Phylogenetic foot-printing and family studies to identify genes essential for enamel formation

Georgios Nikolopoulos

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

The University of Leeds

Leeds Institute of Medical Research at St James's

School of Dentistry

August, 2022

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

Acknowledgements

Starting my journey in the field of the genetic study of amelogenesis imperfecta and then investigating the evolutionary history of the associated genes I would not expect to come to the other side feeling so overwhelmed with emotion. My work with the group of Prof Chris Inglehearn under the guidance of Dr Alan Mighell, Dr Mary O'Connell and Dr Claire Smith, who gave me the opportunity to learn and evolve as a person and as a biologist. I want to thank all of them for their support and patience during the course of this project, it is clearly a team effort and without them I wouldn't be able to make it.

I want to thank Chris for his guidance when I started with this project and his support and enthusiasm through all the stages of it and mostly for his patience and understanding. His critical eye and constructive criticism have pushed me to understand genetics better than I thought I could, and his optimism and inquisitive view of research helped me look at research as an outlet for my curiosity and creativity. I want to thank Alan for his insights in the clinical aspect of this project. Without his drive for recruiting affected families this project would not be possible and without his input in phenotypical matters and guidance when interpreting our results. I also want to thank Mary for introducing me to evolutionary biology and for guiding me on my first steps in this field, steps of which I hope that more will follow. Her energy and optimism, as well as her pushing me to improve and look at all levels of evolution and to not be satisfied with the initial results have made an impression which will stay with me.

Most of all I would like to thank Claire for being present when I needed her, as a mentor and as a friend, to help, guide and occasionally ~~chastise~~ encourage me, to be inquisitive and to always search for more answers. She taught me to never give up in days of disappointment and to always Be Curious, to look for and find out about how everything works.

I would like to thank the School of Dentistry, a scholarship from which funded this project and our collaborators for their contribution in time and knowledge, Steve Brookes for the analysis of the tooth microstructure, Dr Sarah Harris for the molecular dynamics simulations and Prof Per Palsboll for the help with the sequencing of minke whale samples.

A very big thank you goes to my lab mates during these four years, Jo, Sunayna, Ben, Ummey, James and the other members of the VRG and level 8 for making the lab such a warm and welcoming place. I will truly miss WTBB level 8. And of course Ewa, without her tough love we would all literally fall apart. Also, I would like to thank the members of the O'Connell lab, Peter, David, Ali, Isabel and Bede, for their insight that helped me overcome so many problems when conducting my evolutionary analyses and for a couple pints of beer we shared that also had the same effect.

I would finally like to thank my family who introduced me to research and the scholarly ways and were always enthusiastically supporting my endeavors. I know I inherited my hunger for knowledge from them and I intend to make them proud.

Last but not least, I would like to thank my partner, Katerina, for always being by my side, through thick and thin, to support me with her kindness and for being my home away from home.

Victoria Concordia Crescit

Abstract

Amelogenesis Imperfecta (AI) is a Mendelian inherited rare disease, affecting the formation of enamel during tooth development. Mutations in various genes have been implicated in AI, however, even after whole exome sequencing (WES) the genetic basis of the disease remains unexplained in between a third and half of AI patients. This project aimed to identify new genes mutated in non-syndromic AI by combining WES in unsolved AI families with the study of the evolutionary history of genes associated with AI. Specifically, whether selective pressure during the genes' evolution created a pattern of selection shared among them, that could provide further evidence to support new candidate genes and variants implicated in AI.

The WES of individuals from 33 families, presenting with various AI phenotypes resulted in the identification of 25 variants causing AI. Unrelated families sharing variants were further examined for phenotype overlap and where unrelated families shared variants, these were tested to determine whether there was a common ancestral founder. The unexpectedly high level of diagnostic success obtained in the cohort of families was probably due to the selection of families with clear patterns of dominant or consanguineous recessive inheritance.

It is proposed that genes that co-evolved and cooperate to form tooth enamel in mammals should exhibit similar substitution patterns. Loss of enamel or teeth occurred independently multiple times in placental mammals, so representatives of both character states across the phylogeny were included in the analysis. Using molecular evolutionary tools (i.e.: codeml and SLAC), the variation in selective pressure was calculated by examining the substitution patterns in protein-coding sequences of these genes. Positive selection was found to act on genes active during amelogenesis, while genes with broader functions did not exhibit detectable levels of positive selection. Toothless and enamel-less species showed signatures of species-specific positive selection in AI associated genes, indicating a potential shift of function in these species. Surprisingly these genes had been considered pseudogenised, yet a high level of sequence conservation is evident, suggestive of a selective constraint. Future work on these regions is required to determine whether they are producing functional protein sequences.

Table of contents

List of tables

List of figures

Abbreviations

ACP4	:	Acid Phosphatase 4
AD	:	Autosomal Dominant
AI	:	Amelogenesis Imperfecta
AMBN	:	Ameloblastin
AMELX	:	Amelogenin, X-linked
AMELY	:	Amelogenin, Y-linked
AMTN	:	Amelotin
AR	:	Autosomal Recessive
AQP4	:	Aquaporin 4
CADD	:	Combined Annotation Dependent Depletion
CMT1A	:	Charcot-Marie-tooth disease type 1A
CNV	:	Copy Number Variation
COL17A1	:	Collagen Type XVII Alpha 1 chain
DLX3	:	Distal-less Homeobox 3
DNA	:	deoxyribonucleic acid
dNTPs	:	deoxynucleotide triphosphates
ddNTPs	:	dideoxynucleotide triphosphates
DSTNP2	:	Destrin-2 Pseudogene
EBP	:	Earth Biogenome Project
EDTA	:	Ethylenediaminetetraacetic acid
ENAM	:	Enamelin
ERS	:	Enamel Renal Syndrome
ExoSAP-IT	:	Exonuclease I and Shrimp Alkaline Phosphatase in buffer
FAM20A	:	Family with sequence similarity 20, member A
FAM20C	:	Family with sequence similarity 20, member C
FAM83H	:	Family with sequence similarity 83, member H
FDR	:	False Discovery Rate
FTR	:	Functional Translational Readthrough
GPR68	:	G protein-coupled receptor 68
HA	:	Hydroxy(l)apatite
HS	:	Heimler Syndrome

HyPhy	:	Hypothesis testing using Phylogenies
IGV	:	Integrative Genomics Viewer
ITGB6	:	Integrin subunit beta 6
JEB	:	Junctional Epidermolysis Bullosa
KLK4	:	Kallikrein related peptidase 4
LAMA3	:	Laminin subunit alpha 3
LAMB3	:	Laminin subunit beta 3
LAMC2	:	Laminin subunit gamma 2
LOVD	:	Leiden Open Variation Database
LRT	:	Likelihood Ratio Test
MCS	:	Multiple Cloning Site
MGI	:	Mouse Genome Informatics database
MIH	:	Molar-Incisor Hypomineralisation
MK	:	McDonald and Kreitman test
ML	:	Maximum Likelihood
MMP20	:	Matrix Metallopeptidase 20
MOI	:	Mode of Inheritance
MP	:	Maximum Parsimony
MSA	:	Multiple Sequence Alignment
NAP1L4P1	:	Nucleosome Assembly Protein 1 Like 4, Pseudogene 1
NGS	:	Next Generation Sequencing
NJ	:	Neighbour Joining
ODAPH	:	Odontogenesis Associated Phosphoprotein
Ori	:	Origin of replication
PAML	:	Phylogenetic Analysis by Maximum Likelihood
PCR	:	Polymerase Chain Reaction
PEX	:	Peroxin gene
PROVEAN	:	Protein Variation Effect Analyzer
PTCs	:	Premature Termination Codons
RELT	:	Receptor Expressed in Lymphoid Tissues
RHO	:	Rhodopsin
RNA	:	ribonucleic acid

