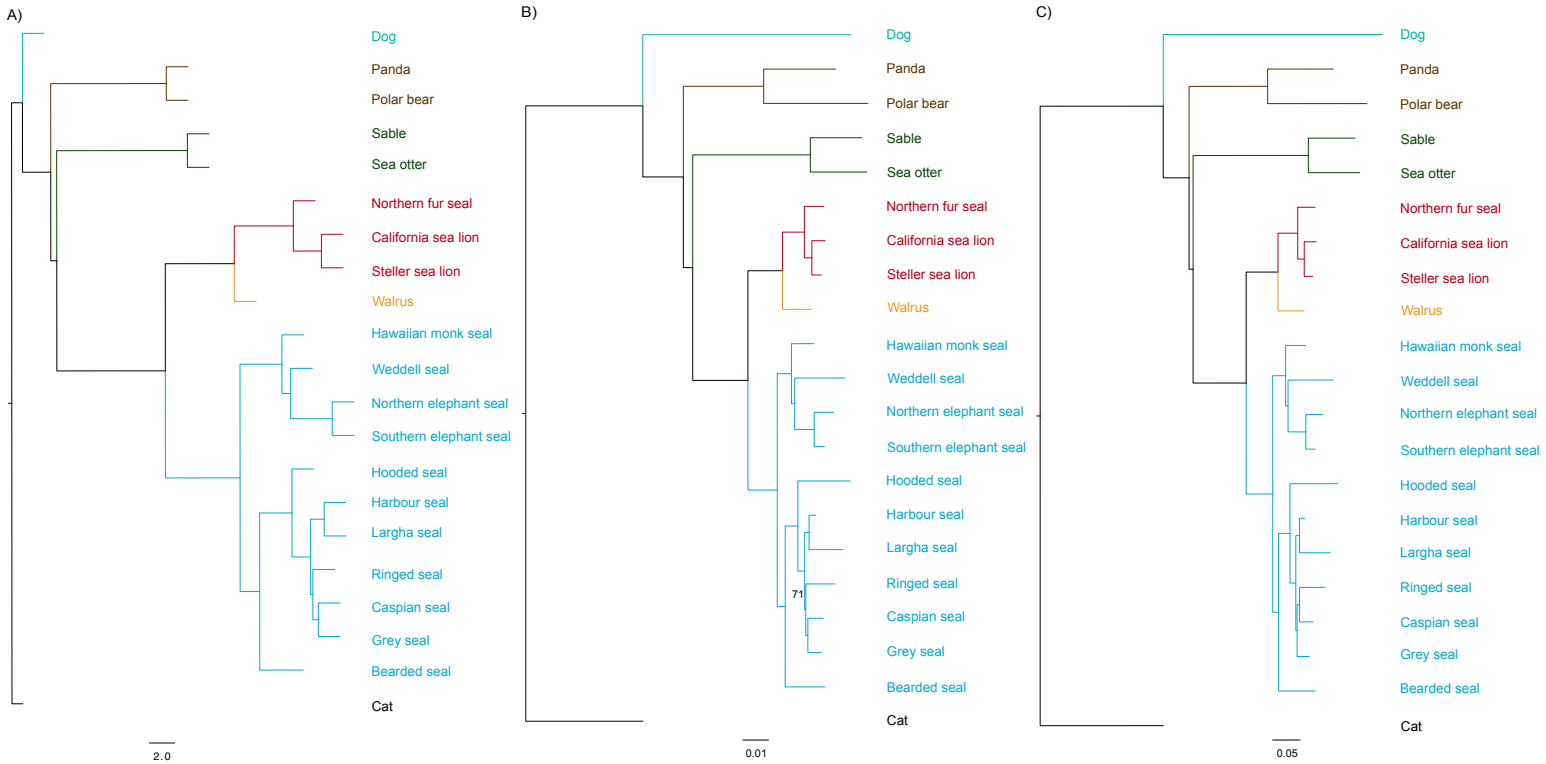


Figure 3.13 Phylogenetic reconstructions of 402AA dataset. Phylogenies produced in: A) a supertree framework using ASTRAL-III (Zhang et al., 2018) substitution model assessment and gene trees for each orthologous group were performed individually in IQ-Tree (Nguyen et al., 2015), B) Phylogenetic reconstruction of selected Carnivora species using the concatenated amino acid sequences (402AA_cat) in a Maximum Likelihood framework using IQ-Tree (Nguyen et al., 2015), using the optimal substitution model JTT+F+R6, C) a Bayesian framework using PhyloBayes (Lartillot et al., 2009), using the substitution model CATGTR, Branch support is represented by posterior probabilities or ultra-fast bootstrap supports from 1000 replicates, only values less than 1 or 100 are shown for the PPs and bootstrap supports respectively.

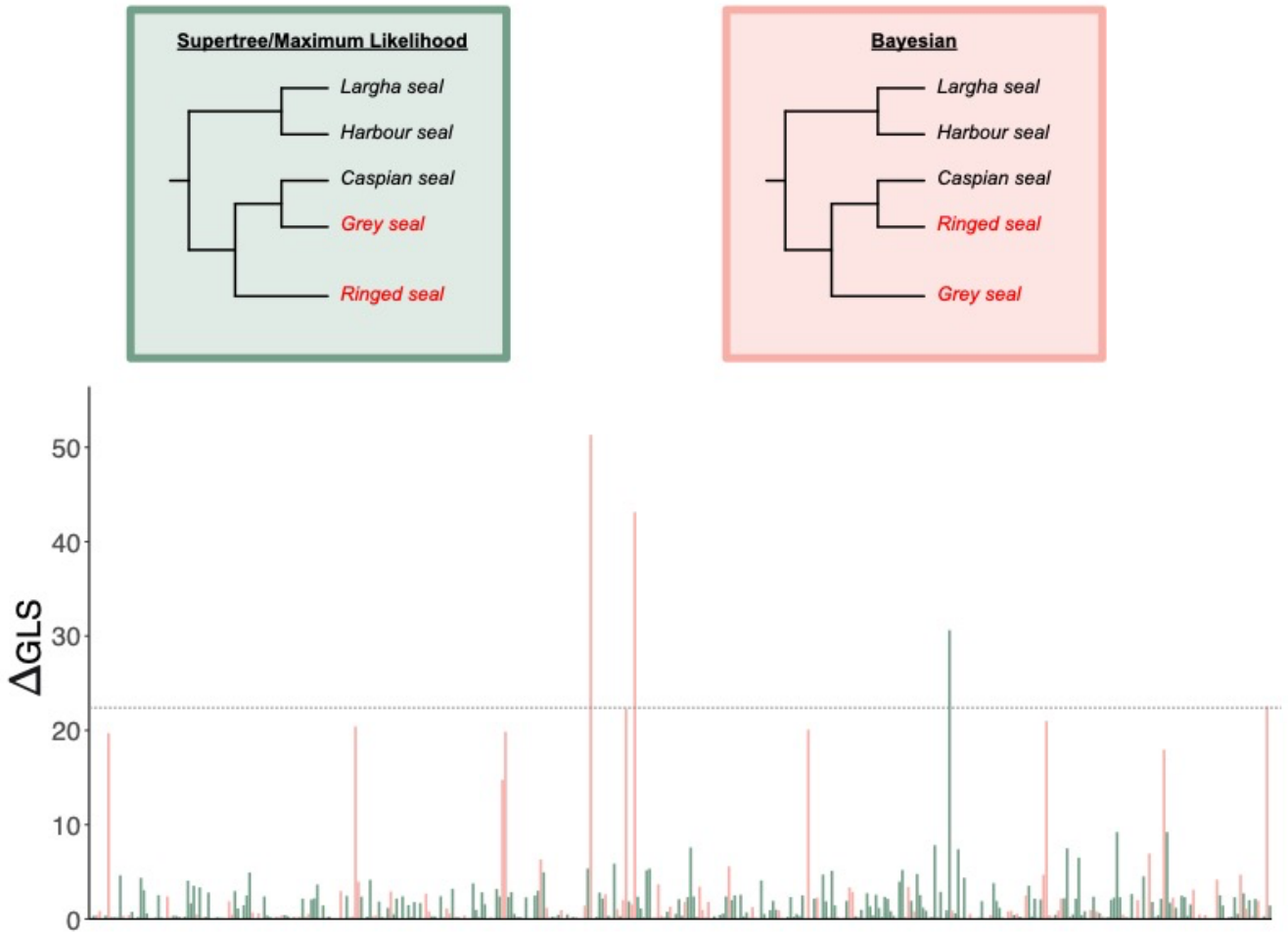


Figure 3.14. Phylogenetic signal differences (ΔGLS) between the best fitting topology and alternate topologies from the three different methods of phylogenetic reconstructions, using 402 data set. Each bar represents a MSA with colours indicating the best fitting topology: red bars are for the topology resolved using a Bayesian approach, and green bars indicate the identical topologies generated from a supertree/Maximum Likelihood methodology.

OG0008883/FANCM (*Δ*GLS = 51.31), OG0009035/SIK3 (*Δ*GLS = 43.14), OG0009912/PEPL (*Δ*GLS = 30.62), and OG0014259/THRSP (*Δ*GLS = 22.54).

Congruent phylogenies were derived from the 398AA data set when run under the three different phylogenetic approaches. With the final topology resembling T1, with Grey seal as sister to Caspian seal and Ringed seal as sister to this Caspian seal – Grey seal clade. The ML approach retained the Grey seal as sister to Caspian seal with an improved bootstrap support of 100%. The Bayesian approach and the supertree approach both resolved an identical phylogeny with posterior probabilities of 1. To further inspect the confidence of the phylogeny, the corresponding nucleotide format of the 398 orthologous groups from 398AA were interrogated to investigate whether they also resolve the same topology. The ML and supertree approaches both resolved the T1 topology with maximum bootstrap values and posterior probabilities respectively. The Bayesian approach resolved the same phylogeny for the Pinnipedia, although the chains failed to converge after 10,000 iterations. From further analysis of the topologies generated by the chains throughout the run, the Mustelidae and Ursidae were responsible for the lack of convergence. Pinnipedia remains fixed after 10,000 iterations, with Mustelidae and Ursidae constantly switching as sister to the pinniped clade, suggesting the uncertainty from the dataset is derived from the Ursidae and Mustelidae sequences rather than that of pinnipeds. Ursidae as sister to Pinnipedia was also resolved when running the Bayesian analysis with the 406NUC dataset, suggesting that the genes with high phylogenetic signal were genes that only affected the pinniped clade and a different set of genes or sites are responsible for this incongruence.

The AU test is an accurate test that was developed to test the confidence in regions of a phylogeny. Here, the AU test is used to determine two factors that were still unresolved from the initial reconstruction analysis: (1) can the 398NUC dataset significantly reject the hypothesis of Ursidae as sister to Pinnipedia, in favour of Mustelidae as sister to Pinnipedia, (2) can the T1 topology seen as favoured by the 398AA dataset reject topologies T2 and T3, and how many individual genes in the 398AA dataset can also significantly reject the two alternate topologies. The 398NUC alignment was run in an AU test using the two differing pinniped sister hypotheses, Ursidae-sister and Mustelidae-sister. The results of the AU test could significantly reject the Ursidae-sister topology (*p-value* 4.8×10^{-6}), supporting the Mustelidae sister hypothesis. The Bayesian analysis for the 398NUC was then rerun specifying a fixed tree of Mustelidae-sister, this converged within 4425 cycles with a burn in of 1000.

To assess the confidence of T1 over T2 and T3 an AU test was also run on a gene-by-gene basis. I generated idealised gene trees for the 398 genes in the 398AA dataset, before running an AU test on the three alternate topologies to assess whether two of the alternate topologies could be significantly rejected in favour of the third topology. Using the 398 genes in the 398AA dataset, 109/398 could significantly reject two out of the three trees, with 95/109 (87.16%) rejecting all but T1, 4/109 (3.67%) rejecting all but T2, and 10/109 (9.17%) rejecting all but T3.

3.3.3 Tests of compositional heterogeneity

The accuracy of modelling evolutionary processes from the data is crucial for confidence in the resulting phylogenies. To assess the fit of models used in the phylogenetic reconstructions in comparison to the resulting phylogenies PPAs were used to perform model adequacy tests. PPA-DIV calculates the mean amino acid diversity per site. This test requires MCMC chains resulting from Bayesian analyses, for this analysis I ran Bayesian analyses in *Phylobayes* for all my datasets i.e., 398AA, 398NUC, 406AA, 406NUC using both a heterogeneous model (CATGTR) and a site homogeneous model (JTT + 6Γ), specified as the optimal model from ModelTest, which does not account for heterogeneity over the data. PPAs had z-scores ranging between -1.636 and -0.169 across all chains and both datasets using the homogeneous JTT model whereas using the CATGTR heterogeneous model z-scores indicated a bad fit for the model, ranging from -18.264 to -26.607 (Table 3.5). This indicates that despite using a range of genes within my dataset, which would be expected to be subjected to various evolutionary rates, the JTT model captures the variation between the rates to a degree to sufficiently to recreate the model in simulations. Thus,
the resolved topology of the 398AA dataset, modelled with the JTT + 6Γ was chosen as the final topology used for downstream analyses (Figure 3.15).

3.3.4 Can timings of speciation events be determined?

Although the purpose of this analysis was to attempt to derive the evolutionary relationships within Arctoidea and specifically Phocidae, I attempted to infer the timings of speciation events to link with paleogeographic events.

The data and topology used to produce the final phylogeny (398AA) was used as an input for the time calibrated chronogram (Figure 3.16), with prior age estimates for nodes across the phylogeny. Divergence dating ran for 4000 cycles across 2 chains, reaching convergence with a burn in of 3300. The root of

Table 3.5 Z-scores for each model fit on the different datasets. Calculated using PPA-DIV function in Phylobayes v4.1 (Lartillot et al., 2009), to 2 significant figures.

Dataset	**Homogeneous model (JTT/GTR)**				**Homogeneous model (CATGTR)**			
	Chain 1		Chain 2		Chain 1		Chain 2	
	Z-score	P-value	Z-score	P-value	Z-score	P-value	Z-score	P-value
398NUC	-29.42	1	-40.30	1	-83.26	1	-80.21	1
398AA	-0.60	0.69	-1.64	0.94	-18.26	1	-26.61	1
406NUC	-33.90	1	-33.24	1	-83.62	1	-81.00	1
406AA	-0.65	0.77	-0.17	0.54	-19.22	1	-22.51	1

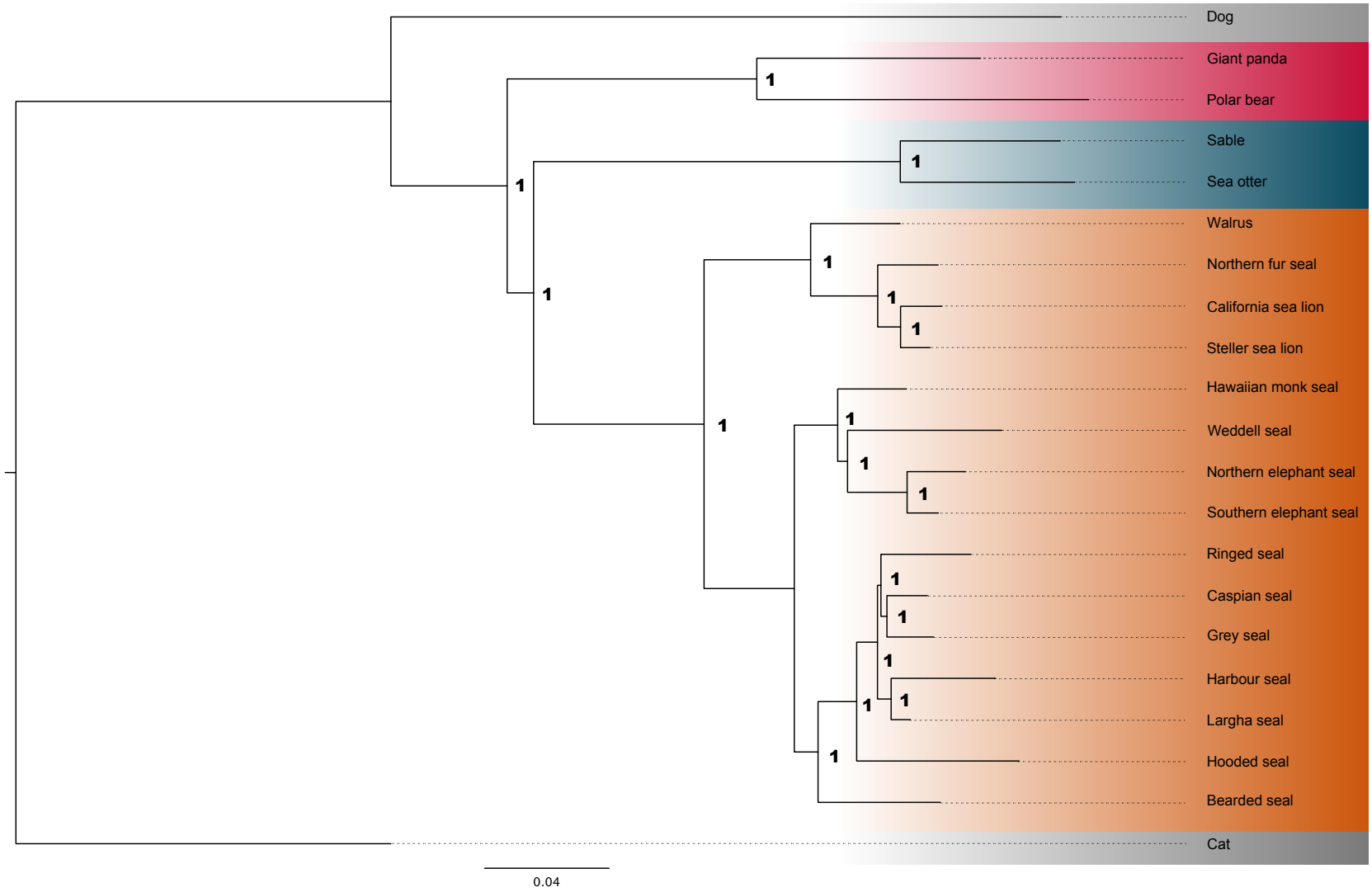


Figure 3.15 Final congruent phylogeny using 398AA. Passing model fit tests, using JTT + 6Γ substitution model in a Bayesian framework. Phylogeny is shaded by clades, grey for Felidae outgroup, green for Canidae, Ursidae in pink, Mustelidae in blue, and Pinnipedia in orange.

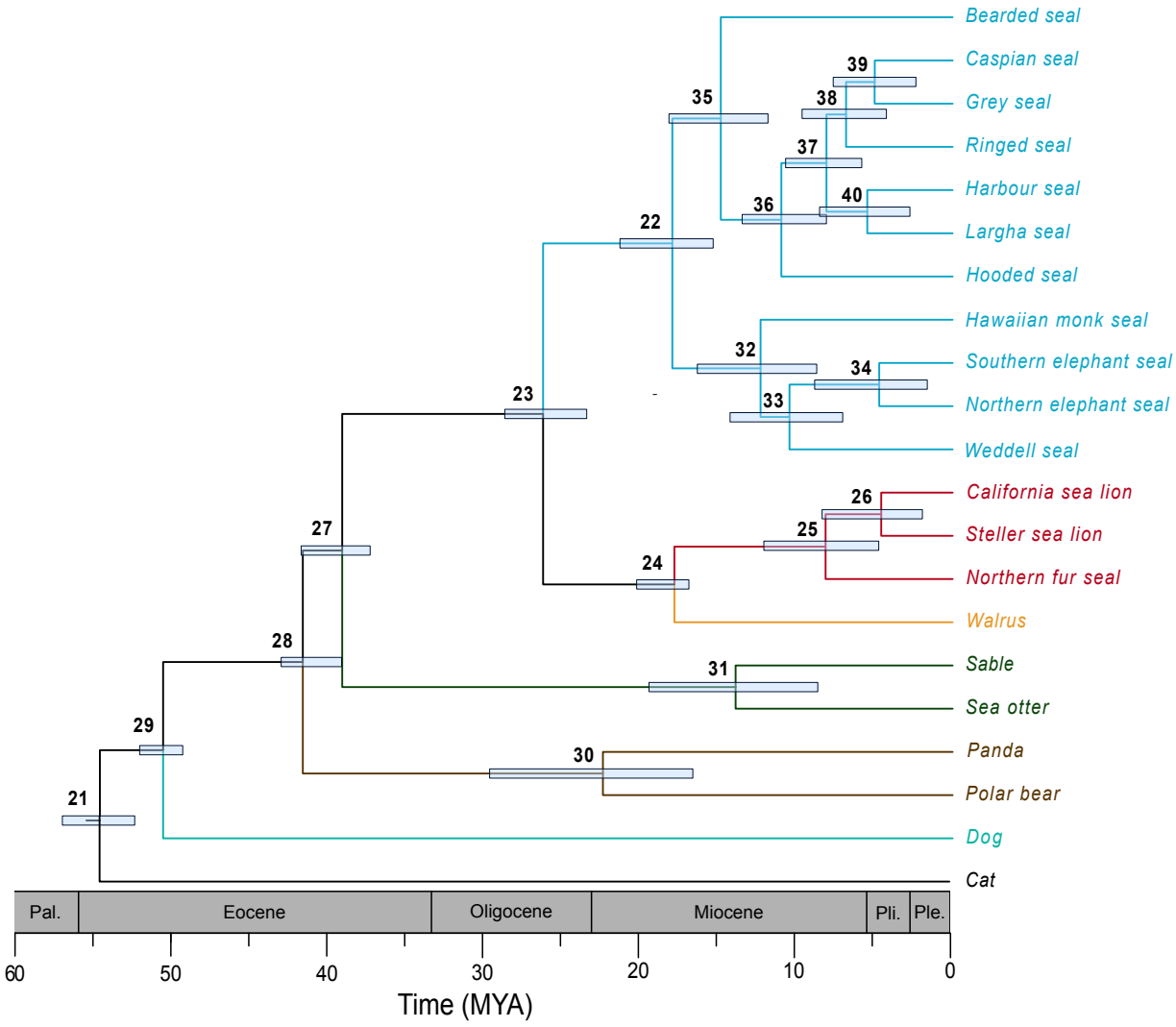


Figure 3.16. Time calibrated species tree of Pinnipedia. Estimations performed in Phylobayes, using 398 AA dataset. Blue bars represent the 95% confidence intervals. Numbers correspond to the date overview in Table 3.5.

Table 3.6. Calibrated dating intervals of nodes. Node number correspond to node numbers in Figure 3.16**.**

Node number	Node Name	Mean date	95% Lower Confidence Limit	95% Upper Confidence Limit
21	Feliformia - Carniformia	54.51	52.22	56.90
22	Monachinae - Phocinae	17.92	15.26	21.25
23	Phocidae - Otariidae/Odobenidae	26.18	23.35	28.65
24	Otariidae - Odobenidae	17.79	16.82	20.24
25	Northern fur seal - Otariinae	8.12	4.67	12.13
26	California sea lion - Steller sea lion	4.57	1.74	8.29
27	Pinnipedia - Musteloidea	39.02	37.14	41.62
28	Ursidae - Musteloidea/Pinnipedia	41.54	38.99	42.95
29	Canidae - Arctoidea	50.47	49.10	51.92
30	Ursidae spp.	22.36	16.47	29.67
31	Musteloidea spp.	13.88	8.45	19.45
32	Monachini - Lobodontini/ Miroungini	12.28	8.52	16.38
33	Miroungini - Lobodontini	10.43	6.88	14.24
34	Miroungini	4.70	2.10	7.86
35	Bearded seal	14.82	11.73	18.14
36	Hooded seal	10.95	8.11	13.51
37	Phoca - Pusa	8.07	5.78	10.75
38	Harbour seal - Pusa	6.81	4.13	9.66
39	Caspian seal – Grey seal	4.98	2.29	7.73
40	Harbour seal – Largha seal	5.46	2.65	8.59

Carnivora node, Caniformia-Feliformia, is estimated to split 54.51 MYA (52.22 – 56.90 MYA, 95% confidence), this is more recent than the 65 MYA split reported by Nyatakura and Bininda-Emonds, (2012), but is consistent with the 54.7 MYA split reported by Meridith et al. (2011). The divergence of the Pinnipedia and Mustelid lineages is estimated to have occurred 39.02 Mya (37.14 – 41.62 MYA, 95% confidence), separating from the Ursidae lineage 41.51 Mya (38.99 - 42.59 MYA, 95% confidence). From my estimates, Grey seal and Caspian seal separated 4.98 MYA (2.29 – 7.73 MYA, 95% confidence) this is significantly older than reported in previous estimates (Higdon et al., 2007; Nyakatura and Bininda-Emonds, 2012; Arnason et al., 2006) which have all estimated a divergence of approximately 1-2 MYA.

The history of the effective population size (N_e) was modelled from the distribution of heterozygous sites across the genome using a PSMC analysis (Figure 3.17). I performed this analysis separately for each of the *Pusa, Phoca* and Grey seal species for which I had genome data (Table 3.2). I found that Largha seal reaches coalescence earlier than the other species and has a lower N_e (< 50,000) compared to the other Phocidae species. This is consistent with low heterozygosity within the genome and may reflect the decrease in population sizes that the Largha seal may have suffered in recent years (Boveng et al., 2009). Caspian seal and Grey seal, followed a similar trajectory of N_e from 10 Mya to 1 Mya, suggesting the same population variations effects could have been as a result of experiencing similar climatic events. From the PSMC, Caspian seal had a very high N_e in the past 10,000 years and could be reflective of the high population sizes seen estimated to be >1,000,000, prior to the population decrease in the past century (Harkonen et al. 2012).

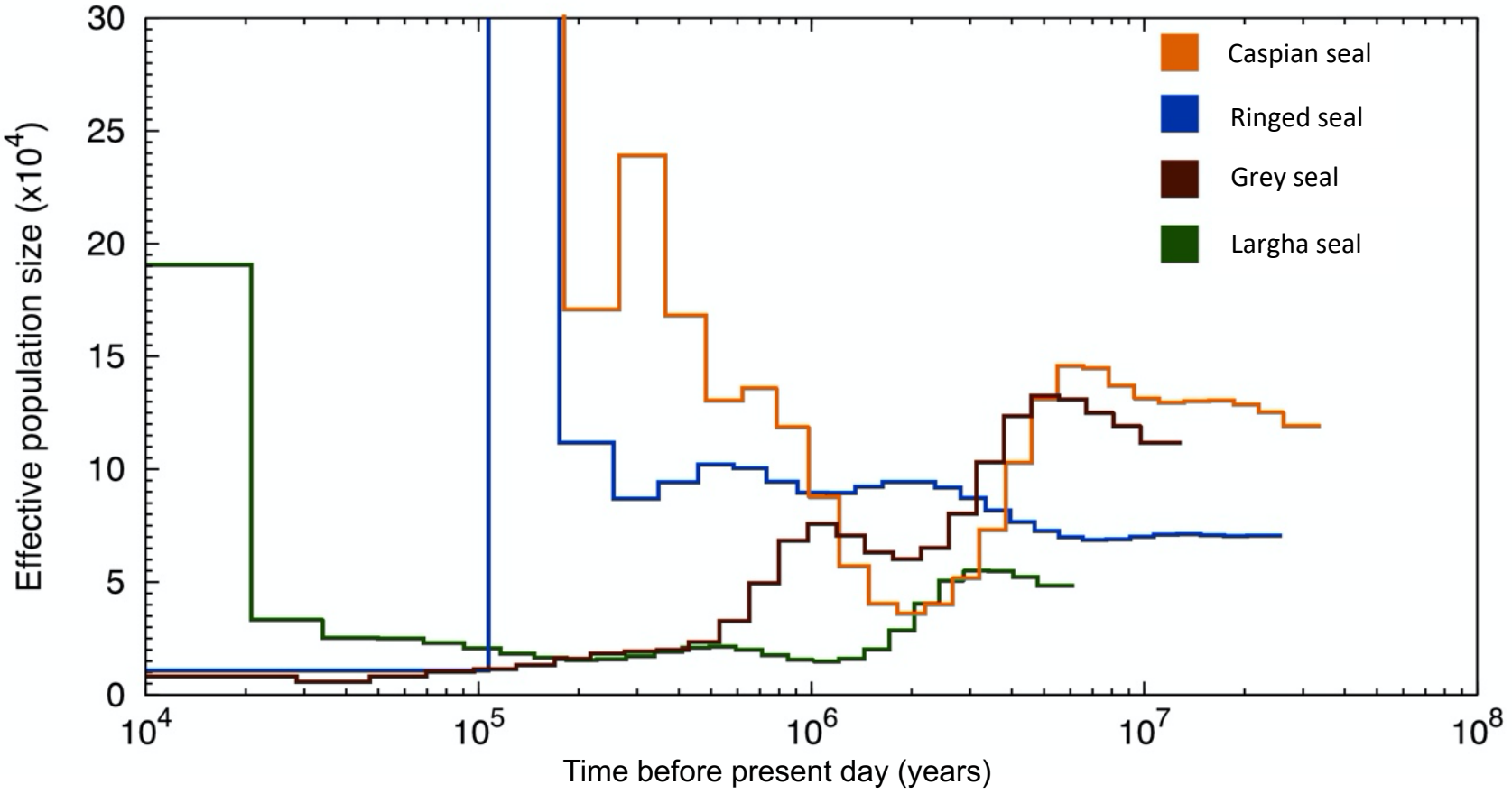


Figure 3.17. Historical N_e using PSMC analysis. Performed for available Pusa/Phoca species. Plots were scaled using a mutation rate (μ) of 0.4 x 10^8 substitutions per nucleotide per generation and species-specific generation times were extracted from the IUCN red list species profile (https://www.iucnredlist.org/). The y axis shows N_e and the x axis shows time. Effective population sizes are cut off at 30x10^4 for comparisons between species.

3.4 Discussion

I present the species tree of Pinnipedia using whole genomes, identifying 398 genes that are capable of consistently recovering concordant phylogeny across multiple statistically distinct phylogenetic inference methods: Maximum Likelihood, Bayesian, and a super tree coalescent based method. With this dataset I was able to resolve the ambiguous placement of pinnipeds within the Arctoidea, confirming pinnipeds as sister clade to Mustelidae and addressing uncertainties over the topology of the *Pusa* clade. My phylogeny confidently resolved Grey seal as the sister taxa to Caspian seal. This leaves, the placement of the Baikal seal as a remaining ambiguity within *Pusa* (a Baikal seal genome sequence was not available at the time the phylogenetic analyses were conducted) to be addressed with full genome derived data. Previous smaller scale molecular studies suggest *P. siberica* clusters within the *Pusa* clade, but its position as either sister to Ringed seal or Caspian seal is still unresolved (Higdon et al., 2007; Nyakatura and Binida-Emonds, 2012). Genome assemblies are still lacking for 7 Phocidae species, the Harp seal, Ribbon seal (*Histriophoca fasciata*), Crabeater seal *(Lobodon carcinophaga*), Ross seal (*Ommatophoca rossii*), Leopard seal *(Hydrurga leptonyx*, Mediterranean monk seal (*Monachus monachus*), and Baikal seal. The Baikal seal genome assembly has been published since this analysis (Yuan et al., 2021) and consortia such as pinniped consortia should prioritise the assembly of the other Phocidae species to fully resolve all ambiguities in Phocidae evolution.

Musteloidea as sister clade to Pinnipedia

My results strongly support Musteloidea as the sister clade to Pinnipedia, in contrast to some studies that supported Ursidae, or an Ursidae-Musteloidea branch as sister to Pinnipedia (Delisle and Strobeck, 2005; Feijoo and Parada, 2017). Disconcordance occurring between topologies whilst using concatenated and coalescent based methods are well recognised within Arctoidea (Gatesy et al., 2016; Edwards et al., 2016; Feijoo and Parada, 2017). Although we experience discordance in my study, this was overcome through stringent filtering. From my stringent filtering I was able to distinguish a set of 406 genes that provided sufficient phylogenetic signal and were predicted to be truly orthologous. This was further reduced to 398 genes as eight genes were identified as causing incongruence between the phylogenetic methods. Resolving Musteloidea as sister to Pinnipedia contrasted with the most recent molecular phylogeny of Arctoidea (Feijoo and Parada, 2017), which produced a Musteloidea-Urisidae clade as sister to pinnipeds. Feijoo and Parada (2017) did not perform any homology checks of the raw data and were reliant on the

coalescent-based supertree methods to overcome inconsistencies accumulated through “noisy genes”. Springer and Gatesy (2017) reanalysed this 29 gene dataset and found evidence of homology and contamination errors in 11 genes. By providing a thorough filter to remove possible homology errors I can be confident that my topology, which is identical to that of Springer and Gatesy (2017), is a more accurate representation of the species tree.

Grey seal as sister species to Caspian seal

My phylogeny resolved Grey seal within the Phocini clade and as sister to Caspian seal, the placement of Grey seal within the Phocini tribe has been suggested since early pinniped genetic phylogenetic analyses (Davis et al., 2004; Delisle and Strobeck, 2005; Fulton and Strobeck, 2006; Palo and Väinölä, 2006; Árnason et al., 2006; Higdon et al., 2007). However, it is only in the most recent coalescent super tree analyses (Nyakatura and Binida-Emonds, 2012) that Grey seal has been suggested to be sister to Caspian seal, rather than all *Pusa spp*., with strong support. I find further support for this relationship and congruence between concatenation and coalescent methods after the removal of a small proportion of genes with exceptionally high phylogenetic signal. This relationship between *Pusa* and Grey seal has been supported by several studies, and a taxonomy change should be considered by the Society of Marine Mammalogy Taxonomic committee. I propose that the Grey seal be subsumed into the *Pusa* genus as suggested previously (Arnason et al, 1995; Arnason et al., 2006; Nyakatura and Binida-Emonds, 2012).

The rest of my pinniped phylogeny is consistent with other recent studies (Árnason et al., 2006; Higdon et al., 2007; Nyakatura and Binida-Emonds, 2012). I have sought to use all the genomic data of Pinnipedia publicly available at the time of analysis, which passed my quality control criteria. With genomic data being produced at a rapid rate it is expected that new genomic data will be consistently produced before the analyses can be run in their entirety. A lack of publicly available Otariidae species made the thorough analysis of the clade difficult, with good quality genome level data available for only three from 15 species. Since finalising the dataset for this study, a well resolved Otariid phylogeny has been produced, using mapped genomic data, rather than SGOs (Lopes et al., 2020). Two additional genomes have also become publicly available since the conclusion of this study, an Otariidae genome, *Arctocephalus townsendi*, and Phocidae genome, Baikal seal, though annotations for these assemblies have not been released (Dudchenko et al., 2018; Yuan et al., 2021).

Visual inspection of the 8 genes that were removed from the final phylogenetic alignment found that some of these sequences obtain high phylogenetic signal due to annotation artifacts (Electronic appendix 3.4). Mixing genome annotation methods has been known to introduce errors in comparative analyses even across some of the most accurately curated assembly sets (Weisman et al., 2022). Performing phylogenetic signal tests across genes was essential to analyse annotation data accumulated from different public resources, generated using differing methods of annotation. In addition to annotation artifacts, incongruence between phylogenetic methods may have been produced because of introgression or lineage sorting between species. Both ILS and introgression have been noted to cause phylogenetic incongruence and conflicting gene trees in recent analyses of pinnipeds and other marine mammals (Árnason et al., 2018; Lammers et al., 2019; Lopes et al., 2020). With further time and resources, this analysis would be of interest to investigate for the Phocinae. The relationships within the Phocini clade appear to be complex, with uncertainties around the periods during which gene flow between species could occur.

Phylogeny divergence dating

My dating of divergence times is mostly consistent with previous estimates, with my results supporting a divergence between Ursidae, Mustelodiea and Pinnipedia occurring in a relatively small evolutionary timeframe, with all 3 lineages diverging within 2.5 Mya (Figure 3.17) (Nyakatura and Binida-Emonds, 2012). My results support a split between Phocidae and Otarioidea approximately 26 Mya, and between Otariidae and Odobenidae approximately 18 Mya, which is consistent with previous molecular analyses (Nyakatura and Binida-Emonds, 2012; Lopes et al., 2020; Berta et al., 2018; Yuan et al., 2021). My results suggest a slightly older divergence time within the Phocini, previous studies have suggested a divergence between *Phoca* and *Pusa* approximately 5-6 Mya (Árnason et al. 2006; Nyakatura and Binida-Emonds, 2012; Berta et al. 2018), whereas my results estimate this divergence to occur approximately 8 Mya. Within the *Pusa* clade I also estimate the split between Grey seal/Caspian seal to occur approximately 5 Mya, whereas previous studies which have resolved Grey seal - Caspian seal as sister taxa have estimated a much more recent divergence 1.6-3 Mya (Árnason et al. 2006; Nyakatura and Binida-Emonds, 2012). A recent study on the paleogeography of the Caspian Sea predicts a marine inclusion event from the Arctic water from 2.75-2.35 Mya (Hoyle et al., 2021; Lazarev et al., 2021). This is in the lower boundaries with my results, and in alignment to previous dating of previous phylogenetic analyses (Palo and Väinölä, 2006; Nyakatura and Bininda-Emonds, 2012).

My results support the findings of Palo and Vainola (2006), with regards to phocid diversity being a product of marine radiations from an arctic basin. Previous hypotheses have speculated that *Pusa spp*., or even all Phocini, originated in the Paratethys sea with either a single or multiple migrations into the North Atlantic, with Caspian seal being the single extant species that remained in Caspian basin of the Paratethys (Árnason, 1974; Árnason et al., 1994). I find that in agreement with previous analyses that Caspian seal is not basal to *Pusa spp*. or Phocini, thus radiation of Phocini is likely to have occurred in a North Atlantic basin in a post Miocene radiation. An arctic radiation is supported by the synapomorphic white lanugo coat that is shared between all species of Phocids, which is an adaptation to an ice habitat, rather than the 'tropical' environment of the Miocene Paratethys (Deméré et al., 2003; Koretsky, 2001). My results show that Caspian seal and Grey seal are more closely related to each other, than Caspian seal is to the Ringed seal*,* diverging 6.81 Mya (Table 3.6). I show support for the hypothesis of colonisation of the Caspian Sea, with a divergence of Caspian seal and Grey seal occurring between 2.29 – 7.73 Mya.

The dating of divergence times within was calibrated using previous phylogenetic analyses. Thus, any errors in those analysis would cause critical compounding factors in my analysis. It would be imperative for this analysis to be reperformed with the use of fossil calibrations. The rich source of fossil data of Carnivora data (Faurby et al., 2021) in combination would allow lower limits to be set in relation to the diversification date. Ancient DNA of past species of pinniped would improve the confidence of dating, ancient DNA has been extracted from species of ancient samples of harp seal, grey seal, and walrus (Bro-Jørgensen et al., 2021). These samples could be used to calibrate molecular clocks which set mutation rates.

3.5 Conclusion

In summary, I have further strengthened the evidence that the Grey seal resides in the *Pusa* clade of Phocidae, whilst also resolving pinnipeds of a sister lineage to Mustelidae. I found that evolutionary relationships, especially within the Phocidae, were sensitive to the effects of small numbers of genes that introduced bias into reconstruction analyses, mostly as a result methodological artefact. By applying stringent filters to reduce paralogy across samples, and using multiple methods of phylogenetic reconstruction, I was able to produce a highly supported phylogeny of pinnipeds within Carnivora.

Chapter 4: Analysis of selective pressure variation across pinniped lineages and the identification of lactation associated coding regions under positive selection.

4.1 Introduction

Pinnipeds have undergone a suite of adaptations during the secondary colonisation of the marine environment, see Chapter 1.1. It is likely that the major morphological and physiological variation across, and within families, is driven by alteration to both protein coding regions and regulatory elements. Phocidae, Otariidae and Odobenidae have evolved diverse lactation strategies, which are adaptations to niches characterised by temporal and spatial partitioning of marine foraging resources, versus obligate terrestrial parturition and lactation (Berta, 2018). Primarily, Phocidae follow a capital breeding strategy, whereas Otariidae and Odobenidae use an income breeding and nursing strategy respectively (Chapter 1). These lactation strategies have driven lactation associated traits across the groups within Pinnipedia (Figure 4.1). I will focus on a specific subset of life history trait variation in this chapter, using selective pressure analyses to identify genomic regions under positive selection, and attempt to understand how they may contribute to lactation related phenotypic variation. Combining the resolved pinniped phylogeny produced in Chapter 3, with a genome-wide selective pressure analysis across five lineages (pinnipeds, Otariidae, Phocidae, Hooded seal, and Caspian seal), I sought to determine if gene orthologs exhibit patterns from developed evolutionary models (Chapter 1.5) that are indicative of positive selection. Next, I use overrepresentation analyses to assess whether these gene orthologs are associated with biological processes with known involvement in lactation trait phenotypes in other mammals, before assessing the functionality of the areas under selective pressure variation. By assessing areas of positive selection in genes with a known lactation associated function in closely related clades, it is possible to make inferences about functional shifts in these proteins that in turn may drive phenotypic differences seen across the clade.