RNS	:	Raine Syndrome
sdH2O	:	sterile – distilled water
SDM	:	Site Directed Mutator
SECIS	:	Selenocysteine insertion sequence element
SEM	:	Scanning Electron Microscope
SLAC	:	Single-Likelihood Ancestor Counting
SLC24A4	:	Solute carrier family 24 (sodium/potassium/calcium exchanger) member A4
smMIPs	:	single-molecule Molecular Inversion Probes
SNP	:	Single Nucleotide Polymorphism
SP6	:	SP6 transcription factor
SV	:	Structural Variation
TDO	:	Trichodentoosseous syndrome
TR	:	Translational Recoding
TUBA4A	:	Tubulin Alpha 4a
vcf	:	Variable Call Format
Vep	:	Variant effect predictor
WDR72	:	WD Repeat Domain 72
WES	:	Whole Exome Sequencing
WGS	:	Whole Genome Sequencing

Chapter 1 - Introduction

1.1 Introduction to the project

The purpose of this project is to study the genetics of amelogenesis imperfecta (AI), a rare inherited disease of tooth enamel. As detailed in section 1.2.6.2, AI is a heterogeneous group of enamel phenotypes that affect all the teeth of a person, in both deciduous and permanent dentitions. It results in pain, with difficulty in mastication, poor social aesthetics and is a cause of distress (Coffield et al., 2005), with a significant impact on the quality of life of the affected. Improving our understanding of the genetics of AI allows better options for counselling of patients, helping them understand their condition better. Additionally, knowing the genetic basis of the condition aids in providing more specialised treatment options to the patients. Even after utilising next generation sequencing (NGS) techniques, only about 50 % of the patients can get a result, while the rest remain unsolved (Gadhia et al., 2012; Smith, Poulter, et al., 2017). To that end, a cohort of AI patients, recruited by collaborators in dental clinics across the UK and abroad, were examined and selected samples were used for whole exome sequencing, to attempt to identify the genetic variant that is the cause of the observed AI phenotype.

Concurrently with the genetic analysis, an analysis of the molecular evolution of genes that have been associated with AI was conducted. By comparing the signatures of natural selection on the sequences of these genes across the mammalian lineage it is possible to examine the selective pressure acting on them and attempt to discern a pattern of evolution shared among them. We hypothesise that genes that cooperate during amelogenesis have co-evolved, being affected by similar evolutionary pressures that would lead to them sharing signatures of natural selection that would allow us to distinguish candidate variants and candidate genes that are associated with amelogenesis and with AI from the other variants and genes identified during the NGS analyses. The evolutionary history of all genes associated with non-syndromic AI is examined, taking into consideration representative species from all major clades of the mammalian lineage, as will be discussed in section 1.4.

1.2 Formation of teeth and the enamel layer

1.2.1 Evolutionary origins of teeth

Teeth are a specialised mineralised structure and are the hardest and most mineralised tissue in human and other mammalian bodies (Smith, 1998). They are primarily used for mastication, while in carnivorous animals they are also used to grab and cut prey and are also tools for defence in both herbivores and carnivores. The evolution of teeth is not yet clearly defined and there are two major competing hypotheses known as the “outside-in” and the “inside-out” hypothesis for the origin of teeth.

The “outside-in” hypothesis proposes that teeth originated as dermal denticles on the skin close to the jaws that slowly migrated to the oral cavity and became specialised structures for mastication (Ørvig, 1967; Blais et al., 2011). This theory is supported by the morphological similarity of teeth to the placoid scales of sharks and other chondrichthyans, as well as the broader homology that is suggested among odontodes, denticles and teeth (Sire and Huysseune, 2003). A significant lack of fossil evidence of the expected gradual transitional

forms of the first migrating denticles and teeth has been the primary argument against the "outside-in" hypothesis, although structures that can be considered as transitional between the two forms have been identified in both chondrichthyans (Miller et al., 2003) and osteichthyans (Botella et al., 2007).

On the contrary, the "inside-out" hypothesis states that teeth originated from denticles appearing in the oro-pharyngeal cavity and migrating to the jaws, independently from the dermal denticles that differentiated to scales in fish (Smith and Coates, 1998). This theory is supported by paleontological data, showing the presence of internal arrays of denticles in the fossil of a jawless vertebrate, the thelodont, *Loganellia scotica* (Der Brugghen and Janvier, 1993), as well as from studies that show that mutant zebrafish (*Danio rerio*) that do not develop branchial arches, do develop teeth (Schilling et al., 1996). Branchial arches are found in all vertebrate embryos and in jawed fish. The first arch develops into the jaws, so mutants that do not form branchial arches are effectively jawless. The formation of teeth in jawless mutants supports the theory that jaws and teeth emerged separately in vertebrates, with the denticles appearing in the pharyngeal cavity, independent of the presence or absence of the jaw.

1.2.2 Comparative biology of mammalian teeth

Most mammals, including human, have diphyodont dentitions, with two successive sets of teeth, where the deciduous or primary teeth are replaced by the permanent or adult teeth. The deciduous teeth are developed during embryonic development and erupt during infancy. Exceptions to this are elephants (species of the family Elephantidae), kangaroos (species of the family Macropodidae) and manatees (species of the genus *Trichechus*), which are polyphyodonts, with new teeth replacing the damaged older teeth. Elephants are able to use up to six sets of molars, per quadrant through their lifetime, while manatees have a seemingly unlimited number (Steenkamp, 2021).

The classes of teeth that are found in the mammalian dentition are the incisors, canines, premolars and molars, with each tooth class being differentiated to better adapt for a specific need (Figure 1.1). For example, the incisors were adapted to be used primarily for cutting food, while canines were developed to better grip food prior to cutting. Molars were adapted for chewing and premolars are an intermediary type between canines and molars.

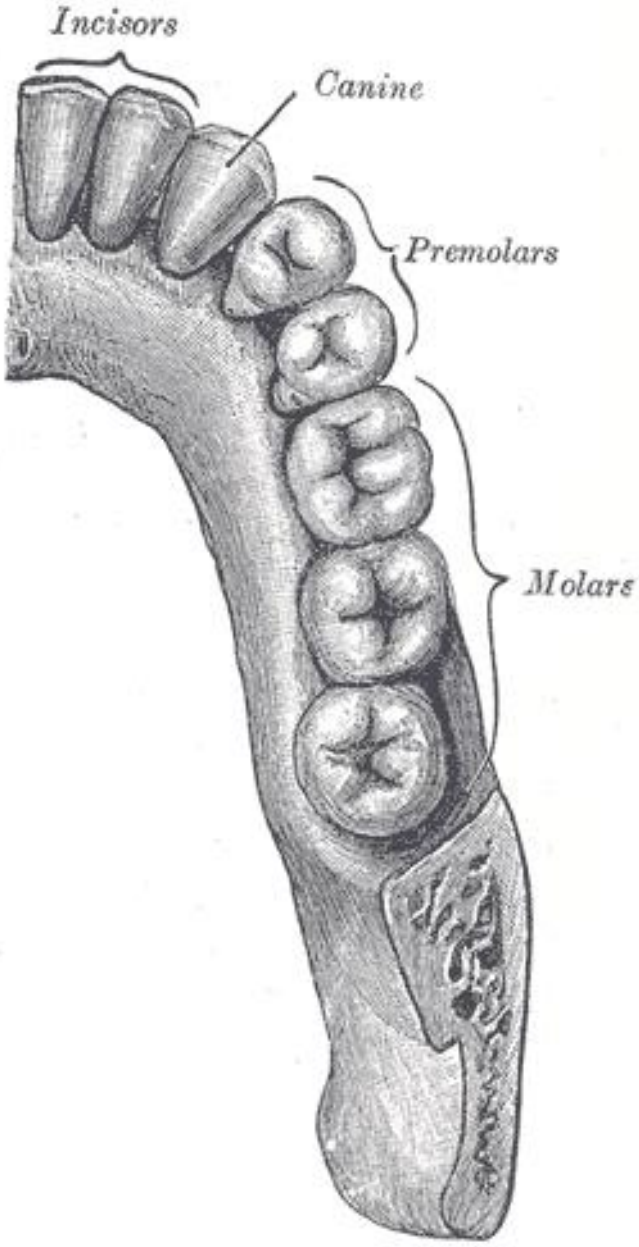


Figure 1.1: Permanent teeth on the human jaw.
Each quadrant of the human dentition typically contains two incisors, a canine, two premolars and up to three molars in that order, starting from the front of the mouth. Image source: Gray, H. (1918). Anatomy of the human body, plate 997.

The dentitions of different species differ in their makeup and number of each tooth class, depending on the needs and the diet of the species. As a result, not all tooth classes are found in all mammalian dentitions. For example, in nine-banded armadillos (*Dasypus novemcinctus*) incisors are formed in the embryo, but degenerate and are absent at birth, while cows (*Bos taurus*) only have incisors on the bottom jaw. Rodents lack canines and premolars. In elephants, the incisors are elongated forming tusks, while in pigs (*Sus scrofa domesticus*) and walruses (*Odobenus rosmarus*) the tusks are formed by elongated canines (Nasoori, 2020). Among Cetacean dentitions great variability is found in the number and function of teeth; narwhals (*Monodon monoceros*) have only two teeth, in males one of them elongating and piercing the lip to form an external tusk (Nweeia et al., 2012), sperm whales (*Physeter macrocephalus*) have up to forty teeth in their bottom jaw and no functional teeth in the upper jaw and the common dolphin (*Delphinus delphis*) can have up to 268 teeth. Some species have lost their teeth altogether, e.g.: anteaters (suborder Vermilingua), the pangolins (suborder Eupholidota) and the baleen whales (suborder Mysticeti), have alternative means of feeding with the tongue being the primary tool for anteaters and pangolins and of course the baleen in the Mysticeti.

1.2.3 Tooth Development

Tooth development can be roughly divided in five successive stages, the initiation stage, the bud stage, the cap stage, the bell stage and the advanced bell stage when tissue hardening and root formation occurs (Figure 1.2). These stages are then followed by the tooth eruption.

During the initiation stage signalling molecules such as the bone morphogenetic protein (BMP), wingless, notch, sonic hedgehog (SHH) and fibroblast growth factor (FGF), that determine the position and type of the tooth, gather at the oral epithelium (Li et al., 2013).

During the bud stage the dental lamina becomes invaginated into the underlying mesenchyme (Figure 1.2a).

The cap stage that follows is characterised by the formation of a cap derived from the epithelial cells that moved into the mesenchyme in the bud stage, this cap will become the enamel organ (Figure 1.2b). Within the enamel organ, some of the epithelial cells differentiate to the star-shaped cells that form the stellate reticulum. Enclosed by the newly formed cap is a condensed group of mesenchymal cells that are now called the dental papilla. The structure of the cap and dental papilla is surrounded by the dental follicle and an enamel knot is formed, which is a cluster of non-dividing epithelial cells that will control the signals that guide the formation of the developing tooth (Catón and Tucker, 2009).

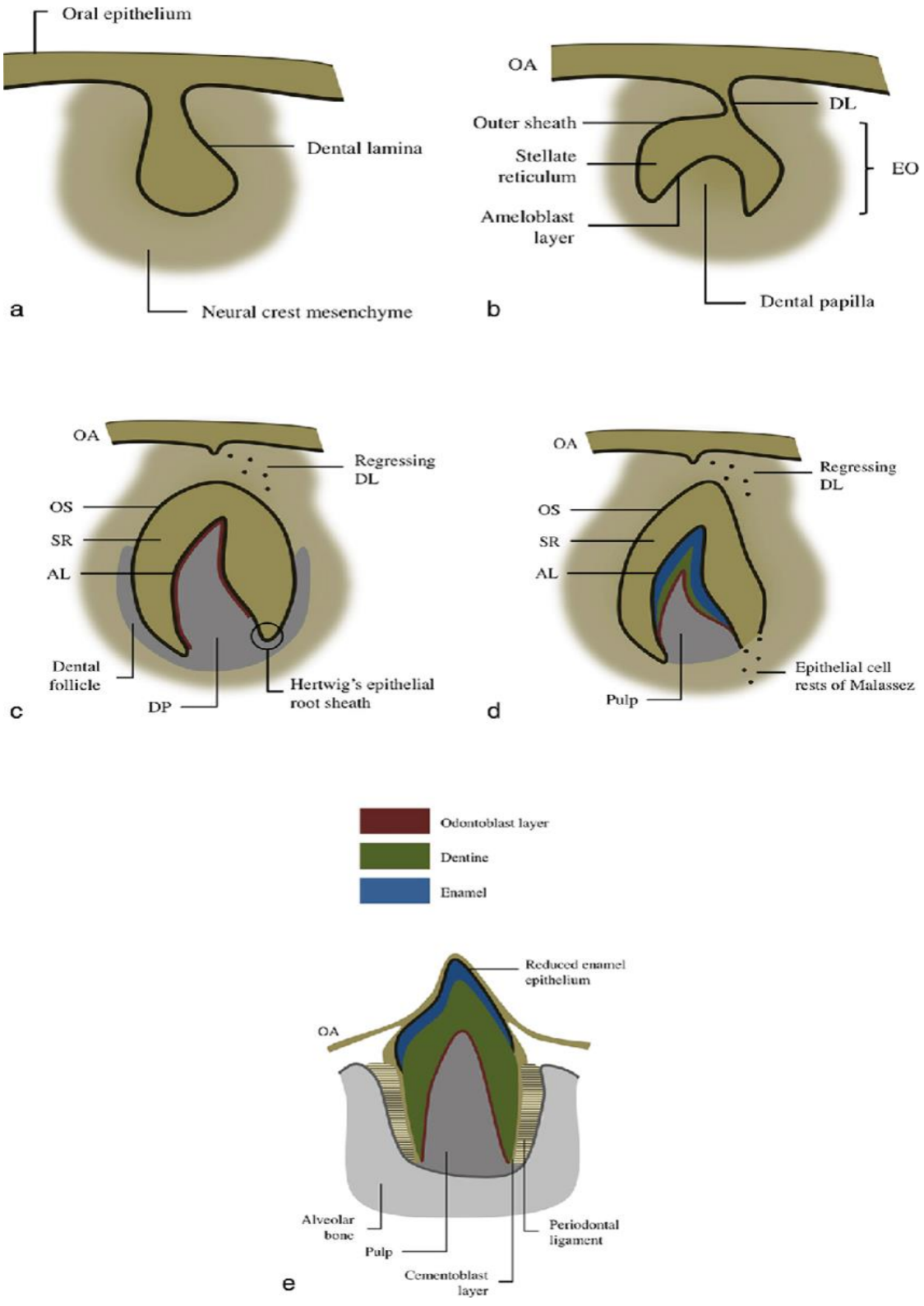


Figure 1.2: The stages of tooth development.