Predicting genotype-phenotype interactions for non-model organisms can be unreliable, as small numbers of amino acid alterations can dramatically alter function, and so sequence homology between species may not relate to functional homology (Philippe et al., 2003). Furthermore, many traits are polygenic, with small changes across numerous genes driving phenotypic responses (Foll et al., 2014). Genome-wide association studies (GWAS) can be used to make inferences to the molecular underpinnings of novel phenotypes. GWAS have been repeatedly performed in dogs, for instance single nucleotide polymorphisms (SNPs) that are attributed to >90% of body size variance across breeds have been characterised (Karlsson et al., 2013; Plassais et al., 2019). Candidate genes from these studies have been

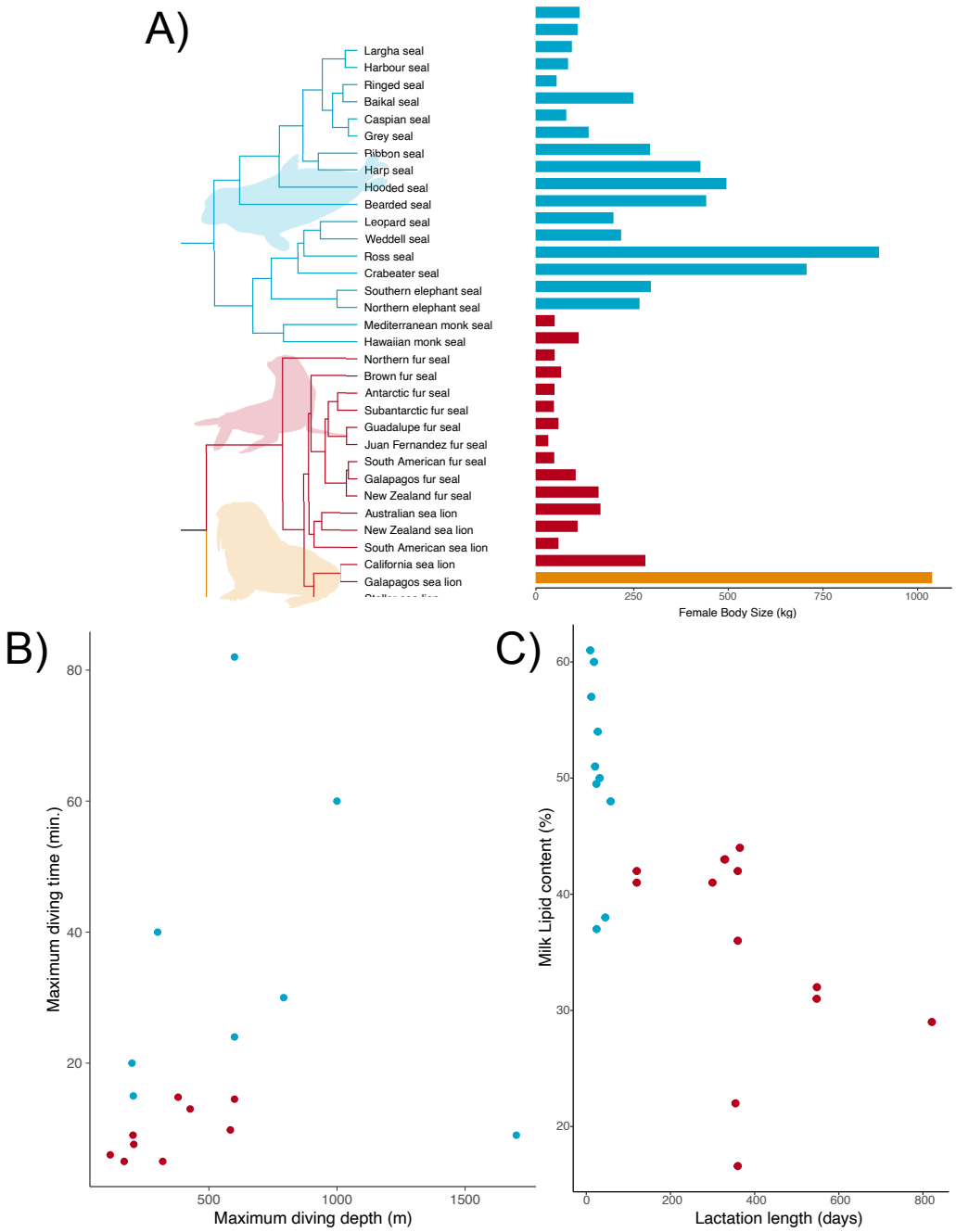


Figure 4.1. Lactation associated traits across the different groups in pinnipeds (Phocidae, Otariidae, and Odobenidae). A) Average maternal body size across phyogeny. This phylogeny is produced from the analysis in Chapter 3, with Otariidae evolutionary relationships from Lopes et al. (2019). B) a comparison of diving depth and times across pinniped species, C) a comparison of lactation length and fat content of milk.

compiled in publicly available databases (Machiela and Chanock, 2014; Tang et al., 2019). Candidate genes involved in lactation have received a significant amount of interest, due to their medical and economic benefits in the dairy industry. Domestic cattle (*Bos taurus*) have been bred to maximise different milk related traits including yield and fat content, the advent of genomics has facilitated the rapid progress of novel trait selective breeding and the understanding of the genetic factors responsible (Ogorevc et al., 2009; Peñagaricano, 2020). Cross comparisons from this comprehensive literature sources can be used to make inferences into genotype-phenotype associations in pinnipeds.

4.1.1 Physiology of lactation strategy variation in pinnipeds

Despite large variations in diet, all species of Phocidae produce lipid-rich milk (Costa, 1993). The demands of producing such a high energy milk whilst undergoing a prolonged fasting period create complex challenges for metabolic regulation around storage and mobilisation of large lipid reserves. The internal layer of blubber is enriched with short fatty acid (SFA) chains and long chain monounsaturated fatty acids (MUFA). These have been shown to be heavily metabolised during lactation to provision the young and meet “maternal overheads” - the energy spent on self-maintenance by a lactating female whilst nursing on land (Fedak and Anderson, 1982; Noren et al., 2003; Strandberg et al., 2008). The milk of Phocid seals lacks short-medium chain lipids, which are a marker of *de novo* mammary gland lipid synthesis (Neville et al., 1997), and instead contains high concentrations of MUFAs (Fowler et al., 2014; Iverson et al., 1995). This shows that the mobilisation of lipids from the blubber is integral for Phocidae, with the MUFAs being used to provision the offspring whilst SFAs and polyunsaturated fatty acids (PUFAs) being retained for maternal energy requirements. As the lactation period increases the proportion of long chain MUFAs in milk also increases, this is possibly due to facilitate the establishment of a thermoregulatory blubber layer in the pup (Reidman and Ortiz, 1979; Arriola et al., 2013; Wheatley et al., 2003).

The metabolism and mobilisation of lipids, from adipose tissue into milk lipids by the mother and milk fats into adipose tissue by the pups, are key pathways of interest in Phocidae due to their extreme nature in comparison to other mammals. Lipids exist in the plasma as non-esterified fatty acids (NEFAs) or are converted to triacylglycerol (TAG)/ lipoprotein complexes ready for tissue uptake (Figure 1.2). TAG/lipoprotein complex tissue uptake is thought to be facilitated by lipoprotein lipase (*LPL*), which remains at a constantly low level in some species of *Monachinae* but has shown to increase with correlation to increasing milk fat levels in *Phocini* species (Iverson et al., 1995; McDonald et al., 2006;

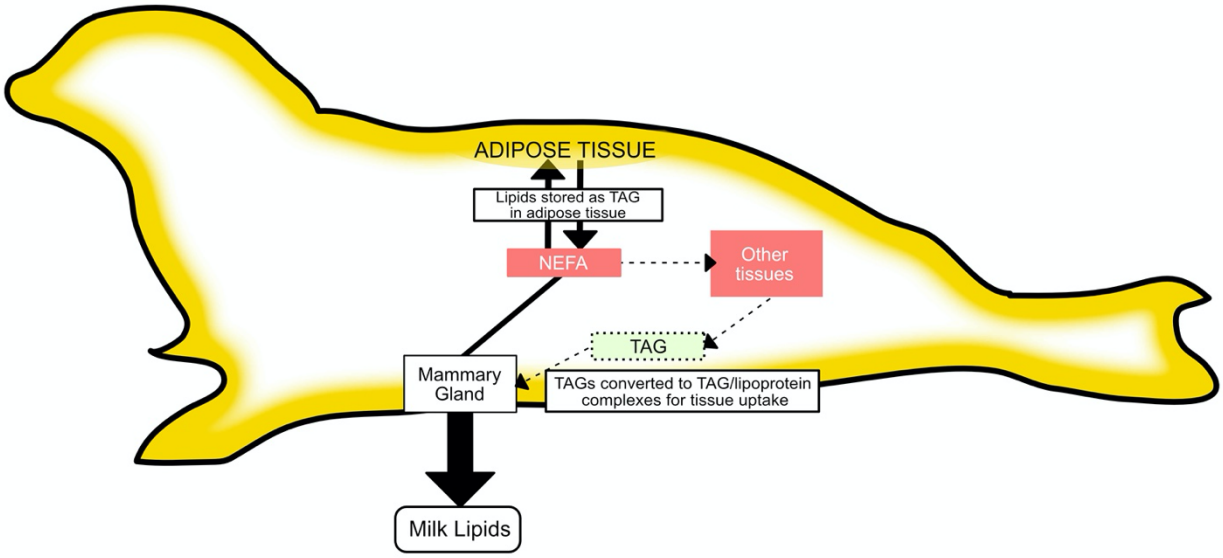


Figure 4.2. Processes of how fasting and lactation simultaneously occur in Northern elephant seals. It is thought that LPL is mainly responsible for hydrolysing circulating TAGs into FAs for tissue uptake. TAG = triglycerol, NEFA = non-esterised fatty acids. Adapted from Fowler et al., 2018.

Mellish et al., 1999). In most mammals, hormone sensitive lipase E (LIPE) has been shown to play a major role in lipid mobilisation. Although, LIPE has been shown to remain at low expression levels in blubber throughout lactation, with adipose triglyceride lipase (PNPLA2) appearing to play a much more prominent role (Crocker et al., 2014). High cortisol, and subsequently low insulin levels, have been seen to be strongly linked to high levels of plasma circulating NEFAs and milk lipid contents in Northern elephant seals. It is thought these levels are maintained to prevent the re-esterification of NEFAs and increase lipolysis (Fowler et al., 2014, Fowler et al., 2016). Despite participating in energy intensive activities whilst fasting, seals avoid catabolising protein and are able to conserve muscle tissue (McCue, 2012). Low insulin levels are negatively correlated with glucagon levels, with glucagon being shown to stimulate both gluconeogenesis and lipolysis in mammals (Perea et al., 1995). Glucagon levels remain low during fasting and lactation in seals, most likely due to the undesired protein catabolism that is initiated by glucagon. This suggests that regulation of lipid metabolism and mobilisation are somewhat dissimilar to that seen in other mammals, with a prioritisation of lipolysis and protein sparing.

Complex changes occur in the mammary gland during the lactation process, with cell proliferation and differentiation enabling milk secretion during initiation of mammary gland involution after weaning. Otariidae have the evolved a unique ability to go undergo long bouts of lactation disruption without inducing involution. This novel adaptation allows Otariidae mothers to temporarily abandon pups at shore for prolonged periods, whilst embarking on long distance foraging trips (Bonner et al., 1984). In mammals it is known that lactation involution is a 2-phase process that is thought to be initiated by the local mechanoreceptor response to an absence of suckling (Green and Streuli, 2004). In the initial phase, milk genes are downregulated which results in a decrease in milk volume, the second phase involves apoptosis of 50-80% of epithelial cells from the gland and a re-initiation of the gland is not possible after the second phase has begun (Sharp et al., 2007). Otariidae can pause involution at the second phase, which Is thought to be facilitated by a mutation in the lactose protein Alpha-lactalbumin (*LALBA*) gene (Reich and Arnould, 2007). Lactose is responsible for the water content in milk, as it promotes an influx of water across the cell membrane into the mammary alveoli via osmosis. Pinnipeds have very low water levels in their milk, 30-50% in pinnipeds compared to 80-90% in cows (Rasheed et al., 2016; Aspariza, 2019). This is reinforced by lactose only being found in trace amounts in Phocidae milk and is completely absent in Otariidae. Reich and Arnould (2007) and Sharp et al. (2008), found that the *LALBA* gene is truncated in Otariidae, which leads to an absence of *LALBA* protein in the milk. A lack of lactose would increase viscosity of the milk and would reduce the distensive mechanistic force during lactation. This could lead to involution when absent,

although the effects of autocrine factors that are responsible for regulation of lactation secretion in terrestrial mammals are unknown (Wilde et al., 1998).

Body size plays a direct role in pinniped lactation strategies, facilitating the fasting and foraging techniques unique to capital breeding Phocidae. Pinnipeds from all families have representatives which have extremely large body sizes, relative to other terrestrial carnivores. An increase in body size is not unique to pinnipeds and many marine mammals have experienced shifts towards increased body sizes driven by thermoregulatory factors (Gearty et al., 2018). Separate foraging and provisioning of young in pinnipeds means that while selective pressure towards a larger body size would be advantageous for thermoregulation in water (due to the high heat capacity of water), heat also needs to be off loaded whilst on land (Schultz and Bowen, 2005). The rate at which energy required for metabolism is lower than that of fat stores as mass increases (Kleiber, 1961). Thus, species which strive to reduce their metabolic overheads, for example when fasting, are more efficient if they have a larger body mass. The effects of this predicted relationship have been observed empirically across species of capital breeding Phocidae. Increases in body size result in lower metabolic overheads for the mother, allowing species to divert a higher proportion of energy to pup development through milk energy. For example, the Elephant seal (which is the largest Phocid species) has a lower maternal overhead of 40%, when compared to the intermediate sized Grey seal, which has an maternal overhead of 45%. Species such as the Hooded seal (27% maternal overheads) achieve reduced maternal overheads through the reduction of lactation length (Costa et al., 1986; Mellish et al., 1999a; Mellish et al., 1999b). Greater body size enables capital breeding species to engage in longer fasting and provisioning periods, where ecological factors such as predation and substrate allow. For instance, the Elephant seals occupy stable breeding grounds devoid of predation, whereas the Hooded seal occupies unstable pack ice and aims reduce visibility to avoid Polar bear predation (Schultz and Bowen, 2004).

Capital breeding Phocids are released from the constraint of provisioning the pup during foraging. Instead, making use of distant foraging grounds to build reserves, with an increase in body mass facilitating longer transit times and prolonged dive times. Foraging in a single long trip is more economical than shorter more frequent trips, and as transit time is reduced, this increases the proportion of time spent foraging (Adachi et al., 2014; Costa, 1991; Maresh et al., 2015). Diving depth and duration is physiologically constrained in air breathing marine mammals by the oxygen stored in tissues, compared to the metabolic rate of oxygen used (Ponganis, 2016). Like the relationship between fat stores and maternal overheads,

the rate in which stored oxygen increases with body mass is greater than of oxygen metabolised as body mass increases, thus, increased body mass allows for longer, deeper dives (Ponganis, 2016). In addition to increase oxygen stores through body mass, capital breeding Phocidae, and other pinnipeds, have undergone additional adaptive modifications to allow long, deep dives. Pinnipeds have traits that are also seen in other mammals which exhibit diving behaviours e.g., cetaceans. These traits include increased body size, bradycardia, spleen mass, blood volume, and haemoglobin and myoglobin concentrations (Lincoln et al., 1973; Påsche and Krog, 1980; Ponganis et al., 1992; Mirceta et al., 2013). Decreases in blubber haemoglobin concentrations have been seen during routines dives in Harbour seals. However, despite this action to reduce blood flow to the peripheral tissues, central oxygen stores are still depleted during dives (Allen and Vázquez-Medina, 2019). Constant depletion of oxygen stores will cause even the constantly perfused tissues, such as the brain, to experience hypoxia. . Pinniped brains have been shown counteract hypoxic effects by increasing neuroglobin levels during dives, and by an increased tolerance to lactate and increased glycogen stores (Czech-Damal, et al., 2014). It has also been shown, through transcriptomic analyses, that major stress response pathways are upregulated during dives to reduce hypoxic injury to the brain (Hoff et al., 2017).

With such variation within across the different lactation strategies, I will look to interrogate protein coding regions across the whole genome for signatures of selection. Thus, determining genes under positive selection, in an attempt uncover some of the genetic underpinnings in different phenotypic traits. I will use pinnipeds, Phocidae, Otariidae, Hooded seal and Caspian seal lineages as focal lineages in five selective pressure variation scans, revealing genes uniquely under selection in each. I would expect pinnipeds as a lineage to have acquired signatures of selection, as with other diving mammals, in pathways associated with diving physiology, morphology and body size, immunity, and lactation. In capital breeding Phocidae I might expect genes with involvement in fat mobilisation and metabolism to be under high selective pressures. The Hooded seal line has an extreme form of the capital breeding system ancestral to Phocidae. If this variation is controlled by protein coding elements, rather than regulatory elements I would expect to see genes that are related to increasing milk fat content under positive selection. Although Caspian seals still have large fat reserves and high milk fat in comparison to other mammals, they display a hybrid income-capital breeding strategy. Thus, I would expect to see fewer genes related to a capital breeding strategy, such as fat mobilisation and metabolism. In income breeding Otariidae I would expect to see genes involved in lactation involution, and mammary gland morphology genes.

4.2 Methods and Materials

4.2.1 Functional annotation of filtered gene families under positive selection in pinniped lineages

4.2.1.1 Assembly of species dataset

Multiple species comparative analysis requires comprehensive sampling from many lineages. Although, when using larger phylogeny sets, large evolutionary distances between species can lead to saturation for silent substitutions, computational limitations, and Type I errors (Anisimova et al., 2003). Taking these factors into account I gathered dataset of 23 species exclusively from the Carnivora clade, I aimed to alleviate these issues by not having an excess of representatives from clades but also minimising evolutionary distances between clades (Table 4.1). The species in these datasets were chosen on the basis of which genomes were available as of 01/12/2020, this was a reduced species set of the dataset used for the phylogenetic reconstruction in Chapter 3 (Table 3.2), retaining only Carnivora species. I had intended to use the whole dataset from Chapter 3, including non-Carnivora species, but the excessive run times of the selection variation analyses when using more than 20 species prevented this. This contained 13 pinniped genomes with CDS data in Ensembl (v102), Ensembl Rapid (accessed on 10/11/20) or RefSeq (accessed on 10/11/20), 2 *de novo* pinniped genomes (assembled in Chapter 2), in addition to 2 unpublished pinniped genomes for the Saimaa Ringed seal (pers. comms Jernvall et al., 2020) and Mediterranean monk seal (Gaughran, 2020) obtained through collaborators. In total, this species set contained 12 Phocidae species, 4 Otariidae species, and walrus, which covered most of the variation of lactation strategies across pinnipeds. Six outgroup species were chosen from Ensembl (v102) and Ensembl Rapid, with two species each representing Ursidae (Polar bear and Panda) and Mustelidae (Sable and sea otter), Canis (Dog) and Feliformia (Cat) (Figure 4.1).

Ensembl, Ensembl Rapid and RefSeq annotated nucleotide CDS were downloaded from their corresponding databases. Headers for each sequence in each species were edited to the format "SPECIES|GENEID" with the scripts *preclean_ensembl.py* (Electronic appendix, 3.1), *preclean_refseq.py* (Electronic appendix, 3.2) or *preclean_inhouse.py* (Electronic appendix, 3.3). Sequences from annotations produced in Chapter 2, had flags of "partial" or "low-quality" were removed. A quality control filter was applied, with the criterion that sequences must contain complete codons (sequence length is divisible by 3) to be retained for further analysis, using the *clean* function within the *Vespa.py* package (Webb et al., 2017). Filtered nucleotide sequences were then translated to amino acid sequences with the *translate*

Table 4.1. Species chosen for selective pressure analysis from their corresponding sources

Common Name	Family	Database	Genome version
Antarctic fur seal	Otariidae	NCBI Refseq	arcGaz3
Bearded seal	Phocidae	DNA Zoo	Erignathus_barbatus_HiC
California sea lion	Otariidae	Ensembl	mZalCal1.pri
Caspian seal	Phocidae	Orr et al., 2021	puscas4
Cat	Felidae	Ensembl	Felis_catus_9.0
Dog	Canidae	Ensembl	CanFam3.1
Giant panda	Ursidae	Ensembl Rapid Release	GCA_002007445.2
Grey seal	Phocidae	NCBI Refseq	Tufts_HGry_1.1
Harbour seal	Phocidae	Ensembl Rapid Release	GCA_004348235.1
Hawaiian monk seal	Phocidae	Ensembl Rapid Release	GCA_002201575.1
Hooded seal	Phocidae	Orr et al., 2021	cyscri1.1
Largha seal	Phocidae	DNA Zoo	Phoca_largha_HiC
Mediterranean monk seal	Phocidae	per comms	MMS_114
Northern elephant seal	Phocidae	DNA Zoo	Mirounga_angustirostris_HiC
Northern fur seal	Otariidae	Ensembl Rapid Release	GCA_003265705.1
Polar bear	Ursidae	Ensembl	UrsMar_1.0
Ringed seal	Phocidae	Saimaa Ringed seal Genome Project	Pushis
Sable	Mustelidae	Ensembl Rapid Release	GCA_012583365.1
Sea otter	Mustelidae	Ensembl Rapid Release	GCA_002288905.2
Southern elephant seal	Phocidae	Ensembl Rapid Release	GCA_011800145.1
Steller sea lion	Otariidae	Ensembl Rapid Release	GCA_004028035.1
Walrus	Odobenidae	NCBI Refseq	Oros_1.0
Weddell seal	Phocidae	NCBI Refseq	LepWed1.0

function of *Vespa.py* (Webb et al., 2017). Original nucleotide sequences were retained within a database, to allow corresponding nucleotide alignments to be generated further down the analysis pipeline.

4.2.1.2 Orthologous Group Generation and Filtering

I used DIAMOND to search sequence similarity across species (Buchfink, 2015) and orthologous group inference was achieved using *Orthofinder2* (Emms and Kelly, 2019), using default parameters. Orthologous clusters (OCs) with <6 species representatives were discarded due to the low power of likelihood ratio tests (LRTs) conducted with <6 taxa (Anisimova et al., 2001). Leaving 16,919 OCs in MSAs. These OCs were aligned using *MAFFT* (Katoh and Standley, 2013), *MUSCLE* (Edger, 2004), *T-COFFEE* (Notredame et al., 2000), and *Clustal-Omega* (Sievers et al., 2011) and *norMD* was used to assess resultant alignments and the highest scoring alignment chosen, as in Chapter 3.

To ensure the comparison of true orthologous genes, I filtered single gene orthologs (SGOs), with a Robinson-Foulds score < 0.5, as in chapter 3. Of the 16,919 MSAs from *Orthofinder*, 6,323 were SGOs and 10,596 multi-gene orthologous families (MGFs). To further refine the MGFs a tree-based orthology inference method was employed, using information from gene trees to partition MSAs into their single orthologous groupings. There are multiple methods that can be employed to perform tree-based orthology inference: largest subtree (LS) searches the tree, from root to tip, to find the largest subtree that contains single representatives of each taxa, when this subtree is found it is detached from the rest of tree and retained as a SGO; maximum inclusion (MI) uses the same principle as the LS methods but once the a subtree is detached the process is repeated on the remaining tree, and is iterated until no sequences remain; monophyletic outgroup (MO) method uses defined outgroups to root a tree before performing the LS method. *Phylopypruner* (https://gitlab.com/fethalen/phylopypruner/-/wikis/home) can perform all of these techniques on gene trees. *Orthofinder2* uses a duplication-loss-coalescence (DLC) approach to infer the most parsimonious reconciliation between a gene tree and species tree. Reconciliation trees were provided as an input to *Phylopypruner* using the MI tree-based orthology inference method, specifying a minimum of 7 taxa per orthologous group. If there were any species-specific duplications within the orthologous groups the sequence that has the shortest pairwise distance to the clade's sister node was retained. Using a MI method ensured I retained as many SGOs as possible, from the 10,596 MGFs I obtained 12,041 SGOs.

As hidden paralogy is a major concern when determining ortholog relationships, OGs were tested for hidden paralogy. I generated gene trees for the 6,323 SGOs identified directly from the *Orthofinder2* approach, and I compared the resultant gene trees against the species tree produced in Chapter 3, using a Robinson-Foulds (RF) test within Clan (Creevey and McInerney, 2005). I retained SGOs that had a cut off score < 0.5 meaning that the distance between the species tree and gene tree is not too large, whilst accounting for some deviation. After this step 3,533 SGOs remained from the *Orthofinder2* group of 6,323 SGOs.

From my analysis of the pinniped phylogeny (Chapter 3), the evolutionary history of genes with pinnipeds are possibly subject to evolutionary dynamics such as ILS or hybridisation. Thus, some deviation from the species tree is expected. Therefore, performing an assessment of putative SGOs using gene tree species tree distances may not be suitable. The SGOs generated from MGFs using Phylopypruner (Thalen, 2021) were tested for hidden paralogy using *Clan_check* (Siu-Ting et al., 2019). The approach uses incontrovertible "Clans" (Creevey and McInerney, 2005) to assess whether a gene family can recapitulate known and uncontroversial regions of the tree. In this case the clans were assigned at a family level. In total there were 6,171/12,041 SGOs that had no clan violations and were retained for further analysis. Gene trees for the 6,171 SGOs were pruned to retain single representatives at the family level, these family level genes trees were compared with the species phylogeny acquired in Chapter 3, a RF test was conducted within Clan (Creevey and McInerney, 2005). RF scores of 0, which indicate identical phylogenies, were passed resulting in 4,088/6,171 SGOs being retained. Compiling the 3,533 filtered SGOs from *Orthofinder2* and 4,088 filtered SGOs from pruned MGFs I generated 7,621 SGOs in total for use in the selective pressure analysis stage. If a gene tree was under selection it could result in a different topology to that of the species tree. Although, SGOs generated through pruning were susceptible to the inclusion of paralogous gene inclusion and so the decision was made to enforce stringent filtering, even though this could possibly result in some truly orthologous genes being removed.

4.2.1.3 Selective Pressure Analysis

The aligned 7,621 SGOs were converted back into nucleotide format using the "map_alignments" function in *Vespa.py* (Webb et al., 2017), using the original retained nucleotide sequences. The genes were mapped on to the resolved Carnivora phylogeny produced in Chapter 3, using the "*infer-gene-tree*" function from the Vespasian pipeline (Constantinides, *in prep.*). The selective pressure analyses were conducted using *CodeML* within the *PAML*4.9 package (Yang, 2007). This requires a specific directory structure, and the

directory structure was generated using the "*codeml-setup*" function within Vespasian (Constantinides, *in prep.*), providing the alignments at the nucleotide level for all SGOs and the inferred gene trees. Analysis of selective pressure variation was to be investigated on a lineage-specific level for multiple lineages and so *"branches.yaml*" was also provided under "-b" flag, specifying the Pinniped, Phocid, Otariid, Caspian seal, and Hooded seal lineages (Figure 4.3).

To assess selective pressure variation in an alignment the frequency of non-synonymous substitutions per non-synonymous site (dN) is compared against the frequency of synonymous substitutions per synonymous site (dS), calculating the dN/dS ratio (ω). $\omega < 1$ putatively indicates purifying selection, $\omega = 1$ indicates an absence of selection (or neutral evolution), and a $\omega > 1$ indicates positive selection (Zhang et al., 2005; Yang, 2007). Variation in selective pressure is estimated using site-specific and lineage-specificcodon models of evolution, as described in section 1.4.4. *CodeML* analyses were run on the Leeds High Performance Computing Cluster (ARC3) using the command "*snakemake -k --jobs 100 --cluster "qsub -cwd -V" --max-status-checks-per-second 0.1*". The "*report"* function within *Vespasian* (Constandinides, *in prep.*) was used to extract all tables of model parameters and LRT results for each SGO. To identify the model of best fit, simpler models are compared to more complex extensions of themselves to assess for significant fit of the model to the data. The models I was most interested in comparing were the null model M1neutral and its lineage specific extension Model A. Model M1neutral (df = 2) is a site model that allows for ω to fit 2 site classes, $\omega_0 < 1$ or $\omega_1 = 1$, modelling a neutrally evolving set of sequences, whereas Model A is a lineage-site model. Model A (df = 4) separates the phylogeny into background and foreground (lineage of interest): in background branches ω is fit in to 2 classes, $\omega_0 < 1$ and $\omega_1 = 1$, and in the foreground branches ω is fit in to one class, $\omega_2 > 1$ (Table 4.2). LRTs determine the whether the more complex/extension of the null model is a more significant fit to the data than the null. It does this by calculating 2(lnl) between models that are extensions of one another, with the null model being rejected if $p < 0.05$ in a X^2 distribution with degrees of freedom equal to the number of additional parameters. If Model A is determined to be a better fit that M1neutral, then Model A is then compared to Model A null. Model A null (df = 3) has the same background parameters as Model A but in the foreground branch $\omega_2 = 1$, i.e., it is a model that allows for lineage specific variation but only in purifying selective pressure and proportion of sites evolving under neutrality.

In SGOs where Model A is determined to be significant, the Bayes Empirical Bayes (BEB) method (Zhang et al., 2005) is used to calculate the probability of each codon being under positive selection ($\omega > 1$). Only

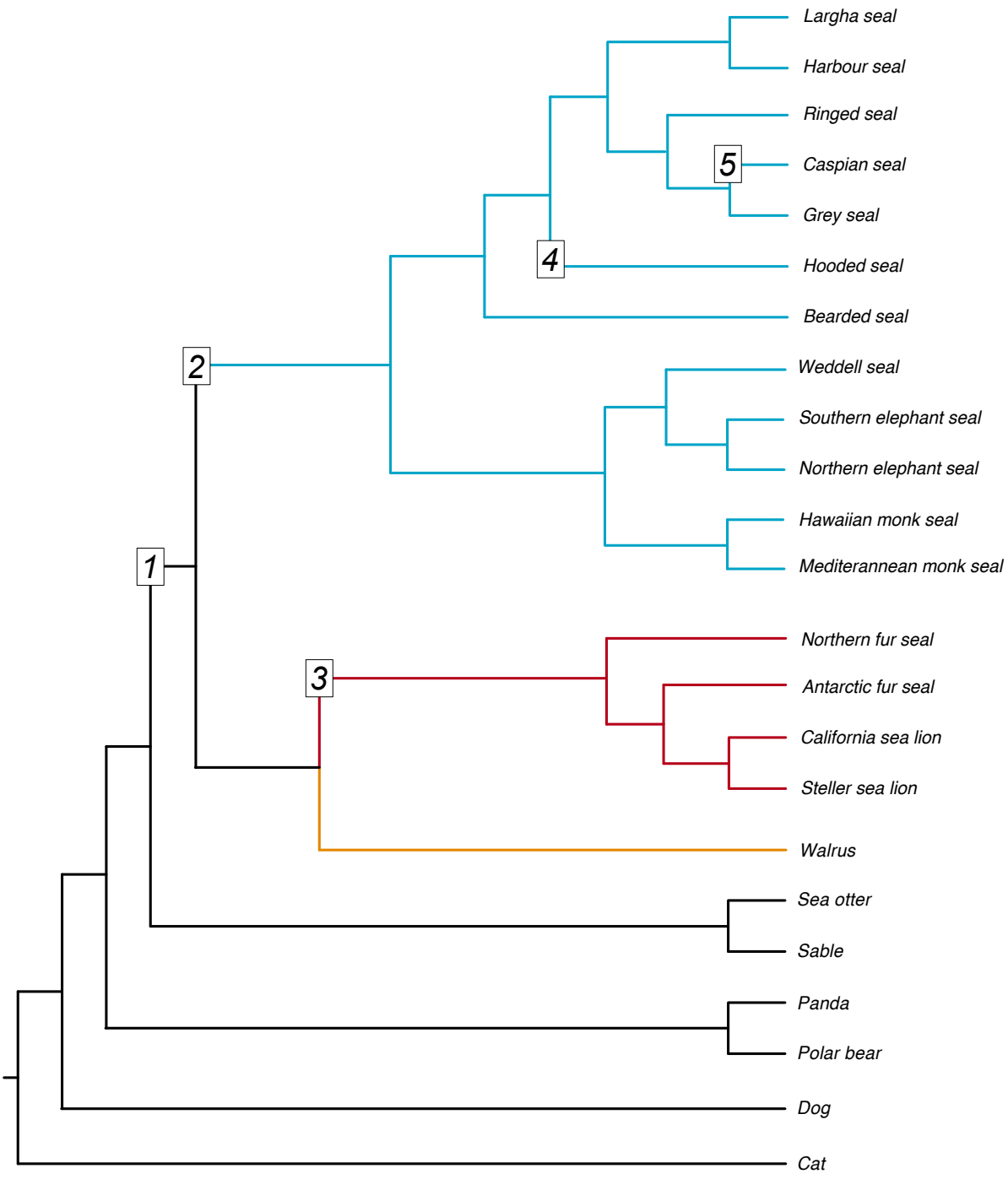


Figure 4.3. Evolutionary relationships of the different species used in the selective pressure variation analysis. The topology of the tree is based on the results on the phylogeny produced in Chapter 3. Colours represent the different families within Pinnipedia (blue = Phocidae, red = Otariidae, and yellow = Odobenidae). Labels 1-5 represent the different lineages investigated: 1 = pinnipeds, 2 = Phocidae, 3 = Otariidae, 4 = Hooded seal, and 5 = Caspian seal.

Table 4.2 Lineage specific models used within the CodeML analysis. Adapted from Yang, 2007.

Model	Parameters	Foreground	Background
M1Neutral	$p_0 : \omega_0 < 1$	n/a	n/a
modelA	$p_0 : \omega_0, p_1 : \omega_1, p_2 : \omega_0, \omega_2, p_3 : \omega_1, \omega_2$	$\omega_2 > 1$	$0 < \omega_0 < 1, \omega_1 = 1$
modelAnull	$p_0 : \omega_0, p_1 : \omega_1, p_2 : \omega_0, \omega_1$	$\omega_1 = 1$	$0 < \omega_0 < 1, \omega_1 = 1$

those sites with >=0.50 Posterior probability (PP) of being positively selected are reported by CodeML, and I only consider those with >0.80PP in this chapter. *Vespasian_postreaderV5.py (*Electronic appendix 4.4) was used to extract information from the Vespasian *report* results to identify the sites putatively under positive selection. Positive sites were called where BEB > 0.8 and codons are identical across all the species in the foreground lineage (i.e., I focused on pinnipeds, Phocidae, Otariidae, Caspian seal lineage, and Hooded seal lineage). Some SGOs did not have sufficient outgroup representatives to their foreground lineages so were not analysed by CodeML (Yang, 2007) (Table 4.3).

4.2.1.4 Analysis of putative positively selected genes

Recent studies have claimed the requirement of performing false discovery rate tests on selective pressure analyses is inappropriate for branch-site selective pressure analyses, and instead a 5% threshold on unhalved P-values is sufficient (Potter et al., 2021; Zou et al., 2021). Potter et al. (2021) performed simulations examining the effect of model specification on LRT results, empirically showing the need for false discovery rate tests on branch-site tests being fundamentally flawed due to model misspecification. Most proteins will be evolutionary conserved (Eyre-Walker 2006; Kimura 1983), hence there will be a lack of neutrally evolving sites ($\omega = 1$), which is used as the null model in model A null (see section 1.4.4).

I found similar properties regarding the distributions of my p-values, with an excess of p-values close to, or above, 1 (55-68%) across all lineages (Figure 4.4). P-value distribution analyses suggests my empirical p-value distribution is highly non-uniform. Displaying a U-shape pattern with high frequency of p-values close to 0, and an increase in frequency as P becomes close to 1, even when p-values close to 1 (>0.98) are removed (Figure 4.5). Suggesting that my data follows similar patterns to the simulated data in Potter et al. (2021), and as a result applying FDR tests would remove true positives from my analyses. Thus, I opted to use a filter removing any orthologous groups with unhalved p-value cut off > 0.05 produced through the LRT.

Annotation heterogeneity, an artifact introduced using differing annotation methods, can have a significant impact on the outcome of comparative analyses (Eddy et al., 2022). In this analysis, I found erroneous regions at the ends of coding sequences and large indels present in a single species coding sequence, seen in species from all annotation methods. Alignment and annotation errors can cause many sites within small regions of a sequence to marked as under positive selection (Tsagkogeorga et al., 2015; Davies et al., 2018). Tools have been previously developed to circumvent the errors caused by annotation

Table 4.3 Number of SGOs analysed in selective pressure variation analysis for each lineage. Selection pressure varation analysis performed in CodeML (Yang, 2003).

Lineage	No. Species	SGOs with ≥ 7 species (inc. outgroup)
pinnipeds	*17*	6,892
Phocidae	*12*	7,426
Otariidae	*4*	6,840
Caspian seal	*1*	5,022
Hooded seal	*1*	4,257

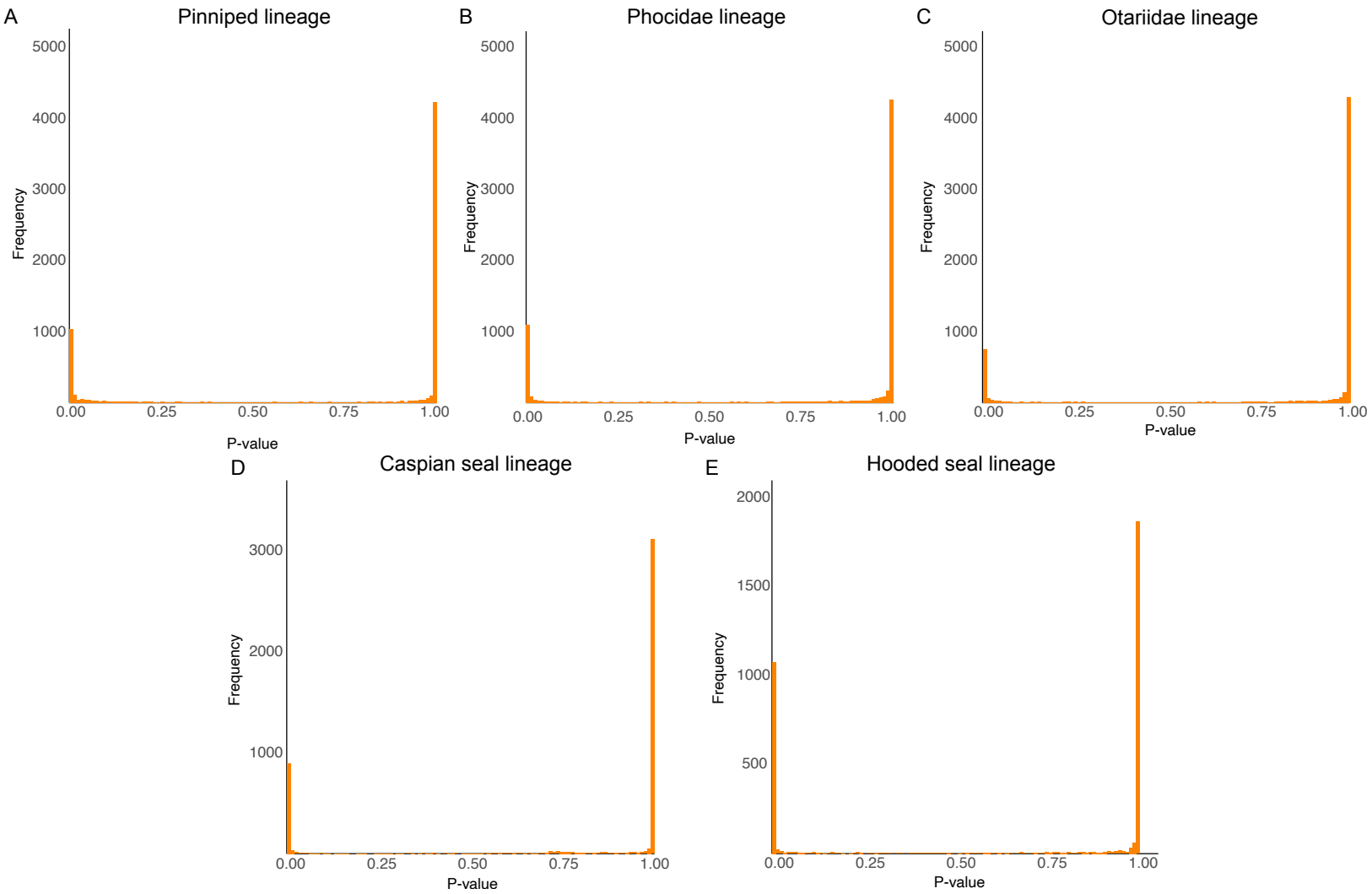


Figure 4.4. Empirical distribution of P-values from branch-site tests by lineage. Distribution of p-values from the from LRT tests from each lineage tested. Plots A-C show are of the same Y-axis range whereas D and E have differing ranges.