(a) bud stage, (b) cap stage, (c) bell stage, (d) advanced-bell stage, (e) maturation stage. AL: ameloblast layer; DL: dental lamina; DP: dental papilla; EO: enamel organ; OA: oral epithelium; OS: outer sheath; SR: stellate reticulum. Image credit: Lopes Dias et al. (2016), CC-BY-SA 4.0

At the bell stage the enamel organ develops further, pushing into the dental papilla, while the dental lamina starts to degrade (Figure 1.2c, d). The layer of epithelial cells that borders with the dental papilla is called the inner enamel epithelium (IEE). Cells from IEE will differentiate further to form the stratum intermedium (SI), while other cells prompted by signals from the enamel knot and the dental papilla will differentiate to become the preodontoblasts and preameloblasts (Tompkins, 2006). The two cell types start exchanging signals that lead to their final differentiation to odontoblasts and ameloblasts. The odontoblasts start migrating towards the middle of the dental papilla while secreting a pre-dentine matrix, and the ameloblasts move away from that secreted matrix, while they also start to secrete the proteins that will form the enamel matrix (Nanci, 2017). The boundary that is formed among the newly secreted dentine and enamel layers is now called the dentine-enamel junction.

Root formation starts after the formation of the dentine and the enamel of the tooth crown (Figure 1.2e) and it in turn initiates tooth eruption (Nanci, 2017). At eruption, any remaining ameloblasts fuse with the surrounding oral epithelium and degenerate, making room for the new tooth to erupt. The remaining oral epithelium forms the junctional epithelium (Bosshardt and Lang, 2005).

1.2.4 Tooth Structure

Human teeth, like most mammal teeth, consist of four layers of tissues, the mineralised layers of the enamel, the dentine and the cementum and the non-mineralised pulp tissue (Jheon et al., 2013).

Most of the tissues composing a tooth are mineralised, as a result of an important process in the formation of teeth, called biomineralisation (Simmer and Fincham, 1995; Sharma et al., 2021). This term describes the process of forming tissues that are hardened by depositing inorganic minerals in an orderly manner, interlaced with organic material in varying amounts, guided by specialised signalling molecules. The minerals commonly found in biomineralized tissues consist of calcium ions and phosphates, forming hydroxyapatite (HA), also sometimes called hydroxylapatite, with the chemical formula: $Ca_5(PO_4)_3OH$. However, HA crystals are found in units formed by dyads of HA so the formula is more commonly written as $Ca_{10}(PO_4)_6(OH)_2$ to denote that.

When describing the structure of a typical mammalian tooth, the top of the tooth that protrudes from the gums is called the crown, the part that is set in the jawbone is called the root of the tooth and the part that connects the two extremities and is surrounded by the gingiva, or the gum, is called the cervical region (Figure 1.3)(Ungar, 2010). The raised points at the crown of the tooth are called cusps and can range from one on canines, to five on the first molar, with only incisors having no cusps.

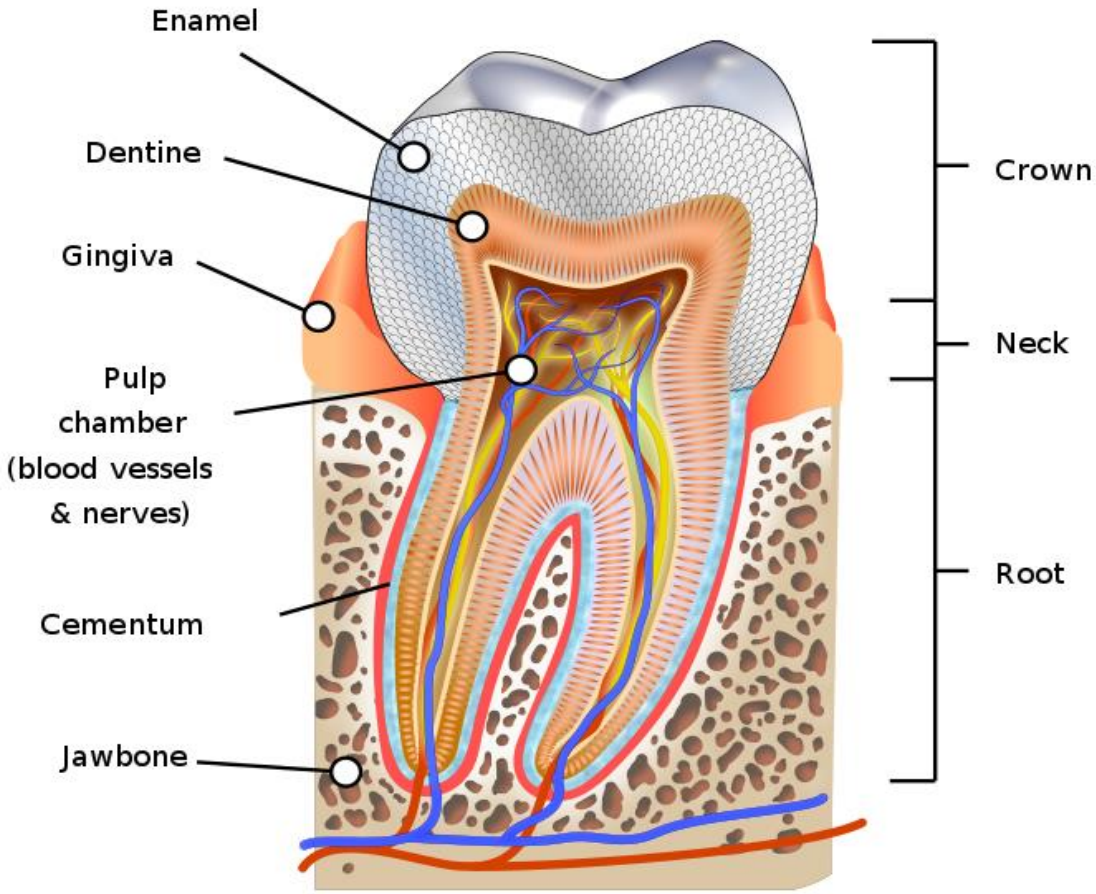


Figure 1.3: Diagram of a human tooth embedded in the jaw.
The different parts of the tooth and the tissues that form it are indicated. Image credit: Human tooth diagram-en.svg from Wikimedia Commons by K. D. Schroeder, CC-BY-SA 4.0

Of the tissues involved, the pulp tissue, or pulp chamber, is the innermost part of a tooth and it contains the blood vessels and nerves that keep the tooth alive. These blood vessels are responsible for the supply of oxygen and nutrients to the tooth, and the nerves provide sensory information for the feeling of pain, the perception of hot or cold on the tooth and provide sensory feedback to determine appropriate biting strength. The pulp chamber is surrounded and protected by the mineralised dentine, an avascular tissue which consists of 45 % HA, 33 % organic material and 22 % water (Tjäderhane et al., 2009). At the crown of the tooth, the dentine is covered by a layer of enamel and at the root it is covered by the cementum. The cementum has similar consistency to the dentine layer and its main role is to provide the surface that the periodontal ligaments will bind to, to secure the stability of the tooth. The enamel that covers the crown of the tooth is the most mineralised tissue in the human body, consisting of 96 % HA by weight, 3 % water and 1 % organic material (Smith, 1998; Avery et al., 2002), forming a highly organised and complex structure of enamel prisms that is able to withstand the mechanical and chemical stress associated with eating, as in human and other diphyodont species there is no capacity for cellular repair of enamel, so enamel at the point of formation needs to have properties that allow it to last for a lifetime. This is by contrast with dentine where their capacity for cellular deposition of new, reparative dentine at the interface with the dental pulp (Sloan and Smith, 2007).