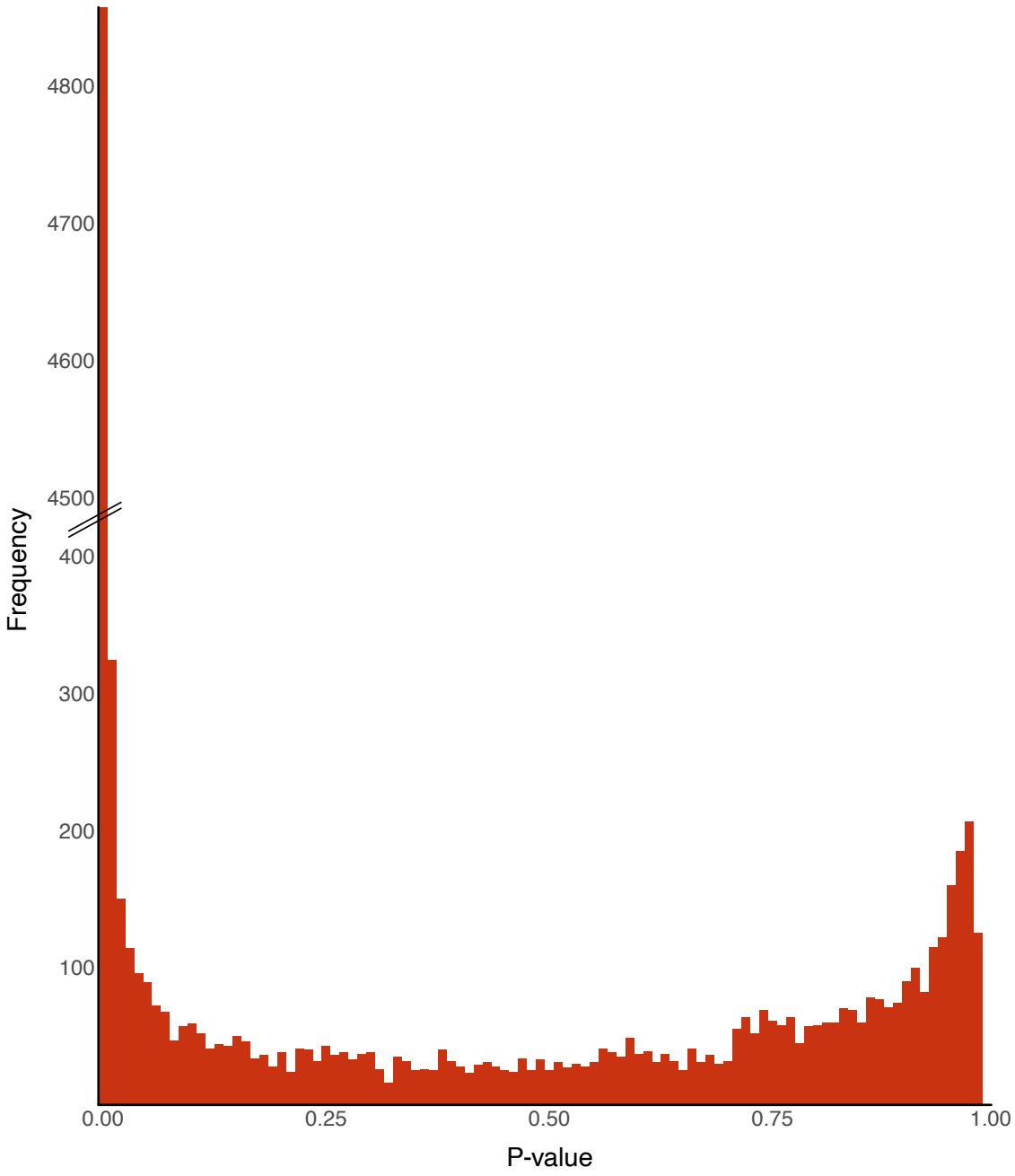


Figure 4.5. Overall empirical distribution of P-values from branch-site tests. 10,337 p-values <0.98, with a non-uniform distribution and U-shape. Y-axis is broken between 400 and 4500.

artifacts or by misalignments in the alignment process (Penn et al., 2010; Talavera and Castresana, 2007), although the performance of these tools can result in an increase of Type II errors (Goldman and Jordan, 2012). To reduce the presence of Type I errors, positively selected genes (PSGs) underwent post-hoc filtering.

Possible erroneous sites marked as putatively under positive selection were identified using a custom script *find_falsepos.py* (Electronic appendix 4.5), this identified PSGs with > 5 positively selected sites, or positively selected sites that were located at the first or last 10% of the alignment. Any sites called by *find_falsepos.py* were then visually inspected to ensure they did fall in areas of poor-quality with in alignments, such as unconserved regions or large insertions.

SGOs that passed this filter were then analysed to ensure they were unique to their lineage. Firstly, positively selected genes (PSGs) from the pinniped lineage were analysed to ensure they contained a mix of species from at least two of the pinniped families (Phocidae, Otariidae, and Odobenidae), using a custom script *lineage_specific.py* (Electronic appendix 4.6). To ensure I focused on cases of positive selection that were unique to a lineage, PSGs were cross referenced across lineages and retained on a hierarchical basis. For example, if a gene was under positive selection in both Phocidae/Otariidae and pinnipeds, albeit at different sites, it would only be analysed in the highest clade - pinnipeds).

Ensembl ID was used to extract gene information, thus only SGOs with representatives with Ensembl species (Cat, Dog, Giant Panda, and Polar Bear) were analysed downstream. The R library "biomaRt 3.13" (Durink et al., 2009) in the script "*ensembl_conversion.R"* (Electronic appendix 4.7) was used to extract gene and gene ontology information ("external_gene_name", "go_id") for each PSG.

4.2.1.5 GO Term enrichment

To determine whether functions as summarised by GO terms were enriched in particular lineages, Gene ontology (GO) IDs and GO slim terms (Ashburner et al., 2000; Gene Ontology Consortium, 2021) were assessed. GO slim terms are a subset of GO IDs that provide a broad overview of the fine grain GO IDs. For each lineage, a Fisher's exact test was performed, using an edited R script "*positiveSel.Rmd"* (Herrera-Álvarez et al., 2020), iteratively on counts of each Biological Process from the positively selected genes against the full set of SGOs for each lineage (Table 4.3). GO terms were defined as enriched if they had a significant p-value after a traditional Bonferroni correction for multiple testing. GO terms for lineages of

interest were investigated using similarity search in a REVIGO (Supeket al., 2011). GO terms were supplied to REVIGO, which removes redundant GO IDs and groups terms using their relatedness, producing a Treemap of semantic uniqueness of biological process GO terms in PSGs. Similarity is performed using SimRel algorithm (Schlicker et al., 2006) and GO terms from the entire Uniprot database (The UniProt Consortium, 2021) were used, with redundancy set on the 'small' setting to reduce the outputted list size. For the analysis of single species as foreground (Caspian seal and Hooded seal), PSGs were analysed for rapidly evolving genes (REGs) - PSGs that have the highest raw p-distance difference in comparison to their sister species since divergence from the MRCA. This is to identify the most significantly evolving genes from the large amount of positively selected genes identified. MSAs were run in *HmmCleaner.pl* (Di Franco et al., 2019) which use hidden Markov models to identify areas of the alignments with poor fit and remove them from the alignment.

Protein trios were identified and extracted for the PSG MSAs for Hooded seal, and Caspian seal. This included Southern elephant seal, Hooded seal, Harbour seal for the Hooded seal evalutation and were Harbour seal, Caspian seal, Grey seal for the Caspian seal evaluation. The Southern elephant seal and Harbour seal serve as outgroups to calculate the difference in distance for the Hooded seal and Caspian seal respectively against their MRCA. The difference in protein distance (PD_{diff}) was calculated from the p-distance between the species of interest to the outgroup (PD_{ac}) and sister species and outgroup (PD_{bc}), with a PD_{diff} of zero indicating the lineages had evolved at equal rates since divergence from their MRCA. PD index (PD_i) was calculated by dividing the PD_{diff} by PD_{ac} to account for differences in protein evolution between PSGs (Equation 4.1). The top 5% of positive PD_i values were classified as REGs.

Equation 4.1 Calculation of PDi from the raw protein distances between species of interest (a) and outgroup species (c) in comparison to that between sister species (b) and outgroup.

$$PDI = \frac{PD_{ac} - PD_{bc}}{PD_{ac}}$$

4.2.1.6 Gene function analysis

I attempted to derive possible gene functions of positively selected elements, through their role in pathways, and their role in other mammalian species. Thus, I assessed my dataset through two methods: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis and database search of phenotype-genotype associations. Firstly, KEGG pathway analysis was run to determine whether genes were significantly enriched for a particular pathway. Genes were converted from HUGO Gene Nomenclature

Committee (HGNC) gene names to a synonymous Entrez gene ID for human genes, before performing a KEGG enrichment in *ClusterProfiler* 4.0 (Wu et al., 2021). The total set of SGOs used in the *CodeML* analysis for each lineage (e.g., 6,892 SGOs for pinnipeds) were used as a background set of genes, to assess enrichment.

I tested the hypotheses that genes related to certain phenotypic traits will be under positive selection in species that exhibit differentiated lactation strategies. I used pre-existing mammalian gene sets from managed database sources, if no publicly available database was available, and I identified published GWAS or expression analyses with the most evolutionary related species. Extracted gene lists are presented in supplementary dataset 4.10. I predicted that capital breeding Phocidae will have genes under positive selection that are involved in the mobilisation of high quantities of lipids from blubber to milk during lactation and genes associated with sequestration of lipids into blubber. To investigate the underlying molecular changes of these processes I gathered a list of candidate genes associated with lipid metabolism and mobilisation from the mouse genome informatics (MGI) database, which collates functional associations from a large cohort of laboratory mice studies (Bult et al., 2019).

Capital breeding Phocidae generally have larger body mass than income breeding Otariidae (Figure 4.1a), this larger body mass equates to larger blubber reserves, and thus larger reserves allow for an extended period of self-provisioning and offspring provisioning during lactation. Larger body mass also permits the increased storage of oxygen during diving which facilitates dives that have durations of up to 1.5 hours and depths of more than 1700 metres (covered in more detail in section 4.1). To investigate the genetic basis of the increase of body size with capital feeding Phocidae, I collated information of genes with relationships to body size difference in canine species or *Canidae lupus familiaris* breeds (Table 4.4). In addition, I identified genes associated with cardiovascular control and cell-level hypoxia, which are key phenotypes exhibited in marine mammals that undertake deep dives with extended submergence times. These were extracted from a manually curated, experimentally related catalogue (Khurana et al., 2013). I also extracted genes with expression-regulated relationships to fasting in humans (Couto Alves et al., 2018).

One of the most prominent differences between lactation strategies of capital breeding phocids and income breeding Otariids is the increase of milk fat content of the Phocidae, which enables rapid and efficient energy transfer between the mother and offspring. GWAS analyses into milk contents are of an

economical value to the dairy industry and so I exploited these analyses to generate a candidate gene set of milk lipid level associated genes (Table 4.4). The genetic basis of the physiology of lactational apparatus of the pinnipeds was also of great interest. With the ability of Otariidae to undergo lactational pauses facilitating the pronounced foraging trips without undergoing lactation involution. In addition, the production and delivery of lipid-rich and water-sparse milk has been shown to have large effects on the physiology and composition of mammary glands in pinnipeds (Tedman, 1983). Again, I chose to use bovine derived candidate gene sets, which may also be involved in the lactation physiology of pinnipeds. PSGs for the different lineages were searched against the candidate gene set lists for each functional category (Table 4.4) to find genes of relevance from gene trait association studies and other selective pressure variation analysis in mammals that could be used to provide further evidence of function. In all cases I are aware that assigning function across species is challenging, and therefore the trends I identify would require functional analyses to be confirmed, through either cellular assays or transcriptomic analyses (covered in more deail in chapter 5). To further understand the role of the PSGs, any gene with a representative from a functional category was subjected to a literature search to find more detailed information on its putative function within closely related mammalian species.

Table 4.4 Candidate gene lists based on functional properties

Functional category	No. genes	Source	Reference
Lipid metabolism / mobilisation	*1,369*	GO terms: 'lipid metabolism' and 'lipid transport' from MGI (Mouse Genome Informatics) database	*Bult et al. 2019*
Diving/hypoxia	*2,289*	Hypoxia related genes from HypoxiaDB	*Khurana et al. 2013*
Body size/ obesity	*589*	GWAS of candidate genes associated with obesity-related traits in Canines	*Sheet et al. 2020*
Lactation physiology: mammary function	*2,423*	"Lactation-mammary" gene set produced using bovine EST libraries	*Lamay et al., 2009*
Lactation physiology: mammary gland involution	*684*	"Involution-mammary" gene set produced using bovine EST libraries	*Lamay et al., 2009*
Fasting	*610*	Candidate gene set derived from expression analysis of genes in skin and fat tissues during fasting	*Couto Alves et al. 2018*
Milk properties	*316*	GWAS and candidate genes from 4 studies using Bovine species	*Bionaz and Loor, 2008; Dadousis et al., 2017; Littlejohn et al., 2014; Nayeri and Stothard, 2016*

4.3 Results

4.3.1 Genes that have undergone positive selection across different lineages

A total of 3,266 uniquely PSGs were identified among the 5 focal lineages, pinnipeds, Phocidae, Otariidae, Hooded seal, and Caspian seal. PSGs were filtered for fixed residues at the selected site, LRT p-value < 0.05, visual inspection of alignment quality in selected region, and genes uniquely under positive selection in their lineage, leaving 1,562 PSGs across the five lineages (Table 4.5).

Pinnipedia lineage

From the 338 SGOs putatively under positive selection in the pinniped lineages, 262 PSGs had a representative in the Ensembl database. No biological process GO IDs or GO slim terms were overrepresented after performing a functional enrichment analysis with corrections for multiple testing. The resulting treemap returned 43 categories (Figure 4.6), the largest in size being involved in lipid level regulation ("regulation of plasma lipoprotein particle levels"). Other terms with possible association to phenotypic adaptations seen in pinnipeds, such as cellular communication and adhesion ("collagen catabolic process", "cell communication by chemical coupling", and "cell adhesion"), and lipid metabolism ("lipid metabolic process") are overrepresented.

Phocidae lineage

290 SGOs putatively positively selected in Phocidae had representation in the Ensembl database, from this no GO IDs or GO slim terms were significantly enriched after corrections for multiple testing. The treemap of GO IDs clustered in REVIGO (Figure 4.7), found 39 categories revealing semantic similarities in GO IDs possibly involved in lipid metabolism and storage ("glucose-6-phosphate transport", "lipid metabolism process"), as well as adaptations to marine life and deep diving ("collagen catabolic process", "UV protection", "nitrogen compound metabolic process", and 'fluid transport").

Table 4.5 The number of PSGs passing each stage of the analysis filters.

Lineage	Vespasian out	Fixed residue	P-value < 0.05	Alignment checked	Unique to lineage
Pinnipedia	415	398	350	338	338
Phocidae	561	543	492	497	312
Otariidae	465	455	413	428	260
Caspian seal	610	610	608	337	188
Hooded seal	1215	1215	1210	701	464

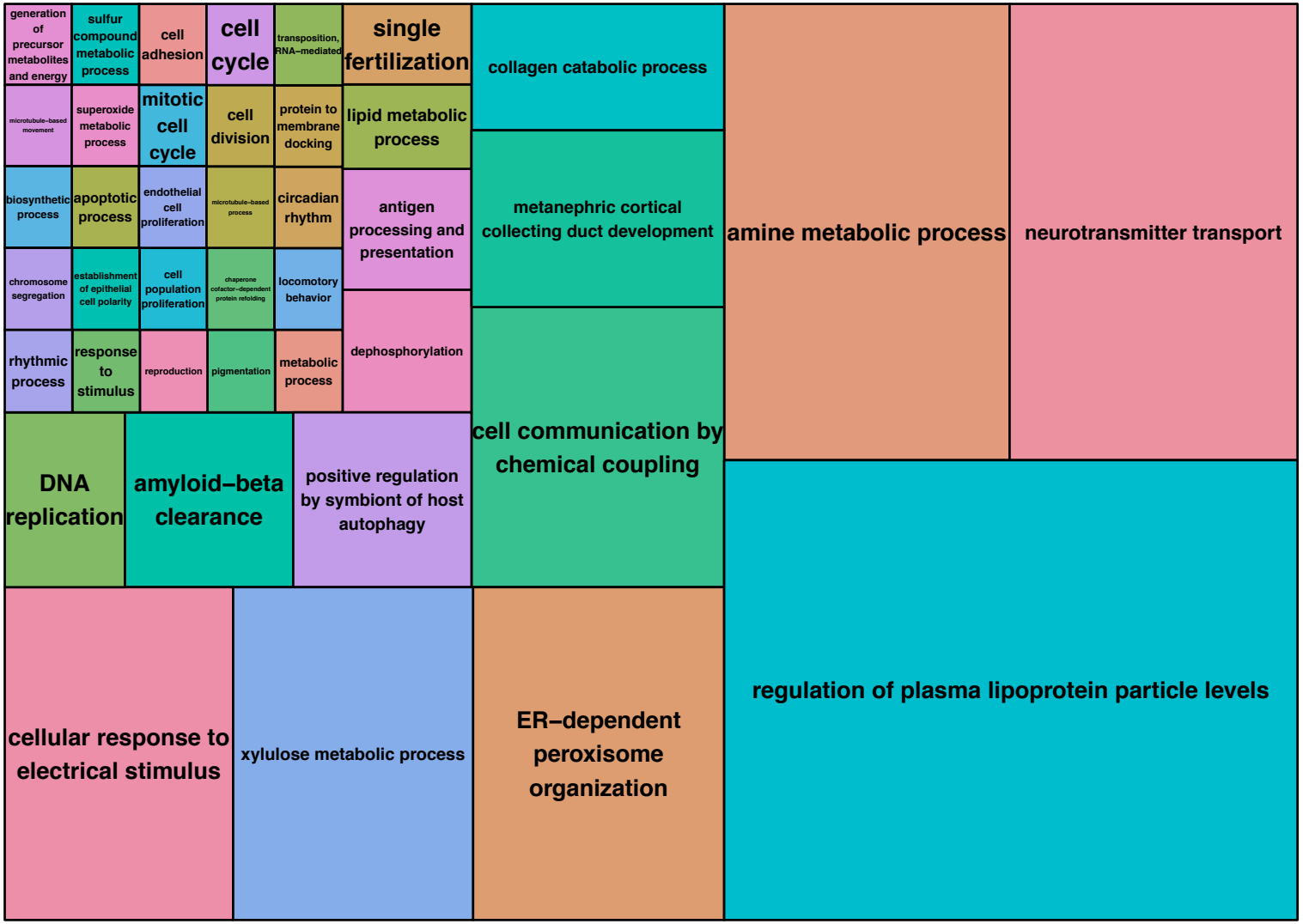


Figure 4.6. TreeMap of GO biological process terms for pinniped lineage. GO terms present in unique gene sets evolving under positive selection in the pinniped lineage. Rectangle size reflects semantic uniqueness of GO term, which measures the degree to which the term is an outlier when compared semantically with the list of terms present in the UniProt database (Bateman et al., 2021).

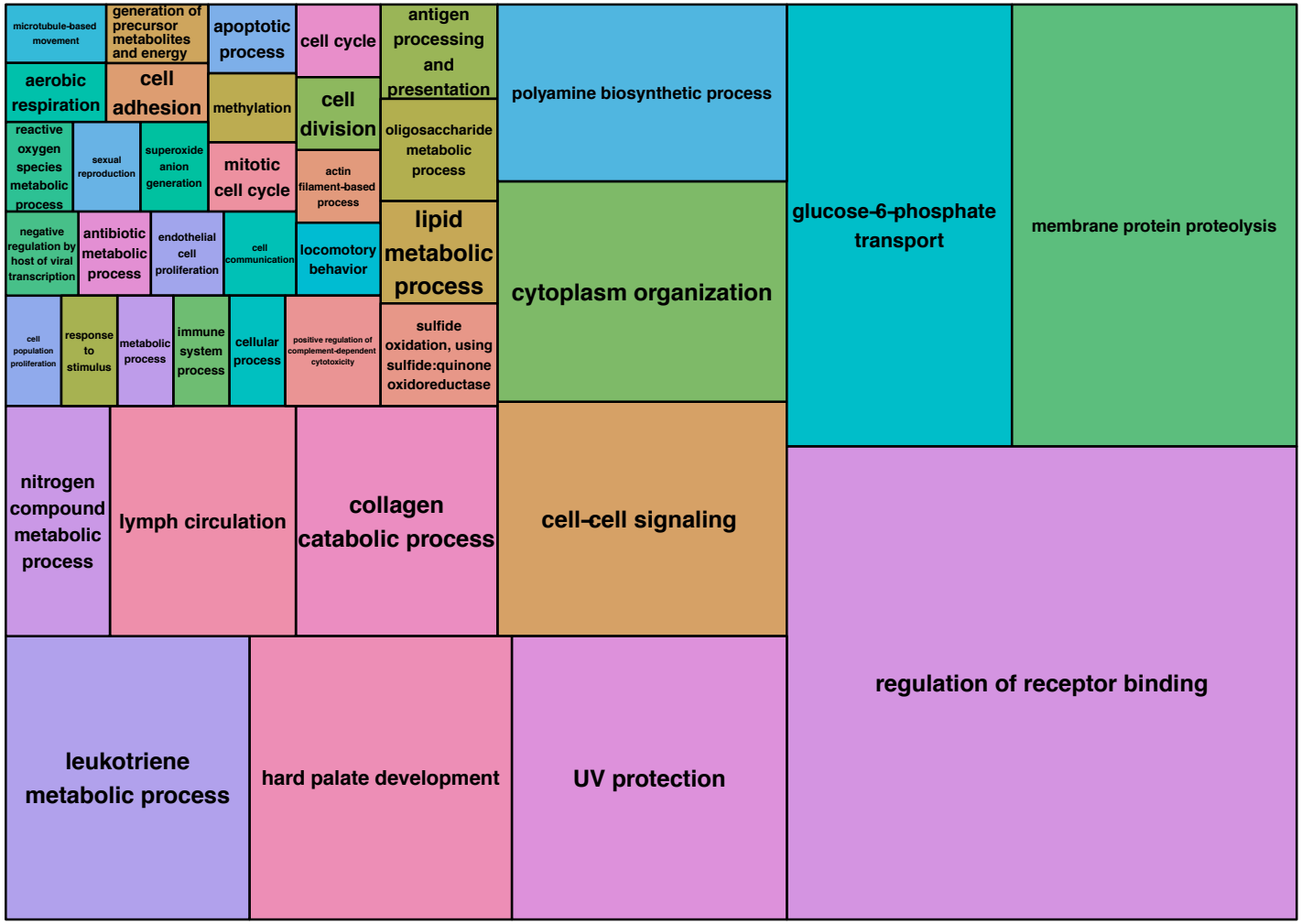


Figure 4.7. TreeMap of GO biological process terms for Phocidae lineage. GO terms present in unique gene sets evolving under positive selection in the Phocidae lineage. Rectangle size reflects semantic uniqueness of GO term, which measures the degree to which the term is an outlier when compared semantically with the list of terms present in the UniProt database (Bateman et al., 2021).

Otariidae lineage

From analysis of the 260 SGOs putatively under positive selection in the Otariidae lineage with a representative in the Ensembl database, no GO IDs or GO slim terms were significantly enriched. The resulting treemap (Figure 4.8) revealed 36 related GO term categories related to functions in immune response ("regulation of viral entry into host cell" and "immune system process"), lipid metabolism ("glycogen metabolic process" and "lipid metabolism process"), in addition to functions possibly related to a return to marine lifestyle ("response to salt stress" and "mastication").

Caspian seal lineage

There are no GO IDs or GO terms significantly enriched from the 188 SGOs classified as putatively positively selected from the selective pressure analysis. The REVIGO analysis revealed patterns of semantic similarity from the GO terms analysed, producing a treemap with 33 categories (Figure 4.9). Some of these terms have possible involvement to adaptations to anatomical morphology ("roof of mouth development", "multi-organism membrane organisation") as well as associations with lipid metabolism ("negative regulation of lipoprotein lipase activity" and "lipid phosphorylation"). I also found from this analysis, no evidence of selection for functions related to osmotic or thermal regulation, which might be expected as responses to unique features of the Caspian Sea environment.

Hooded seal lineage

There are no GO IDs or GO terms significantly enriched from the 464 SGOs classified as putatively positively selected from the selective pressure analysis. A treemap with 52 categories GO ID categories was returned from the REVIGO analysis (Figure 4.10). I found associations with "endoplasmic reticulum Golgi intermediate compartment organisation" with the Golgi apparatus responsible for the synthesis of lactose, and the "sensory perception of umami taste", which has previously been shown to have been reduced or lost in all lineages of Pinnipedia and Cetaceans (Sato and Wolsan, 2012; Zhu et al., 2014).

I set out to identify whether I could identify any broad signatures of functions from genes under positive selective pressure, with a special reference to lactation. I found all focal lineages had PSG GO terms with associations to lipid metabolism, but also found evidence for selective pressure on genes affecting additional adaptations of pinnipeds relevant to recolonisation of the marine environment.

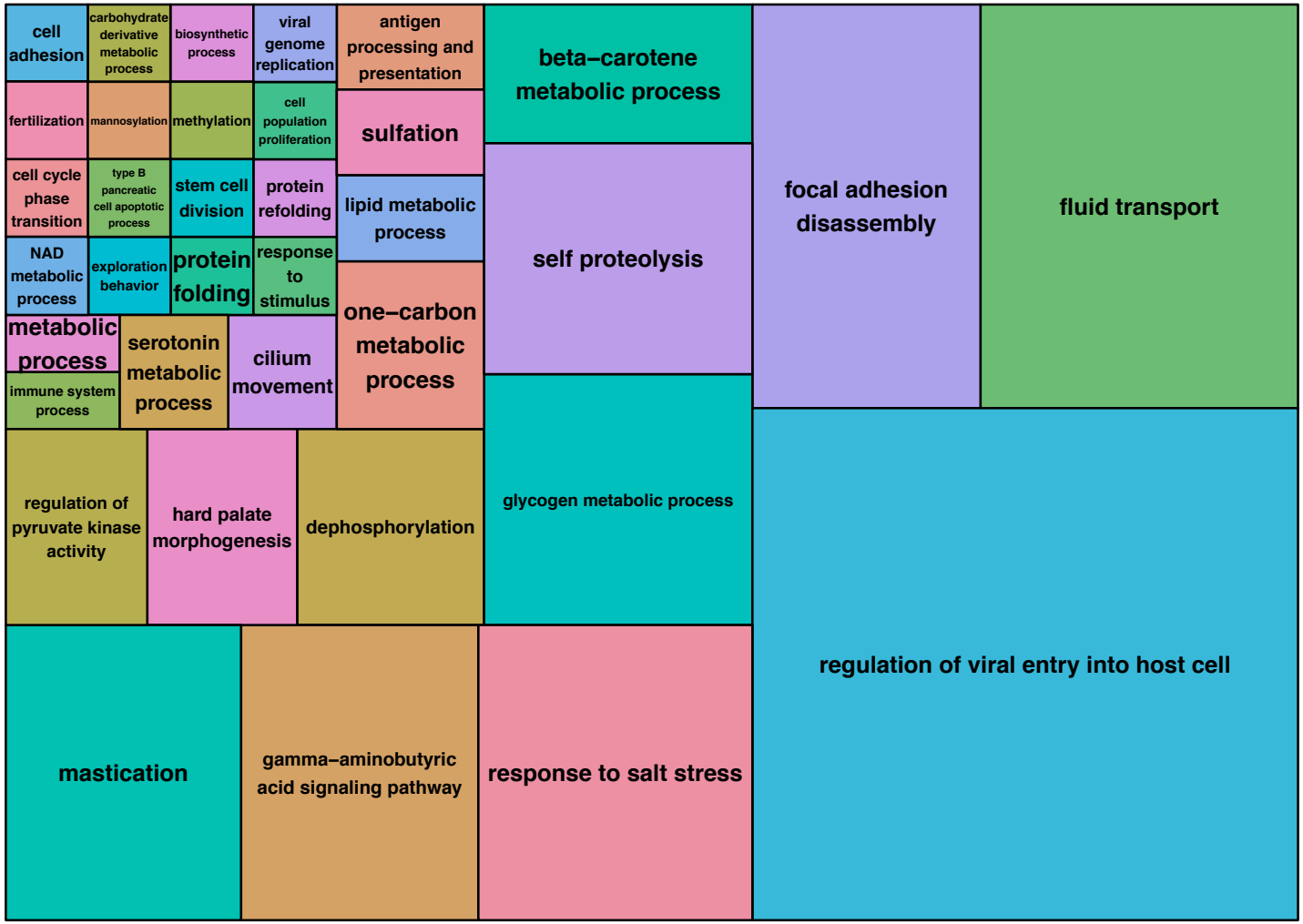


Figure 4.8. TreeMap of GO biological process terms for Otariidae lineage. GO terms present in unique gene sets evolving under positive selection in the Otariidae lineage. Rectangle size reflects semantic uniqueness of GO term, which measures the degree to which the term is an outlier when compared semantically with the list of terms present in the UniProt database (Bateman et al., 2021).

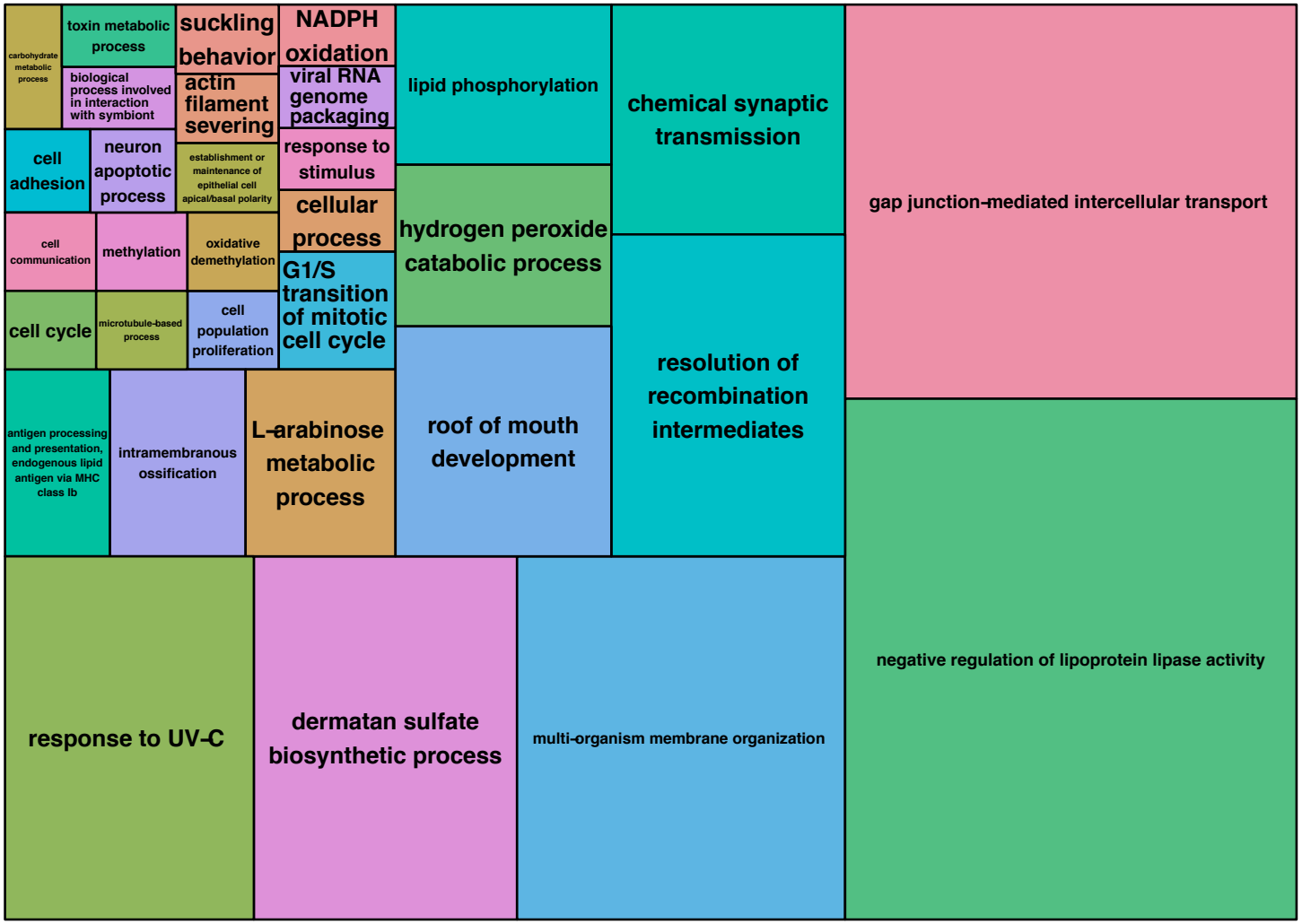


Figure 4.9. TreeMap of GO biological process terms for Caspian seal lineage. GO terms present in unique gene sets evolving under positive selection in the Caspian seal lineage. Rectangle size reflects semantic uniqueness of GO term, which measures the degree to which the term is an outlier when compared semantically with the list of terms present in the UniProt database (Bateman et al., 2021).

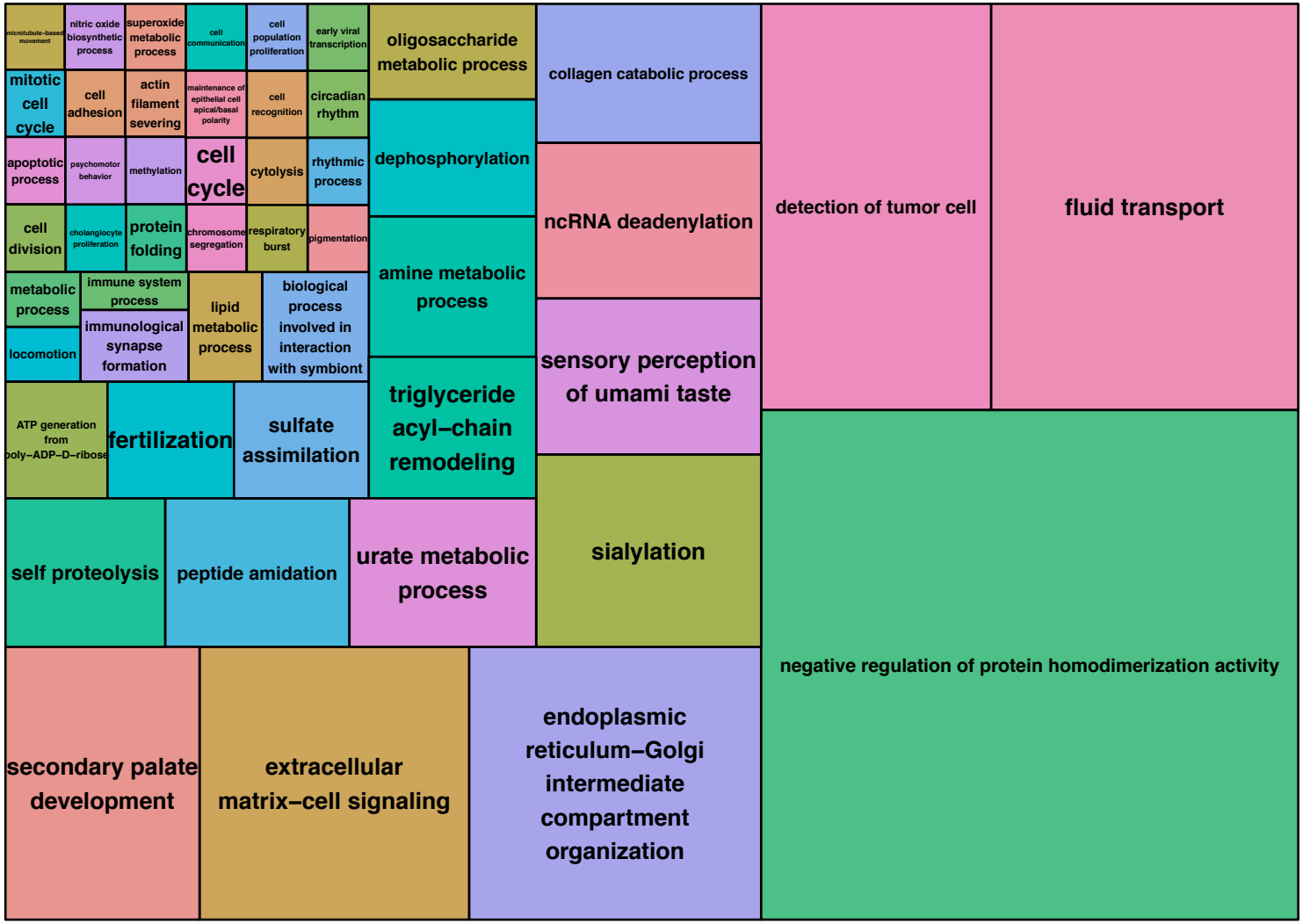


Figure 4.10. TreeMap of GO biological process terms for Hooded seal lineage. GO terms present in unique gene sets evolving under positive selection in the Hooded seal lineage. Rectangle size reflects semantic uniqueness of GO term, which measures the degree to which the term is an outlier when compared semantically with the list of terms present in the UniProt database (Bateman et al., 2021).

I found evidence for adaptations of palate/mouth development across within all lineages, this could be a consequence of the very important feeding adaptations, which vary across species and lineage (Kienle et al., 2021).

4.3.2 Rapidly evolving genes in Caspian seal lineage and Hooded seal lineage

4.3.2.1 Caspian seal lineage

From the 5,022 Caspian seal SGOs included in the branch-site selective pressure analysis, 4,567 were also represented in the closet sister taxon, the Grey seal. The protein distance comparison retrieved 1,586 SGOs with a PD_I of > 0, with 77 genes falling within the top 5 percentile of rapidly evolving genes. Of the 77 genes classed as REGs, 4 of these, protein crumbs homolog 1 (CRB1), thioredoxin domain-containing protein 3 (NME8), Kinesin family member 21A (KIF21A), and Peroxisomal Biogenesis Factor 12 (PEX12), were also putatively under positive selection (Figure 4.11).

There were no significantly enriched GO terms or GO slim terms returned when analysing the rapidly evolving positively selected genes. Although, 41 GO terms and 4 GO slim terms were significantly enriched before correction for multiple testing (Table 4.6). Significant enriched GO terms were analysed for their hierarchical function using the PantherDB v16 (Mi et al., 2020) and it was found that CRB1 has links to “blood vessel remodeling” and “post-embryonic retina morphogenesis in camera-type eye” and NME8 has also found to be involved in “flagellated sperm motility” and ‘cellular response to reactive oxygen species’.

4.3.2.2 Hooded seal lineage

3,864 of 4,257 Hooded seal SGOs had representatives in harbour seal, the closest sister species in the phylogeny. The protein distance analysis found 2,537 SGOs with a $PD_I > 0$, with 127 rapidly evolving genes of which 21 were putatively under positive selection (Figure 4.11).

There were no significantly enriched GO terms or GO slim terms after multiple test correction, despite 128 GO terms and 2 GO slim terms, being significantly enriched. The functions of these genes were investigated using PantherDB v16 (Mi et al., 2020) and several of these genes returned interesting associations. LDL receptor related protein 5 (LRP5) was shown to be associated with “cell signalling in mammary gland development”, “Norrin signalling”, and “blood vessel morphogenesis”; latent

transforming growth factor beta binding protein 3 (LTBP3) associated to “Growth factor beta signalling”; WNK lysine deficient protein kinase 3 (WNK3) associated to “osmotic cellular stress”; Diacylglycerol kinase theta (DGKQ) associated with lipid metabolic process through “progesterone biosynthetic process” and “Ketone biosynthetic process”; Sacsin molecular chaperone (SACS) was also found to be a chaperone of Heat shock protein 70.

4.3.3 Lactation associated positively selected genes

4.3.3.1 Gene function association through KEGG pathway

Despite multiple KEGG pathways being overrepresented within the different database lists, no pathways or traits were significantly enriched after correcting for multiple testing from the KEGG pathway and database overrepresentation analysis, most likely due to low gene counts. However, I investigated the pathways that had the highest positively selected genes to background gene ratio from all KEGG pathways.

Two KEGG pathways linked to the immune system were found to have a high number of representatives in pinnipeds (Table 4.7). The was due to several cytokine genes being found to be under positive selection, interleukin family genes such as IL-17 have previously been found under positive selection in pinnipeds (Park et al., 2018) and other marine mammals (Tollis et al., 2018), due to their role in immune responses to pathogen infections. Two genes involved in ferroptosis were found to be under positive selection in pinnipeds, TFRC and LPCAT3. Ferroptosis is a mode of cell death caused by lipid peroxidation, adaptive responses to reduce ferroptosis are expected in a clade of species that generate high reactive oxygen species created by regular ischemia events, whilst possessing a large lipid store (Piotrowski et al., 2021). Ferroptosis an iron dependent process and TFRC is an essential iron transporter involved in Ferroptosis (Lu et al., 2021). Although as TFRC is also involved in iron regulation and responds to low cellular iron levels (Penso-Dolfin, 2020), selection on TFRC could also be due to the need for iron during the synthesis of respiratory pigments that are essential for oxygen transport and storage in blood and muscle (Lenfant et al., 1970). LPCAT3 is a lipid metabolism enzyme that is a key promotor of ferroptosis (Li and Li, 2020), but LPCAT3 could play an addition role in the pinniped adaptations due to its function reducing lipid plasma levels through in lipid absorption (Li et al., 2015).

Table 4.6. Significantly enriched GO slim terms from positively selected REGs.