Other than the crown and root, other terms can be used to describe a specific side of a tooth and are particularly useful when describing the condition of a tooth, as well as various pathogenic phenotypes. Surface descriptors depend on which tooth is being considered, a brief summary of the terminology used to describe a particular surface on a tooth is presented on Table 1.1 below.

Table 1.1: Surface descriptors of teeth.

The tooth and orientation of the described surface determines its description.

Tooth Surface Descriptor	Orientation of tooth surface
Labial Surface	Surface of the tooth towards the lips
Buccal Surface	Surface of the tooth towards the buccal mucosa
Facial Surface	The labial and buccal surfaces are collectively called facial surfaces
Palatal Surface	Surface of the tooth towards the palate
Lingual Surface	Surface of the tooth towards the tongue
Mesial Surface	Surface of tooth towards the midline
Distal Surface	Surface of tooth away from midline
Proximal Surface	The mesial and distal surface are collectively called proximal surface
Incisal Surface	The cutting surface of anterior teeth
Occlusal Surface	The grinding surface of posterior teeth

1.2.5 Odontogenesis

1.2.5.1 Amelogenesis

Each layer of tissue that composes the complete tooth is formed by the activity of a different type of cell during the ontogenesis of the tooth. The formation of the enamel layer is performed by the activity of specialised cells, called ameloblasts, during a process called amelogenesis (Smith, 1998). Amelogenesis is characterised by three distinct stages, the secretion, transition and maturation stages, spanning the bell stage of tooth formation and continuing until the remaining ameloblasts apoptose before the eruption of the tooth (Smith, Poulter, et al., 2017).

The secretion stage is characterised by the increased secretion of proteins of the enamel matrix from the ameloblasts, as they move away from the dentine layer. Guided by signals from the enamel knot, the dental papilla and the odontoblasts, the ameloblasts will start to elongate and the nucleus of each cell moves towards the SI, reversing the cell polarity (Matsuo et al., 1992), while at the other end of the cell, the basal side neighbouring the dentine layer, an extension of the cytoplasm is formed, which will become the Tomes' process. Each cell needs to act coordinated with the others whilst moving and during the secretion of the enamel matrix proteins (EMPs), so that the enamel layer is formed in an organised and connected manner. The EMPs are secreted by exocytosis in the extracellular matrix, over the dentine layer (Kallenbach, 1977), from both the proximal part, later forming the interprismatic enamel after maturation, and the distal part of the Tomes' process, later forming prismatic enamel (Nanci, 2017).

During the transition stage, the secretion of EMPs by the ameloblasts is reduced, the cells become smaller, while the nucleus moves closer to the centre of the cell and the Tomes' process is lost (Reith, 1970). The cells of the SI and the outer enamel epithelium form the papillary cell layer (PL), which cooperate with ameloblasts for the transport of mineral ions and the removal of degraded or cleaved protein products and water from the enamel layer during the maturation stage. The enamel layer reaches its full thickness during this stage and starts to mineralise in a process that will conclude in the maturation stage. Consequently, the transition stage is not clearly separated from the maturation stage, as most of the events that characterise it continue into the maturation stage. The number of ameloblasts is reduced starting from the transition stage, with an estimated 50 % of the total ameloblast population having undergone apoptosis by the end of this stage (Smith and Warshawsky, 1977).

The maturation stage sees the conclusion of the events started in the secretory stage, with the cleavage of the EMPs by the enamel matrix proteases, the completion of the mineralisation of the enamel layer, the growth of the HA crystals and the apoptosis of the remaining ameloblasts (Robinson et al., 1995). Maturation of the enamel can take years to be completed, although the enamel of deciduous teeth matures faster than that of permanent teeth (Smith, 1998). As mentioned earlier, the majority of ameloblasts apoptose and the remaining ameloblasts degenerate and join with the surrounding oral epithelium to form the junctional epithelium (Bosshardt and Lang, 2005).

1.2.5.2 Dentinogenesis

The dentine layer is formed by the odontoblasts during a process known as dentinogenesis, during the late bell stage of tooth development (Linde and Goldberg, 1994). Dentine is a layer

of mineralised tissue consisting of 70 % mineral content (Tjäderhane et al., 2009), It is formed in a process like amelogenesis. The odontoblasts secrete proteins that will form the dentine into the extracellular matrix and then the layer undergoes maturation and mineralisation to become the final dentine layer. However, unlike enamel, dentine can be repaired after eruption as new layers can be added, the amount of secondary dentine in a tooth can also be quantified and used as an indicator of the age of the individual (Tziafas, 1994).

1.2.6 Disease Phenotypes – Enamelopathies

1.2.6.1 Enamelopathies

As mentioned in section 1.2.3, the development of the tooth and the formation of enamel are complex processes that are strictly controlled at every stage, from the positioning and movement of the cells involved, to the ion transport and HA crystal growth. As a consequence, any disruption of amelogenesis could lead to permanent defects of the enamel layer, leading to a diverse group of conditions that are collectively called enamelopathies. Phenotypically enamelopathies resemble the effects produced by dental caries but are differentiated from caries as they are not caused by the normal use of the teeth, but due to factors acting during amelogenesis.

Fluorosis is a condition caused by overexposing the body to fluoride while the teeth are still developing. This leads to variable symptoms, including yellow or brown teeth and pitted hypomineralised enamel with white spots (Fejerskov et al., 1990). The effect of the fluoride on the enamel is cumulative, meaning that the severity of the symptoms is determined by the fluoride intake during the tooth formation and can range from being only cosmetic, e.g.: discolouration, in mild cases or lead to pain during mastication due to fragile hypomineralised enamel, with a chalky appearance in severe cases (Aoba and Fejerskov, 2002). The discolouration occurs from the post-eruption incorporation of exogenous ions within the hypomineralised enamel (Alvarez et al., 2009). The severity of the fluorosis phenotype is also dependent on the developmental stage of each affected tooth. Fluoride can affect the ameloblasts at all stages of amelogenesis, regardless of the source of intake, with the severity of the fluorosis phenotype magnified when the exposure is constant throughout all stages of enamel development (Bronckers et al., 2009).

Another such disease is molar incisor hypomineralisation (MIH), which presents with teeth that are more susceptible to caries due to brittle hypomineralised enamel (Almuallem and Busuttil-Naudi, 2018). It was first described by Weerheijm et al. (2001) as affecting the incisors and the first permanent molars of patients, but has since been shown to affect any tooth of the primary or permanent dentition. Phenotypically, MIH presents with brittle and discoloured enamel. The cause of MIH is thought to be multifactorial, with both genetic and environmental factors acting during early childhood, such as illness and intake of antibiotics (Taylor, 2017) being linked to the phenotype, without any of them being definitively proven (Crombie et al., 2009).

1.2.6.2 Amelogenesis Imperfecta

1.2.6.2.1 Mode of Inheritance and prevalence

Amelogenesis imperfecta (AI) is a heterogeneous group of enamel defects that are of genetic origin and are typically inherited via a Mendelian inheritance pattern that can be autosomal dominant (AD), autosomal recessive (AR) or X-linked. Additionally, AI can be differentiated as syndromic AI, where the AI phenotype presents along with other non-tooth related symptoms, or as non-syndromic AI, where the phenotype presents as an isolated affliction, a distinction which is discussed in section 1.2.3 below. The enamel defects, a characteristic of the AI phenotype, are observed in all teeth of an affected person in both the deciduous and permanent dentitions. The prevalence of AI has been shown to vary, ranging from 1 : 233 in Turkey (Altug-Atac and Erdem, 2007), 1 : 700 in Sweden (Bäckman and Holm, 1986), 1 : 1000 in Argentina (Sedano, 1975), 1 : 8000 in Israel (Chosack et al., 1979), to 1 : 14,000 in the USA (Witkop and Sauk, 1976).

AI results in poor dental aesthetics, social avoidance and distress, as well as pain and difficulty in eating for those affected (Coffield et al., 2005), lowering quality of life. Enamel affected by AI can be thin and/or very fragile, making it more susceptible to caries and chipping due to normal wear than healthy enamel is. As the adult teeth in humans are not replaced during our lifetime and there is no capability of repair, the treatment options for people affected by AI are very limited. Protection of the damaged teeth and replacing them are the only options to alleviate the negative effects on quality of life. Bridges and crowns are regularly used to stall the degradation of affected enamel, with removal and replacement with dental implants being the final solution when the tooth can no longer perform its regular functions.

1.2.6.2.2 Clinical diagnosis

The diagnosis of AI is generally made by paediatric dentists in dental clinics. The diagnosis takes into consideration both the appearance of a person's dentition and the radiometric density and thickness of the tooth enamel. This is to ensure that people affected by AI and their families receive proper counselling and treatment. This also distinguishes individuals with AI from those affected by diseases that show a very similar tooth phenotype but have a completely different underlying cause (Section 1.2.6.1).

1.2.6.2.3 AI phenotypes

Enamel defects can be caused by failure of one or more stages of amelogenesis. The enamel phenotype differs depending on the stage affected. A failure of amelogenesis at the secretion stage manifests as thin mineralised enamel or even in the total absence of enamel. This is referred to as hypoplastic AI. A failure at the transition or the maturation stage results in enamel that has normal thickness but is weak and brittle. This is referred to as hypomineralised AI. This phenotype can be further divided into hypocalcified AI, soft enamel that wears off easily due to insufficient mineralisation, and hypomaturation AI, frail enamel produced by incomplete degradation and removal of the matrix proteins. Cases of mixed phenotypes have been reported and any post-eruption changes further complicate the phenotyping of affected individuals.

Table 1. 2: AI phenotypes

Phenotype	Characteristics	Stage of amelogenesis affected
1) Hypoplastic	• Abnormal HA crystal formation and elongation • thin enamel layer • pitting and grooves • normal radiographic density	Caused by defects during the secretory stage
2) Hypomineralised	• Irregular enamel mineralisation • Further classified as Hypomature or Hypocalcified	Commonly caused by defects during the transition or maturation stage .
2 a) Hypocalcified	• Weak and fragile enamel • Opaque or chalky appearance • Teeth become stained • Normal thickness of enamel • Radiographically less opaque than dentine	
2 b) Hypomature	• Soft enamel • Mottled appearance • Vulnerable to tooth wear • Less severe than hypocalcified AI • Normal enamel thickness • Radiographically similar appearance as dentine	

1.2.6.2.4 Mouse models in the study of AI

The study of the phases of the progression of AI phenotypes on human teeth becomes challenging, because amelogenesis halts post-eruption in humans, with the apoptosis of the remaining ameloblasts (Section 1.2.5). Obtaining unerupted human teeth is not practical and ethically questionable and although tissue cultures of teeth can be grown, they have been shown to offer an altered expression pattern compared to natural teeth (Boskey and Roy, 2008). As a result mouse models have been traditionally used to study the tooth formation and amelogenesis processes and to examine the effect of disease causing mutations on the tooth phenotype (Iwasaki et al., 2005; Poulter, Murillo, et al., 2014; Smith, Poulter, et al., 2017; Lu et al., 2018; Dubail et al., 2018). Mouse models are preferred due to how closely related rodents are to humans, as well as their unique trait of a continuously erupting incisor (Zhao et al., 2014; An et al., 2018) that allows the study of all stages of amelogenesis even after the eruption of the tooth.

1.2.7 Genotypes of AI

Genes that have been associated with an AI phenotype can be categorised as causative for syndromic or non-syndromic AI by whether the pathogenic variants give rise to abnormal phenotypes in organs outside the oral cavity, or solely in the enamel respectively.

1.2.7.1 Non syndromic AI

With regards to the non-syndromic AI phenotypes, a variety of processes are involved in amelogenesis and many of the genes involved have been associated with isolated or non-syndromic AI. Pathogenic variants in more than 20 genes have been shown to cause non syndromic AI (Smith, Poulter, et al., 2017; Kim et al., 2019; Smith et al., 2020), variants documented in the Leeds database of mutations causing AI (LOVD, http://dna2.leeds.ac.uk/LOVD/), date accessed: June 2022. These genes can be broadly categorised based on their function as genes encoding for protein of the enamel matrix, enamel matrix proteases, cell adhesion proteins, ion transport proteins and others and are summarised in Table 1.3.

Table 1.3: Genes associated with non-syndromic AI.

The genes that have been reported to carry variants causative for an AI phenotype are shown here, along with the respective AI phenotype and the first instance that it was referenced in the literature. Only phenotypes associated with AI are presented here. Genes that are also linked with non-tooth related phenotypes are denoted with an asterisk.

Gene	Gene Name	Gene OMIM ID	Associated Phenotype	MOI	Phenotype OMIM ID	Reference
ACP4	Acid phosphatase 4	606362	Hypoplastic AI	AR	617297	(Seymen et al., 2016)
AMBN	Ameloblastin	601259	Hypoplastic AI	AR	616270	(Poulter, Murillo, et al., 2014)
AMELX	Amelogenin	300391	Hypoplastic / Hypomaturation AI	X-linked	301200	(Gibson et al., 2001)
AMTN	Amelotin	610912	Hypomineralised AI	AD	617607	(Iwasaki et al., 2005)
*COL17A1**	Collagen type XVII, alpha 1	113811	Hypoplastic AI	AD	104530	(McGrath et al., 1996)
*DLX3**	Distal-less homeobox 3	600525	Hypoplastic / Hypomaturation AI	AD	104510	(Price et al., 1998)
ENAM	Enamelin	606585	Hypoplastic AI	AD	104500	(Mårdh et al., 2002)
			Hypoplastic AI	AR	204650	(Hart et al., 2003)
*FAM20A**	Family with sequence similarity 20, member A	611062	Hypoplastic AI	AR	204690	(O'Sullivan et al., 2011)
FAM83H	Family with sequence similarity 83, member H	611927	Hypomineralised AI	AD	130900	(S.K. Lee et al., 2008)
GPR68	G-protein coupled receptor 68	601404	Hypomaturation AI	AR	617217	(Parry, Smith, et al., 2016)
ITGB6	Integrin beta 6	147558	Hypoplastic / Hypomineralised AI	AR	616221	(Poulter, Brookes, et al., 2014)
KLK4	Kallikrein related peptidase 4	603767	Hypomaturation AI	AR	204700	(Hart et al., 2004)
*LAMA3**	Laminin alpha 3	600805	Hypoplastic AI	AD	104530	(Yuen et al., 2012)
*LAMB3**	Laminin beta 3	150310	Hypoplastic AI	AD	104530	(Kim et al., 2013)