GO slim term	Fisher test P value	Bonferroni corrected P value
Caspian seal		
cytoskeletal protein binding	0.027093802	0.867001668
plasma membrane organization	0.031768909	1
peroxisome	0.038958711	1
protein targeting	0.043722856	1
Hooded seal		
kinase activity	0.026759098	1
transferase activity, transferring acyl groups	0.006248469	0.487380544

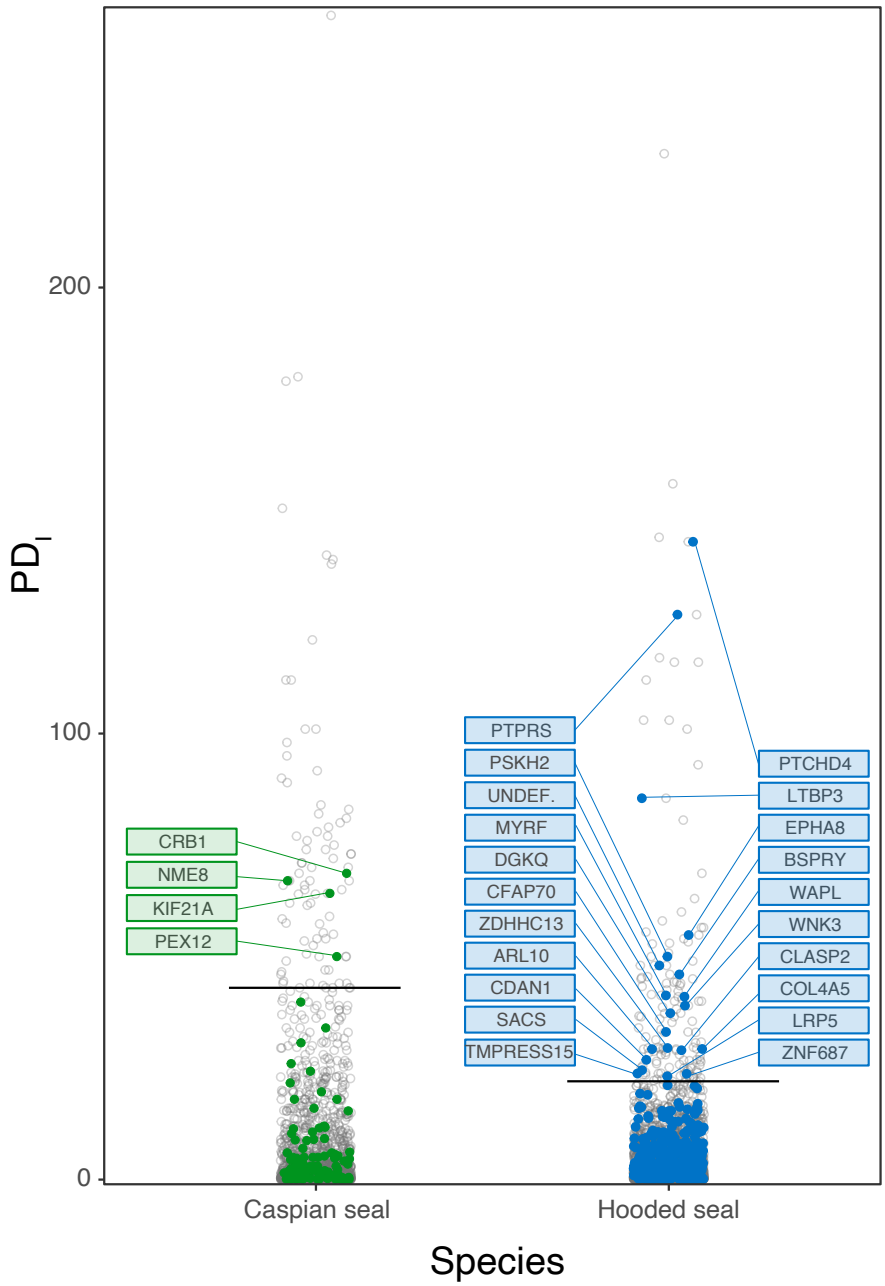


Figure 4.11. Protein distance analysis of Caspian and Hooded seal lineages. The PD_I of SGOs from the Caspian (N=4,567) and Hooded seal (N=3,864) lineages. Grey unfilled dots represent the PD_I of an SGO with no signature of positive selction, whereas the coloured dot signals that the SGO was classified as putatively under positive selection through the branch-site selective pressure analysis. The solid lines represent the 5% cut-off of PD_I scores, points above the line are described as 'rapidly evolving genes' (REGs). The HGNC gene names of the positively selected REGs are highlighted, 4 PSGs were rapidly evolving in the Caspian lineage, whereas 21 PSGs were rapidly evolving in the Hooded seal lineage (1

of the positively selected REGs in the Hooded seal lineage had no HGNC gene name, labelled as UNDEF.).

Two KEGG pathways in Phocidae returned an association to carbohydrate metabolism, this could reflect of modifications Phocids have acquired as part of fasting processes, which facilitate sustained use of lipids whilst avoiding ketoacidosis (Crocker ad Champagne, 2018). One of the genes, GYG2, found in the pathway 'starch and sucrose metabolism' also plays a role in glycogen biosynthesis, some species of Phocid have been found to have larger stores of glucose in the brain that terrestrial mammals, allowing neurons to survive for extensive periods during hypoxia (Czech-Daniel et al., 2014). Genes related to mitophagy may also have a role in the deep diving behaviour of Phocids, since hypoxia caused by deep dives can lead to mitochondrial oxidative stress (Wu and Chen, 2015), mitophagy would be a critical process required to remove damaged mitochondria.

The top KEGG pathways for Otariidae showed no obvious relationship to lactation. Although within the biosynthesis of cofactors I found two genes responsible for lipid metabolism, EPRS1 and NADK. EPRS1 gene contributes to adiposity, with mutations in EPRS1 leading to reduced adipose tissue mass (Arif et al., 2017), whilst knockouts of NADK causes lipid storage defects due to its role in providing NADPH, which is an essential reducing agent in lipogenesis (Xu et al., 2021). Phagosome and EGFR tyrosine kinase inhibitor resistance relate to immune responses.

Within the species-specific lineages, Caspian seal and Hooded seal, the top scoring KEGG pathways showed no direct relation to lactation strategy, with genes mostly being related to immune and other cellular processes. Within the Caspian seal I saw a proportion of genes selected for protein digestion and absorption, five out of the six PSGs in this pathway were collagen genes, suggesting this may be related to collagen rich blubber organisation. All the top scoring KEGG pathways in the Hooded seal PSGs are related to immune response and cancer preventative pathways.

4.3.3.2 Lactation associated positively selected genes

From cross-referencing against the candidate gene lists (Table 4.4), 181 genes were positively selected genes within the pinniped branch, 94 within the Phocidae branch, 84 within the Otariidae branch, 190 within the Hooded seal lineage, and 78 within the Caspian seal lineage (Table 4.8). PSGs with representation were subjected to a literature search to provide putative roles in mammalian adaptation.

It can be quite common for genes to have no known function in the literature. Thus, genes with no literature describing their function in mammals are not discussed in detail below. Leaving 15 PSGs in the

Table 4.7. Top five positively selected gene to background genes ratio from all human KEGG pathways, for each lineage of interest.

ID	Description	No. genes	PSG/Pathway ratio	KEGG Class
pinnipeds				
hsa05321	Inflammatory bowel disease	5	0.217	Immune disease
hsa04623	Cytosolic DNA-sensing pathway	4	0.190	Immune system
hsa04977	Vitamin digestion and absorption	2	0.182	Digestive system
hsa04216	Ferroptosis	2	0.167	Cell growth and death
hsa04929	GnRH secretion	2	0.154	Endocrine system
Phocidae				
hsa00052	Galactose metabolism	2	0.200	Carbohydrate metabolism
hsa03420	Nucleotide excision repair	2	0.200	Replication and repair
hsa00500	Starch and sucrose metabolism	3	0.188	Carbohydrate metabolism
hsa04977	Vitamin digestion and absorption	2	0.182	Digestive system
hsa04137	Mitophagy – animal	2	0.143	Transport and catabolism
Otariidae				
hsa00983	Drug metabolism - other enzymes	2	0.200	Xenobiotics biodegradation and metabolism
hsa01240	Biosynthesis of cofactors	5	0.152	n/a
hsa04145	Phagosome	5	0.147	Transport and catabolism
hsa04970	Salivary secretion	2	0.125	Digestive system
hsa01521	EGFR tyrosine kinase inhibitor resistance	2	0.105	Drug resistance: antineoplastic
Caspian seal				
hsa04540	Gap junction	3	0.250	Cellular community - eukaryotes
hsa05212	Pancreatic cancer	2	0.182	Cancer: specific types

hsa04725	Cholinergic synapse	3	0.150	Nervous system
hsa04974	Protein digestion and absorption	6	0.143	Digestive system
hsa05130	Pathogenic E. coli infection	4	0.138	Infectious disease: bacterial
Hooded seal				
hsa05224	Breast cancer	6	0.375	Cancer: specific types
hsa04610	Complement and coagulation cascades	8	0.320	Immune system
hsa04350	TGF-beta signalling pathway	3	0.300	Signal transduction
hsa05322	Systemic lupus erythematosus	3	0.300	Immune disease
hsa01522	Endocrine resistance	5	0.294	Drug resistance: antineoplastic

Table 4.8. Number of positively selected genes found in each functional category, generated from candidate gene sets. L = Lipid metabolism and mobilisation, D = Diving, BS = Body size, LM = Lactation: mammary function, LI = Lactation: involution, F = Fasting, M = Milk properties. Candidate genes sets are mutually inclusive.

Lineage	L	D	BS	LM	LI	F	M
pinnipeds	31	26	8	28	14	9	2
Phocidae	20	23	9	18	8	12	4
Otariidae	17	27	5	22	6	4	3
Hooded seal	29	54	17	48	17	18	7
Caspian seal	7	20	7	18	9	14	3

Table 4.9. 22 Lactation-associated PSGs within the pinniped branch.

Gene name	Related function	No. sites w>1
CD93	hypoxia	2
PRCP	hypoxia	2
CHI3L1	lactation: involution	1
HERC6	lactation: mammary function	2
NOSTRIN	lactation: mammary function	11
CD74	lactation: mammary function; lipid metabolism/ transport	2
APOA1	lipid metabolism/ transport	3
APOC4	lipid metabolism/ transport	1
ASXL3	lipid metabolism/ transport	5
CYP8B1	lipid metabolism/ transport	2
HACD3	lipid metabolism/ transport	1
LPCAT3	lipid metabolism/ transport	7
MOGAT1	lipid metabolism/ transport	2
NCEH1	lipid metabolism/ transport	2
PLCD3	lipid metabolism/ transport	1
FASN	lipid metabolism/ transport; milk properties	2

Table 4.10. 15 lactation associated PSGs within the Phocidae branch.

Gene name	Related function	No. sites w>1
F13A1	fasting	2
SIGLEC1	hypoxia	1
STAT6	lactation: involution	1
ACSL5	lipid metabolism/ transport	5
ELOVL2	lipid metabolism/ transport	1
GLA	lipid metabolism/ transport	1
INPPL1	lipid metabolism/ transport	1
PLIN2	lipid metabolism/ transport	1
GHITM	milk properties	4
SLC37A1	milk properties	1
SORCS1	milk properties	1

Table 4.11. Lactation associated PSGs within the Otariidae branch.

Gene name	Related function	No. sites w>1
CERK	fasting	1
JAK1	hypoxia; lactation: mammary function	1
CEP63	lactation: mammary function	1
CGN	lactation: mammary function	1
MCL1	lactation: mammary function	4
ADIPOR1	lipid metabolism/ transport	10
STAT5B	lipid metabolism/ transport	2
CD36	lipid metabolism/ transport; milk properties	1
MGST1	milk properties	1

Table 4.12. Lactation associated PSGs within the Hooded seal lineage.

Gene name	Related function	No. sites w>1
CAMKK1	fasting	2
KLF11	fasting	14
SERPINE1	fasting	1
ENPEP	hypoxia	1
EPAS1	hypoxia	29
NCF1	hypoxia	2
NOS3	hypoxia	1
GBF1	lactation: mammary function	22
ABCA7	lipid metabolism/ transport	15
DGKQ	lipid metabolism/ transport	19
LIPC	lipid metabolism/ transport	17
LRP5	lipid metabolism/ transport	7
PNPLA3	lipid metabolism/ transport	1
SLC51B	lipid metabolism/ transport	2
THBS1	lipid metabolism/ transport	6
CACNA1E	milk properties	1
RPAP3	milk properties	3
SEMA4G	milk properties	2
TMIGD3	milk properties	4
XDH	milk properties	13
ZC3H3	milk properties	11

T**able 4.13. Lactation associated PSGs within the** Caspian seal **lineage.**

Gene name	Related function	No. sites w>1
EEF2K	fasting	1
VAV3	hypoxia	3
NEK11	lactation: mammary function	1
ALMS1	lipid metabolism/ transport	3
FAM83H	milk properties	3
MROH1	milk properties	1
SPHK2	milk properties	1

4.3.3.3 Pinniped branch - Lactation associated positively selected genes

Previous studies of selection analyses in Cetaceans have also identified FASN, MOGAT1, CYP8B1, and PLCD3 as under positive selection, suggesting possible convergent evolution of some genes (Endo et al., 2018; Park et al., 2018; Wang et al., 2015). CYP8B1 encodes a member of the cytochrome P450 superfamily of enzymes that are responsible for the catalysis of primary bile acids that play important roles in glucose tolerance and insulin sensitivity (Kaur et al., 2015). FASN encodes fatty-acid synthase, which catalyses the conversion of acetyl-CoA and malonyl-CoA to long-chain saturated fatty acids with key roles in obesity (Bernt et al., 2007; Mayas et al., 2010). I found three genes related to TAG metabolism and transport (MOGAT1, PLCD3, and LPCAT3). MOGAT1 and PLCD3 are essential enzymes involved in the TAG biosynthesis pathway (Sankella et al., 2016). LPCAT3 encodes a phospholipid remodelling enzyme which modulates lipogenesis (Rong et al., 2017), and is involved in the transport of TAGs, it is also involved in the production of lipoproteins in humans (Hashidate-Yoshida et al., 2015). NCEH1 is a key regulator of cholesterol transport and involved in hydrolysis of cholesterol esters to a comparable degree as LIPE (Igarashi et al., 2010). Two apolipoprotein genes were also found to be under positive selection in pinnipeds, APOA1 and APOC4. APOA1 produces apolipoprotein A1 which is major constituent of high-density lipoprotein and involved in reverse cholesterol transport (Liao et al., 2015), whereas APOC4 is involved in the regulation of plasma lipid levels (Xu et al., 2015). HACD3 was also under positive selection in pinnipeds which produces a 3-hydroxyacyl-CoA dehydratase which is involved in very-long fatty acid chain synthesis (Ikeda et al., 2008).

Genes related to mammary gland function, were related to immunity responses of the mammary gland or (CD74 and CHI3L1) (Breyne et al., 2017; dos Santos et al., 2013). I found two genes that have been shown to influence lipid levels of milk (HERC6 and ASXL3) in cattle. HERC6 has significant effects on fat and protein yield in breeds of *Bos taurus* (Cohen-Zinder et al., 2005) and ASXL3 has been found to interact with many genes that directly affect milk composition in cattle (Sanchez et al., 2019), FASN has also been shown to be a well-known candidate effecting milk fatty acid levels in cattle (Pegolo et al., 2017).

I also identified two genes in the pinniped lineage that could influence the effects of diving and hypoxia: PRCP has been shown to influence blood pressure, vascular anticoagulation and contribute to cell proliferation and angiogenesis (Adams et al., 2013); NOSTRIN plays a significant role in nitric oxide synthesis and trafficking, which is a key mediator of vasodilatory responses to hypoxia in terrestrial mammals (Nuñez et al., 2014).

4.3.3.4 Phocidae branch - Lactation associated positively selected genes

One gene, F13A1, has an annotated function consistent with involvement in fasting. Tthrough regulating insulin resistance and has been identified as a gene that may contribute to obesity in humans (Myneni et al., 2016).

Two genes were shown to have associations with TAGs: ACSL5, which is primarily expressed in adipose tissue is involved in the synthesis of TAGs from fatty acids (Bu and Mashek, 2010), ACSL5 has also been suggested to contribute to increased blubber synthesis in cetaceans and other carnivores (Zhao et al., 2019). The process of which ELOVL2 increases TAG synthesis is not well understood, although it has been shown to increase TAG from extracellular fatty acids (Kobayashi et al., 2007). The mechanisms of other lipid associated genes are not fully detailed, INPPL1 knockouts in mice show decreased weight gain when on a high fat diet (Sleeman et al., 2005); GLA mutations are a cause of Fabry disease in humans, a condition with increased lipid building up in blood vessels and organs (Song et al., 2009).

STAT6 has been classified as a 'FAT STAT' due to its activation by LEPTIN (Ghilardi et al., 1996), with functions involved with immune and inflammatory responses, in addition to being shown to mediate development of the mammary epithelial cells during mammary gland development (Wu et al., 2021). Four of the identified genes are involved in fat transport or synthesis during lactation: PLIN2 is a cytoplasmic lipid droplet binding protein which could be involved with lipid droplet membrane wrapping of cytoplasmic lipids, it has also been shown to be involved in milk fat synthesis in goats (Mu et al., 2021; Zhu et al., 2015). SORCS1 was found by SNP associations with fatty acids in *Bos taurus* (Li et al., 2014; Palombo et al., 2018) and SLC37A1 is a solute carrier gene that has associations with mineral content of milk in *Bos taurus* (Neyeri et al., 2016; Sanchez et al., 2021). GHITM is associated with increased polyunsaturated and saturated SFAs in *Bos taurus* milk (Palombo et al.,2018).

I also found a putative involvement for hypoxia related traits through SIGLEC1, which has been linked to genetic adaptation to high altitudes in the Tibetan cashmere goat (Song et al., 2016).

4.3.3.5 Otariidae branch - Lactation associated positively selected genes

Within the Otariidae lineage ADIPOR1 was found to be under positive selection; ADIPOR1 is a receptor for adiponectin. Adiponectin is a key adipocyte specific hormone involved in glucose regulation and fatty acid oxidisation (Siitonen et al., 2011), with a regulatory role of mammary epithelial cells (Zhao et al., 2021).

CERK has a role in adipocyte cell differentiation, with a putative role in adipogenesis regulation (Ordoñez et al., 2017). CD36 is used in facilitating and regulating the transport of fatty acids across the plasma membrane and has been found to be highly expressed in the mammary gland of *Bos taurus* (Ganguly et al., 2017). CD36 was also identified as under positive selection in cetaceans, suggesting possible convergent evolution of genes related to increased body mass (Wang et al., 2015). STAT5B has been shown to be involved in the expression of growth hormone genes, which have been shown to be sexually dimorphic in rats and mice (Kaltenecker et al., 2019), with humans with STAT5B missense variants showing extreme growth deficiencies (Hwa et al., 2005).

I found genes with a putative role in mammary gland involution. In addition to roles in cell growth and development JAK1 provides a crucial link between cytokines and STAT3 activation which activates cell death and mammary gland remodelling during mammary gland involution (Jena et al., 2019), STAT5B also links to the STAT3 pathway with indirect induction of the LIF-STAT3 pathway with initiates involution (Jena et al., 2019). MCL1 expression levels have been shown to fall with the onset of mammary gland involution, suggesting its down regulation is essential for initiation of involution (Fu et al., 2015). From the 3 genes with a relation to milk production, only CD36 and MGST1 (Nayeri and Stothard, 2016) were shown to influence milk lipids levels, while CEP63 is associated with milk yield (Chen et al., 2018). CGN is an immune related gene that has been shown to be expressed in mid-lactation in mammals (Arora et al., 2019).

4.3.3.6 Hooded seal lineage - Lactation associated positively selected genes

For the Hooded seal I found nine genes with functions related to lipid metabolism or transport in mammals. LIPC is a key enzyme responsible for TAG hydrolysis as well as metabolism of high-density lipoproteins (Teng et al., 2018). LIPC also plays a role in plasma lipid level regulation and has knock out studies in mice have shown a reduction in body fat (Chiu et al., 2010). DGKQ is also an essential enzyme involved in triglycerol metabolism, which has additionally been found under selection in cetaceans (Wang et al., 2015). I found lipid metabolism related genes that could also be involved in fasting. PNPLA3 has been shown to reduce the activity of PNPLA2 (Dong, 2019). PNPLA2 is an enzyme involved in the initial steps of TG hydrolysis in lipid droplets. PNPLA2 expression has been shown to increase during lactation in Phocidae without an increase in enzyme activity (Fowler et al., 2018), and PNPLA3 could be a candidate as a potential regulator for this reaction. FGFR1 is a receptor which effects expression of genes involved in lipid metabolism in the liver, which could have possible function reulating lipid metablism during fasting (Yang et al., 2012). Two genes of interest, THBS1 and LRP5, have been observed in humans to contribute

to non-alcoholic fatty liver disease (NAFLD) and obesity risk respectively, suggesting putative roles in lipid metabolism or liver protection (Bai et al., 2020; Loh et al., 2015).

Of the 7 genes I found with associations to milk properties (Table 4.8), 5 were related to milk fat levels. XDH is an enzyme involved with lipid droplet formation which has been shown to be highly expressed in lactating *Bos taurus* as well as being shown to regulate milk lipid secretion in humans (Ganguly, 2017; Zhao et al., 2020). CACN1E, SEMA4G TMIGD3, and ZC3H3 have associations with an increase in milk fat in gene association analyses in *Bos taurus* or *Bubalus bubalis* (Du et al., 2019; Palombo et al., 2018). SLC51B and GBF1 show an association to milk fat percentage o*f Bos taurus* (Neyeri and Stothard, 2016). ABCA7 was found through its association to lipid metabolism/transportation but through further investigation was realised to be a lipid transporter that has an inverse relationship with blood cholesterol levels in *Bos taurus* mammary glands (Mani et al., 2009). There is evidence of RAP3 being related to protein levels of milk in *Bos taurus* (Raven et al., 2014)

From genes related to fasting, I found that a mutation in SERPINE1 in humans is found to decrease insulin levels whilst fasting (Khan et al., 2017); KLF11 knockouts in mice have increased insulin sensitivity (Mathison et al., 2015). Whilst, CAMKK1 has been observed to undergo methylation in response to weight reduction (Crujeiras et al., 2021).

I found 4 genes with a potential association to diving-induced hypoxia in the Hooded seal lineage. NOS3 has been found to be highly expressed in the brain of Weddell seal, another deep diving Phocid species, and is potentially linked with enhanced nitric oxide production and vasodilation (Hindle et al., 2019). ENPEP is also shown to be under positive selection in Weddell seal (Noh et al., 2022). ENPEP is an aminopeptidase which is a part of the renin-angiotensin signaling pathway, other aminopeptidases, such as ANPEP or LNPEP, have been shown to be under positive selection in species of cetaceans (Kishida et al., 2007; Lui et al., 2019). NCF1 has been shown to be involved in reactive oxygen species (ROS) mediation (Hultqvist et al., 2004) and EPAS1 has been shown to be involved with hypoxia adaptations in *Canis lupus familiaris* and humans (von Holdt et al., 2017).

4.3.3.7 Caspian seal lineage - Lactation associated positively selected genes

Only 7 genes were found with known associations to lactation trait in mammals. ALSM1 was the only gene with an association to lipid metabolism, with mutations in this gene causing Alström disease, a symptom

of this disease is obesity in humans (Hearn et al., 2002). EEF2K has been shown to preserve cellular ATP during long periods of nutrient deprivation (Cope et al., 2018). VAV3 has been shown to be involved in cardiovascular regulation and angiogenesis, and is under selection in human and *Ovis aries* high altitude populations (Edea et al., 2019; Scheinfeldt et al. 2012).

I found 4 genes with a relation to milk composition. NEK11, FAM83H and MROH1 have associations with protein and cholesterol composition of milk and milk yield respectively (Atashi et al., 2020; Bhat et al., 2019; Do et al., 2018; Illa et al., 2021). SPHK2 is upregulated in the late stages of lactation in *Bos taurus* and is involved in sphingolipid fat synthesis (Bionaz and Loor, 2008).

4.3.3.8 Tests of synonymous saturation

dS saturation tests were performed for each of the 90 LAPSGs, measuring dS using across the entire gene, except for gapped regions. The variance of dS in each LAPSG was calculated (Figure 4.12) and the LAPSGs in the top quartile of variance ranges were further inspected to ensure the sites under positive selection did not fall in a region that was had a relatively high dS (Figure 4.13).

Seventeen LAPSGs fell in the top quartile of variance levels, with a σ >35.2. Of these 17 LAPSGs, no LAPSGs had all their sites under positive selection in such an area (Figure 4.12). Three LAPSGs (OG0002425_pruned.aln, OG0005009_pruned.aln, OG0006917_pruned.aln) had at least one selected site in relatively high dS region, but in these genes, there were at least one positively selected site that was in a relatively low region (Figure 4.13).

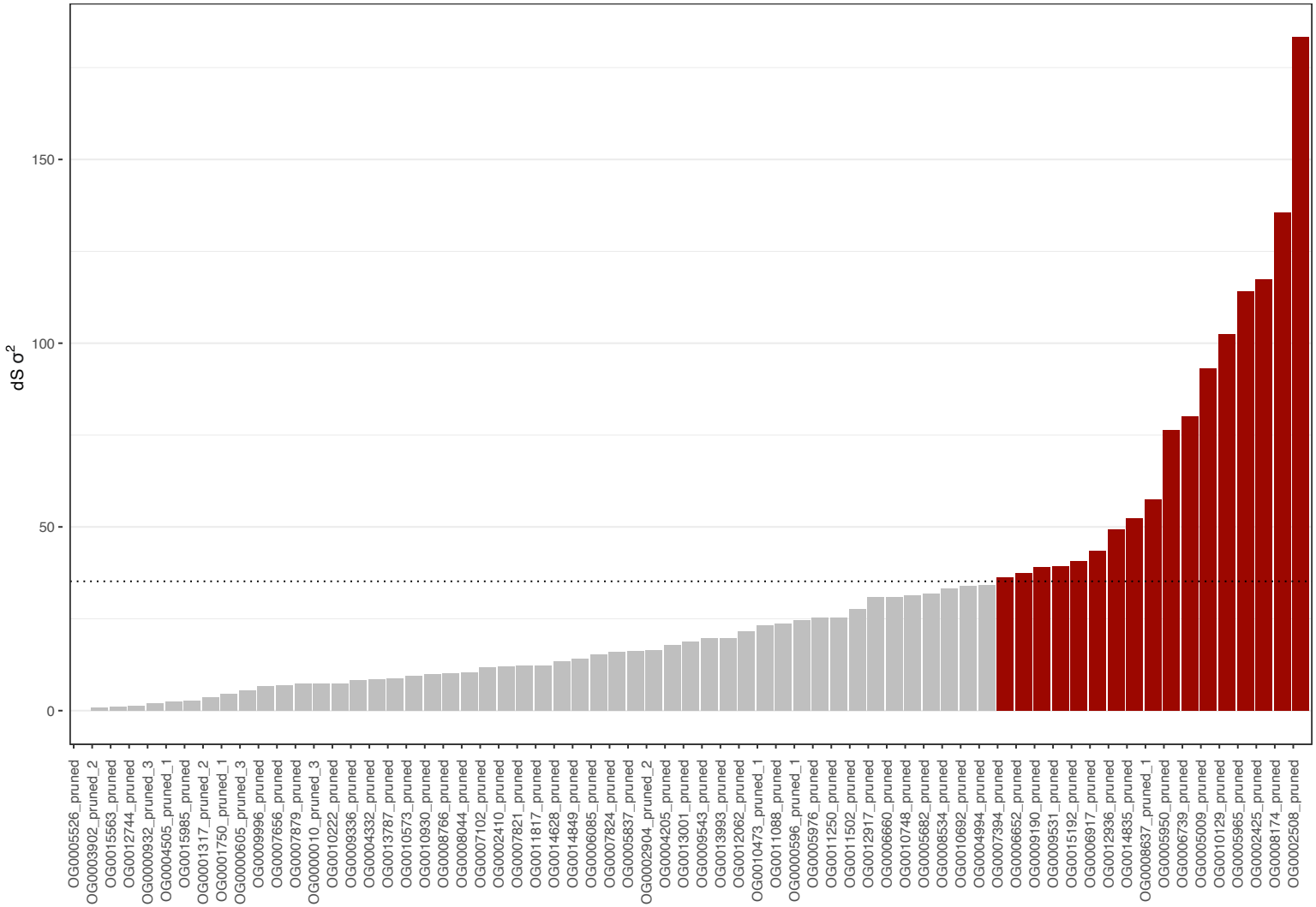
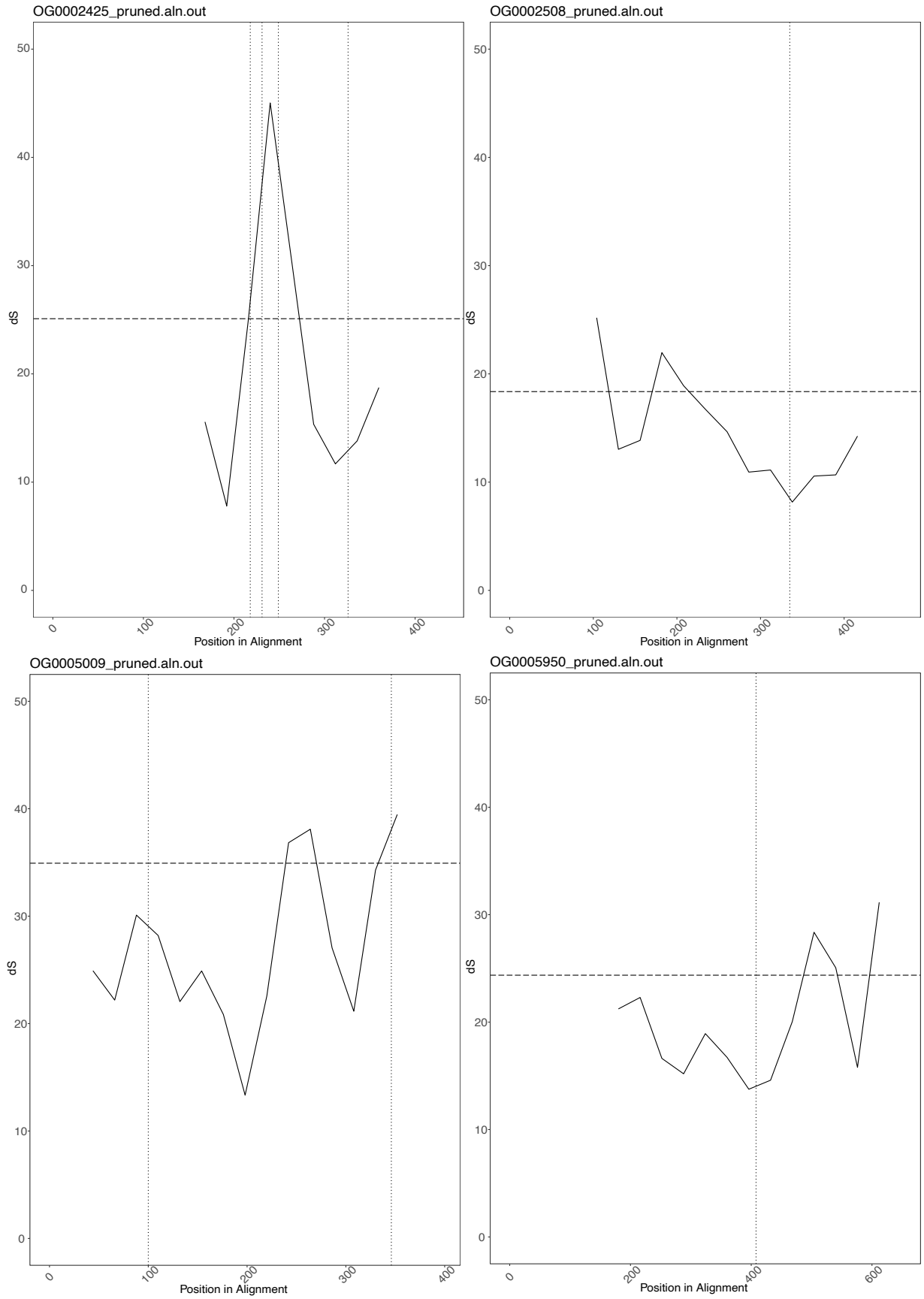
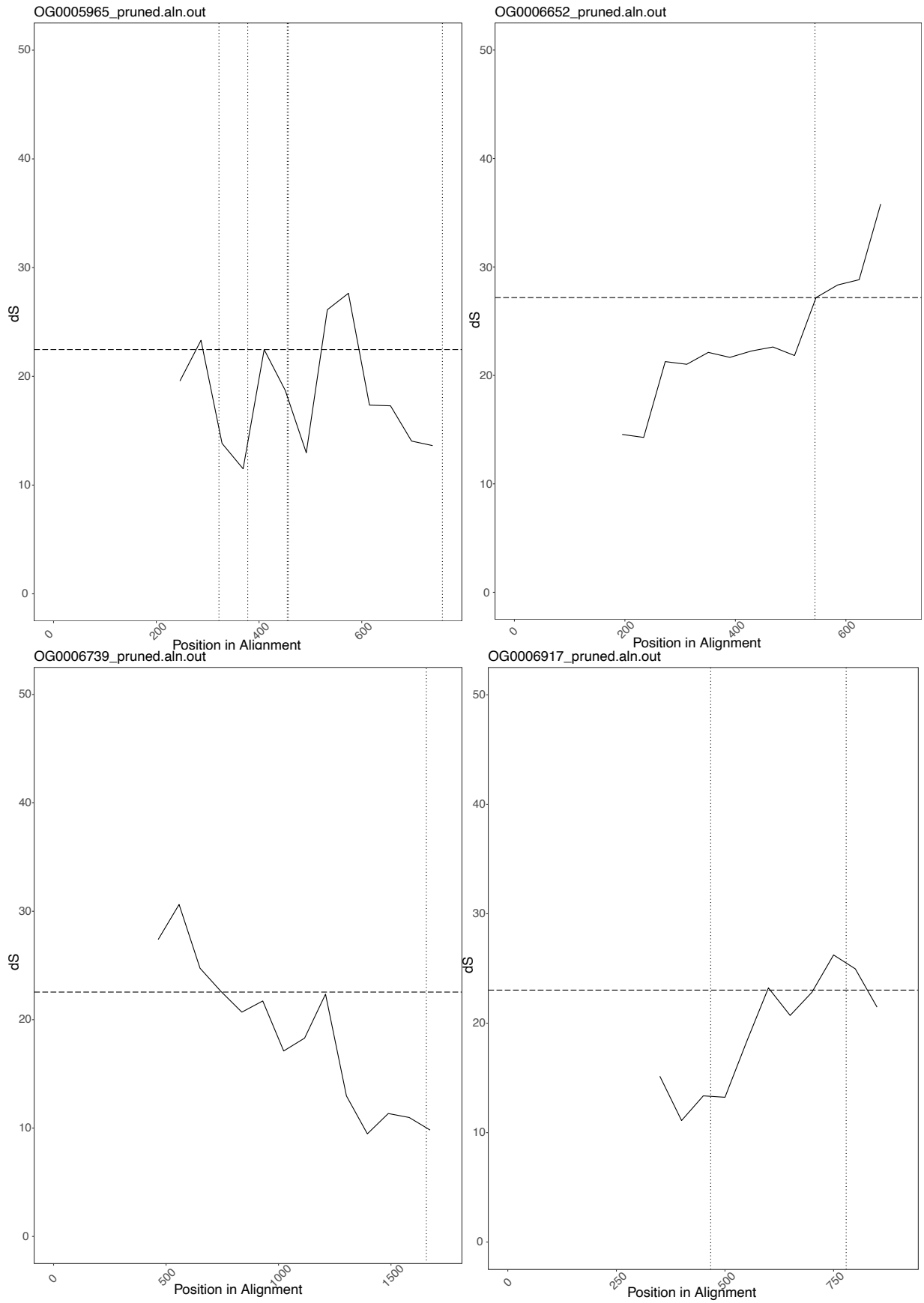
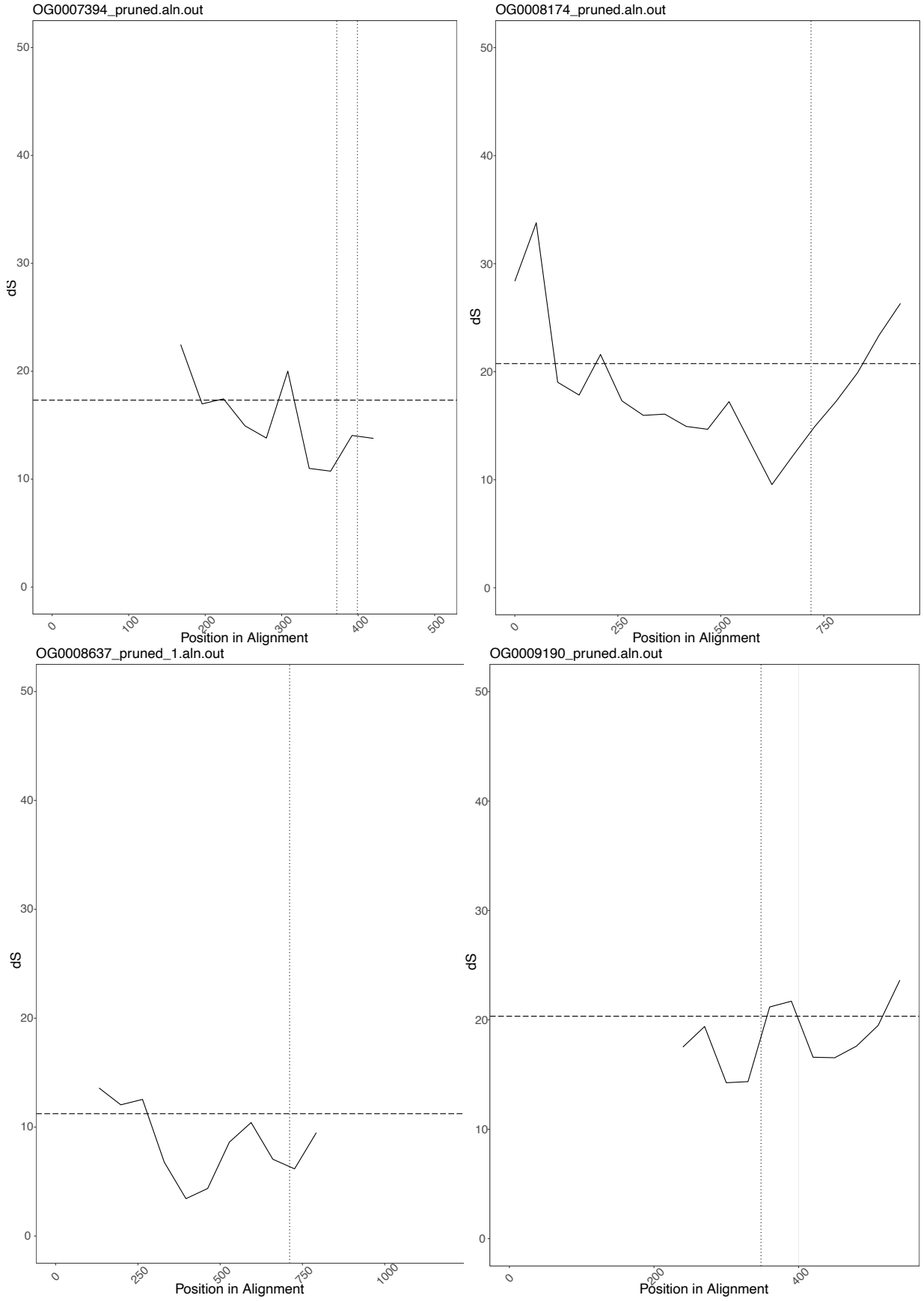
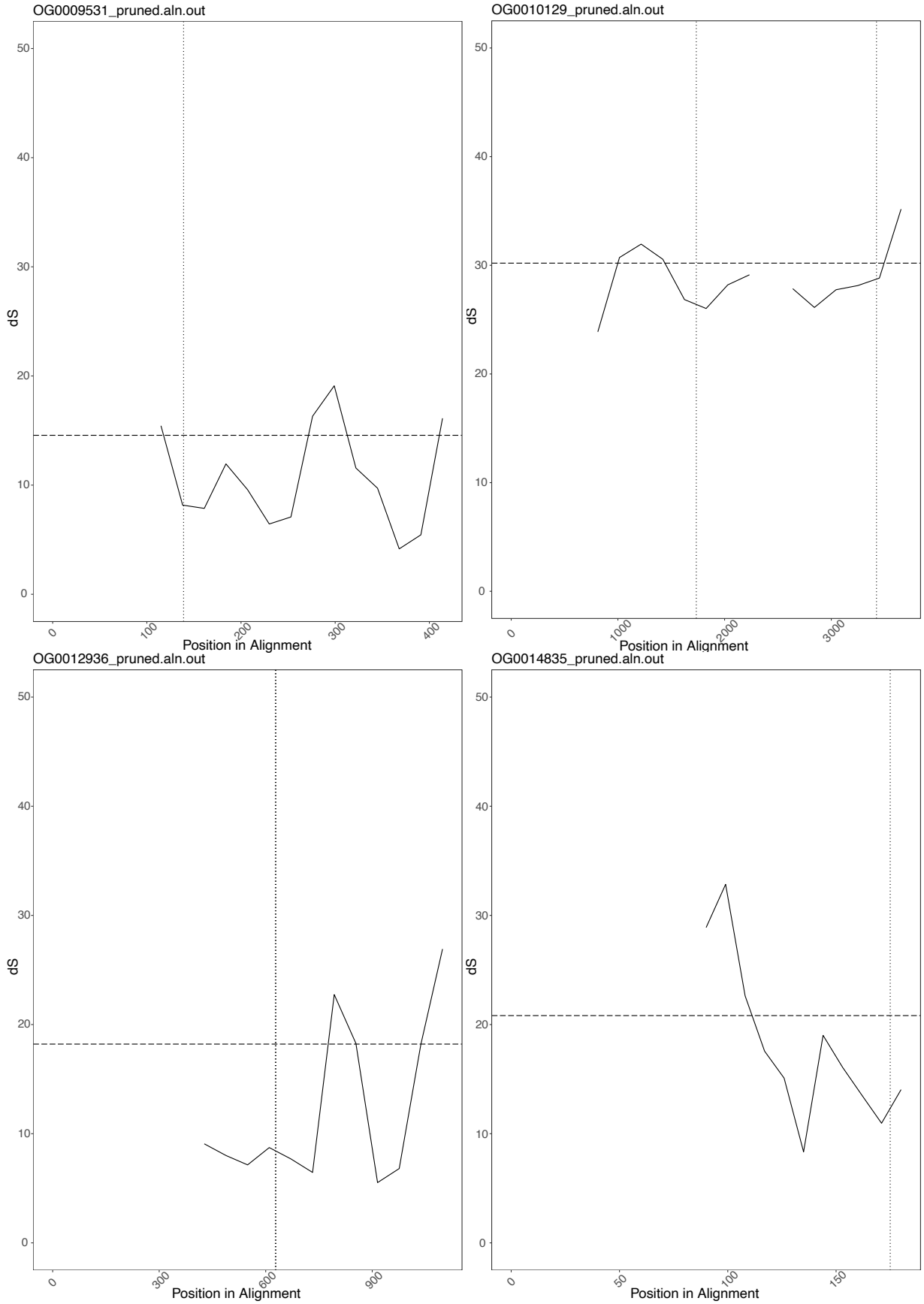


Figure 4.12. dS variance across all lactation associated positively selected genes. The line corresponds to the difference levels of variance of dS within that gene. The dotted line represents the value of the 3rd quartile of values of variance, the genes that high a higher variance than the 3rd quartile are shown in red.