MMP20	Matrix metalloproteinase 20	604629	Hypomaturation AI	AR	612529	(Kim et al., 2005)
ODAPH (*C4orf26*)	Odontogenesis associated phosphoprotein	614829	Hypomaturation AI	AR	614832	(Parry et al., 2012)
RELT	Receptor expressed in lymphoid tissues	611211	Hypocalcification AI	AR	618386	(Kim et al., 2019)
*SLC24A4**	Solute carrier family 24, member 4	609840	Hypomaturation AI	AR	615887	(Parry et al., 2013)
SP6	Specificity protein 6	-	Hypoplastic AI	AD	-	(Smith et al., 2020)
WDR72	WD repeat domain 72	613214	Hypomaturation AI	AR	613211	(El-Sayed et al., 2009)

1.2.7.1.1 Enamel Matrix Proteins

The EMPs are proteins secreted from the ameloblasts that make up the organic enamel matrix. The genes encoding these proteins are amelogenin (*AMELX* OMIM * 300391 and *AMELY*, OMIM * 410000), ameloblastin (*AMBN*, OMIM * 610259) and enamelin (*ENAM*, OMIM * 606585). These genes along with amelotin (*AMTN*, OMIM * 610912) and odontogenic, ameloblast associated (*ODAM*, OMIM * 614843), which are also expressed by ameloblasts, have been shown to originate from gene duplication events. *AMELX* and *AMELY* are derived from *AMBN* and were subsequently transposed to the X and Y chromosomes respectively, while *AMBN*, *AMTN* and *ODAM* are derived from duplications of *ENAM*, forming a cluster of enamel genes on chromosome 4 (Sire et al., 2007).

The three EMPs are secreted on the surface of the dentine-enamel junction and are cleaved by the enamel matrix proteases (Section 1.2.7.1.2) to produce the functional forms of the peptides, before getting degraded further during maturation (Gadhia et al., 2012). The uncleaved proteins are only found in the outer layer of the enamel matrix, guiding the organisation of the HA crystal into enamel prisms.

EMPs were regarded as tooth specific proteins, but recent studies on *AMBN* report that variants of the protein are expressed in human adipose tissue, being involved in adipocyte differentiation (Stakkestad et al., 2018). In the ameloblasts, the canonical *AMBN* transcript, which is 447 amino acid residues in length, is secreted and is critical for the formation of the developing enamel matrix. Stakkestad et al (2018) report that in adipocytes in addition to the canonical *AMBN* mRNA, an alternatively spliced, short variant is also detected. The specific function of the shorter *AMBN* variant has not been determined.

1.2.7.1.2 Enamel Matrix Proteases

The enamel matrix proteases are enzymes that specifically cleave the EMPs and include matrix metalloproteinase 20 (*MMP20*, OMIM * 604629) acting mainly during the secretion stage of amelogenesis and kallikrein related peptidase 4 (*KLK4*, OMIM * 603767) *MMP20* is expressed in the ameloblasts from the secretory stage till the early maturation stage, while *KLK4* is expressed during the transition and maturation stages of enamel development (Simmer and Hu, 2002). In addition to ameloblasts, *MMP20* was also detected in odontoblasts and in tissues of the large intestine, albeit in very low quantities (Bartlett, 2013). *MMP20* has also been associated with kidney aging in a 2009 report (Wheeler et al., 2009), but any studies attempting to find a link between *MMP20* and kidney function failed to detect any expression of the protease in kidney tissues and no kidney problems have been reported in knock-out mouse models or people carrying *MMP20* mutations (Bartlett, 2013). *KLK4* is overwhelmingly detected in developing teeth, although expression assays in mice have detected low quantities of KLK4 in salivary gland and prostate epithelial tissues (Simmer et al., 2011). Both proteases cleave the EMPs, with MMP20 also cleaving the KLK4 pro-peptide to produce the catalytically active KLK4 protease and KLK4 inactivating MMP20 (Bartlett, 2013).

1.2.7.1.3 Cell Matrix Adhesion Proteins

To achieve their normal function the ameloblasts need to be coordinated in their action while also maintaining their connection to the dentine surface while retreating from it during the secretion stage of amelogenesis. The cells are anchored to the developing enamel matrix with the involvement of integrins, collagen and laminins, forming desmosomes and hemidesmosomes. Any mutation that reduces the stability or function of these proteins will result in defects in the enamel layer.

Such mutations that have been associated with AI phenotypes have been found in the laminins: laminin alpha 3 (*LAMA3*, OMIM * 600805) and laminin beta 3 (*LAMB3*, OMIM * 150310), which along with laminin gamma 2 (*LAMC2*, OMIM * 150292) form the laminin 332 subunit which is a part of the hemidesmosomes (Matsui et al., 1995). Collagen type XVII, alpha 1 (*COL17A1*, OMIM * 113811) is a ligand to laminin 332 that has also been associated with AI. Another cell adhesion protein that has been similarly associated with AI is the integrin beta 6 (*ITGB6*, OMIM * 147558) which is prominently expressed in maturation stage ameloblasts and is involved in the adhesion of the cells to the extracellular matrix (Wang et al., 2014). *AMTN* which is secreted by and interacts with transition and maturation stage ameloblasts, by binding to the basal lamina forms aggregates that enable the binding of the ameloblasts to the enamel layer (Moffatt et al., 2014). A gene shown to be mutated in AI and involved in cell-cell adhesion via the formation of desmosomes is family with sequence similarity 83, member H (*FAM83H*, OMIM * 611927) (Kuga et al., 2016). Interestingly, all reported mutations of *FAM83H* are located on the last exon, leading to truncated peptides that are unlikely to undergo degradation by nonsense mediated decay (NMD), and are theorised to lead to a dominant gain of function effect that causes the AI phenotype (Smith, Poulter, et al., 2017).

1.2.7.1.4 Ion Transport Proteins

During the formation of enamel a large amount of organic materials needs to be secreted and also mineral ions which are necessary for HA crystal growth. Variants associated with an AI phenotype have been found in genes that encode proteins that are ion transporters, specifically WD repeat domain 72 (*WDR72*, OMIM * 613214)(El-Sayed et al., 2009) and solute carrier family 24, member 4 (*SLC24A4*, OMIM * 609840)(Parry et al., 2013). Defects in these proteins inhibit the maturation and mineralisation of the enamel layer.

1.2.7.1.5 Other Proteins

Defects in proteins with more diverse or unknown functions have been identified as causative for non-syndromic AI. The odontogenesis associated phosphoprotein (ODAPH, OMIM * 614829) has been shown to induce HA nucleation and crystal growth *in vitro* but its explicit role in amelogenesis has not been discovered (Parry et al., 2012). Acid phosphatase 4 (*ACP4*, OMIM * 606362)(Seymen et al., 2016) is expressed during the secretion stage but its exact role as a phosphatase in amelogenesis has not been identified. Similarly, receptor expressed in lymphoid tissues (*RELT*, OMIM * 611211) (Kim et al., 2019)and specificity protein 6 (*SP6*) (Smith et al., 2020) are associated with AI phenotypes, but their role in amelogenesis remains unclear.