OG0002425_pruned.aln.out
OG0002508_pruned.aln.out
OG0005009_pruned.aln.out
OG0005950_pruned.aln.out
dS
Position in Alignment


OG0005965_pruned.aln.out
OG0006652_pruned.aln.out
OG0006739_pruned.aln.out
OG0006917_pruned.aln.out
dS
Position in Alignment


OG0007394_pruned.aln.out
dS
Position in Alignment
OG0008174_pruned.aln.out
dS
Position in Alignment
OG0008637_pruned_1.aln.out
dS
Position in Alignment
OG0009190_pruned.aln.out
dS
Position in Alignment


OG0009531_pruned.aln.out
dS
Position in Alignment
OG0010129_pruned.aln.out
dS
Position in Alignment
OG0012936_pruned.aln.out
dS
Position in Alignment
OG0014835_pruned.aln.out
dS
Position in Alignment

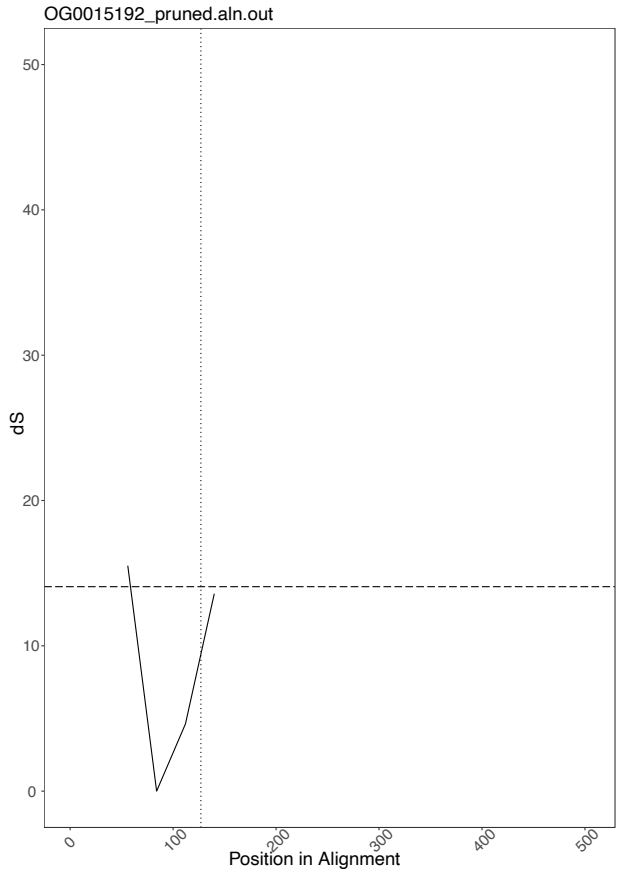


Figure 4.13. Sliding window of dS values across high variance genes. 17 genes that fall in the top quartile of variance in dS values, showing dS values across the genes using a sliding window approach. The horizontal dotted line represents the top quartile of dS values found across the gene where all species are present, and the vertical lines represents the position of positively selected sites found within the gene. Partial regions of the gene have an uncalculatable dS due to Genes that have all positively selected sites in areas of high dS values are be considered as false positives.

4.4 Discussion

In summary, I found 1,562 genes showing signatures of positive selection across the five different pinniped lineages evaluated. On further refinement of this list, I have been able to make inferences into how these genes might be involved in the phenotypic variation that occurs across present day species. The separation of provisioning and feeding has fundamentally shaped pinniped evolution, leading to the diverse phenotypic traits and variation in lactation strategies observed in present day Phocidae and Otariidae species. In this study I have tried to unravel some of the molecular underpinnings of the adaptations that influence these traits. Several phenotypic traits interact in the different lactation strategies and as a result drive extreme phenotypic adaptation. I found hundreds of genes across several lineages of pinnipeds, which have the potential to explain some of the variation seen across extant Pinnipedia.

I found 1,562 genes under positive selection from the 30,437 orthologous groups that underwent selective pressure analysis in CodeML, suggesting 5.13% of genes under positive selection. This agrees with previous publications that have found between 0.2-15% of genes under positive selection in Mammalia (Eyre-Walker, 2006; Hawkins et al., 2019). Two recent selective pressure analyses of pinniped species had found 0.25% and 1.16% of orthologous groups under positive selection (Foote et al., 2015; Park et al., 2019), these studies used filtering using correction for multiple testing and so were expected to have removed possible false negatives from their final datasets. My analyses could also be affected by the quality of the genome annotation or assembly, although I have attempted to mitigate this by including least one highly sequenced genome in all my clades, with positive selected sites required to be identical in all representatives of the clade. Using a method for testing for correction in multiple testing, which although may remove some false negatives, may also bring down the proportion of genes under positive selection in my analysis. My analyses may have produced a higher number of positively selected orthologous groups than previous analyses, but my results fall in the range of other mammalian selective pressure analyses.

<u>Lipid metabolism and mobilisation in pinniped clades</u>

Utilising blubber as an energy reserve is ubiquitous across all pinniped species. Capital breeding Phocids have the largest adipose stores, and typically exploit rich pelagic feeding grounds far away from breeding

locations. Large adipose stores can then be mobilised during lactation, sufficiently provisioning young whilst meeting their own metabolic requirements. The evolutionary constraints of no, or low, provisioning during lactation has created pressures for Phocidae to acquire larger adipose stores whilst compressing the length of lactation, facilitating rapid energy transfer from mother to pup. Converting fat reserves into an energy rich milk requires Phocids to be able to mobilise large quantities of lipids to the mammary gland. Digested lipids from prey are converted to TAGs and stored as blubber in adipocytes (fat cells). These fat reserves are later mobilised to meet maternal and pup energy requirements throughout lactation. Utilising energy reserves from lipids requires 3 main pathways: lipolysis of non-esterised fatty acids from triglycerol; lipid transport, which involves the use of lipoproteins to transport the lipids in plasma; and tissue uptake of lipids, in which TAGs are hydrolised ready for uptake by tissues, including the mammary gland.

Lipid metabolism was expected to be an integral process in the life history traits of pinnipeds, this was reflected through the GO term associations with 'lipid metabolism process' appearing in my semantic similarity analysis across all my lineages of interest, except for the Caspian seal lineage. I found many genes with established roles in lipid metabolism and mobilisation in other marine mammals, including FASN, suggesting possible convergent evolution of adipose function related genes (Wang et al., 2015). I was expecting to see genes with previously established roles in energy balance, such as Leptin (LEP) or its receptor LEPR, which has already been investigated to show evidence of positive selection in some Phocid species (Hammond et al., 2012). LEP was not included in my positive selection scans as it was discarded during my filtering process, failing the Robinson-Foulds cut off. Adiponectin (ADIPOQ) has been suggested as a possible mediator of metabolic effects in blubber (Khudyakov et al., 2019). This was found to be under positive selection in both the Phocidae and Otariidae lineages but not in the overall pinniped lineage. As I was interested in genes uniquely under selection in lineages this was filtered from my downstream analysis. I found many genes across all lineages that could have a putative function in the synthesis of blubber reserves with established roles in lipogenesis or biosynthesis of TAGs. Also, I found genes for which the mechanisms are not fully understood but phenotypic differences are seen in knock out experiments. This analysis has expanded the scope of genes that may be contributing to blubber function. Previous analyses have focused on well-studied genes in a limited number of Phocidae, I have expanded on this and now have a multitude of genes that may affect phenotypic differences across many species of pinnipeds.

The mobilisation of lipids is a key mechanism in pinniped species that allows them to utilise established energy reserves for the provisioning of young and their own metabolic requirements. This trait is extreme in capital breeding Phocids, which fast during their lactation period and so solely rely on reserves to maintain metabolic processes whilst provisioning young with extremely high percentage fat milk. Plasma lipid levels are significantly raised during lactation of Elephant seals, displaying how a capital breeding Phocids utilises these reserves (Fowler et al., 2018). Although the exact processes of how lipids are mobilised from blubber reserves, then transported to relevant tissues are not fully understood in Phocids. LPL has been shown to be involved in clearing lipid plasma levels in Phocids, with TAG plasma levels being inversely correlated with LPL activity in lactating Grey seals, and LPL activity being increased during lactation (Iverson et al., 1995). 'Regulation of plasma lipoprotein particles' and 'Negative regulation of lipoprotein lipase activity' had the largest area in my semantic similarity analyses for pinnipeds and Caspian seal lineages respectively, suggesting the regulation of plasma lipid level regulation is under selection in pinnipeds and Phocidae lineages. I found apolipoproteins and other genes with putative roles in the transport of lipids across all clades of pinnipeds that I investigated. Although, in the Otariid lineage I did not find many genes with a putative function in plasma lipid transport, but did find CD36 under positive selection, which is integral in transmembrane transport of fatty acids (Xu et al., 2013) and so may be imperative in the tissue uptake of lipids rather than their transport through plasma.

LIPE has been shown to have a key role in lipid mobilisation in other mammals, and some studies have found it expression to be positively correlated with milk fat levels of some Phocidae (Fowler et al., 2013), but in Elephant seals it remains at low expression levels in blubber throughout lactation (Crocker et al., 2014). It has been suggested that PNPLA2 might serve as a primary lipase in blubber lipolysis in Elephant seals (Fowler and Crocker. 2018). Expression or activity levels are not correlated with milk lipid levels, instead other enzymes such as Perilipin (PLIN1) may contribute to regulation of lipolysis (Fowler et al., 2015). I found PLIN2, a paralogue of PLIN1, to be under selection in Phocids, consistent with a putative role in lipolysis regulation in Phocidae. PNPLA3, a paralogue and regulator of PNPLA2, was under positive selection in the Hooded seal lineage, possibly contributing to the extreme lipid mobilisation that allows the Hooded seal to direct large amounts of lipids toward to mammary gland during lactation. In addition to this I have also found genes that have roles in non-alcoholic fatty liver disease in humans, suggesting they may have a function in protecting the liver from high levels of circulating lipids in the plasma. Despite being a potential avenue for drug therapeutics and study candidates for diseases that affect lipolysis, there is little research into the underlying molecular details of genes that permit capital breeding Phocids to

thrive despite high lipid levels in the plasma. Here I have uncovered a selection of candidate genes that might be suitable avenues of future research.

Through observation of milk fatty acid chain composition, lipids in pinniped milk fat are primarily mobilised and from adipose tissue and up taken by mammary gland tissue, rather than *de novo* lipid synthesis by the mammary gland (Neville et al., 1997). I found some established genes involved in milk composition such as CSN1S1 (Balia et al., 2013), but this was filtered out due to being under selection in both Otariidae and Phocidae lineages (although at different sites). The resources established from GWAS of high milk bovine species gave us a unique opportunity to see whether another mammalian species had any comparative gene modifications that could uncover some candidate genes that allow capital breeding Phocids to have higher milk lipid levels than other pinnipeds. I detected positive selection in several genes with an association to milk lipid levels in cattle across all my lineages, especially in the Hooded seal lineage. This could possibly infer that despite the main regulatory network dictating lipid composition being blubber mobilisation in pinnipeds, some additional regulatory mechanisms, possibly involving lipid transportation or milk lipid secretion may be important for some of the highest fat milk producing species.

Fasting in capital breeding Phocidae

The access to, and predictability, of hauling out site along with their extreme nature of the post-partition fasting cycles has resulted in studies on the mobilisation of energy reserves during fasting being primarily performed on Elephant seals (*Mirounga spp*.). Elephant seals experience fasting periods lasting months during which individuals losing up to 40% of their total mass. Blubber tissue expression and enzymatic analyses and have uncovered gene networks that may promote lipolysis and adipogenesis (the process in which adipocytes can accumulate into adipose tissues) (Crocker et al.,2014), but not much is understood about the networks that regulate hormonal factors such as cortisol and insulin, and whether these systems are identical across the difference capital breeding Phocidae. Elephant seals have low ketone levels and avoid ketoacidosis during fasting, unlike other fasting mammals, which can occur when organisms increase dependence on lipid catabolism for energy (Champagne et al., 2006; Crocker et al., 2014; Houser et al., 2007). I found DGKQ to be both under rapid and positive selection in the Hooded seal lineage and found that HMGCS2, a key enzyme in ketogenesis, had an omega value > 1 when the Phocidae lineage was used as foreground, although the p-value of the LRT was > 0.05. I only found five genes related to fasting under positive selection with corresponding mammalian functional evidence in the literature.

Thus, generating a small number of avenues of investigation for fasting properties of capital breeding Phocidae. However, this might also suggest that the fasting properties of the Phocidae may differ across lineages and could even be differentiated through regulatory changes rather than amino acid substitutions.

Lactation involution suppression in income breeding Otariidae

Income breeding Otariidae and nursing strategy Odobenidae have a much longer lactation lengths than that of capital breeding Phocidae, and this is facilitated through their unique ability to prevent involution during long foraging trips. Otariids can leave their young for weeks to build their reserves in local low-productive feeding areas, before returning to their young to continue lactation. This has decreased selective pressure on high fat milk and extended their lactation in comparison to capital breeding Phocids. The process by which Otariidae can prevent the involution of mammary glands in the absence of a suckling stimulus is still not fully resolved but it has been suggested that that the modification of the LALBA gene, which would reduce or eliminate the level of lactose in milk, may reduce the apoptosis in the mammary gland. I expected to see several genes under selection that were involved in mammary gland remodelling or immunity during lactation, although LALBA was not included in my selective pressure analysis, due to failing the Robinson-Foulds cut off. I found genes under positive selection in Otariids that possibly could be associated with the ability to prevent involution, such as JAK1 and MCL1. Both JAK1 and MCL1 have a functional role in involution and mammary gland remodelling that occurs during involution in other mammal species (Fu et al., 2015; Jena et al., 2019). I found comparable levels of genes involved with mammary function across the different lineage in showing that lactation related genes are important source of adaptation across pinnipeds.

Body size in Pinnipedia

It is thought that a larger body size of Phocidae was an adaptation to the arctic climate in which they diversified. A larger body size has then served an additional function of accumulating large energy reserves. Capital breeding Phocidae are some of the deepest and longest diving mammals, with Northern elephant seal capable of reaching depths of 1,700 metres and over 90 minutes submerged. These incredible feats allow Phocidae to reach productive pelagic waters which in turn help fill their large energy reserves, thus makes the diving capabilities of these seals instrumental to their lactation strategy and their

subsequent success. Many comparative analyses have concentrated on genes that allow marine mammals, including pinnipeds, to forage at deep levels whilst avoiding severe negatives such as hypoxia. I found genes under selection in all lineages that have putative functions in vasodilation, blood coagulation and antioxidation. I found many genes from my hypoxia-related candidate gene set that have also been found under positive selection in other marine mammal comparative analyses, suggesting possible convergent evolution between the two clades in relation to diving physiology (Foote et al., 2015; Zhou et al., 2015; Chikina et al., 2016; Yuan et al., 2021).

The purpose of my analyses was to try and find genes that may contribute to the variation in lactation strategies across Phocidae and Otariidae, but not produce an exhaustive list of genes. There are many genes that I expected to find, such as LEPTIN and LALBA not under selection. In addition, phenotypes of interest, such as fasting, were found to be lacking in my analysis. My analysis has highlighted a breadth of genes that are under putative positive selection across Pinnipedia, many which have a possible involvement in lactation strategies.

Caveats in selective pressure variation analyses

In this analysis, I have used genes with previous evidence of roles in lactation, or lactation related traits, inferring that the selective pressure is a result of adaptation to lactation related traits. It is feasible that, despite a known role in lactation, the positive selection in this gene could be due to other adaptations. Further differentiation between the events of adaptation and the selection in the genes would be achievable though an ancestral reconstruction of the lactation states in pinnipeds. By comparing the ancestral states of old pinnipeds with the evolutionary timings of variation within these genes, it would be possible to assess if the variations under positive selection occurred at a time when the phenotypic properties also evolved.

Results of large-scale genomic analyses are dependent on the quality of the data and the fit of the models used to fit such data. The annotation of genomes in eukaryotic species is a complex problem, which is reliant on the quality of the underpinning genome assembly. Ensuring the gene models are supported by cutting edge RNA sequencing methods and high-quality DNA sequencing is paramount to attaining a complete and error free genome assembly and annotation (Rhie et al., 2021). Genomes assemblies for previously unsequenced species are being produced at a rapid rate, meaning that comprehensive

annotation processes involving manual gene curation are not always feasible with time and monetary constraints (Salzberg, 2019). A practical alternative it to use quicker computational methods, although these can be prone to errors, thus it is important to identify possible errors during analysis. Common problems causing annotation errors include incorrect assignment of exon and gene boundaries, fragmentation or fusion of gene models, and missing gene models (Meyer et al., 2020). The rate of these errors is often influenced by the methodology used to produce the annotations (Scalzitti et al., 2020) and thus, the impact of errors is reduced when all annotations in comparative analysis are produced using the same methodology. However, this can be difficult to ensure when comparing newly assembled genomes or genomes that have not been publicly released (Eddy et al., 2022).

Errors in annotations become prominent when the gene models from different species are aligned in a MSA. It has been shown that errors introduced from misalignments can produce false positives in branch-site test models of selection, and false negatives can be introduced through misalignments from highly divergent sequences (Fletcher and Yang, 2010; Jordan and Goldman, 2012). Due to their importance in the preparation of the data, the accuracy of alignment tools has been compared continuously, with different aligners performing better with different sequences (Pais et al., 2014; Thompson et al., 1999). Thus, multiple alignment tools should be compared to attain the optimal alignment for sequences between different species (Anisimova and Liberies, 2010). In this analysis, I used an optimised version of *AQUA* (Muller et al., 2010), using four different alignment approaches: *MAFFT* (Katoh and Standley, 2013), *MUSCLE* (Edger, 2004), *T-COFFEE* (Notredame et al., 2000), and *Clustal-Omega* (Sievers et al., 2011). The outputs of these alignment tools were refined using *RASCAL* (Thompson et al., 2003) software, before all non-refined and refined outputs were compared using the *NORMD* (Thompson et al., 2001) assessment tool. This method sought to alleviate any errors in the data caused by alignment tools. In addition, I used a filter to visually inspect alignments that showed signatures of false positives, with previous studies suggesting that alignment errors can produced a high number of sites under selection being called in small area of sequence space (Potter et al., 2021; Zou et al., 2021). My filter listed the alignments containing greater than 5 BEB sites across the alignment, which then were visually checked for alignment issues, assessing whether the regions around the site under selection were conserved. Another issue that had been seen when first visually checking the alignments was poorly aligned regions at the start and end of the alignments, this a product of putatively missing exon boundaries (Guigó et al., 2000) or retention of noncoding sequences in coding regions (Drăgan et al., 2016). As these occurred at the start and end of the alignments all MSAs with BEB sites occurring in the first and last 10% of the sequence were also subjected

to visual alignment assessment. I also only retained BEB sites for which all residues were identical between all species within the foreground lineage, this would help to improve confidence of a site by reducing the possibility of an assembly error contributing to a false positive (Mallick et al., 2009).

Alignments can be more prone to errors when the divergence between species is high (Fletcher and Yang, 2010), for this reason I my species was chosen in a phylogenetic informative manner, attempting to reduce divergence times between selected species, and so attempting to limit divergence between sequences. It has been shown in areas which have high rates non-synonymous substitutions can be susceptible to false positives, to circumvent this possibility my final set of genes, that were putatively involved in lactation associated functions and under positive selection, were subjected to a synonymous substitution analysis. Genes were removed from the final set if they displayed a relatively high rate of synonymous substitutions in the region of the positively selected site.

When performing a selective pressure analysis in CodeML (Yang, 2007) some omega values are fixed, i.e., $\omega_0 = 0$ in ModelA, whereas other omega values are variable and estimated from the data. In this analysis I use Vespasian (Constantinides et al., 2021) to generate the configuration files for the CodeML analysis. Vespasian creates several starting values for omega, this reduces the risk of reporting the estimated omega values from areas of local minima on the likelihood plane.

<u>False positive detection in selective pressure analyses</u>

Correcting for multiple testing is an important issue, especially in the field of genomics where investigators will test thousands to millions of hypotheses in single analysis. These methods, such as Bonferroni corrections (Bonferroni, 1936) and the Benjamini and Hochberg step up procedure (Benjamini and Hochberg, 1995), are implemented to reduce the probability of a Type I error being reported (Goeman and Solari, 2014). Tests performed in absence of biological hypotheses, e.g., thousands of genes from a genome scanned for positive selection, or testing using all branches of a phylogeny iteratively as foreground, will raise the probability of falsely rejecting a hypothesis (Anisimova and Yang, 2007). Although in the presence of a biological hypothesis, e.g., testing for positive selection in a single foreground branch, would not require corrections for multiple testing as there is an expectation for positive results, and the power of a LRT being sufficient (Anisimova and Yang, 2007; Zhang et al., 2005). Recent studies using branch-site models for single, or many branches, have often blindly implemented

corrections for multiple testing, usually through Benjamini and Holberg (BH) false discovery rates (FDR) (Derious et al., 2021; Foote et al., 2015; Huang et al., 2021; McGowen et al., 2020). Recently it has been shown that the use of these tests could possibly be inappropriate due to the underlying assumptions of the models being violated and that implementing the tests could be too conservative and would be discarding true positives (Potter et al., 2021). I tested my p-values to the same extent of Potter et al. (2021) and found similar results: an excess of p-values equal to one and a non-uniform distribution of p-values less than 1. I instead used additional filtering methods to reduce the possibility of false positives: requiring a p-value from the LRT to be greater than 0.05, a BEB site posterior probability of > 0.8, and visual assessment of alignments.

4.5 Conclusion

In this chapter, I reveal genes that have been subject to positive selection in different pinniped lineages. I attempt to link these genes with physiological and phenotypic adaptations, attempting to find connections from the genes to the unique lactation strategy of the lineages. I find genes that may contribute to the unique lactation strategies seen in pinnipeds, such as mammary gland remodelling genes in Otariid species which can supress lactation involution; many lipid mobilising, synthesising, and lipolysis genes in all species, including the high energy milk producing Phocidae. I do not provide clear mechanisms for these genes but future analyses exploring these genes in more detail different lineages may explain what genes have driven the extreme lactation associated phenotypes in pinnipeds.

Chapter 5: General discussion

Molecular evolution has progressed from a predominantly theoretical field to an empirical one due to the advancement of experimental molecular genetics. Decreasing costs and technological advancements have reduced the difficulty of generating genome-scale data for non-model organisms, transforming the molecular evolutionary ecology field (Liberles, et al., 2020). In this thesis, I took advantage of recently developed methods and techniques to further my understanding of the molecular evolutionary dynamics in Pinnipedia. This semi-aquatic clade is one of the few mammal lineages that have successfully recolonised a marine environment, evolving a suite of adaptations that have allowed them to be thrive in a marine environment whilst retaining a dependency on solid substrates for birth and lactation. For this study, I have focused on the unique traits that have arisen as consequence of overcoming spatial-temporal separation of foraging and breeding, with reference to the distinct lactation strategies that have evolved between families. The unique traits observed between species have the potential to be driven through changes in protein coding regions. Conducting genomic comparative analyses across pinnipeds offers a fantastic opportunity to investigate the role of selective pressure in the development of novel traits involved in lactation.

The capacity to generate novel genomic data for pinniped species is what facilitated this study. When this study was initiated, there were only 5 annotated pinniped genomes that were publicly available. In Chapter 2 I sought to expand this with two extra species Caspian seal and Hooded seal. I chose these species as they offered insight to either phylogenetic ambiguity, conservational relevance, or extreme lactation related adaptations. Having access to state-of-the-art long read sequencing technology allowed us to try and produce a less fragmented assembly than what was currently available, such as the Walrus, Hawaiian monk seal and Weddell seal assemblies, whist remaining at moderate sequencing depth and cost. I was able to produce assemblies of excellent contiguity and completeness, with gene content assessment tools finding my assemblies being close or superior to other consortium produced Carnivora assemblies. My assemblies were still quite fragmented, in comparison to some of the assemblies that made use of long-range sequencing technologies, although my assemblies had no ambiguous bases. It would be advantageous in the future to apply long range sequencing to my assemblies to further decrease the contiguity, and allow further investigation into chromosomal rearrangements and synteny, for which pinnipeds are an interesting clade due to their variability in chromosomal number within and between families (Beklemisheva et al., 2016; Beklemisheva et al., 2020).

An important facet of Chapter 2 was the annotation of the produced assemblies, which would be integral for my investigations in Chapters 3 and 4. I initially used pre-developed pipelines to produce my gene models, but these resulted in unexpectedly high gene number counts with low BUSCO scores. Instead, I developed an in-house pipeline based on that of recent studies (Jebb et al., 2020), this produced gene counts in line with other publicly available pinniped annotations. The quality of the gene models produced is lower than what is produced using some of the advanced pipelines produced by larger organisations such as Ensembl annotation (Aken et al., 2016) or NCBI eukaryote annotation (Thibaud-Nissen et al., 2016), but it is only with public release of data that these pipelines could be utilised. Once these assemblies are in the public domain it would be beneficial to submit these for annotation using a more advanced pipeline.

From my gene models produced in Chapter 2 I was able to apply these in a comparative phylogenetic reconstruction analysis in Chapter 3. To date, some of phylogenetic relationships within Pinnipedia have been contentious, with the placement of some of species, especially within the Phoca*/Pusa* clade changing as data is increased. I used strict filters based on assess homologous relationships to create my data reducing paralogy. I investigated evolutionary relationships using three different phylogenetic techniques with varying underlying principles, for which I was able to confidently produce a congruent phylogeny for Pinnipedia. I found that a small number of genes were disproportionately influencing the phylogeny and were creating discordance between phylogenies produced through Bayesian inference and Maximum Likelihood methods. I found that the vastly increased phylogenetic signal in some genes was a result of erroneous parts of gene models, rather than introgression or incomplete lineage sorting which had caused similar effects in a previous analysis of Otariids (Lopes et al., 2021) and other marine mammals (Árnason et al., 2018). From my data it appeared differing annotation methods possibly contributed to increased similarity between distantly related species. My results from this chapter highlight the importance of data quality and filtering in phylogenetic analyses such as these, where few differences exist between coding regions of recently diverged species.

Through my resulting phylogeny I confidently estimated that pinnipeds are the sister taxon to Mustelidae. I estimate that a Mustelidae/Pinnipedia lineage diverged from Ursidae, approximately 41.5, before Pinnipedia lineage diverged from Mustelidae approximately 39 Mya. I also conclusively demonstrated that the Grey seal is sister to the Caspian seal, following its contentious placements variously within the Pusa clade (Nyakatura and Binida-Emonds, 2012); as sister to the Pusa clade (Fulton and Strobeck, 2010; Fulton

and Strobeck, 2011); and as sister to Pusa/Phoca clade (Burns and Fay, 1970; de Muizon, 1982; Perry et al., 1995; Binida-Emonds et al., 1995). This placement of Grey seals conflicts with the taxonomical naming conventions currently in place for Grey seal and raises questions about whether the Grey seal re-evolved a larger body size or did multiple lineages independently evolve smaller body sizes. This analysis would have greatly benefitted from the inclusion Baikal seal, which at the time of analysis was not publicly available. Therefore, it would be of great interest if I was to repeat my phylogenetic analysis pipeline, after performing the annotation of the recently released Baikal seal assembly (Yuan et al.,2021).

In Chapter 4, I then investigated signatures of positive selection that are present in pinnipeds. I found abundant evidence of positive selection in Phocidae and Otariidae, in addition to the whole of Pinnipedia. I sought to identify genes with a putative role in the key features of lactation strategies that exist between different clades, such as fasting, lipid metabolism, lactation physiology, and milk composition.

I found hundreds of genes under selection from each clade, with many genes with putative functions in lipid metabolism and mobilisation, reflecting the expected importance of developing and subsequent utilisation of a thick blubber layer ubiquitous in Pinnipedia. Many genes with related functions to the synthesis and transport of triglycerol were found in the overall pinniped lineage in addition to additional candidates in only the Phocidae lineage, highlighting the central role of storage and subsequent mobilisation of lipids in the capital breeding species. I also identified two genes that may influence key lipolysis regulatory pathways, with evidence suggesting their paralogues play a role in the milk fat levels in Phocidae (Fowler et al., 2013; Fowler et al., 2015). I find a high number of genes that share a putative function in increasing milk lipid levels with cattle also have signatures of positive selection unique to the Hooded seal, the species with the highest milk fat content across Mammalia (Oftedal et al., 1988). Thus, suggesting convergence between human driven selective breeding for traits in agricultural species and natural selection for traits related to life history strategies in wild taxa**.** For income breeding Otariids I found that modifications in genes that have known functions in mammary gland physiology, including gland remodelling, possibly playing a role in the mammary gland involution prevention that is unique to Otariidae (Sharp et al., 2005). Finally, I also investigated any genes that could be associated with deep diving, finding genes with putative roles in hypoxia. Some of these hypoxia-related genes were also found in studies investigating hypoxia prevention in other marine mammals. Again, identifying addition possible avenues for convergent evolution the molecular level (Foote et al., 2015; Zhou et al., 2015; Chikina et al.,

2016; Yuan et al., 2021). In summary, I have identified many candidate pathways that could be subject to positive selection in pinnipeds and could contribute to trait that define their diverse lactation strategies.

By only using coding regions of the genome, this analysis only makes use of a small portion of available features of which selection can act. Non-coding regions including regulatory regions, transposable elements have been shown to have impacts on fitness and adaptive features in various organisms (McLean et al., 2011; Mack et al., 2018; Lewis and Reed, 2019). Within Pinnipedia, it has been observed that novel regulatory microRNAs could have possible roles in the adaptations to diving and preventing hypoxia at deep depths (Penso-Dolfin et al., 2020). Improving the microRNA catalogue with representatives of Pinnipedia (Kozomara et al., 2019) would improve links between microRNA sequences and functions. In addition to this conducting more analyses that make use of non-coding regions, i.e., GWAS, with Pinnipedia models could have a great impact on the understanding of non-coding factors that contribute to phenotypic variation within Pinnipedia.

It would be imperative to further analyse the candidates, mapping the residues under positive selection on to predicted 3D proteins and performing experimental residue replacement analyses. It would be advantageous in future studies to not only increase the sampling of available pinniped species but also increase the number of individuals from a representative species, helping to determine whether molecular adaptations are fixed in populations. Given more time, I would have liked to have evaluated patterns of milk expression from different species of pinniped which differing lactation strategies. Comparing the profiles of expression in the candidate genes I have identified could help verify that the selection pressure experienced at a site has had a significant impact on the lactation-related phenotypes seen across pinnipeds.

Additional work could be done in terms of achieving a comprehensive understanding of lactation within pinnipeds. It would be important to perform whole genome sequencing of the remaining pinnipeds without a representative genome assembly and establishing a co-ordinated effort within a pinniped consortium would increase efficiency. There has been some *in vitro* work using transcriptomic analyses that help investigate different physiological traits in pinnipeds (Khudyakov et al., 2015; Martinez et al., 2018). Producing cell lines of a pinniped species would increase our knowledge on the mechanisms of genes specific to pinnipeds. Although, producing pinniped-derived organoids would hugely increase our

knowledge of cell-cell interactions within tissues and would bypass the limitations of needing a captive population of a pinniped species, which would be impossible due to the size and range of the species.

In this thesis I set out to identify molecular underpinnings that have in contributed in the to the different lactation strategies that are present across Pinnipedia. I explored the patterns that are important in evolutionary biology, in which adaptive molecular processes can occur and contribute to complex phenotypic phenotypes. With the increase in available genetic data of non-model species it is increasingly possible to use wild populations to test hypotheses of genotype-phenotype associations and underlying mechanisms of phenotypes. But it would be an avenue of great interest to understand the evolutionary history of phenotypic traits associated with lactation, comparing the timings of the emergence of such phenotypic traits with the ancestral sequences of candidate genes found in this analysis. This would help resolve the critical genes that help bring about diverse changes to lactation strategies. With reference to pinnipeds, I show that they are excellent models for investigating the evolution of novel adaptations their drivers in mammals.

There were several constraints in time or current knowledge that, as a result certain analyses could not be performed. It is of great disappointment that the genome assembly of the Baikal seal was not available at the initial stages of the phylogenetic reconstruction. The inclusion of this species would have greatly improved the resolution understanding the evolution of the *Pusa* clade. It would have been beneficial to also have investigated alternative methods of dating the timings of speciation events in Chapter 3. In addition, the use of previously published estimates of calibration times and use of familiar tools to investigate the divergence dates may have improved my analysis. Using fossil calibrations and established tools such as BEAST (Bouckaert et al., 2019), could possibly have improved the confidence of my results. One additional analysis, that would also have complemented Chapter 4 would have been a gene family expansion analysis, combining genes under selection and genes that have undergone reductions or expansions.

References

Abdellah, Z. et al., 2004. Finishing the euchromatic sequence of the human genome. *Nature*. [Online]. **431**(7011), pp.931–945. Available from: https://doi.org/10.1038/nature03001.

Adachi, T., Maresh, J.L., Robinson, P.W., Peterson, S.H., Costa, D.P., Naito, Y., Watanabe, Y.Y. and Takahashi, A. 2014. The foraging benefits of being fat in a highly migratory marine mammal. *Proceedings of the Royal Society B: Biological Sciences*. **281**(1797), p.20142120.

Adachi, T., Takahashi, A., Costa, D.P., Robinson, P.W., Hückstädt, L.A., Peterson, S.H., Holser, R.R., Beltran, R.S., Keates, T.R. and Naito, Y. 2021. Forced into an ecological corner: Round-the-clock deep foraging on small prey by elephant seals. *Science Advances*. [Online]. **7**(20), p.eabg3628. Available from: https://doi.org/10.1126/sciadv.abg3628.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F.J., Murphy, D.N., Nag, R., Ruffier, M., Schuster, M., Tang, Y.A., Vogel, J.H., White, S., Zadissa, A., Flicek, P. and Searle, S.M.J. 2016. The Ensembl gene annotation system. *Database : the journal of biological databases and curation*. **2016**.

Albouy, C., Delattre, V., Donati, G., Frölicher, T.L., Albouy-Boyer, S., Rufino, M., Pellissier, L., Mouillot, D. and Leprieur, F. 2020. Global vulnerability of marine mammals to global warming. *Scientific Reports*. [Online]. **10**(1), p.548. Available from: https://doi.org/10.1038/s41598-019-57280-3.

Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M. and Gerstein, M.B. 2010. Annotating non-coding regions of the genome. *Nature Reviews Genetics*. [Online]. **11**(8), pp.559–571. Available from: https://doi.org/10.1038/nrg2814.

Allen, K.N. and Vázquez-Medina, J.P. 2019. Natural Tolerance to Ischemia and Hypoxemia in Diving Mammals: A Review. *Frontiers in Physiology*. **10**, p.1199.

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. and Gouil, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*. [Online]. **21**(1), p.30. Available from: https://doi.org/10.1186/s13059-020-1935-5.

Andrews, S. and others 2010. FastQC: a quality control tool for high throughput sequence data. 2010. *Https://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*., http://www.bioinformatics.babraham.ac.uk/projects/.

Anisimova, M., Bielawski, J.P. and Yang, Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution*. [Online]. **19**(6), pp.950–958. Available from: https://doi.org/10.1093/oxfordjournals.molbev.a004152.

Anisimova, M., Bielawski, J.P. and Yang, Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution*. [Online]. **18**(8), pp.1585–1592. Available from: http://dx.doi.org/10.1093/oxfordjournals.molbev.a003945.

Anisimova, M. and Yang, Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Molecular Biology and Evolution*. [Online]. **24**(5), pp.1219–1228. Available from: https://doi.org/10.1093/molbev/msm042.

Arcila, D., Ortí, G., Vari, R., Armbruster, J.W., Stiassny, M.L.J., Ko, K.D., Sabaj, M.H., Lundberg, J., Revell, L.J. and Betancur, R.R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology and Evolution*. **1**(2), pp.1–10.

Arenas, M. 2015. Trends in substitution models of molecular evolution. *Frontiers in Genetics*. [Online]. **6**(OCT), p.319. Available from: https://www.frontiersin.org/article/10.3389/fgene.2015.00319.

ÁRNASON, Ú. 1974. Comparative chromosome studies in Cetacea. *Hereditas*. **77**(1), pp.1–36.

Arnason, U., Gullberg, A., Janke, A., Kullberg, M., Lehman, N., Petrov, E.A. and Väinölä, R. 2006. . Pinniped phylogeny and a new hypothesis for their origin and dispersal. *Molecular Phylogenetics and Evolution*. [Online]. **41**(2), pp.345–354. Available from: http://dx.doi.org/10.1016/j.ympev.2006.05.022.

Arora, R., Sharma, A., Sharma, U., Girdhar, Y., Kaur, M., Kapoor, P., Ahlawat, S. and Vijh, R.K. 2019. Buffalo milk transcriptome: A comparative analysis of early, mid and late lactation. *Scientific Reports*. [Online]. **9**(1), p.5993. Available from: https://pubmed.ncbi.nlm.nih.gov/30979954.

Arriola, A., Biuw, M., Walton, M., Moss, S. and Pomeroy, P. 2013. Selective blubber fatty acid mobilization in lactating gray seals (Halichoerus grypus). *Physiological and Biochemical Zoology*. [Online]. **86**(4), pp.441–450. Available from: http://www.jstor.org/stable/10.1086/671446.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. 2000. Gene ontology: Tool for the unification of biology. *Nature Genetics*. **25**(1), pp.25–29.

Atashi, H., Salavati, M., De Koster, J., Ehrlich, J., Crowe, M., Opsomer, G., Hostens, M., McLoughlin, N., Fahey, A., Matthews, E., Santoro, A., Byrne, C., Rudd, P., O'Flaherty, R.,

Hallinan, S., Wathes, C., Cheng, Z., Fouladi, A., Pollott, G., Werling, D., Bernardo, B.S., Wylie, A., Bell, M., Vaneetvelde, M., Hermans, K., Moerman, S., Bogaert, H., Vandepitte, J., Vandevelde, L., Vanranst, B., Hoglund, J., Dahl, S., Ostergaard, S., Rothmann, J., Krogh, M., Meyer, E., Gaillard, C., Ettema, J., Rousing, T., Signorelli, F., Napolitano, F., Moioli, B., Crisà, A., Buttazzoni, L., McClure, J., Matthews, D., Kearney, F., Cromie, A., McClure, M., Zhang, S., Chen, X., Chen, H., Zhao, J., Yang, L., Hua, G., Tan, C., Wang, G., Bonneau, M., Pompozzi, A., Pearn, A., Evertson, A., Kosten, L., Fogh, A., Andersen, T., Lucey, M., Elsik, C., Conant, G., Taylor, J., Gengler, N., Georges, M., Colinet, F., Pamplona, M.R., Hammami, H., Bastin, C., Takeda, H., Laine, A., Van Laere, A.S., Schulze, M., Vera, S.P., Ferris, C. and Marchitelli, C. 2020. Genome-wide association for milk production and lactation curve parameters in Holstein dairy cows. *Journal of Animal Breeding and Genetics*. **137**(3), pp.292–304.

Bai, J., Xia, M., Xue, Y., Ma, F., Cui, A., Sun, Y., Han, Y., Xu, X., Zhang, F., Hu, Z., Liu, Z., Liu, Y., Cai, G., Su, W., Sun, X., Wu, H., Yan, H., Chang, X., Hu, X., Bian, H., Xia, P., Gao, J., Li, Y. and Gao, X. 2020. Thrombospondin 1 improves hepatic steatosis in diet-induced insulin-resistant mice and is associated with hepatic fat content in humans. *EBioMedicine*. [Online]. **57**. Available from: https://doi.org/10.1016/j.ebiom.2020.102849.

Baldwin, M.W., Toda, Y., Nakagita, T., O'Connell, M.J., Klasing, K.C., Misaka, T., Edwards, S. V. and Liberles, S.D. 2014. Evolution of sweet taste perception in hummingbirds by transformation of the ancestral umami receptor. *Science*. [Online]. **345**(6199), pp.929–933. Available from: https://doi.org/10.1126/science.1255097.

Balia, F., Pazzola, M., Dettori, M.L., Mura, M.C., Luridiana, S., Carcangiu, V., Piras, G. and Vacca, G.M. 2013. Effect of CSN1S1 gene polymorphism and stage of lactation on milk yield and composition of extensively reared goats. *Journal of Dairy Research*. **80**(2), pp.129–137.

Bao, W., Kojima, K.K. and Kohany, O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*. **6**(1), p.11.

Barnes, L.G., Ray, C.E. and Koretsky, I.A. 2006. Mesozoic and Cenozoic Vertebrates and Paleoenvironments: Tributes to the Career of Prof. Dan Grigorescu.

Barnes, L.G. 1989. A new enaliarctine pinniped from the Astoria Formation, Oregon, and a classification of the Otariidae (Mammalia: Carnivora). *Contributions in science*. **403**, pp.1–26.

Barrón-Ortiz, C.I., Avilla, L.S., Jass, C.N., Bravo-Cuevas, V.M., Machado, H. and Mothé, D. 2019. What is Equus? Reconciling taxonomy and phylogenetic analyses. *Frontiers in Ecology and Evolution*. [Online]. **7**(SEP), p.343. Available from: https://www.frontiersin.org/article/10.3389/fevo.2019.00343.

Barton, N.H. 2000. Genetic hitchhiking. *Philosophical Transactions of the Royal Society B: Biological Sciences*. **355**(1403), pp.1553–1562.

Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., da Silva, A., Denny, P., Dogan, T., Ebenezer, T.G., Fan, J., Castro, L.G., Garmiri, P., Georghiou, G., Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Jokinen, P., Joshi, V., Jyothi, D., Lock, A., Lopez, R., Luciani, A., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Menchi, M., Mishra, A., Moulang, K., Nightingale, A., Oliveira, C.S., Pundir, S., Qi, G., Raj, S., Rice, D., Lopez, M.R., Saidi, R., Sampson, J., Sawford, T., Speretta, E., Turner, E., Tyagi, N., Vasudev, P., Volynkin, V., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.C., Bolleman, J., Boutet, E., Breuza, L., Casals-Casas, C., de Castro, E., Echioukh, K.C., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Keller, G., Kerhornou, A., Lara, V., Le Mercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T.B., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Pozzato, M., Pruess, M., Rivoire, C., Sigrist, C., Sonesson, K., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Garavelli, J.S., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.S., Zhang, J., Ruch, P. and Teodoro, D. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. [Online]. **49**(D1), pp.D480–D489. Available from: https://doi.org/10.1093/nar/gkaa1100.

Beklemisheva, V.R., Perelman, P.L., Lemskaya, N.A., Kulemzina, A.I., Proskuryakova, A.A., Burkanov, V.N. and Graphodatsky, A.S. 2016. The ancestral carnivore karyotype as substantiated by comparative chromosome painting of three pinnipeds, the walrus, the steller sea lion and the Baikal seal (Pinnipedia, Carnivora). *PLoS ONE*. **11**(1), p.e0147647.

Beklemisheva, V.R., Perelman, P.L., Lemskaya, N.A., Proskuryakova, A.A., Serdyukova, N.A., Burkanov, V.N., Gorshunov, M.B., Ryder, O., Thompson, M., Lento, G., O'brien, S.J. and Graphodatsky, A.S. 2020. Karyotype evolution in 10 pinniped species: Variability of heterochromatin versus high conservatism of euchromatin as revealed by comparative molecular cytogenetics. *Genes*. **11**(12), pp.1–26.

Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'Ang, L.Y., Huang, W., Liu, B., Shen, Y., Tam, P.K.H., Tsui, L.C., Waye, M.M.Y., Wong, J.T.F., Zeng, C., Zhang, Q., Chee, M.S., Galver, L.M., Kruglyak, S., Murray, S.S., Oliphant, A.R., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Phillips, M.S., Verner, A., Duan, S., Lind, D.L., Miller, R.D., Rice, J., Saccone, N.L., Taillon-Miller, P., Xiao, M., Sekine, A., Sorimachi, K., Tanaka, Y., Tsunoda, T., Yoshino, E., Bentley, D.R., Hunt, S., Powell, D., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D.R., Suda, E., Rotimi, C., Adebamowo, C.A., Aniagwu, T., Marshall, P.A., Matthew, O., Nkwodimmah, C., Royal, C.D.M., Leppert, M.F., Dixon, M., Cunningham, F., Kanani, A., Thorisson, G.A., Chen, P.E., Cutler, D.J., Kashuk, C.S., Donnelly, P., Marchini, J., McVean, G.A.T., Myers, S.R., Cardon, L.R., Morris, A., Weir, B.S., Mullikin, J.C., Feolo, M., Daly, M.J., Qiu, R., Kent, A., Dunston, G.M., Kato, K., Niikawa, N., Watkin, J., Gibbs, R.A., Sodergren, E., Weinstock, G.M.,

Wilson, R.K., Fulton, L.L., Rogers, J., Birren, B.W., Han, H., Wang, H., Godbout, M., Wallenburg, J.C., L'Archevêque, P., Bellemare, G., Todani, K., Fujita, T., Tanaka, S., Holden, A.L., Collins, F.S., Brooks, L.D., McEwen, J.E., Guyer, M.S., Jordan, E., Peterson, J.L., Spiegel, J., Sung, L.M., Zacharia, L.F., Kennedy, K., Dunn, M.G., Seabrook, R., Shillito, M., Skene, B., Stewart, J.G., Valle, D.L., Clayton, E.W., Jorde, L.B., Chakravarti, A., Cho, M.K., Duster, T., Foster, M.W., Jasperse, M., Knoppers, B.M., Kwok, P.Y., Licinio, J., Long, J.C., Ossorio, P., Wang, V.O., Rotimi, C.N., Spallone, P., Terry, S.F., Lander, E.S., Lai, E.H., Nickerson, D.A., Abecasis, G.R., Altshuler, D., Boehnke, M., Deloukas, P., Douglas, J.A., Gabriel, S.B., Hudson, R.R., Hudson, T.J., Kruglyak, L., Nakamura, Y., Nussbaum, R.L., Schaffner, S.F., Sherry, S.T., Stein, L.D. and Tanaka, T. 2003. The international HapMap project. *Nature*. [Online]. **426**(6968), pp.789–796. Available from: https://doi.org/10.1038/nature02168.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. **57**(1), pp.289–300.

Berta, A. and Demere, T.A. 1986. Callorhinus gilmorei n.sp., (Carnivora: Otariidae) from the San Diego Formation (Blancan) and its implications for otariid phylogeny. *In*: *Transactions - San Diego Society of Natural History*. San Diego Society of Natural History, pp.111–126.

Berta, A. and Wyss, A.R. 1994. . Pinniped phylogeny. *Proceedings - San Diego Society of Natural History*. [Online]. **29**, pp.33–56. Available from: All Papers/B/Berta and Wyss 1994 - pinnipedPhylogeny.pdf.

Berta, A. 1991. New Enaliarctos* (Pinnipedimorpha) from the Oligocene and Miocene of Oregon and the Role of 'Enaliarctids' in pinnipedPhylogeny. *Smithsonian Contributions to Paleobiology*. (69), pp.1–33.

Berta, A. and Churchill, M. 2012. . Pinniped taxonomy: Review of currently recognized species and subspecies, and evidence used for their description. *Mammal Review*. [Online]. **42**(3), pp.207–234. Available from: http://dx.doi.org/10.1111/j.1365-2907.2011.00193.x.

Berta, A., Sumich, J.L. and Kovacs, K.M. 2005. *Marine Mammals: Evolutionary Biology* [Online]. Elsevier. Available from: https://market.android.com/details?id=book-4sWbuL0hM1kC.

Bhat, S.A., Ahmad, S.M., Ibeagha-Awemu, E.M., Bhat, B.A., Dar, M.A., Mumtaz, P.T., Shah, R.A. and Ganai, N.A. 2019. Comparative transcriptome analysis of mammary epithelial cells at different stages of lactation reveals wide differences in gene expression and pathways regulating milk synthesis between Jersey and Kashmiri cattle. *PLoS ONE*. [Online]. **14**(2), pp.e0211773–e0211773. Available from: https://pubmed.ncbi.nlm.nih.gov/30721247.

BININDA-EMONDS, O.R.P., GITTLEMAN, J.L. and PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora

(Mammalia). *Biological Reviews of the Cambridge Philosophical Society*. **74**(2), pp.143–175.

Bionaz, M. and Loor, J.J. 2008. Gene networks driving bovine milk fat synthesis during the lactation cycle. *BMC Genomics*. [Online]. **9**, p.366. Available from: http://dx.doi.org/10.1186/1471-2164-9-366.

Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G.A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D.H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A. and Finn, R.D. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*. [Online]. **49**(D1), pp.D344–D354. Available from: https://doi.org/10.1093/nar/gkaa977.

Boessenecker, R.W. and Churchill, M. 2015. The oldest known fur seal. *Biology Letters*. **11**(2), p.20140835.

Bonner, W.N. 1984. Lactation strategies in pinnipeds: Problems for a marine mammalian group. *Physiological Strategies in Lactation*. **51**, pp.253–272.

Booker, T.R., Jackson, B.C. and Keightley, P.D. 2017. Detecting positive selection in the genome. *BMC Biology*. [Online]. **15**(1), p.98. Available from: http://dx.doi.org/10.1186/s12915-017-0434-y.

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F.K., Müller, N.F., Ogilvie, H.A., Du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M.A., Wu, C.H., Xie, D., Zhang, C., Stadler, T. and Drummond, A.J. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*. [Online]. **15**(4), p.e1006650. Available from: https://doi.org/10.1371/journal.pcbi.1006650.

Bowen, W.D., Oftedal, O.T. and Boness, D.J. 1985. Birth to weaning in 4 days: remarkable growth in the Hooded seal, Cystophora cristata. *Canadian Journal of Zoology*. [Online]. **63**(12), pp.2841–2846. Available from: https://doi.org/10.1139/z85-424.

Brandley, M.C., Warren, D.L., Leaché, A.D. and McGuire, J.A. 2009. Homoplasy and clade support. *Systematic Biology*. [Online]. **58**(2), pp.184–198. Available from: https://doi.org/10.1093/sysbio/syp019.

Brozek, J. 1962. The Fire of Life. An Introduction to Animal Energetics. Max Kleiber . *The Quarterly Review of Biology*. **37**(1), pp.55–55.

Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M. and Borodovsky, M. 2021. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported

by a protein database. *NAR Genomics and Bioinformatics*. [Online]. **3**(1), pp.1–11. Available from: https://doi.org/10.1093/nargab/lqaa108.

Buchfink, B., Xie, C. and Huson, D.H. 2014. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*. **12**(1), pp.59–60.

Buchfink, B., Xie, C. and Huson, D.H. 2014. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*. [Online]. **12**(1), pp.59–60. Available from: https://doi.org/10.1038/nmeth.3176.

Buchholz, R. 2000. *the Handicap Principle: a Missing Piece of Darwin'S Puzzle*. Oxford University Press.

Burnham, K.P. and Efron, B. 1983. *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM.

Burns, J.J. and Fay, F.H. 1970. Comparative morphology of the skull of the Ribbon seal, Histriophoca fasciata, with remarks on systematics of Phocidae. *Journal of Zoology*. [Online]. **161**(3), pp.363–394. Available from: https://doi.org/10.1111/j.1469-7998.1970.tb04519.x.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. [Online]. **25**(15), pp.1972–1973. Available from: https://pubmed.ncbi.nlm.nih.gov/19505945.

Casillas, S. and Barbadilla, A. 2017. *Molecular population genetics*. Oxford University Press.

Casillas, S. and Barbadilla, A. 2017. Molecular population genetics. *Genetics*. [Online]. **205**(3), pp.1003–1035. Available from: http://dx.doi.org/10.1534/genetics.116.196493.

Challis, R., Richards, E., Rajan, J., Cochrane, G. and Blaxter, M. 2020. BlobToolKit - interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*. **10**(4), pp.1361–1374.

Chapskii, K.K. 1955. Contribution to the problem of the history of development of the Caspian and Baikal seals. *Trudy Zool. Inst. Akad. Nauk, SSSR*. **17**, pp.200–216.

Charlesworth, J. and Eyre-Walker, A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution*. [Online]. **25**(6), pp.1007–1015. Available from: http://dx.doi.org/10.1093/molbev/msn005.

Chatzou, M., Magis, C., Chang, J.M., Kemena, C., Bussotti, G., Erb, I. and Notredame, C. 2016. Multiple sequence alignment modeling: Methods and applications. *Briefings in Bioinformatics*. [Online]. **17**(6), pp.1009–1023. Available from: https://doi.org/10.1093/bib/bbv099.

Chen, M.Y., Liang, D. and Zhang, P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: A case study of jawed vertebrate backbone phylogeny. *Systematic Biology*. **64**(6), pp.1104–1120.

Chen, P. and Zhang, J. 2020. Antagonistic pleiotropy conceals molecular adaptations in changing environments. *Nature Ecology and Evolution*. [Online]. **4**(3), pp.461–469. Available from: https://doi.org/10.1038/s41559-020-1107-8.

Chen, Y., Nie, F., Xie, S.Q., Zheng, Y.F., Dai, Q., Bray, T., Wang, Y.X., Xing, J.F., Huang, Z.J., Wang, D.P., He, L.J., Luo, F., Wang, J.X., Liu, Y.Z. and Xiao, C. Le 2021. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications*. **12**(1), pp.1–10.

Chen, Z., Yao, Y., Ma, P., Wang, Q. and Pan, Y. 2018. Haplotype-based genome-wide association study identifies loci and candidate genes for milk yield in Holsteins. *PLoS ONE*. [Online]. **13**(2), p.e0192695. Available from: https://doi.org/10.1371/journal.pone.0192695.

Chikhi, R. and Medvedev, P. 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. [Online]. **30**(1), pp.31–37. Available from: https://doi.org/10.1093/bioinformatics/btt310.

Chikhi, R. and Medvedev, P. 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. [Online]. **30**(1), pp.31–37. Available from: https://doi.org/10.1093/bioinformatics/btt310.

Chikina, M., Robinson, J.D. and Clark, N.L. 2016. Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. *Molecular Biology and Evolution*. [Online]. **33**(9), pp.2182–2192. Available from: https://doi.org/10.1093/molbev/msw112.

Chira, A.M. and Thomas, G.H. 2016. The impact of rate heterogeneity on inference of phylogenetic models of trait evolution. *Journal of Evolutionary Biology*. [Online]. **29**(12), pp.2502–2518. Available from: https://doi.org/10.1111/jeb.12979.

Chiu, H.K., Qian, K., Ogimoto, K., Morton, G.J., Wisse, B.E., Agrawal, N., McDonald, T.O., Schwartz, M.W. and Dichek, H.L. 2010. Mice lacking hepatic lipase are lean and protected against diet-induced obesity and hepatic steatosis. *Endocrinology*. **151**(3), pp.993–1001.

Churchill, M. and Clementz, M.T. 2016. The evolution of aquatic feeding in seals: Insights from Enaliarctos (Carnivora: Pinnipedimorpha), the oldest known seal. *Journal of Evolutionary Biology*. [Online]. **29**(2), pp.319–334. Available from: http://dx.doi.org/10.1111/jeb.12783.

Churchill, M., Boessenecker, R.W. and Clementz, M.T. 2014. Colonization of the Southern Hemisphere by fur seals and sea lions (Carnivora: Otariidae) revealed by combined evidence phylogenetic and Bayesian biogeographical analysis. *Zoological Journal of the Linnean Society*. [Online]. **172**(1), pp.200–225. Available from: https://academic.oup.com/zoolinnean/article-abstract/172/1/200/3797006.

Clark, C.T., Horstmann, L. and Misarti, N. 2020. Evaluating tooth strontium and barium as indicators of weaning age in Pacific walruses. *Methods in Ecology and Evolution*. [Online]. **11**(12), pp.1626–1638. Available from: https://doi.org/10.1111/2041-210X.13482.

Cohen-Zinder, M., Seroussi, E., Larkin, D.M., Loor, J.J., Everts-Van Der Wind, A., Lee, J.H., Drackley, J.K., Band, M.R., Hernandez, A.G., Shani, M., Lewin, H.A., Weller, J.I. and Ron, M. 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research*. **15**(7), pp.936–944.

Coorg, M. 2012. Extracting DNA Using Phenol-Chloroform Reagents Needed. *Pacific Bioscience Protocol*.

Cope, E.R., Voy, B.H., Whitlock, B.K., Staton, M., Lane, T., Davitt, J. and Mulliniks, J.T. 2018. Beta-hydroxybutyrate infusion identifies acutely differentially expressed genes related to metabolism and reproduction in the hypothalamus and pituitary of castrated male sheep. *Physiological Genomics*. **50**(6), pp.468–477.

Costa, D.P., Boeuf, B.J.L., Huntley, A.C. and Ortiz, C.L. 1986. The energetics of lactation in the Northern elephant seal, Mirounga angustirostris. *Journal of Zoology*. **209**(1), pp.21–33.

Costa, D.P. 1993. The relationship between reproductive and foraging energetics and the evolution of the Pinnipedia. *Symposia of the Zoological Society of London*. [Online]. **66**(October), pp.293–314. Available from: https://www.researchgate.net/publication/228328880_The_relationship_between_reproductive_and_foraging_energetics_and_the_evolution_of_Pinnipedia.

Costa, D.P. and Maresh, J.L. 2022. *Reproductive Energetics of Phocids* [Online]. Springer International Publishing. Available from: http://dx.doi.org/10.1007/978-3-030-88923-4_8.

Couto Alves, A., Glastonbury, C.A., El-Sayed Moustafa, J.S. and Small, K.S. 2018. Fasting and time of day independently modulate circadian rhythm relevant gene expression in adipose and skin tissue. *BMC Genomics*. [Online]. **19**(1), p.659. Available from: https://pubmed.ncbi.nlm.nih.gov/30193568.

Creevey, C.J. and McInerney, J.O. 2005. Clann: Investigating phylogenetic information through supertree analyses. *Bioinformatics*. [Online]. **21**(3), pp.390–392. Available from: https://doi.org/10.1093/bioinformatics/bti020.

Crocker, D.E., Champagne, C.D., Fowler, M.A. and Houser, D.S. 2014. Adiposity and fat metabolism in lactating and fasting northern elephant seals. *Advances in Nutrition*. [Online]. **5**(1), pp.57–64. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24425723.

Crujeiras, A.B., Izquierdo, A.G., Primo, D., Milagro, F.I., Sajoux, I., Jácome, A., Fernandez-Quintela, A., Portillo, M.P., Martínez, J.A., Martinez-Olmos, M.A., de Luis, D. and Casanueva, F.F. 2021. Epigenetic landscape in blood leukocytes following ketosis and

weight loss induced by a very low calorie ketogenic diet (VLCKD) in patients with obesity. *Clinical Nutrition*. [Online]. **40**(6), pp.3959–3972. Available from: https://www.sciencedirect.com/science/article/pii/S0261561421002600.

Cummings, M.P. 2004. PHYLIP (PHYLogeny Inference Package). *Dictionary of Bioinformatics and Computational Biology*.

Czech-Damal, N.U., Geiseler, S.J., Hoff, M.L.M., Schliep, R., Ramirez, J.M., Folkow, L.P. and Burmester, T. 2014. The role of glycogen, glucose and lactate in neuronal activity during hypoxia in the Hooded seal brain. *Neuroscience*. [Online]. **275**, pp.374–383. Available from: https://www.sciencedirect.com/science/article/pii/S0306452214005090.

Dadousis, C., Pegolo, S., Rosa, G.J.M., Gianola, D., Bittante, G. and Cecchinato, A. 2017. Pathway-based genome-wide association analysis of milk coagulation properties, curd firmness, cheese yield, and curd nutrient recovery in dairy cattle. *Journal of Dairy Science*. [Online]. **100**(2), pp.1223–1231. Available from: https://doi.org/10.3168/jds.2016-11587.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G. and Durbin, R. 2011. The variant call format and VCFtools. *Bioinformatics*. [Online]. **27**(15), pp.2156–2158. Available from: http://dx.doi.org/10.1093/bioinformatics/btr330.

Darriba, D., Weiß, M. and Stamatakis, A. 2016. Prediction of missing sequences and branch lengths in phylogenomic data. *Bioinformatics*. [Online]. **32**(9), pp.1331–1337. Available from: https://doi.org/10.1093/bioinformatics/btv768.

Darwin, C. 2014. *Origin of species by means of natural selection, or the preservation of favored races in the struggle for life*. AL Burt New York.

Dasmeh, P., Serohijos, A.W.R., Kepp, K.P. and Shakhnovich, E.I. 2013. Positively Selected Sites in Cetacean Myoglobins Contribute to Protein Stability. *PLoS Computational Biology*. [Online]. **9**(3), p.e1002929. Available from: https://doi.org/10.1371/journal.pcbi.1002929.

Davies, J.L. 1958. The Pinnipedia: An Essay in Zoogeography. *Geographical Review*. **48**(4), p.474.

Davies, K.T.J., Bennett, N.C., Faulkes, C.G. and Rossiter, S.J. 2018. Limited evidence for parallel molecular adaptations associated with the subterranean niche in mammals: A comparative study of three superorders. *Molecular Biology and Evolution*. [Online]. **35**(10), pp.2544–2559. Available from: https://doi.org/10.1093/molbev/msy161.

Davis, C.S., Delisle, I., Stirling, I., Siniff, D.B. and Strobeck, C. 2004. A phylogeny of the extant Phocidae inferred from complete mitochondrial DNA coding regions. *Molecular Phylogenetics and Evolution*. [Online]. **33**(2), pp.363–377. Available from: https://www.sciencedirect.com/science/article/pii/S1055790304001940.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M. and Van Broeckhoven, C. 2018. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics*. **34**(15), pp.2666–2669.

de Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A. and Pollock, D.D. 2011. Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*. [Online]. **7**(12), p.e1002384. Available from: https://doi.org/10.1371/journal.pgen.1002384.

Debier, C., Kovacs, K.M., Lydersen, C., Mignolet, E. and Larondelle, Y. 1999. Vitamin E and vitamin A contents, fatty acid profiles, and gross composition of harp and Hooded seal milk through lactation. *Canadian Journal of Zoology*. **77**(6), pp.952–958.

Delisle, I. and Strobeck, C. 2005. A phylogeny of the Caniformia (order Carnivora) based on 12 complete protein-coding mitochondrial genes. *Molecular Phylogenetics and Evolution*. [Online]. **37**(1), pp.192–201. Available from: https://www.sciencedirect.com/science/article/pii/S1055790305001624.

Derous, D., Sahu, J., Douglas, A., Lusseau, D. and Wenzel, M. 2021. Comparative genomics of cetartiodactyla: Energy metabolism underpins the transition to an aquatic lifestyle. *Conservation Physiology*. **9**(1), p.coaa136.

Di Franco, A., Poujol, R., Baurain, D. and Philippe, H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary Biology*. [Online]. **19**(1), p.21. Available from: https://doi.org/10.1186/s12862-019-1350-2.

Di Genova, A., Buena-Atienza, E., Ossowski, S. and Sagot, M.F. 2021. Efficient hybrid de novo assembly of human genomes with WENGAN. *Nature Biotechnology*. [Online]. **39**(4), pp.422–430. Available from: https://doi.org/10.1038/s41587-020-00747-w.

Do, D.N., Schenkel, F.S., Miglior, F., Zhao, X. and Ibeagha-Awemu, E.M. 2018. Genome wide association study identifies novel potential candidate genes for bovine milk cholesterol content. *Scientific Reports*. [Online]. **8**(1), p.13239. Available from: https://doi.org/10.1038/s41598-018-31427-0.

Dong, X.C. 2019. PNPLA3—A Potential Therapeutic Target for Personalized Treatment of Chronic Liver Disease. *Frontiers in Medicine*. [Online]. **6**. Available from: https://www.frontiersin.org/article/10.3389/fmed.2019.00304.

Donohue, M.J., Costa, D.P., Goebel, E., Antonelis, G.A. and Baker, J.D. 2002. Milk intake and energy expenditure of free-ranging northern fur seal, Callorhinus ursinus, pups. *Physiological and Biochemical Zoology*. **75**(1), pp.3–18.

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science*. **284**(5423), pp.2124–2128.

Dragan, M.A., Moghul, I., Priyam, A., Bustos, C. and Wurm, Y. 2016. GeneValidator: Identify problems with protein-coding gene predictions. *Bioinformatics*. [Online]. **32**(10), pp.1559–1561. Available from: https://pubmed.ncbi.nlm.nih.gov/26787666.

Du, C., Deng, T.X., Zhou, Y., Ghanem, N. and Hua, G.H. 2020. Bioinformatics analysis of candidate genes for milk production traits in water buffalo (Bubalus bubalis). *Tropical Animal Health and Production*. **52**(1), pp.63–69.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P. and Aiden, E.L. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*. [Online]. **356**(6333), pp.92–95. Available from: http://science.sciencemag.org/content/356/6333/92.abstract.

Dudchenko, O., Shamim, M.S., Batra, S.S., Durand, N.C., Musial, N.T., Mostofa, R., Pham, M., Hilaire, B.G.S., Yao, W., Stamenova, E., Hoeger, M., Nyquist, S.K., Korchina, V., Pletch, K., Flanagan, J.P., Tomaszewicz, A., McAloose, D., Estrada, C.P., Novak, B.J., Omer, A.D., Aiden, E.L. and Nathaniel, T. 2018. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv*. [Online], p.254797. Available from: https://www.biorxiv.org/content/early/2018/01/28/254797%0Ahttps://www.biorxiv.org/content/early/2018/01/28/254797?utm_content=buffer421d5&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer.

Dunham, I. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**(7414), pp.57–74.

Duret Laurent 2008. Neutral Theory: The Null Hypothesis of Molecular Evolution | Learn Science at Scitable. *Nature Education*. [Online]. **1**(2008), pp.1–7. Available from: https://www.nature.com/scitable/topicpage/neutral-theory-the-null-hypothesis-of-molecular-839/.

Durinck, S., Spellman, P.T., Birney, E. and Huber, W. 2009. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nature Protocols*. [Online]. **4**(8), pp.1184–1191. Available from: https://doi.org/10.1038/nprot.2009.97.

Edea, Z., Dadi, H., Dessie, T. and Kim, K.S. 2019. Genomic signatures of high-altitude adaptation in Ethiopian sheep populations. *Genes and Genomics*. [Online]. **41**(8), pp.973–981. Available from: https://doi.org/10.1007/s13258-019-00820-y.

Edgar, R.C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. [Online]. **5**(1), p.113. Available from: https://doi.org/10.1186/1471-2105-5-113.

EDGE 2021. https://www.edgeofexistence.org/science/.

Efron, B., Halloran, E. and Holmes, S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America*. **93**(23), pp.13429–13434.

Emms, D.M. and Kelly, S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*. [Online]. **20**(1), p.238. Available from: https://doi.org/10.1186/s13059-019-1832-y.

Emms, D.M. and Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*. [Online]. **16**(1), p.157. Available from: https://doi.org/10.1186/s13059-015-0721-2.

Endo, Y., Kamei, K. ichiro and Inoue-Murayama, M. 2018. Genetic signatures of lipid metabolism evolution in Cetacea since the divergence from terrestrial ancestor. *Journal of Evolutionary Biology*. **31**(11), pp.1655–1665.

Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. *Trends in Ecology and Evolution*. [Online]. **21**(10), pp.569–575. Available from: http://dx.doi.org/10.1016/j.tree.2006.06.015.

Eyre-Walker, A. and Keightley, P.D. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution*. [Online]. **26**(9), pp.2097–2108. Available from: https://doi.org/10.1093/molbev/msp119.

Favilla, A.B. and Costa, D.P. 2020. Thermoregulatory Strategies of Diving Air-Breathing Marine Vertebrates: A Review. *Frontiers in Ecology and Evolution*. [Online]. **8**. Available from: https://www.frontiersin.org/article/10.3389/fevo.2020.555509.

Feijoo, M. and Parada, A. 2017. Macrosystematics of eutherian mammals combining HTS data to expand taxon coverage. *Molecular Phylogenetics and Evolution*. [Online]. **113**, pp.76–83. Available from: https://www.sciencedirect.com/science/article/pii/S1055790316303669.

Felsenstein, J. and Kishino, H. 1993. Is there something wrong with the bootstrap on phylogenies a reply to hillis and bull. *Systematic Biology*. **42**(2), pp.193–200.

Felsenstein, J. and Kishino, H. 1993. Is there something wrong with the bootstrap on phylogenies a reply to hillis and bull. *Systematic Biology*. **42**(2), pp.193–200.

Fernández, R., Kallal, R.J., Dimitrov, D., Ballesteros, J.A., Arnedo, M.A., Giribet, G. and Hormiga, G. 2018. Phylogenomics, Diversification Dynamics, and Comparative Transcriptomics across the Spider Tree of Life. *Current Biology*. [Online]. **28**(9), pp.1489-1497.e5. Available from: https://www.sciencedirect.com/science/article/pii/S0960982218304226.

Fisher, K.I. and Stewart, R.E.A. 1997. Summer foods of Atlantic walrus, Odobenus rosmarus rosmarus, in northern Foxe Basin, Northwest Territories. *Canadian Journal of Zoology*. **75**(7), pp.1166–1175.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Systematic Zoology*. **19**(2), pp.99–113.

Fletcher, K. and Michelmore, R. 2018. From short reads to chromosome-scale genome assemblies *In*: W. Ma and T. Wolpert, eds. *Methods in Molecular Biology* [Online]. New York, NY: Springer New York, pp.151–197. Available from: https://doi.org/10.1007/978-1-4939-8724-5_13.

Fletcher, W. and Yang, Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular Biology and Evolution*. **27**(10), pp.2257–2267.

Flynn Neff, N.A., Tedford, R.H., J.J. 1987. Phylogeny of the Carnivora. *The phylogeny and classification of the Tetrapods, volume 2: Mammals*., pp.73–116.

Flynn, J.J., Finarelli, J.A., Zehr, S., Hsu, J. and Nedbal, M.A. 2005. Molecular phylogeny of the Carnivora (Mammalia): Assessing the impact of increased sampling on resolving enigmatic relationships. *Systematic Biology*. **54**(2), pp.317–337.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. and Smit, A.F. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*. [Online]. **117**(17), pp.9451–9457. Available from: http://www.pnas.org/content/117/17/9451.abstract.

Foley, N.M., Springer, M.S. and Teeling, E.C. 2016. Mammal madness: Is the mammal tree of life not yet resolved? *Philosophical Transactions of the Royal Society B: Biological Sciences*. [Online]. **371**(1699), p.20150140. Available from: https://doi.org/10.1098/rstb.2015.0140.

Folkow, L.P., Mårtensson, P.E. and Blix, A.S. 1996. Annual distribution of Hooded seals (Cystophora cristata) in the Greenland and Norwegian seas. *Polar Biology*. **16**(3), pp.179–189.

Foll, M., Gaggiotti, O.E., Daub, J.T., Vatsiou, A. and Excoffier, L. 2014. Widespread signals of convergent adaptation to high altitude in Asia and America. *American Journal of Human Genetics*. [Online]. **95**(4), pp.394–407. Available from: https://pubmed.ncbi.nlm.nih.gov/25262650.

Foote, A.D., Liu, Y., Thomas, G.W.C., Vinař, T., Alföldi, J., Deng, J., Dugan, S., Van Elk, C.E., Hunter, M.E., Joshi, V., Khan, Z., Kovar, C., Lee, S.L., Lindblad-Toh, K., Mancia, A., Nielsen, R., Qin, X., Qu, J., Raney, B.J., Vijay, N., Wolf, J.B.W., Hahn, M.W., Muzny, D.M., Worley, K.C., Gilbert, M.T.P. and Gibbs, R.A. 2015. Convergent evolution of the

genomes of marine mammals. *Nature Genetics*. [Online]. **47**(3), pp.272–275. Available from: http://dx.doi.org/10.1038/ng.3198.

Foster, K.R., Wenseleers, T. and Ratnieks, F.L.W. 2006. Kin selection is the key to altruism. *Trends in Ecology and Evolution*. [Online]. **21**(2), pp.57–60. Available from: https://www.sciencedirect.com/science/article/pii/S0169534705003836.

Fowler, M.A., Debier, C., Mignolet, E., Linard, C., Crocker, D.E. and Costa, D.P. 2014. Fatty acid mobilization and comparison to milk fatty acid content in northern elephant seals. *Journal of Comparative Physiology B: Biochemical, Systemic, and Environmental Physiology*. **184**(1), pp.125–135.

Fowler, M., Champagne, C. and Crocker, D. 2018. Adiposity and fat metabolism during combined fasting and lactation in elephant seals. *Journal of Experimental Biology*. [Online]. **121**(Pt Suppl 1). Available from: http://dx.doi.org/10.1242/jeb.161554.

Fu, N.Y., Rios, A.C., Pal, B., Soetanto, R., Lun, A.T.L., Liu, K., Beck, T., Best, S.A., Vaillant, F., Bouillet, P., Strasser, A., Preiss, T., Smyth, G.K., Lindeman, G.J. and Visvader, J.E. 2015. EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nature Cell Biology*. [Online]. **17**(4), pp.365–375. Available from: https://doi.org/10.1038/ncb3117.

Fulton, T.L. and Strobeck, C. 2006. Molecular phylogeny of the Arctoidea (Carnivora): Effect of missing data on supertree and supermatrix analyses of multiple gene data sets. *Molecular Phylogenetics and Evolution*. **41**(1), pp.165–181.

Fulton, T.L. and Strobeck, C. 2010. Multiple fossil calibrations, nuclear loci and mitochondrial genomes provide new insight into biogeography and divergence timing for true seals (Phocidae, Pinnipedia). *Journal of Biogeography*. [Online]. **37**(5), pp.814–829. Available from: http://dx.doi.org/10.1111/j.1365-2699.2010.02271.x.

Fulton, T.L. and Strobeck, C. 2010. Multiple markers and multiple individuals refine true seal phylogeny and bring molecules and morphology back in line. *Proceedings of the Royal Society B: Biological Sciences*. [Online]. **277**(1684), pp.1065–1070. Available from: http://www.jstor.org/stable/25676673.

Furbish, R. 2015. Something Old, Something New, Something Swimming in the Blue: an Analysis of the pinnipedFamily Desmatophocidae, Its Phylogenetic Position and Swimming Mode. *San Diego: University of California. MSc dissertation*., p.75.

Ganguly, B., Ambwani, T.K. and Rastogi, S.K. 2017. Electronic northern analysis of genes and modeling of gene networks underlying bovine milk fat production N. A. Doggett, ed. *Genetics Research International*. [Online]. **2017**, p.1910530. Available from: https://doi.org/10.1155/2017/1910530.

Gearty, W., McClain, C.R. and Payne, J.L. 2018. Energetic tradeoffs control the size distribution of aquatic mammals. *Proceedings of the National Academy of Sciences of the United States of America*. [Online]. **115**(16), pp.4194–4199. Available from: https://doi.org/10.1073/pnas.1712629115.

Gene Ontology Consortium 2021. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*. **49**(D1), pp.D325–D334.

Goeman, J.J. and Solari, A. 2014. Multiple hypothesis testing in genomics. *Statistics in Medicine*. **33**(11), pp.1946–1978.

Goldman, N., Anderson, J.P. and Rodrigo, A.G. 2000. Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*. **49**(4), pp.652–670.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*. **11**(5), pp.725–736.

Goodman, S. and Dmitrieva, L. 2016. Pusa caspica. The IUCN Red List of Threatened Species 2016: e. T41669A45230700.

Goodwin, S., McPherson, J.D. and McCombie, W.R. 2016. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*. [Online]. **17**(6), pp.333–351. Available from: https://doi.org/10.1038/nrg.2016.49.

Grabherr, M.G.., Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., N. and Friedman, and A.R. 2013. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*. **29**(7), pp.644–652.

Gray, J. and Boucot, A.J. 1979. Historical biogeography, plate tectonics, and the changing environment. *Historical biogeography, plate tectonics, and the changing environment*.

Green, K.A. and Streuli, C.H. 2004. Apoptosis regulation in the mammary gland. *Cellular and Molecular Life Sciences*. [Online]. **61**(15), pp.1867–1883. Available from: https://doi.org/10.1007/s00018-004-3366-y.

Greenleaf, W.J. and Sidow, A. 2014. The future of sequencing: Convergence of intelligent design and market Darwinism. *Genome Biology*. [Online]. **15**(3), p.303. Available from: https://doi.org/10.1186/gb4168.

Gregory, W.K. 1910. *The Orders of Mammals*. The Trustees.

Gremme, G. 2013. GenomeThreader Gene Prediction Software. Available from: http://www.genomethreader.org/doc/gthmanual.pdf.

Guigó, R., Agarwal, P., Abril, J.F., Burset, M. and Fickett, J.W. 2000. An assessment of gene prediction accuracy in large DNA sequences. *Genome Research*. [Online]. **10**(10), pp.1631–1642. Available from: https://pubmed.ncbi.nlm.nih.gov/11042160.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Robin, C.R. and Wortman, J.R. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*. [Online]. **9**(1), p.R7. Available from: https://doi.org/10.1186/gb-2008-9-1-r7.

Haley, B.A., Frank, M., Spielhagen, R.F. and Fietzke, J. 2008. Radiogenic isotope record of Arctic Ocean circulation and weathering inputs of the past 15 million years. *Paleoceanography*. **23**(1).

Hall, B.G. 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Molecular Biology and Evolution*. **22**(3), pp.792–802.

Hamilton, W.D. 1972. Altruism and Related Phenomena, Mainly in Social Insects. *Annual Review of Ecology and Systematics*. [Online]. **3**(1), pp.193–232. Available from: https://doi.org/10.1146/annurev.es.03.110172.001205.

Hammond, J.A., Hauton, C., Bennett, K.A. and Hall, A.J. 2012. Phocid seal leptin: Tertiary structure and hydrophobic receptor binding site Preservation during Distinct leptin Gene Evolution. *PLoS ONE*. [Online]. **7**(4), pp.e35395–e35395. Available from: https://pubmed.ncbi.nlm.nih.gov/22536379.

Hancock, J.M., Worthey, E.A. and Santibáñez-Koref, M.F. 2001. A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Molecular Biology and Evolution*. [Online]. **18**(6), pp.1014–1023. Available from: https://doi.org/10.1093/oxfordjournals.molbev.a003873.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G.R., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S.E. and Guigo, R. 2006. GENCODE: producing a reference annotation for ENCODE. *Genome biology*. [Online]. **7 Suppl 1**(1), p.S4. Available from: https://doi.org/10.1186/gb-2006-7-s1-s4.

Hashidate-Yoshida, T., Harayama, T., Hishikawa, D., Morimoto, R., Hamano, F., Tokuoka, S.M., Eto, M., Tamura-Nakano, M., Yanobu-Takanashi, R., Mukumoto, Y., Kiyonari, H., Okamura, T., Kita, Y., Shindou, H. and Shimizu, T. 2015. Fatty acid remodeling by LPCAT3 enriches arachidonate in phospholipid membranes and regulates triglyceride transport. *eLife*. [Online]. **4**, p.e06328. Available from: https://pubmed.ncbi.nlm.nih.gov/25898003.

Hearn, T., Renforth, G.L., Spalluto, C., Hanley, N.A., Piper, K., Brickwood, S., White, C., Connolly, V., Taylor, J.F.N., Russell-Eggitt, I., Bonneau, D., Walker, M. and Wilson, D.I. 2002. Mutation of ALMS1, a large gene with a tandem repeat encoding 47 amino acids,

causes Alström syndrome. *Nature Genetics*. [Online]. **31**(1), pp.79–83. Available from: https://doi.org/10.1038/ng874.

Heather, J.M. and Chain, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics*. [Online]. **107**(1), pp.1–8. Available from: https://pubmed.ncbi.nlm.nih.gov/26554401.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M. and Kumar, S. 2015. Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*. [Online]. **32**(4), pp.835–845. Available from: https://pubmed.ncbi.nlm.nih.gov/25739733.

Higdon, J.W., Bininda-Emonds, O.R.P., Beck, R.M.D. and Ferguson, S.H. 2007. Phylogeny and divergence of the pinnipeds (Carnivora: Mammalia) assessed using a multigene dataset. *BMC Evolutionary Biology*. **7**(1), p.216.

Hillis, D.M. and Bull, J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*. **42**(2), pp.182–192.

Hindell, M.A., Slip, D.J. and Burton, H.R. 1991. The diving behaviour of adult male and female southern elephant seals, Mirounga leonina (Pinnipedia: Phocidae). *Australian Journal of Zoology*. [Online]. **39**(5), pp.499–508. Available from: https://doi.org/10.1071/ZO9910595.

Hindle, A.G., Allen, K.N., Batten, A.J., Hückstädt, L.A., Turner-Maier, J., Schulberg, S.A., Johnson, J., Karlsson, E., Lindblad-Toh, K., Costa, D.P., Bloch, D.B., Zapol, W.M. and Buys, E.S. 2019. Low guanylyl cyclase activity in Weddell seals: implications for peripheral vasoconstriction and perfusion of the brain during diving. *American journal of physiology. Regulatory, integrative and comparative physiology*. [Online]. **316**(6), pp.R704–R715. Available from: https://doi.org/10.1152/ajpregu.00283.2018.

Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q. and Vinh, L.S. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*. [Online]. **35**(2), pp.518–522. Available from: https://doi.org/10.1093/molbev/msx281.

Hodcroft, E.B., De Maio, N., Lanfear, R., MacCannell, D.R., Minh, B.Q., Schmidt, H.A., Stamatakis, A., Goldman, N. and Dessimoz, C. 2021. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature*. **591**(7848), pp.30–33.

Hoff, K.J., Lomsadze, A., Stanke, M. and Borodovsky, M. 2018. BRAKER2: incorporating protein homology information into gene prediction with GeneMark-EP and AUGUSTUS. *Plant and Animal Genomes XXVI*. (January).

Hoff, K.J., Lomsadze, A., Borodovsky, M. and Stanke, M. 2019. Whole-genome annotation with BRAKER. *Methods in Molecular Biology*. [Online]. **1962**, pp.65–95. Available from: https://pubmed.ncbi.nlm.nih.gov/31020555.

Hoff, K.J. and Stanke, M. 2019. Predicting Genes in Single Genomes with AUGUSTUS. *Current Protocols in Bioinformatics*. **65**(1), p.e57.

Holt, C. and Yandell, M. 2011. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. [Online]. **12**(1), p.491. Available from: http://dx.doi.org/10.1186/1471-2105-12-491.

Hon, T., Mars, K., Young, G., Tsai, Y.C., Karalius, J.W., Landolin, J.M., Maurer, N., Kudrna, D., Hardigan, M.A., Steiner, C.C., Knapp, S.J., Ware, D., Shapiro, B., Peluso, P. and Rank, D.R. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*. [Online]. **7**(1), p.399. Available from: https://doi.org/10.1038/s41597-020-00743-4.

Hooker, S.K., Fahlman, A., Moore, M.J., Aguilar de Soto, N., Bernaldo de Quirós, Y., Brubakk, A.O., Costa, D.P., Costidis, A.M., Dennison, S., Falke, K.J., Fernandez, A., Ferrigno, M., Fitz-Clarke, J.R., Garner, M.M., Houser, D.S., Jepson, P.D., Ketten, D.R., Kvadsheim, P.H., Madsen, P.T., Pollock, N.W., Rotstein, D.S., Rowles, T.K., Simmons, S.E., van Bonn, W., Weathersby, P.K., Weise, M.J., Williams, T.M. and Tyack, P.L. 2012. Deadly diving? physiological and behavioural management of decompression stress in diving mammals. *Proceedings of the Royal Society B: Biological Sciences*. [Online]. **279**(1731), pp.1041–1050. Available from: https://doi.org/10.1098/rspb.2011.2088.

Hopkins, S.R. and Powell, F.L. 2001. Common themes of adaptation to hypoxia: Insights from comparative physiology. *Advances in Experimental Medicine and Biology*. **502**, pp.153–167.

Hoyle, T.M., Leroy, S.A.G., López-Merino, L., Miggins, D.P. and Koppers, A.A.P. 2020. Vegetation succession and climate change across the Plio-Pleistocene transition in eastern Azerbaijan, central Eurasia (2.77–2.45 Ma). *Palaeogeography, Palaeoclimatology, Palaeoecology*. [Online]. **538**, p.109386. Available from: https://www.sciencedirect.com/science/article/pii/S0031018218308769.

Huang, X., Sun, D., Wu, T., Liu, X., Xu, S. and Yang, G. 2021. Genomic insights into body size evolution in Carnivora support Peto's paradox. *BMC Genomics*. [Online]. **22**(1), p.429. Available from: https://doi.org/10.1186/s12864-021-07732-w.

Huelsenbeck, J.P. and Rannala, B. 2004. Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*. **53**(6), pp.904–913.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C. and Bork, P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution*. **34**(8), pp.2115–2122.

Hultqvist, M., Olofsson, P., Holmberg, J., Bäckström, B.T., Tordsson, J. and Holmdahl, R. 2004. Enhanced autoimmunity, arthritis, and encephalomyelitis in mice with a reduced oxidative

burst due to a mutation in the Ncf1 gene. *Proceedings of the National Academy of Sciences of the United States of America*. [Online]. **101**(34), pp.12646–12651. Available from: https://doi.org/10.1073/pnas.0403831101.

Hunt, R.M. and Barnes, L.G. 1994. Basicranial evidence for ursid affinity of the oldest pinnipeds *In*: *Proceedings - San Diego Society of Natural History*., pp.57–67.

Hurst, L.D. and Pál, C. 2001. Evidence for purifying selection acting on silent sites in BRCA1. *Trends in Genetics*. [Online]. **17**(2), pp.62–65. Available from: https://doi.org/10.1016/S0168-9525(00)02173-9.

Hwa, V., Little, B., Adiyaman, P., Kofoed, E.M., Pratt, K.L., Ocal, G., Berberoglu, M. and Rosenfeld, R.G. 2005. Severe growth hormone insensitivity resulting from total absence of signal transducer and activator of transcription 5b. *Journal of Clinical Endocrinology and Metabolism*. **90**(7), pp.4260–4266.

Igarashi, M., Osuga, J.I., Uozaki, H., Sekiya, M., Nagashima, S., Takahashi, M., Takase, S., Takanashi, M., Li, Y., Ohta, K., Kumagai, M., Nishi, M., Hosokawa, M., Fledelius, C., Jacobsen, P., Yagyu, H., Fukayama, M., Nagai, R., Kadowaki, T., Ohashi, K. and Ishibashi, S. 2010. The critical role of neutral cholesterol ester hydrolase 1 in cholesterol removal from human macrophages. *Circulation Research*. [Online]. **107**(11), pp.1387–1395. Available from: https://doi.org/10.1161/CIRCRESAHA.110.226613.

Illa, S.K., Mukherjee, S., Nath, S. and Mukherjee, A. 2021. Genome-Wide Scanning for Signatures of Selection Revealed the Putative Genomic Regions and Candidate Genes Controlling Milk Composition and Coat Color Traits in Sahiwal Cattle. *Frontiers in Genetics*. [Online]. **12**, p.699422. Available from: https://pubmed.ncbi.nlm.nih.gov/34306039.

Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution*. [Online]. **40**(2), pp.190–226. Available from: https://doi.org/10.1007/BF00167113.

Iverson, S.J., Hamosh, M. and Bowen, W.D. 1995. Lipoprotein lipase activity and its relationship to high milk fat transfer during lactation in Grey seals. *Journal of Comparative Physiology B*. [Online]. **165**(5), pp.384–395. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8576451.

Iverson, S.J., Oftedal, O.T., Bowen, W.D., Boness, D.J. and Sampugna, J. 1995. Prenatal and postnatal transfer of fatty acids from mother to pup in the Hooded seal. *Journal of Comparative Physiology B*. [Online]. **165**(1), pp.1–12. Available from: https://www.ncbi.nlm.nih.gov/pubmed/7601954.

Iverson, S.J. 2009. Blubber *In*: W. F. Perrin, B. Würsig and J. G. M. B. T.-E. of M. M. (Second E. Thewissen, eds. *Encyclopedia of Marine Mammals* [Online]. London: Academic Press,

pp.115–120. Available from: https://www.sciencedirect.com/science/article/pii/B9780123735539000328.

Jacobs, L.L., Fiorillo, A.R., Nishida, Y. and Fitzgerald, E.M.G. 2009. Mid-cenozoic marine mammals from Alaska. *Papers on Geology, Vertebrate Paleontology, and Biostratigraphy in Honor of Michael O. Woodburne. Museum of Northern Arizona Bulletin*. **65**(6193 m), pp.171–184.

Jain, M., Olsen, H.E., Paten, B. and Akeson, M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*. **17**(1), p.239.

Jebb, D., Huang, Z., Pippel, M., Hughes, G.M., Lavrichenko, K., Devanna, P., Winkler, S., Jermiin, L.S., Skirmuntt, E.C., Katzourakis, A., Burkitt-Gray, L., Ray, D.A., Sullivan, K.A.M., Roscito, J.G., Kirilenko, B.M., Dávalos, L.M., Corthals, A.P., Power, M.L., Jones, G., Ransome, R.D., Dechmann, D.K.N., Locatelli, A.G., Puechmaille, S.J., Fedrigo, O., Jarvis, E.D., Hiller, M., Vernes, S.C., Myers, E.W. and Teeling, E.C. 2020. Six reference-quality genomes reveal evolution of bat adaptations. *Nature*. [Online]. **583**(7817), pp.578–584. Available from: https://doi.org/10.1038/s41586-020-2486-3.

Jeffares, D.C., Tomiczek, B., Sojo, V. and dos Reis, M. 2015. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome *In*: *Methods in Molecular Biology*. Springer, pp.65–90.

Jeffroy, O., Brinkmann, H., Delsuc, F. and Philippe, H. 2006. Phylogenomics: the beginning of incongruence? *Trends in Genetics*. [Online]. **22**(4), pp.225–231. Available from: https://www.sciencedirect.com/science/article/pii/S0168952506000515.

Jena, M.K., Jaswal, S., Kumar, S. and Mohanty, A.K. 2019. Molecular mechanism of mammary gland involution: An update. *Developmental Biology*. [Online]. **445**(2), pp.145–155. Available from: https://www.sciencedirect.com/science/article/pii/S0012160618302677.

Jordan, G. and Goldman, N. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular Biology and Evolution*. [Online]. **29**(4), pp.1125–1139. Available from: https://doi.org/10.1093/molbev/msr272.

Jukes TH & Cantor CR 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press; *Journal of Human Evolution*. [Online]. **9**(2), pp.21–132. Available from: http://linkinghub.elsevier.com/retrieve/pii/0047248480900822.

Kajiwara, N., Watanabe, M., Wilson, S., Eybatov, T., Mitrofanov, I. V., Aubrey, D.G., Khuraskin, L.S., Miyazaki, N. and Tanabe, S. 2008. Persistent organic pollutants (POPs) in Caspian seals of unusual mortality event during 2000 and 2001. *Environmental Pollution*. **152**(2), pp.431–442.

Kaltenecker, D., Themanns, M., Mueller, K.M., Spirk, K., Suske, T., Merkel, O., Kenner, L., Luís, A., Kozlov, A., Haybaeck, J., Müller, M., Han, X. and Moriggl, R. 2019. Hepatic growth hormone - JAK2 - STAT5 signalling: Metabolic function, non-alcoholic fatty liver disease and hepatocellular carcinoma progression. *Cytokine*. [Online]. **124**, p.154569. Available from: https://www.sciencedirect.com/science/article/pii/S1043466618303983.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A. and Jermiin, L.S. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*. [Online]. **14**(6), pp.587–589. Available from: https://doi.org/10.1038/nmeth.4285.

Kapli, P., Yang, Z. and Telford, M.J. 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*. [Online]. **21**(7), pp.428–444. Available from: https://doi.org/10.1038/s41576-020-0233-0.

Karlsson, E.K., Sigurdsson, S., Ivansson, E., Thomas, R., Elvers, I., Wright, J., Howald, C., Tonomura, N., Perloski, M., Swofford, R., Biagi, T., Fryc, S., Anderson, N., Courtay-Cahen, C., Youell, L., Ricketts, S.L., Mandlebaum, S., Rivera, P., von Euler, H., Kisseberth, W.C., London, C.A., Lander, E.S., Couto, G., Comstock, K., Starkey, M.P., Modiano, J.F., Breen, M. and Lindblad-Toh, K. 2013. Genome-wide analyses implicate 33 loci in heritable dog osteosarcoma, including regulatory variants near CDKN2A/B. *Genome Biology*. [Online]. **14**(12), p.R132. Available from: https://doi.org/10.1186/gb-2013-14-12-r132.

Katoh, K. and Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*. [Online]. **30**(4), pp.772–780. Available from: https://doi.org/10.1093/molbev/mst010.

Keane, M., Semeiks, J., Webb, A.E., Li, Y.I., Quesada, V., Craig, T., Madsen, L.B., van Dam, S., Brawand, D., Marques, P.I., Michalak, P., Kang, L., Bhak, J., Yim, H.S., Grishin, N. V., Nielsen, N.H., Heide-Jørgensen, M.P., Oziolor, E.M., Matson, C.W., Church, G.M., Stuart, G.W., Patton, J.C., George, J.C., Suydam, R., Larsen, K., López-Otín, C., O'Connell, M.J., Bickham, J.W., Thomsen, B. and deMagalhães, J.P. 2015. Insights into the evolution of longevity from the bowhead whale genome. *Cell Reports*. [Online]. **10**(1), pp.112–122. Available from: http://dx.doi.org/10.1016/j.celrep.2014.12.008.

Kemena, C., Dohmen, E. and Bornberg-Bauer, E. 2019. DOGMA: A web server for proteome and transcriptome quality assessment. *Nucleic Acids Research*. **47**(W1), pp.W507–W510.

Kent, W.J. 2002. BLAT—the BLAST-like alignment tool. *Genome research*. **12**(4), pp.656–664.

Kern, A.D. and Hahn, M.W. 2018. The neutral theory in light of natural selection. *Molecular Biology and Evolution*. **35**(6), pp.1366–1371.

Khan, S.S., Shah, S.J., Klyachko, E., Baldridge, A.S., Eren, M., Place, A.T., Aviv, A., Puterman, E., Lloyd-Jones, D.M., Heiman, M., Miyata, T., Gupta, S., Shapiro, A.D. and Vaughan, D.E. 2017. A null mutation in SERPINE1 protects against biological aging in humans.

Science Advances. [Online]. **3**(11), pp.eaao1617–eaao1617. Available from: https://pubmed.ncbi.nlm.nih.gov/29152572.

Khudyakov, J.I., Abdollahi, E., Ngo, A., Sandhu, G., Stephan, A., Costa, D.P. and Crocker, D.E. 2019. Expression of obesity-related adipokine genes during fasting in a naturally obese marine mammal. *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*. [Online]. **317**(4), pp.R513–R520. Available from: https://doi.org/10.1152/ajpregu.00182.2019.

Kienle, S.S. and Berta, A. 2018. The evolution of feeding strategies in phocid seals (Pinnipedia, Phocidae). *Journal of Vertebrate Paleontology*. [Online]. **38**(6), pp.1–13. Available from: https://www.jstor.org/stable/26765803.

Kienle, S.S., Cuthbertson, R.D. and Reidenberg, J.S. 2022. Comparative examination of pinniped craniofacial musculature and its role in aquatic feeding. *Journal of Anatomy*. **240**(2), pp.226–252.

Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*. **37**(8), pp.907–915.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature*. [Online]. **217**(5129), pp.624–626. Available from: https://doi.org/10.1038/217624a0.

Kingman, J.F.C. 1982. The coalescent. *Stochastic Processes and their Applications*. **13**(3), pp.235–248.

Kishida, T., Kubota, S., Shirayama, Y. and Fukami, H. 2007. The olfactory receptor gene repertoires in secondary-adapted marine vertebrates: Evidence for reduction of the functional proportions in cetaceans. *Biology Letters*. **3**(4), pp.428–430.

Kishino, H. and Hasegawa, M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*. [Online]. **29**(2), pp.170–179. Available from: https://doi.org/10.1007/BF02100115.

Kishino, H., Miyata, T. and Hasegawa, M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*. [Online]. **31**(2), pp.151–160. Available from: https://doi.org/10.1007/BF02109483.

Kohno, N. 2006. A new miocene odobenid (Mammalia: Carnivora) from Hokkaido, Japan, and its implications for odobenid phylogeny. *Journal of Vertebrate Paleontology*. **26**(2), pp.411–421.

Kohno, N. 2006. A new miocene odobenid (Mammalia: Carnivora) from Hokkaido, Japan, and its implications for odobenid phylogeny. *Journal of Vertebrate Paleontology*. **26**(2), pp.411–421.

Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P.A. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*. **37**(5), pp.540–546.

Kooyman, G.L. and Ponganis, P.J. 1998. The physiological basis of diving to depth: Birds and mammals. *Annual Review of Physiology*. [Online]. **60**(1), pp.19–32. Available from: https://doi.org/10.1146/annurev.physiol.60.1.19.

Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S., Williams, J.L., Smith, T.P.L. and Phillippy, A.M. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*. [Online]. **36**(12), pp.1174–1182. Available from: https://doi.org/10.1038/nbt.4277.

Koretsky, I.A., Barnes, L.G. and Rahmat, S.J. 2016. Re-evaluation of morphological characters questions current views of pinniped origins. *Vestnik Zoologii*. **50**(4), pp.327–354.

Koretsky, I.A. 2001. Morphology and Systematics of Miocene Phocinae (Mammalia: Carnivora) from Paratethys and the North Atlantic Region. *Geologica Hungarica Series Palaeontologica*. **54**, pp.1–109.

Köster, J. and Rahmann, S. 2012. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*. **28**(19), pp.2520–2522.

Kovacs, K.M. and Lavigne, D.M. 1992. Maternal investment in otariid seals and walruses. *Canadian Journal of Zoology*. **70**(10), pp.1953–1964.

Kovacs, K.M., Aguilar, A., Aurioles, D., Burkanov, V., Campagna, C., Gales, N., Gelatt, T., Goldsworthy, S.D., Goodman, S.J., Hofmeyr, G.J.G., Härkönen, T., Lowry, L., Lydersen, C., Schipper, J., Sipilä, T., Southwell, C., Stuart, S., Thompson, D. and Trillmich, F. 2012. Global threats to pinnipeds. *Marine Mammal Science*. **28**(2), pp.414–436.

Kuhn, C. and Frey, E. 2012. Walking like caterpillars, flying like bats-pinniped locomotion. *Palaeobiodiversity and Palaeoenvironments*. **92**(2), pp.197–210.

Kumar, S. and Gadagkar, S.R. 2001. Disparity index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics*. [Online]. **158**(3), pp.1321–1327. Available from: https://pubmed.ncbi.nlm.nih.gov/11454778.

Kumar, V., Lammers, F., Bidon, T., Pfenninger, M., Kolter, L., Nilsson, M.A. and Janke, A. 2017. The evolutionary history of bears is characterized by gene flow across species. *Scientific Reports*. [Online]. **7**(1), p.46487. Available from: https://doi.org/10.1038/srep46487.

L, F.A. 1901. The South African Museum. *Science*. **14**(363), p.940.

Laidre, K.L., Stirling, I., Lowry, L.F., Wiig, Ø., Heide-Jørgensen, M.P. and Ferguson, S.H. 2008. Quantifying the sensitivity of arctic marine mammals to climate-induced habitat change. *Ecological Applications*. **18**(SUPPL.2), pp.S97–S125.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, Christina, Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Hong, M.L., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., De La Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M. V., Kaul, R., Raymond, Christopher, Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A. and Morgan, M.J. 2001. Initial sequencing and analysis of the human genome. *Nature*. [Online]. **409**(6822), pp.860–921. Available from: https://doi.org/10.1038/35057062.

Lantz, H., Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.F., Vlasova, A., Leskosek, B.L., Soler, L. and Binzer-Panchal, M. 2018. Ten steps to get started in Genome Assembly and Annotation. *F1000Research*. **7**.

Lartillot, N., Rodrigue, N., Stubbs, D. and Richer, J. 2013. Phylobayes mpi: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*. [Online]. **62**(4), pp.611–615. Available from: https://doi.org/10.1093/sysbio/syt022.

Lazarev, S., Kuiper, K.F., Oms, O., Bukhsianidze, M., Vasilyan, D., Jorissen, E.L., Bouwmeester, M.J., Aghayeva, V., van Amerongen, A.J., Agustí, J., Lordkipanidze, D. and Krijgsman, W. 2021. Five-fold expansion of the Caspian Sea in the late Pliocene: New and revised magnetostratigraphic and 40Ar/39Ar age constraints on the Akchagylian Stage *In*: *Global and Planetary Change*., pp.EGU21-15419.

Lee, M.S.Y. and Palci, A. 2015. Morphological phylogenetics in the genomic age. *Current Biology*. [Online]. **25**(19), pp.R922–R929. Available from: https://www.sciencedirect.com/science/article/pii/S096098221500812X.

Lenfant, C., Johansen, K. and Torrance, J.D. 1970. Gas transport and oxygen storage capacity in some pinnipeds and the sea otter. *Respiration Physiology*. [Online]. **9**(2), pp.277–286. Available from: https://www.sciencedirect.com/science/article/pii/0034568770900769.

Lenski, R.E. 2017. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *ISME Journal*. [Online]. **11**(10), pp.2181–2194. Available from: https://doi.org/10.1038/ismej.2017.69.

Levy, S.E. and Myers, R.M. 2016. Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics*. [Online]. **17**(1), pp.95–115. Available from: https://doi.org/10.1146/annurev-genom-083115-022413.

Li, D. and Li, Y. 2020. The interaction between ferroptosis and lipid metabolism in cancer. *Signal Transduction and Targeted Therapy*. [Online]. **5**(1), p.108. Available from: https://doi.org/10.1038/s41392-020-00216-5.

Li, H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**(18), pp.3094–3100.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*. [Online]. Available from: http://arxiv.org/abs/1303.3997.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. **27**(21), pp.2987–2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. [Online]. **25**(16), pp.2078–2079. Available from: http://dx.doi.org/10.1093/bioinformatics/btp352.

Li, J., Cao, F., Yin, H. liang, Huang, Z. jian, Lin, Z. tao, Mao, N., Sun, B. and Wang, G. 2020. Ferroptosis: past, present and future. *Cell Death and Disease*. [Online]. **11**(2), p.88. Available from: https://doi.org/10.1038/s41419-020-2298-2.

Li, W.H., Wu, C.I. and Luo, C.C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*. **2**(2), pp.150–174.

Li, W.H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution*. **36**(1), pp.96–99.

Li, Z., Jiang, H., Ding, T., Lou, C., Bui, H.H., Kuo, M.S. and Jiang, X.C. 2015. Deficiency in Lysophosphatidylcholine Acyltransferase 3 Reduces Plasma Levels of Lipids by Reducing Lipid Absorption in Mice. *Gastroenterology*. [Online]. **149**(6), pp.1519–1529. Available from: https://doi.org/10.1053/j.gastro.2015.07.012.

Lian, M., Castellini, J.M., Kuhn, T., Rea, L., Bishop, L., Keogh, M., Kennedy, S.N., Fadely, B., van Wijngaarden, E., Maniscalco, J.M. and O'Hara, T. 2020. Assessing oxidative stress in Steller sea lions (Eumetopias jubatus): Associations with mercury and selenium concentrations. *Comparative Biochemistry and Physiology Part - C: Toxicology and Pharmacology*. [Online]. **235**(March), p.108786. Available from: https://doi.org/10.1016/j.cbpc.2020.108786.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S. and Dekker, J. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. **326**(5950), pp.289–293.

Lincoln, D.R., Edmunds, D.T., Gribble, T.J. and Schwartz, H.C. 1973. Studies on the hemoglobins of pinnipeds. *Blood*. [Online]. **41**(1), pp.163–170. Available from: https://www.sciencedirect.com/science/article/pii/S0006497120693004.

Littlejohn, M.D., Tiplady, K., Lopdell, T., Law, T.A., Scott, A., Harland, C., Sherlock, R., Henty, K., Obolonkin, V., Lehnert, K., MacGibbon, A., Spelman, R.J., Davis, S.R. and Snell, R.G. 2014. Expression variants of the lipogenic AGPAT6 gene affect diverse milk composition phenotypes in Bos taurus. *PLoS ONE*. [Online]. **9**(1), p.e85757. Available from: https://doi.org/10.1371/journal.pone.0085757.

Liu, A., He, F., Shen, L., Liu, R., Wang, Z. and Zhou, J. 2019. Convergent degeneration of olfactory receptor gene repertoires in marine mammals. *BMC Genomics*. [Online]. **20**(1), p.977. Available from: https://doi.org/10.1186/s12864-019-6290-0.

Liu, L., Anderson, C., Pearl, D. and Edwards, S. V. 2019. Modern phylogenomics: Building phylogenetic trees using the multispecies coalescent model *In*: M. Anisimova, ed. *Methods in Molecular Biology* [Online]. New York, NY: Springer New York, pp.211–239. Available from: https://doi.org/10.1007/978-1-4939-9074-0_7.

Loh, N.Y., Neville, M.J., Marinou, K., Hardcastle, S.A., Fielding, B.A., Duncan, E.L., McCarthy, M.I., Tobias, J.H., Gregson, C.L., Karpe, F. and Christodoulides, C. 2015. LRP5 regulates human body fat distribution by modulating adipose progenitor biology in a dose- and depot-specific fashion. *Cell Metabolism*. [Online]. **21**(2), pp.262–273. Available from: https://pubmed.ncbi.nlm.nih.gov/25651180.

Lopes, F., Oliveira, L.R., Kessler, A., Beux, Y., Crespo, E., Cárdenas-Alayza, S., Majluf, P., Sepúlveda, M., Brownell, R.L., Franco-Trecu, V., Páez-Rosas, D., Chaves, J., Loch, C., Robertson, B.C., Acevedo-Whitehouse, K., Elorriaga-Verplancken, F.R., Kirkman, S.P., Peart, C.R., Wolf, J.B.W. and Bonatto, S.L. 2021. Phylogenomic Discordance in the Eared Seals is best explained by Incomplete Lineage Sorting following Explosive Radiation in the Southern Hemisphere. *Systematic biology*. [Online]. **70**(4), pp.786–802. Available from: https://doi.org/10.1093/sysbio/syaa099.

Lydersen, C., Kovacs, K.M. and Hammill, M.O. 1997. Energetics during nursing and early postweaning fasting in Hooded seal (Cystophora cristata) pups from the Gulf of St Lawrence, Canada. *Journal of Comparative Physiology - B Biochemical, Systemic, and Environmental Physiology*. **167**(2), pp.81–88.

Lynch, M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences of the United States of America*. [Online]. **104**(SUPPL. 1), pp.8597–8604. Available from: http://www.pnas.org/content/104/suppl_1/8597.abstract.

Mable, B.K., Alexandrou, M.A. and Taylor, M.I. 2011. Genome duplication in amphibians and fish: An extended synthesis. *Journal of Zoology*. [Online]. **284**(3), pp.151–182. Available from: https://doi.org/10.1111/j.1469-7998.2011.00829.x.

Mach, N., van Baal, J., Kruijt, L., Jacobs, A. and Smits, M. 2011. Dietary unsaturated fatty acids affect the mammary gland integrity and health in lactating dairy cows. *BMC Proceedings*. [Online]. **5**(S4), p.S35. Available from: https://doi.org/10.1186/1753-6561-5-S4-S35.

Machiela, M.J. and Chanock, S.J. 2014. GWAS is going to the dogs. *Genome Biology*. [Online]. **15**(3), p.105. Available from: https://doi.org/10.1186/gb4166.

Maddison, W.P. 1997. Gene trees in species trees. *Systematic Biology*. **46**(3), pp.523–536.

Majoros, W.H., Pertea, M. and Salzberg, S.L. 2004. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics*. **20**(16), pp.2878–2879.

Mallick, S., Gnerre, S., Muller, P. and Reich, D. 2009. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Research*. **19**(5), pp.922–933.

Mani, O., Sorensen, M.T., Sejrsen, K., Bruckmaier, R.M. and Albrecht, C. 2009. Differential expression and localization of lipid transporters in the bovine mammary gland during the pregnancy-lactation cycle. *Journal of Dairy Science*. [Online]. **92**(8), pp.3744–3756. Available from: https://www.sciencedirect.com/science/article/pii/S0022030209706964.

Marcais, G. and Kingsford, C. 2012. Jellyfish: A fast k-mer counter. *Tutorialis e Manuais*. **1**, pp.1–8.

Marçais, G. and Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. [Online]. **27**(6), pp.764–770. Available from: https://doi.org/10.1093/bioinformatics/btr011.

Maresh, J.L., Adachi, T., Takahashi, A., Naito, Y., Crocker, D.E., Horning, M., Williams, T.M. and Costa, D.P. 2015. Summing the strokes: Energy economy in northern elephant seals during large-scale foraging migrations. *Movement Ecology*. **3**(1), pp.1–16.

Massingham, T. and Goldman, N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics*. [Online]. **169**(3), pp.1753–1762. Available from: https://pubmed.ncbi.nlm.nih.gov/15654091.

Mathison, A., Escande, C., Calvo, E., Seo, S., White, T., Salmonson, A., Faubion, W.A., Buttar, N., Iovanna, J., Lomberk, G., Chini, E.N. and Urrutia, R. 2015. Phenotypic characterization of mice carrying homozygous deletion of KLF11, a gene in which mutations cause human neonatal and MODY VII diabetes. *Endocrinology*. [Online]. **156**(10), pp.3581–3595. Available from: https://doi.org/10.1210/en.2015-1145.

Mathison, A., Escande, C., Calvo, E., Seo, S., White, T., Salmonson, A., Faubion, W.A., Buttar, N., Iovanna, J., Lomberk, G., Chini, E.N. and Urrutia, R. 2015. Phenotypic characterization of mice carrying homozygous deletion of KLF11, a gene in which mutations cause human neonatal and MODY VII diabetes. *Endocrinology*. [Online]. **156**(10), pp.3581–3595. Available from: https://doi.org/10.1210/en.2015-1145.

Matsubara, H. and Yamanaka, T. 1978. Evolution of protein molecules. *Mammalian protein metabolism*. **(I-XII + 412 p.)**; **US$ 38.00**, pp.21–132.

McCormack, J.E. and Faircloth, B.C. 2013. Next-generation phylogenetics takes root. *Molecular Ecology*. **22**(1), pp.19–21.

McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T. and Glenn, T.C. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental

mammal phylogeny when combined with species-tree analysis. *Genome Research*. **22**(4), pp.746–754.

McCue, M.D. 2013. *Comparative physiology of fasting, starvation, and food limitation*. Springer.

McDonald, B.I. and Crocker, D.E. 2006. Physiology and behavior influence lactation efficiency in northern elephant seals (Mirounga angustirostris). *Physiological and Biochemical Zoology*. **79**(3), pp.484–496.

McDonald, J.H. and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*. [Online]. **351**(6328), pp.652–654. Available from: https://doi.org/10.1038/351652a0.

McGowen, M.R., Tsagkogeorga, G., Álvarez-Carretero, S., Dos Reis, M., Struebig, M., Deaville, R., Jepson, P.D., Jarman, S., Polanowski, A., Morin, P.A. and Rossiter, S.J. 2020. Phylogenomic Resolution of the Cetacean Tree of Life Using Target Sequence Capture. *Systematic Biology*. [Online]. **69**(3), pp.479–501. Available from: https://doi.org/10.1093/sysbio/syz068.

McGowen, M.R., Tsagkogeorga, G., Williamson, J., Morin, P.A., Rossiter, A.S.J. and Chang, B. 2020. Positive Selection and Inactivation in the Vision and Hearing Genes of Cetaceans. *Molecular Biology and Evolution*. [Online]. **37**(7), pp.2069–2083. Available from: https://doi.org/10.1093/molbev/msaa070.

McLaren, I.A. 1960. On the origin of the Caspian and Baikal seals and the paleoclimatological implication. *American Journal of Science*. **258**(1), pp.47–65.

Meiklejohn, C.D., Montooth, K.L. and Rand, D.M. 2007. Positive and negative selection on the mitochondrial genome. *Trends in Genetics*. **23**(6), pp.259–263.

Mellish, J.E., Iverson, S.J., Bowen, W.D. and Hammill, M.O. 1999. Fat transfer and energetics during lactation in the Hooded seal: The roles of tissue lipoprotein lipase in milk fat secretion and pup blubber deposition. *Journal of Comparative Physiology - B Biochemical, Systemic, and Environmental Physiology*. **169**(6), pp.377–390.

Mellish, J.A.E., Iverson, S.J. and Bowen, W.D. 1999. Variation in milk production and lactation performance in Grey seals and consequences for pup growth and weaning characteristics. *Physiological and Biochemical Zoology*. **72**(6), pp.677–690.

Meyer, C., Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O. and Thompson, J.D. 2020. Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes. *BMC Bioinformatics*. [Online]. **21**(1), p.513. Available from: https://pubmed.ncbi.nlm.nih.gov/33172385.

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.P., Mushayamaha, T. and Thomas, P.D. 2021. PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*. [Online]. **49**(D1), pp.D394–D403. Available from: https://doi.org/10.1093/nar/gkaa1106.

Mitchell, E. and Tedford, R.H. 1973. The Enaliarctinae: a new group of extinct aquatic Carnivora and a consideration of the origin of the Otariidae. *Bulletin of the American Museum of Natural History*. [Online]. **151**(3), pp.201–284. Available from: http://digitallibrary.amnh.org/dspace/bitstream/2246/1178/1/B151a03.pdf.

Miyashita, T., Coates, M.I., Farrar, R., Larson, P., Manning, P.L., Wogelius, R.A., Edwards, N.P., Anné, J., Bergmann, U., Richard Palmer, A. and Currie, P.J. 2019. Hagfish from the Cretaceous Tethys Sea and a reconciliation of the morphological-molecular conflict in early vertebrate phylogeny. *Proceedings of the National Academy of Sciences of the United States of America*. **116**(6), pp.2146–2151.

Moran, R.J., Morgan, C.C. and O'Connell, M.J. 2015. A guide to phylogenetic reconstruction using heterogeneous models - A case study from the root of the placental mammal tree. *Computation*. **3**(2), pp.177–196.

Morgan, C.C., Foster, P.G., Webb, A.E., Pisani, D., McInerney, J.O. and O'Connell, M.J. 2013. Heterogeneous models place the root of the placental mammal phylogeny. *Molecular Biology and Evolution*. [Online]. **30**(9), pp.2145–2156. Available from: http://dx.doi.org/10.1093/molbev/mst117.

Muller, J., Creevey, C.J., Thompson, J.D., Arendt, D. and Bork, P. 2010. AQUA: Automated quality improvement for multiple sequence alignments. *Bioinformatics*. [Online]. **26**(2), pp.263–265. Available from: http://dx.doi.org/10.1093/bioinformatics/btp651.

Murigneux, V., Rai, S.K., Furtado, A., Bruxner, T.J.C., Tian, W., Harliwong, I., Wei, H., Yang, B., Ye, Q., Anderson, E., Mao, Q., Drmanac, R., Wang, O., Peters, B.A., Xu, M., Wu, P., Topp, B., Coin, L.J.M. and Henry, R.J. 2021. Comparison of long-read methods for sequencing and assembly of a plant genome. *GigaScience*. **9**(12), p.giaa146.

Natsidis, P., Kapli, P., Schiffer, P.H. and Telford, M.J. 2021. Systematic errors in orthology inference and their effects on evolutionary analyses. *iScience*. [Online]. **24**(2), p.102110. Available from: https://www.sciencedirect.com/science/article/pii/S258900422100078X.

Nayeri, S. and Stothard, P. 2016. Tissues, Metabolic Pathways and Genes of Key Importance in Lactating Dairy Cattle. *Springer Science Reviews*. [Online]. **4**(2), pp.49–77. Available from: https://doi.org/10.1007/s40362-016-0040-3.

Nei, M., Suzuki, Y. and Nozawa, M. 2010. The neutral theory of molecular evolution in the genomic era. *Annual Review of Genomics and Human Genetics*. **11**, pp.265–289.

Neville, M.C. and Picciano, M.F. 1997. Regulation of milk lipid secretion and composition. *Annual Review of Nutrition*. **17**(1), pp.159–184.

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. *Molecular Biology and Evolution*. [Online]. **32**(1), pp.268–274. Available from: https://doi.org/10.1093/molbev/msu300.

Noh, H.J., Turner-Maier, J., Schulberg, S.A., Fitzgerald, M.L., Johnson, J., Allen, K.N., Hückstädt, L.A., Batten, A.J., Alfoldi, J., Costa, D.P., Karlsson, E.K., Zapol, W.M., Buys, E.S., Lindblad-Toh, K. and Hindle, A.G. 2022. The Antarctic Weddell seal genome reveals evidence of selection on cardiovascular phenotype and lipid handling. *Communications Biology*. [Online]. **5**(1), p.140. Available from: https://doi.org/10.1038/s42003-022-03089-2.

Nojima, T. 1990. a Morphological Consideration of the Relationships of pinnipeds To Other Carnivorans Based on the Bony Tentorium and Bony Falx. *Marine Mammal Science*. **6**(1), pp.54–74.

Noren, D.P., Budge, S.M., Iverson, S.J., Goebel, M.E., Costa, D.P. and Williams, T.M. 2013. Characterization of blubber fatty acid signatures in northern elephant seals (Mirounga angustirostris) over the postweaning fast. *Journal of Comparative Physiology B: Biochemical, Systemic, and Environmental Physiology*. **183**(8), pp.1065–1074.

Notredame, C., Higgins, D.G. and Heringa, J. 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*. **302**(1), pp.205–217.

Nowak, R.M., Jastrzębski, J.P., Kuśmirek, W., Sałamatin, R., Rydzanicz, M., Sobczyk-Kopcioł, A., Sulima-Celińska, A., Paukszto, Ł., Makowczenko, K.G., Płoski, R., Tkach, V. V., Basałaj, K. and Młocicki, D. 2019. Hybrid de novo whole-genome assembly and annotation of the model tapeworm Hymenolepis diminuta. *Scientific Data*. [Online]. **6**(1), p.302. Available from: https://doi.org/10.1038/s41597-019-0311-3.

Nuñez, C., Victor, V.M., Martí, M. and D'Ocon, P. 2014. Role of endothelial nitric oxide in pulmonary and systemic arteries during hypoxia. *Nitric Oxide - Biology and Chemistry*. **37**(1), pp.17–27.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D. and Pruitt,

K.D. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. **44**(D1), pp.D733–D745.

Oftedal, O.T., Boness, D.J. and Bowen, W.D. 1988. The composition of Hooded seal (Cystophora cristata) milk: an adaptation for postnatal fattening . *Canadian Journal of Zoology*. [Online]. **66**(2), pp.318–322. Available from: https://doi.org/10.1139/z88-047.

Ogden, T.H. and Rosenberg, M.S. 2006. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology*. [Online]. **55**(2), pp.314–328. Available from: https://doi.org/10.1080/10635150500541730.

Ogorevc, J., Kunej, T., Razpet, A. and Dovc, P. 2009. Database of cattle candidate genes and genetic markers for milk production and mastitis. *Animal Genetics*. [Online]. **40**(6), pp.832–851. Available from: https://pubmed.ncbi.nlm.nih.gov/19508288.

Oppert, B., Stoss, S., Monk, A. and Smith, T. 2019. Optimized extraction of insect genomic dna for long-read sequencing. *Methods and Protocols*. **2**(4), pp.1–7.

Ordoñez, M., Presa, N., Trueba, M. and Gomez-Munõz, A. 2017. Implication of Ceramide Kinase in Adipogenesis E. Albi, ed. *Mediators of Inflammation*. [Online]. **2017**, p.9374563. Available from: https://doi.org/10.1155/2017/9374563.

Orlov, J.A. 1933. Semantor macrurus (ordo Pinnipedia, fam. Semantoridae fam. nova) aus den Neogen-Ablagerungen Westsibiriens. *Travaux de l'Institut Paléozoologique, Académie des Sciences, URSS*. **2**, pp.165–268.

Ortiz, R.M., Noren, D.P., Litz, B. and Ortiz, C.L. 2001. A new perspective on adiposity in a naturally obese mammal. *American Journal of Physiology - Endocrinology and Metabolism*. **281**(6 44-6), pp.E1347-51.

Overbeek, R., Fonstein, M., D'Souza, M., Push, G.D. and Maltsev, N. 1999. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*. **96**(6), pp.2896–2901.

Pais, F.S.M., Ruy, P. de C., Oliveira, G. and Coimbra, R.S. 2014. Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*. [Online]. **9**(1), p.4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4015676/.

Palo, J.U. and Väinölä, R. 2006. The enigma of the landlocked Baikal and Caspian seals addressed through phylogeny of phocine mitochondrial sequences. *Biological Journal of the Linnean Society*. [Online]. **88**(1), pp.61–72. Available from: https://doi.org/10.1111/j.1095-8312.2006.00607.x.

Papageorgiou, L., Eleni, P., Raftopoulou, S., Mantaiou, M., Megalooikonomou, V. and Vlachakis, D. 2018. Genomic big data hitting the storage bottleneck. *EMBnet.journal*. [Online]. **24**, p.e910. Available from: https://pubmed.ncbi.nlm.nih.gov/29782620.

Påsche, A. and Krog, J. 1980. Heart rate in resting seals on land and in water. *Comparative Biochemistry and Physiology -- Part A: Physiology*. [Online]. **67**(1), pp.77–83. Available from: https://www.sciencedirect.com/science/article/pii/0300962980904107.

Paterson, R.S., Rybczynski, N., Kohno, N. and Maddin, H.C. 2020. A Total Evidence Phylogenetic Analysis of pinnipedPhylogeny and the Possibility of Parallel Evolution Within a Monophyletic Framework. *Frontiers in Ecology and Evolution*. [Online]. **7**, p.457. Available from: https://www.frontiersin.org/article/10.3389/fevo.2019.00457.

Payne, A., Holmes, N., Rakyan, V. and Loose, M. 2019. Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*. [Online]. **35**(13), pp.2193–2198. Available from: https://doi.org/10.1093/bioinformatics/bty841.

Pearson, T., Hornstra, H.M., Sahl, J.W., Schaack, S., Schupp, J.M., Beckstrom-Sternberg, S.M., O'Neill, M.W., Priestley, R.A., Champion, M.D., Beckstrom-Sternberg, J.S., Kersh, G.J., Samuel, J.E., Massung, R.F. and Keim, P. 2013. When outgroups fail; Phylogenomics of rooting the emerging pathogen, Coxiella burnetii. *Systematic Biology*. [Online]. **62**(5), pp.752–762. Available from: https://pubmed.ncbi.nlm.nih.gov/23736103.

Pegolo, S., Dadousis, C., MacH, N., Ramayo-Caldas, Y., Mele, M., Conte, G., Schiavon, S., Bittante, G. and Cecchinato, A. 2017. SNP co-association and network analyses identify E2F3, KDM5A and BACH2 as key regulators of the bovine milk fatty acid profile. *Scientific Reports*. [Online]. **7**(1), p.17317. Available from: https://doi.org/10.1038/s41598-017-17434-7.

Peñagaricano, F. 2019. Genetics and genomics of dairy cattle *In*: F. W. Bazer, G. C. Lamb and G. B. T.-A. A. Wu, eds. *Animal Agriculture: Sustainability, Challenges and Innovations* [Online]. Academic Press, pp.101–119. Available from: https://www.sciencedirect.com/science/article/pii/B9780128170526000069.

Perea, A., Clemente, F., Martinell, J., Villanueva-Penacarrillo, M.L. and Valverde, I. 1995. Physiological effect of glucagon in human isolated adipocytes. *Hormone and Metabolic Research*. **27**(8), pp.372–375.

Perry, E.A., Carr, S.M., Bartlett, S.E. and Davidson, W.S. 1995. A phylogenetic perspective on the evolution of reproductive behavior in pagophilic seals of the northwest Atlantic as indicated by mitochondrial DNA sequences. *Journal of Mammalogy*. [Online]. **76**(1), pp.22–31. Available from: http://www.jstor.org/stable/1382311.

Peto, H., Roe, F.J.C., Lee, P.N., Levy, L. and Clack, J. 1975. Cancer and ageing in mice and men. *British Journal of Cancer*. **32**(4), pp.411–426.

Pfenninger, M. and Posada, D. 2002. *Phylogeographic history of the land snail Candidula unifasciata (Helicellinae, Stylommatophora): Fragmentation, corridor migration, and secondary contact*.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D.T.J., Manuel, M., Wörheide, G. and Baurain, D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*. [Online]. **9**(3), pp.e1000602–e1000602. Available from: https://pubmed.ncbi.nlm.nih.gov/21423652.

Philippe, H., Casane, D., Gribaldo, S., Lopez, P. and Meunier, J. 2003. Heterotachy and functional shift in protein evolution. *IUBMB Life*. **55**(4–5), pp.257–265.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D.J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G. and Manuel, M. 2009. Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Current Biology*. [Online]. **19**(8), pp.706–712. Available from: https://www.sciencedirect.com/science/article/pii/S0960982209008057.

Philippe, H. and Roure, B. 2011. Difficult phylogenetic questions: More data, maybe; better methods, certainly. *BMC Biology*. [Online]. **9**(1), p.91. Available from: https://doi.org/10.1186/1741-7007-9-91.

Piotrowski, E.R., Tift, M.S., Crocker, D.E., Pearson, A.B., Vázquez-Medina, J.P., Keith, A.D. and Khudyakov, J.I. 2021. Ontogeny of Carbon Monoxide-Related Gene Expression in a Deep-Diving Marine Mammal. *Frontiers in Physiology*. [Online]. **12**. Available from: https://www.frontiersin.org/article/10.3389/fphys.2021.762102.

Plassais, J., Kim, J., Davis, B.W., Karyadi, D.M., Hogan, A.N., Harris, A.C., Decker, B., Parker, H.G. and Ostrander, E.A. 2019. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nature Communications*. [Online]. **10**(1), p.1489. Available from: https://doi.org/10.1038/s41467-019-09373-w.

Pollard, D.A., Iyer, V.N., Moses, A.M. and Eisen, M.B. 2006. Widespread discordance of gene trees with species tree in drosophila: Evidence for incomplete lineage sorting. *PLoS Genetics*. [Online]. **2**(10), pp.1634–1647. Available from: https://doi.org/10.1371/journal.pgen.0020173.

Ponganis, P.J., Kooyman, G.L., Sartoris, D. and Jobsis, P. 1992. . Pinniped splenic volumes. *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*. [Online]. **262**(2 31-2), pp.R322–R325. Available from: https://doi.org/10.1152/ajpregu.1992.262.2.R322.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. 2005. InterProScan: Protein domains identifier. *Nucleic Acids Research*. [Online]. **33**(SUPPL. 2), pp.W116–W120. Available from: https://pubmed.ncbi.nlm.nih.gov/15980438.

Rambaut, A., Drummond, A.J., Xie, D., Baele, G. and Suchard, M.A. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*. [Online]. **67**(5), pp.901–904. Available from: https://doi.org/10.1093/sysbio/syy032.

Raven, L.A., Cocks, B.G. and Hayes, B.J. 2014. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics*. [Online]. **15**(1), p.62. Available from: https://pubmed.ncbi.nlm.nih.gov/24456127.

Ray, C.E. 1976. Geography of phocid evolution. *Systematic Zoology*. **25**(4), pp.391–406.

Reeves, J.H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *Journal of Molecular Evolution*. **35**(1), pp.17–31.

Reich, C.M. and Arnould, J.P.Y. 2007. Evolution of Pinnipedia lactation strategies: A potential role for α-lactalbumin? *Biology Letters*. [Online]. **3**(5), pp.546–549. Available from: http://dx.doi.org/10.1098/rsbl.2007.0265.

Reuter, J.A., Spacek, D. V. and Snyder, M.P. 2015. High-Throughput Sequencing Technologies. *Molecular Cell*. **58**(4), pp.586–597.

Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B.J., Chaisson, M., Gedman, G.L., Cantin, L.J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., Haase, B., Mountcastle, J., Winkler, S., Paez, S., Howard, J., Vernes, S.C., Lama, T.M., Grutzner, F., Warren, W.C., Balakrishnan, C.N., Burt, D., George, J.M., Biegler, M.T., Iorns, D., Digby, A., Eason, D., Robertson, B., Edwards, T., Wilkinson, M., Turner, G., Meyer, A., Kautt, A.F., Franchini, P., Detrich, H.W., Svardal, H., Wagner, M., Naylor, G.J.P., Pippel, M., Malinsky, M., Mooney, M., Simbirsky, M., Hannigan, B.T., Pesout, T., Houck, M., Misuraca, A., Kingan, S.B., Hall, R., Kronenberg, Z., Sović, I., Dunn, C., Ning, Z., Hastie, A., Lee, J., Selvaraj, S., Green, R.E., Putnam, N.H., Gut, I., Ghurye, J., Garrison, E., Sims, Y., Collins, J., Pelan, S., Torrance, J., Tracey, A., Wood, J., Dagnew, R.E., Guan, D., London, S.E., Clayton, D.F., Mello, C. V., Friedrich, S.R., Lovell, P. V., Osipova, E., Al-Ajli, F.O., Secomandi, S., Kim, H., Theofanopoulou, C., Hiller, M., Zhou, Y., Harris, R.S., Makova, K.D., Medvedev, P., Hoffman, J., Masterson, P., Clark, K., Martin, F., Howe, Kevin, Flicek, P., Walenz, B.P., Kwak, W., Clawson, H., Diekhans, M., Nassar, L., Paten, B., Kraus, R.H.S., Crawford, A.J., Gilbert, M.T.P., Zhang, G., Venkatesh, B., Murphy, R.W., Koepfli, K.P., Shapiro, B., Johnson, W.E., Di Palma, F., Marques-Bonet, T., Teeling, E.C., Warnow, T., Graves, J.M., Ryder, O.A., Haussler, D., O'Brien, S.J., Korlach, J., Lewin, H.A., Howe, Kerstin, Myers, E.W., Durbin, R., Phillippy, A.M. and Jarvis, E.D. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. [Online]. **592**(7856), pp.737–746. Available from: https://doi.org/10.1038/s41586-021-03451-0.

Rhie, A., Walenz, B.P., Koren, S. and Phillippy, A.M. 2020. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*. **21**(1), pp.1–27.

Riedman, M. and Ortiz, C.L. 1979. Changes in Milk Composition during Lactation in the Northern Elephant Seal. *Physiological Zoology*. [Online]. **52**(2), pp.240–249. Available from: http://www.jstor.org/stable/30152567.

Roberts, R.J., Carneiro, M.O. and Schatz, M.C. 2013. The advantages of SMRT sequencing. *Genome Biology*. **14**(6), p.405.

Robinson, J.A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B.Y., Vonholdt, B.M., Marsden, C.D., Lohmueller, K.E. and Wayne, R.K. 2016. Genomic Flatlining in the Endangered Island Fox. *Current Biology*. **26**(9), pp.1183–1189.

Rokas, A. and Carroll, S.B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution*. [Online]. **22**(5), pp.1337–1344. Available from: https://doi.org/10.1093/molbev/msi121.

Rokas, A. and Carroll, S.B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution*. [Online]. **22**(5), pp.1337–1344. Available from: https://doi.org/10.1093/molbev/msi121.

Rokas, A. and Carroll, S.B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution*. [Online]. **22**(5), pp.1337–1344. Available from: https://doi.org/10.1093/molbev/msi121.

Rokas, A., Williams, B.I., King, N. and Carroll, S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*. [Online]. **425**(6960), pp.798–804. Available from: https://doi.org/10.1038/nature02053.

Roure, B., Baurain, D. and Philippe, H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution*. [Online]. **30**(1), pp.197–214. Available from: https://doi.org/10.1093/molbev/mss208.

Ruan, J. and Li, H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*. [Online]. **17**(2), pp.155–158. Available from: http://biorxiv.org/content/early/2019/01/26/530972.abstract.

Salkind, N. 2012. *Teoria Statistica Delle Classi e Calcolo Delle Probabilità* [Online]. Seeber. Available from: https://books.google.co.uk/books?id=3CY-HQAACAAJ.

Salkind, N. 2012. Encyclopedia of Research Design. *Encyclopedia of Research Design*. [Online]. Available from: https://methods.sagepub.com/reference/encyc-of-research-design.

Salzberg, S.L. 2019. Next-generation genome annotation: We still struggle to get it right. *Genome Biology*. [Online]. **20**(1), p.92. Available from: https://doi.org/10.1186/s13059-019-1715-2.

Samad, A. 2013. 済無No Title No Title. *Journal of Chemical Information and Modeling*. **53**(9), pp.1689–1699.

Sanchez, M.P., Ramayo-Caldas, Y., Wolf, V., Laithier, C., El Jabri, M., Michenet, A., Boussaha, M., Taussat, S., Fritz, S., Delacroix-Buchet, A., Brochard, M. and Boichard, D. 2019. Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbéliarde cows. *Genetics Selection Evolution*. [Online]. **51**(1), p.34. Available from: https://doi.org/10.1186/s12711-019-0473-7.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M. and Smith, M. 1977. Nucleotide sequence of bacteriophage φx174 DNA. *Nature*. [Online]. **265**(5596), pp.687–695. Available from: https://doi.org/10.1038/265687a0.

Sanger, F. and Coulson, A.R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*. [Online]. **94**(3), pp.441–448. Available from: https://www.sciencedirect.com/science/article/pii/0022283675902132.

Sato, J.J. and Wolsan, M. 2012. Loss or major reduction of umami taste sensation in pinnipeds. *Naturwissenschaften*. [Online]. **99**(8), pp.655–659. Available from: https://pubmed.ncbi.nlm.nih.gov/22777285.

Sato, J.J., Wolsan, M., Minami, S., Hosoda, T., Sinaga, M.H., Hiyama, K., Yamaguchi, Y. and Suzuki, H. 2009. Deciphering and dating the red Panda's ancestry and early adaptive radiation of musteloidea. *Molecular Phylogenetics and Evolution*. **53**(3), pp.907–922.

SAVAGE, R.J.G. 1957. the Anatomy of Potamotherium an Oligocene Lutrine *In*: *Proceedings of the Zoological Society of London*. Wiley Online Library, pp.151–244.

Savriama, Y., Valtonen, M., Kammonen, J.I., Rastas, P., Smolander, O.P., Lyyski, A., Häkkinen, T.J., Corfe, I.J., Gerber, S., Salazar-Ciudad, I., Paulin, L., Holm, L., Löytynoja, A., Auvinen, P. and Jernvall, J. 2018. Bracketing phenogenotypic limits of mammalian hybridization. *Royal Society Open Science*. **5**(11), p.180903.

Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O. and Thompson, J.D. 2020. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics*. [Online]. **21**(1), p.293. Available from: https://pubmed.ncbi.nlm.nih.gov/32272892.

Scheinfeldt, L.B., Soi, S., Thompson, S., Ranciaro, A., Woldemeskel, D., Beggs, W., Lambert, C., Jarvis, J.P., Abate, D., Belay, G. and Tishkoff, S.A. 2012. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biology*. **13**(1), pp.1–9.

Schlicker, A., Domingues, F.S., Rahnenführer, J. and Lengauer, T. 2006. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*. [Online]. **7**(1), p.302. Available from: https://doi.org/10.1186/1471-2105-7-302.

Schlötterer, C., Kofler, R., Versace, E., Tobler, R. and Franssen, S.U. 2015. Combining experimental evolution with next-generation sequencing: A powerful tool to study adaptation from standing genetic variation. *Heredity*. [Online]. **114**(5), pp.431–440. Available from: https://doi.org/10.1038/hdy.2014.86.

Schulz, T.M. and Bowen, W.D. 2004. . Pinniped lactation strategies: Evaluation of data on maternal and offspring life history traits. *Marine Mammal Science*. [Online]. **20**(1), pp.86–114. Available from: http://dx.doi.org/10.1111/j.1748-7692.2004.tb01142.x.

Schumann, G.G., Gogvadze, E. V., Osanai-Futahashi, M., Kuroki, A., Münk, C., Fujiwara, H., Ivics, Z. and Buzdin, A.A. 2010. Unique functions of repetitive transcriptomes *In*: K. W. B. T.-I. R. of C. and M. B. Jeon, ed. *International Review of Cell and Molecular Biology* [Online]. Academic Press, pp.115–188. Available from: https://www.sciencedirect.com/science/article/pii/B9780123810472000037.

Scornavacca, C. and Galtier, N. 2017. Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*. [Online]. **66**(1), pp.112–120. Available from: https://doi.org/10.1093/sysbio/syw082.

Sharp, J.A., Cane, K., Arnould, J.P.Y. and Nicholas, K.R. 2005. The lactation cycle of the fur seal. *Journal of Dairy Research*. **72**(SPEC. ISS.), pp.81–89.

Sharp, J.A., Lefèvre, C. and Nicholas, K.R. 2008. Lack of functional alpha-lactalbumin prevents involution in Cape fur seals and identifies the protein as an apoptotic milk factor in mammary gland involution. *BMC Biology*. [Online]. **6**, p.48. Available from: https://pubmed.ncbi.nlm.nih.gov/18986549.

Shaughnessy, P.D. and Fay, F.H. 1977. A review of the taxonomy and nomenclature of North Pacific Harbour seals. *Journal of Zoology*. **182**(3), pp.385–419.

Sheffield, N.C. 2013. The Interaction between Base Compositional Heterogeneity and Among-Site Rate Variation in Models of Molecular Evolution G. Glöckner, J. Fan, & J. A. Norman, eds. *ISRN Evolutionary Biology*. [Online]. **2013**, pp.1–8. Available from: https://doi.org/10.5402/2013/391561.

Shen, X.X., Hittinger, C.T. and Rokas, A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology and Evolution*. [Online]. **1**(5), p.126. Available from: https://pubmed.ncbi.nlm.nih.gov/28812701.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*. [Online]. **51**(3), pp.492–508. Available from: https://doi.org/10.1080/10635150290069913.

Shimodaira, H. and Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*. [Online]. **16**(8), pp.1114–1116. Available from: https://doi.org/10.1093/oxfordjournals.molbev.a026201.

Shultz, A.J. and Sackton, T.B. 2019. Immune genes are hotspots of shared positive selection across birds and mammals C. R. Landry & D. Tautz, eds. *eLife*. [Online]. **8**, p.e41815. Available from: https://doi.org/10.7554/eLife.41815.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D. and Higgins, D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. **7**, p.539.

Siitonen, N., Pulkkinen, L., Lindström, J., Kolehmainen, M., Eriksson, J.G., Venojärvi, M., Ilanne-Parikka, P., Keinänen-Kiukaanniemi, S., Tuomilehto, J. and Uusitupa, M. 2011. Association of ADIPOQ gene variants with body weight, type 2 diabetes and serum adiponectin concentrations: The Finnish Diabetes Prevention Study. *BMC Medical Genetics*. [Online]. **12**(1), p.5. Available from: https://doi.org/10.1186/1471-2350-12-5.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V. and Zdobnov, E.M. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. [Online]. **31**(19), pp.3210–3212. Available from: https://doi.org/10.1093/bioinformatics/btv351.

Siu-Ting, K., Torres-Sanchez, M., Mauro, D.S., Wilcockson, D., Wilkinson, M., Pisani, D., O'Connell, M.J. and Creevey, C.J. 2019. Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. *Molecular Biology and Evolution*. [Online]. **36**(6), pp.1344–1356. Available from: https://doi.org/10.1093/molbev/msz067.

Smit, A.F.A., Hubley, R. and Green, P. 2016. RepeatMasker Open-4.0. 2015.

Smith, N.G.C. and Eyre-Walker, A. 2002. Adaptive protein evolution in Drosophila. *Nature*. [Online]. **415**(6875), pp.1022–1024. Available from: https://doi.org/10.1038/4151022a.

Springer, M.S. 2013. Phylogenetics: Bats united, microbats divided. *Current Biology*. **23**(22), pp.R999–R1001.

Springer, M.S. and Gatesy, J. 2018. . Pinniped Diphyly and Bat Triphyly: More Homology Errors Drive Conflicts in the Mammalian Tree. *Journal of Heredity*. [Online]. **109**(3), pp.297–307. Available from: https://doi.org/10.1093/jhered/esx089.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. [Online]. **30**(9), pp.1312–1313. Available from: https://doi.org/10.1093/bioinformatics/btu033.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. 2006. AUGUSTUS: A b initio prediction of alternative transcripts. *Nucleic Acids Research*. [Online]. **34**(WEB. SERV. ISS.), pp.W435–W439. Available from: https://doi.org/10.1093/nar/gkl200.

Steenwyk, J.L., Buida, T.J., Labella, A.L., Li, Y., Shen, X.X. and Rokas, A. 2021. PhyKIT: A broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics*. [Online]. **37**(16), pp.2325–2331. Available from: https://doi.org/10.1093/bioinformatics/btab096.

Steenwyk, J.L., Buida, T.J., Li, Y., Shen, X.X. and Rokas, A. 2020. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biology*. [Online]. **18**(12), p.e3001007. Available from: https://doi.org/10.1371/journal.pbio.3001007.

Stoffel, M.A., Humble, E., Paijmans, A.J., Acevedo-Whitehouse, K., Chilvers, B.L., Dickerson, B., Galimberti, F., Gemmell, N.J., Goldsworthy, S.D., Nichols, H.J., Krüger, O., Negro, S., Osborne, A., Pastor, T., Robertson, B.C., Sanvito, S., Schultz, J.K., Shafer, A.B.A., Wolf, J.B.W. and Hoffman, J.I. 2018. Demographic histories and genetic diversity across pinnipeds are shaped by human exploitation, ecology and life-history. *Nature Communications*. [Online]. **9**(1), p.4836. Available from: https://doi.org/10.1038/s41467-018-06695-z.

Stoler, N. and Nekrutenko, A. 2021. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics*. [Online]. **3**(1). Available from: https://doi.org/10.1093/nargab/lqab019.

Stoletzki, N. 2008. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. *BMC Evolutionary Biology*. **8**(1), p.224.

Storer, J., Hubley, R., Rosen, J., Wheeler, T.J. and Smit, A.F. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*. [Online]. **12**(1), p.2. Available from: https://doi.org/10.1186/s13100-020-00230-y.

Strandberg, U., Käkelä, A., Lydersen, C., Kovacs, K.M., Grahl-Nielsen, O., Hyvärinen, H. and Käkelä, R. 2008. Stratification, composition, and function of marine mammal blubber: The ecology of fatty acids in marine mammals. *Physiological and Biochemical Zoology*. **81**(4), pp.473–485.

Strimmer, K. and Von Haeseler, A. 1997. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*. [Online]. **94**(13), pp.6815–6819. Available from: http://www.pnas.org/content/94/13/6815.abstract.

Supek, F., Bošnjak, M., Škunca, N. and Šmuc, T. 2011. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*. [Online]. **6**(7), p.e21800. Available from: https://doi.org/10.1371/journal.pone.0021800.

Susko, E. 2009. Bootstrap support is not first-order correct. *Systematic Biology*. **58**(2), pp.211–223.

Susko, E. and Roger, A.J. 2007. On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution*. **24**(9), pp.2139–2150.

Susko, E. and Roger, A.J. 2007. On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution*. **24**(9), pp.2139–2150.

Swofford, D.L., Waddell, P.J., Huelsenbeck, J.P., Foster, P.G., Lewis, P.O. and Rogers, J.S. 2001. Bias in Phylogenetic Estimation and Its Relevance to the Choice between Parsimony and Likelihood Methods. *Systematic Biology*. [Online]. **50**(4), pp.525–539. Available from: https://doi.org/10.1080/10635150117959.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. [Online]. **123**(3), pp.585–595. Available from: https://pubmed.ncbi.nlm.nih.gov/2513255.

Tang, B., Zhou, Q., Dong, L., Li, W., Zhang, X., Lan, L., Zhai, S., Xiao, J., Zhang, Z., Bao, Y., Zhang, Y.P., Wang, G.D. and Zhao, W. 2019. IDog: An integrated resource for domestic dogs and wild canids. *Nucleic Acids Research*. [Online]. **47**(D1), pp.D793–D800. Available from: https://doi.org/10.1093/nar/gky1041.

Tarver, J.E., Dos Reis, M., Mirarab, S., Moran, R.J., Parker, S., O'Reilly, J.E., King, B.L., O'Connell, M.J., Asher, R.J., Warnow, T., Peterson, K.J., Donoghue, P.C.J. and Pisani, D. 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biology and Evolution*. [Online]. **8**(2), pp.330–344. Available from: http://dx.doi.org/10.1093/gbe/evv261.

Tedford, R.H., Barnes, L.G. and Ray, C.E. 1994. The early Miocene littoral ursoid Carnivoran Kolponomos: systematics and mode of life *In*: *Proceedings - San Diego Society of Natural History*., pp.11–32.

Tedford, R.H. 1976. Relationship of pinnipeds to other carnivores (mammalia). *Systematic Zoology*. **25**(4), pp.363–374.

Tedman, R.A. 1983. Ultrastructural distribution of morphology of the mammary gland with observations on the size fat droplets in milk of the Weddell seal Leptonychotes weddelli (Pinnipedia). *Journal of Zoology*. [Online]. **200**(1), pp.131–141. Available from: https://doi.org/10.1111/j.1469-7998.1983.tb06113.x.

Tenaillon, O., Barrick, J.E., Ribeck, N., Deatherage, D.E., Blanchard, J.L., Dasgupta, A., Wu, G.C., Wielgoss, S., Cruveiller, S., Médigue, C., Schneider, D. and Lenski, R.E. 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*. [Online]. **536**(7615), pp.165–170. Available from: https://pubmed.ncbi.nlm.nih.gov/27479321.

Teng, M.S., Wu, S., Er, L.K., Hsu, L.A., Chou, H.H. and Ko, Y.L. 2018. LIPC variants as genetic determinants of adiposity status, visceral adiposity indicators, and triglyceride-glucose (TyG) index-related parameters mediated by serum triglyceride levels. *Diabetology and Metabolic Syndrome*. [Online]. **10**(1), p.79. Available from: https://doi.org/10.1186/s13098-018-0383-9.

The C. elegans Sequencing Consortium* 1998. Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science*. [Online]. **282**(5396), pp.2012–2018. Available from: http://science.sciencemag.org/content/282/5396/2012.abstract.

Thibaud-Nissen, F., DiCuccio, M., Hlavina, W., Kimchi, A., Kitts, P.A., Murphy, T.D., Pruitt, K.D. and Souvorov, A. 2016. P8008 The NCBI Eukaryotic Genome Annotation Pipeline. *Journal of Animal Science*. [Online]. **94**(suppl_4), pp.184–184. Available from: https://doi.org/10.2527/jas2016.94supplement4184x.

Thomas, G.W.C., Hahn, M.W. and Hahn, Y. 2017. The effects of increasing the number of taxa on inferences of molecular convergence. *Genome Biology and Evolution*. [Online]. **9**(1), pp.213–221. Available from: https://pubmed.ncbi.nlm.nih.gov/28057728.

Thompson, J.D., Thierry, J.C. and Poch, O. 2003. RASCAL: Rapid scanning and correction of multiple sequence alignments. *Bioinformatics*. **19**(9), pp.1155–1161.

Thompson, J.D., Plewniak, F. and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*. [Online]. **27**(13), pp.2682–2690. Available from: https://doi.org/10.1093/nar/27.13.2682.

Thompson, J.D., Plewniak, F., Ripp, R., Thierry, J.C. and Poch, O. 2001. Towards a reliable objective function for multiple sequence alignments. *Journal of Molecular Biology*. **314**(4), pp.937–951.

Tian, R., Yin, D., Liu, Y., Seim, I., Xu, S. and Yang, G. 2017. Adaptive evolution of energy metabolism-related genes in hypoxia-tolerant mammals. *Frontiers in Genetics*. [Online]. **8**(DEC), p.205. Available from: https://pubmed.ncbi.nlm.nih.gov/29270192.

Tollis, M., Robbins, J., Webb, A.E., Kuderna, L.F.K., Caulin, A.F., Garcia, J.D., Bèrubè, M., Pourmand, N., Marques-Bonet, T., O'Connell, M.J., Palsbøll, P.J., Maley, C.C. and Shapiro, B. 2019. Return to the Sea, Get Huge, Beat Cancer: An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). *Molecular Biology and Evolution*. [Online]. **36**(8), pp.1746–1763. Available from: https://doi.org/10.1093/molbev/msz099.

Tsagkogeorga, G., McGowen, M.R., Davies, K.T.J., Jarman, S., Polanowski, A., Bertelsen, M.F. and Rossiter, S.J. 2015. A phylogenomic analysis of the role and timing of molecular adaptation in the aquatic transition of cetartiodactyl mammals. *Royal Society Open Science*. **2**(9), p.150156.

Uhen, M.D. 2007. Evolution of marine mammals: Back to the sea after 300 million years. *Anatomical Record*. [Online]. **290**(6), pp.514–522. Available from: https://doi.org/10.1002/ar.20545.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. and DePristo, M.A. 2013. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*. **43**(SUPL.43), pp.10–11.

Vanderpool, D., Minh, B.Q., Lanfear, R., Hughes, D., Murali, S., Alan Harris, R., Raveendran, M., Muzny, D.M., Hibbins, M.S., Williamson, R.J., Gibbs, R.A., Worley, K.C., Rogers, J. and Hahn, M.W. 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS Biology*. [Online]. **18**(12), p.e3000954. Available from: https://doi.org/10.1371/journal.pbio.3000954.

Vaser, R., Sović, I., Nagarajan, N. and Šikić, M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*. [Online]. **27**(5), pp.737–746. Available from: http://genome.cshlp.org/content/early/2017/01/18/gr.214270.116.abstract.

Velando, A., Beamonte-Barrientos, R. and Torres, R. 2006. Pigment-based skin colour in the blue-footed booby: An honest signal of current condition used by females to adjust reproductive investment. *Oecologia*. [Online]. **149**(3), pp.535–542. Available from: https://doi.org/10.1007/s00442-006-0457-5.

Vinet, L. and Zhedanov, A. 2011. A 'missing' family of classical orthogonal polynomials. *Journal of Physics A: Mathematical and Theoretical*. **44**(8), pp.1689–1699.

von Holdt, B., Fan, Z., Ortega-Del Vecchyo, D. and Wayne, R.K. 2017. EPAS1 variants in high altitude Tibetan wolves were selectively introgressed into highland dogs. *PeerJ*. [Online]. **2017**(7), pp.e3522–e3522. Available from: https://pubmed.ncbi.nlm.nih.gov/28717592.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J. and Schatz, M.C. 2017. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*. [Online]. **33**(14), pp.2202–2204. Available from: https://doi.org/10.1093/bioinformatics/btx153.

Wang, H.C., Minh, B.Q., Susko, E. and Roger, A.J. 2018. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Systematic Biology*. [Online]. **67**(2), pp.216–235. Available from: https://doi.org/10.1093/sysbio/syx068.

Webb, A.E., Walsh, T.A. and O'Connell, M.J. 2017. VESPA: Very large-scale evolutionary and selective pressure analyses. *PeerJ Computer Science*. [Online]. **2017**(6), p.e118. Available from: https://peerj.com/articles/cs-118.

Weber, M. 2011. *Die Säugetiere. Einführung in die Anatomie und Systematik der recenten und fossilen Mammalia.* Fischer.

Weisman, C., Murray, A.W. and Eddy, S.R. 2022. Mixing Genome Annotation Methods in a Comparative Analysis Inflates the Apparent Number of Lineage-Specific Genes. *SSRN Electronic Journal*. [Online], 2022.01.13.476251. Available from: http://biorxiv.org/content/early/2022/01/15/2022.01.13.476251.abstract.

Wheatley, K.E., Nichols, P.D., Hindell, M.A., Harcourt, R.G. and Bradshaw, C.J.A. 2008. Differential mobilization of blubber fatty acids in lactating weddell seals: Evidence for selective use. *Physiological and Biochemical Zoology*. **81**(5), pp.651–662.

Wheeler, T.J. 2009. Large-Scale Neighbor-Joining with NINJA BT - Algorithms in Bioinformatics *In*: S. L. Salzberg and T. Warnow, eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp.375–389.

Wick, R.R., Judd, L.M., Gorrie, C.L. and Holt, K.E. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*. **3**(10), p.e000132.

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., Ruhfel, B.R., Wafula, E., Der, J.P., Graham, S.W., Mathews, S., Melkonian, M., Soltis, D.E., Soltis, P.S., Miles, N.W., Rothfels, C.J., Pokorny, L., Shaw, A.J., De Gironimo, L., Stevenson, D.W., Surek, B., Villarreal, J.C., Roure, B., Philippe, H., De Pamphilis, C.W., Chen, T., Deyholos, M.K., Baucom, R.S., Kutchan, T.M., Augustin, M.M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G.K.S. and Leebens-Mack, J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America*. **111**(45), pp.E4859–E4868.

Wiens, J.J. 2001. Character analysis in morphological phylogenetics: Problems and solutions. *Systematic Biology*. **50**(5), pp.689–699.

Wilberg, E.W. 2015. What's in an Outgroup? the Impact of Outgroup Choice on the Phylogenetic Position of Thalattosuchia (Crocodylomorpha) and the Origin of Crocodyliformes. *Systematic Biology*. [Online]. **64**(4), pp.621–637. Available from: https://doi.org/10.1093/sysbio/syv020.

Wilde, C.J., Addey, C.V.P., Bryson, J.M., Finch, L.M.B., Knight, C.H. and Peaker, M. 1998. Autocrine regulation of milk secretion. *Biochemical Society Symposium*. **63**, pp.81–90.

Wilkinson, M., McInerney, J.O., Hirt, R.P., Foster, P.G. and Embley, T.M. 2007. Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends in Ecology and*

Evolution. [Online]. **22**(3), pp.114–115. Available from: https://doi.org/10.1016/j.tree.2007.01.002.

Williams, T.M. and Davis, R.W. 2021. Physiological resiliency in diving mammals: Insights on hypoxia protection using the Krogh principle to understand COVID-19 symptoms. *Comparative Biochemistry and Physiology -Part A : Molecular and Integrative Physiology*. **253**, p.110849.

Wilson, S.C., Dolgova, E., Trukhanova, I., Dmitrieva, L., Crawford, I., Baimukanov, M. and Goodman, S.J. 2016. Breeding behavior and pup development of the Caspian seal, Pusa caspica. *Journal of Mammalogy*. [Online]. **98**(1), pp.143–153. Available from: https://academic.oup.com/jmammal/article/98/1/143/2525933/Breeding-behavior-and-pup-development-of-the.

Wozencraft, W.C. 2019. 18. The Phylogeny of the Recent Carnivora *In*: *Carnivore Behavior, Ecology, and Evolution*. Springer, pp.495–535.

WOZENCRAFT, W.. 1993. Carnivora : Herpestidae, in WILSON, D.E. and D.M. REEDER, 1993. - Mammal Species of the World, a taxonomic and geographic reference. , 1 206 p.

Wu, H. and Chen, Q. 2015. Hypoxia Activation of Mitophagy and Its Role in Disease Pathogenesis. *Antioxidants and Redox Signaling*. [Online]. **22**(12), pp.1032–1046. Available from: https://doi.org/10.1089/ars.2014.6204.

Wyss, A.R. 1987. The Walrus Auditory Region and the Monophyly of pinnipeds. *American Museum Novitates*. **2871**(2871), pp.1–31.

Xi, Z., Liu, L., Rest, J.S. and Davis, C.C. 2014. Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. *Systematic Biology*. **63**(6), pp.919–932.

Xu, S., Jay, A., Brunaldi, K., Huang, N. and Hamilton, J.A. 2013. CD36 enhances fatty acid uptake by increasing the rate of intracellular esterification but not transport across the plasma membrane. *Biochemistry*. **52**(41), pp.7254–7261.

Yamauchi, T., Kamon, J., Ito, Y., Tsuchida, A., Yokomizo, T., Kita, S., Sugiyama, T., Miyagishi, M., Hara, K., Tsunoda, M., Murakami, K., Ohteki, T., Uchida, S., Takekawa, S., Waki, H., Tsuno, N.H., Shibata, Y., Terauchi, Y., Froguel, P., Tobe, K., Koyasu, S., Taira, K., Kitamura, T., Shimizu, T., Nagai, R. and Kadowaki, T. 2003. Cloning of adiponectin receptors that mediate antidiabetic metabolic effects. *Nature*. [Online]. **423**(6941), pp.762–769. Available from: https://doi.org/10.1038/nature01705.

Yang, C., Wang, C., Ye, M., Jin, C., He, W., Wang, F., McKeehan, W.L. and Luo, Y. 2012. Control of lipid metabolism by adipocyte FGFR1-mediated adipohepatic communication during hepatic stress. *Nutrition and Metabolism*. [Online]. **9**(1), p.94. Available from: https://doi.org/10.1186/1743-7075-9-94.

Yang, H., Li, T., Dang, K. and Bu, W. 2018. Compositional and mutational rate heterogeneity in mitochondrial genomes and its effect on the phylogenetic inferences of Cimicomorpha (Hemiptera: Heteroptera). *BMC Genomics*. [Online]. **19**(1), p.264. Available from: https://doi.org/10.1186/s12864-018-4650-9.

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics*. [Online]. **139**(2), pp.993–1005. Available from: http://www.genetics.org/content/139/2/993.abstract.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*. [Online]. **39**(3), pp.306–314. Available from: https://doi.org/10.1007/BF00160154.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution*. [Online]. **15**(5), pp.568–573. Available from: https://doi.org/10.1093/oxfordjournals.molbev.a025957.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*. [Online]. **11**(9), pp.367–372. Available from: https://www.sciencedirect.com/science/article/pii/0169534796100410.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*. [Online]. **24**(8), pp.1586–1591. Available from: https://academic.oup.com/mbe/article/24/8/1586/1103731.

Yang, Z. and Bielawski, J.R. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*. [Online]. **15**(12), pp.496–503. Available from: https://pubmed.ncbi.nlm.nih.gov/11114436.

Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*. **17**(1), pp.32–43.

Yang, Z., Wong, W.S.W. and Nielsen, R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*. [Online]. **22**(4), pp.1107–1118. Available from: https://doi.org/10.1093/molbev/msi097.

Yang, Z. and Zhu, T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America*. **115**(8), pp.1854–1859.

Yates, A.D., Achuthan, P., Akanni, W., Allen, James, Allen, Jamie, Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J.C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C.G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O.G., Janacek, S.H., Juettemann, T., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J.G., Maurel, T., McDowall, M., McMahon, A., Mohanan, S., Moore, B., Nuhn, M., Oheh, D.N., Parker, A., Parton, A.,

Patricio, M., Sakthivel, M.P., Abdul Salam, A.I., Schmitt, B.M., Schuilenburg, H., Sheppard, D., Sycheva, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Flint, B., Frankish, A., Hunt, S.E., Iisley, G., Kostadima, M., Langridge, N., Loveland, J.E., Martin, F.J., Morales, J., Mudge, J.M., Muffato, M., Perry, E., Ruffier, M., Trevanion, S.J., Cunningham, F., Howe, K.L., Zerbino, D.R. and Flicek, P. 2020. Ensembl 2020. *Nucleic Acids Research*. [Online]. **48**(D1), pp.D682–D688. Available from: https://doi.org/10.1093/nar/gkz966.

Yim, H.S., Cho, Y.S., Guang, X., Kang, S.G., Jeong, J.Y., Cha, S.S., Oh, H.M., Lee, Jae Hak, Yang, E.C., Kwon, K.K., Kim, Y.J., Kim, T.W., Kim, W., Jeon, J.H., Kim, S.J., Choi, D.H., Jho, S., Kim, H.M., Ko, J., Kim, H., Shin, Y.A., Jung, H.J., Zheng, Y., Wang, Z., Chen, Y., Chen, M., Jiang, A., Li, E., Zhang, S., Hou, H., Kim, T.H., Yu, L., Liu, S., Ahn, K., Cooper, J., Park, S.G., Hong, C.P., Jin, W., Kim, H.S., Park, C., Lee, K., Chun, S., Morin, P.A., O'Brien, S.J., Lee, H., Kimura, J., Moon, D.Y., Manica, A., Edwards, J., Kim, B.C., Kim, S., Wang, J., Bhak, J., Lee, H.S. and Lee, Jung Hyun 2014. Minke whale genome and aquatic adaptation in cetaceans. *Nature Genetics*. [Online]. **46**(1), pp.88–92. Available from: http://dx.doi.org/10.1038/ng.2835.

Yonezawa, T., Kohno, N. and Hasegawa, M. 2009. The monophyletic origin of sea lions and fur seals (Carnivora; Otariidae) in the Southern Hemisphere. *Gene*. [Online]. **441**(1–2), pp.89–99. Available from: http://dx.doi.org/10.1016/j.gene.2009.01.022.

Young, A.D. and Gillung, J.P. 2020. Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*. [Online]. **45**(2), pp.225–247. Available from: https://doi.org/10.1111/syen.12406.

Yu, G., Wang, L.G., Han, Y. and He, Q.Y. 2012. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*. [Online]. **16**(5), pp.284–287. Available from: https://doi.org/10.1089/omi.2011.0118.

Yu, L., Jin, W., Zhang, X., Wang, D., Zheng, J. song, Yang, G., Xu, S. xia, Cho, S. and Zhang, Y. ping 2011. Evidence for positive selection on the leptin gene in cetacea and Pinnipedia. *PLoS ONE*. [Online]. **6**(10), p.e26579. Available from: https://doi.org/10.1371/journal.pone.0026579.

Yu, L. and Zhang, Y.P. 2006. Phylogeny of the caniform Carnivora: Evidence from multiple genes. *Genetica*. **127**(1–3), pp.65–79.

Yurchenko, A.A., Deniskova, T.E., Yudin, N.S., Dotsev, A. V., Khamiruev, T.N., Selionova, M.I., Egorov, S. V., Reyer, H., Wimmers, K., Brem, G., Zinovieva, N.A. and Larkin, D.M. 2019. High-density genotyping reveals signatures of selection related to acclimation and economically important traits in 15 local sheep breeds from Russia. *BMC Genomics*. [Online]. **20**(3), p.294. Available from: https://doi.org/10.1186/s12864-019-5537-0.

Zhang, J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution*. [Online]. **21**(7), pp.1332–1339. Available from: https://doi.org/10.1093/molbev/msh117.

Zhang, J., Nielsen, R. and Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*. **22**(12), pp.2472–2479.

Zhang, Y.Z., Akdemir, A., Tremmel, G., Imoto, S., Miyano, S., Shibuya, T. and Yamaguchi, R. 2020. Nanopore basecalling from a perspective of instance segmentation. *BMC Bioinformatics*. [Online]. **21**(3), p.136. Available from: https://doi.org/10.1186/s12859-020-3459-0.

Zhang, Y., Proenca, R., Maffei, M., Barone, M., Leopold, L. and Friedman, J.M. 1995. Positional cloning of the mouse obese gene and its human homologue. *Nature*. **374**(6521), p.479.

Zhao, W., Adjei, M., Wang, H., Yangliu, Y., Zhu, J. and Wu, H. 2021. ADIPOR1 regulates genes involved in milk fat metabolism in goat mammary epithelial cells. *Research in Veterinary Science*. [Online]. **137**, pp.194–200. Available from: https://www.sciencedirect.com/science/article/pii/S0034528821001004.

Zhou, X., Seim, I. and Gladyshev, V.N. 2015. Convergent evolution of marine mammals is associated with distinct substitutions in common genes. *Scientific Reports*. [Online]. **5**(1), p.16550. Available from: https://doi.org/10.1038/srep16550.

Zhou, Y. and Rui, L. 2013. Leptin signaling and leptin resistance. *Frontiers of Medicine*. [Online]. **7**(2), pp.207–222. Available from: https://pubmed.ncbi.nlm.nih.gov/23580174.

Zhu, B.H., Xiao, J., Xue, W., Xu, G.C., Sun, M.Y. and Li, J.T. 2018. P_RNA_scaffolder: A fast and accurate genome scaffolder using paired-end RNA-sequencing reads. *BMC Genomics*. [Online]. **19**(1), p.175. Available from: https://doi.org/10.1186/s12864-018-4567-3.

Zhu, K., Zhou, X., Xu, S., Sun, D., Ren, W., Zhou, K. and Yang, G. 2014. The loss of taste genes in cetaceans. *BMC Evolutionary Biology*. [Online]. **14**(1), p.218. Available from: https://doi.org/10.1186/s12862-014-0218-8.

Zou, Z. and Zhang, J. 2016. Morphological and molecular convergences in mammalian phylogenetics. *Nature Communications*. [Online]. **7**(1), p.12758. Available from: https://doi.org/10.1038/ncomms12758.

Zwick, A., Regier, J.C. and Zwickl, D.J. 2012. Resolving Discrepancy between Nucleotides and Amino Acids in Deep-Level Arthropod Phylogenomics: Differentiating Serine Codons in 21-Amino-Acid Models. *PLoS ONE*. [Online]. **7**(11), p.e47450. Available from: https://doi.org/10.1371/journal.pone.0047450.