The G-protein coupled receptor 68 (*GPR68*, OMIM * 601404) is expressed in ameloblasts during every stage of amelogenesis (Parry, Smith, et al., 2016). It is suggested as responsible for sensing the local pH in the enamel matrix, to help control the microenvironment, as the

efficiency of amelogenesis depends on buffering the acidic conditions of the enamel matrix (Parry, Smith, et al., 2016; Bronckers, 2017). GPR68 is also expressed in other, unrelated to amelogenesis, tissues but has not been associated with a syndrome (Ludwig et al., 2003). Distal-less homeobox 3 (*DLX3*, OMIM * 600525) and family with sequence similarity 20, member A (*FAM20A*, OMIM * 611062) are genes that are considered to be controlling the expression of other proteins during amelogenesis, as well as being involved in protein expression in other non-dental tissues (Nalbant et al., 2005; Zhang et al., 2015). As a result, mutations in either protein can produce a phenotype that, depending on the specific mutation and its position in the sequence, will present with AI in isolation or as a part of a syndrome. The syndromes associated with DLX3 and FAM20A variants are reported in Table 1.4 below.

1.2.7.2 Syndromic AI

Syndromes associated with a syndromic AI phenotype are Heimler syndrome (HS, OMIM # 234580), enamel-renal syndrome (ERS, OMIM # 204690), Raine syndrome (RNS, OMIM # 259775), junctional epidermolysis bullosa (JEB, OMIM # PS226650), trichodentoosseous syndrome (TDO, OMIM # 190320) and Kohlschütter-Tönz syndrome (KTS, OMIM # 226750) and severe combined immunodeficiency caused by *STIM1*/*ORAI1* mutations (OMIM # 612783 and # 612782) summarised in Table 1.4.

Table 1.4: Syndromes presenting with enamel defects.
The genes associated with each syndrome are presented, along with the initial reference describing the syndrome.

Syndrome	OMIM #	Associated Genes	Reference
Heimler Syndrome (HS)	234580	*PEX1, PEX6, PEX26*	Ratbi et al., 2015
Enamel-Renal Syndrome (ERS)	204690	*FAM20A*	Jaureguiberry et al., 2013
Raine Syndrome (RNS)	259775	*FAM20C*	Raine et al., 1989; Simpson et al., 2007
Junctional Epidermolysis Bullosa (JEB)	PS226650	*LAMA3, LAMB3, LAMC2, COL17A1*	Wright et al., 1993; Poulter, El-Sayed, et al., 2014
Trichodentoosseous Syndrome (TDO)	190320	*DLX3*	Price et al., 1998
Kohlschütter-Tönz Syndrome (KTS)	226750	*ROGDI*	Schossig et al., 2012
Severe Combined ImmunoDeficiency	612783, 612782	*STIM1, ORAI1*	(Lacruz and Feske, 2015)

People affected by HS present with a hypoplastic AI phenotype and sensorineural hearing loss, retinal pigmentation and nail abnormalities (Ong et al., 2006). The genes associated with HS are *PEX1*, *PEX6* (Ratbi et al., 2015) and *PEX26* (Neuhaus et al., 2017).

ERS combines a hypoplastic AI phenotype with impaired tooth eruption, gingival overgrowth and nephrocalcinosis and is associated with *FAM20A* mutations (Vogel et al., 2012). However, nephrocalcinosis has a variable age of onset (Ratbi et al., 2015) and mutations in *FAM20A* have been also reported to cause AI without any kidney symptoms (Kantaputra et al., 2017).

JEB is a caused by defective hemidesmosomes between the skin layers, with a variable phenotype that depends on the gene that carries the causative variant, but generally characterised by skin blistering, hair and nail dysmorphia and hypoplastic pitted enamel (Wright et al., 1993). JEB heterozygous carriers can have the AI phenotype without any other symptoms (Poulter, El-Sayed, et al., 2014).

RNS patients have an aggressive phenotype characterised by osteosclerotic bone dysplasia which often results in death within the first weeks of life (Ishikawa et al., 2012). RNS is associated with *FAM20C*, a Golgi kinase that is involved in the modulation of biomineralisation (Ishikawa et al., 2012). *Fam20c*-/- mice display an AI phenotype (Vogel et al., 2012).

TDO is an AD condition caused by mutations in *DLX3* and is characterised by hair, bone and tooth abnormalities, such as thin or pitted enamel, enlarged pulp chambers and taurodontism (Price et al., 1998).

KTS is a neurodegenerative syndrome that presents with hypomineralised AI, seizures, ataxia, and variable other symptoms, with early onset. It is consistent with AR inheritance and carriers are hypothesized to express the AI phenotype without the neurological defects (Guazzi et al., 1994).

Recessive mutations in the *STIM1*/*ORAI1* genes have been shown to cause hypomineralised AI and severe combined immunodeficiency (Lacruz and Feske, 2015; Parry, Holmes, et al., 2016), characterized by recurrent infections due to defective T-cell and/or NK-cell activation, ectodermal dysplasia and hypomineralised enamel (McCarl et al., 2009).

1.3 Identification of gene variants causing a monogenic disease trait

Before the invention of massively parallel sequencing (Rogers and Venter, 2005; Tucker et al., 2009), the study of the genetic basis of monogenic inherited diseases was traditionally conducted by studying families comprised of affected and unaffected people and comparing the genotypes of the two groups. Early family studies used the idea of genetic linkage and looked for a match between the patterns of inheritance of the disease and of other nearby variants (Teare and Barrett, 2005). An alternative strategy was candidate gene sequencing, which utilised understanding of the tissues and processes involved to search for causative mutations within likely candidate genes in families. As a result of the understanding gained in these ways, a large body of knowledge began to build up about the many potential causes of these conditions, leading to the emergence of catalogue-like databases that concentrated this new knowledge (McKusick, 1998). This meant that a set of genes that had already been associated with the disease could be examined first in any new family being studied. By recruiting the affected and the unaffected members of a family, researchers would be able to

sequence the specific gene(s) of interest and compare variant genotypes in additional members of the same family. Sequencing was commonly conducted with the Sanger sequencing method (Section 1.2.1).

1.3.1 Sanger Sequencing

Sanger sequencing was a sequencing technology developed in 1977 by Frederick Sanger (Sanger et al., 1977), which allows us to determine the sequence of the nucleotides that make up our genetic code. This sequencing method is designed so a DNA region is amplified by PCR amplification and the amplification is randomly terminated by the incorporation of fluorescently labelled dideoxyribonucleotides (ddNTPs) at random positions through the region. The polymerase used for the amplification cannot continue the elongation of the amplified fragment after the incorporation of a ddNTP, so the amplicon is released. The random insertion of the ddNTPs ensures that there will be fragments that terminate at all the positions of the DNA sequence and a high quantity of DNA template ensures that there are enough amplicons produced, that are randomly terminated at the same position, so that they can be detected by electrophoresis. A limitation of this method is that the ddNTPs will be introduced more frequently in the first positions of the sequence, with the rate of incorporation in, and the coverage depth of, the later parts of the sequence being accordingly lower. All amplicons are denatured for the capillary electrophoresis, and they run according to their size, with the smaller fragments being read first from the electrophoresis. The results of this electrophoresis will give us the original DNA sequence, also see Figure 1.4 for a summary of the technique.

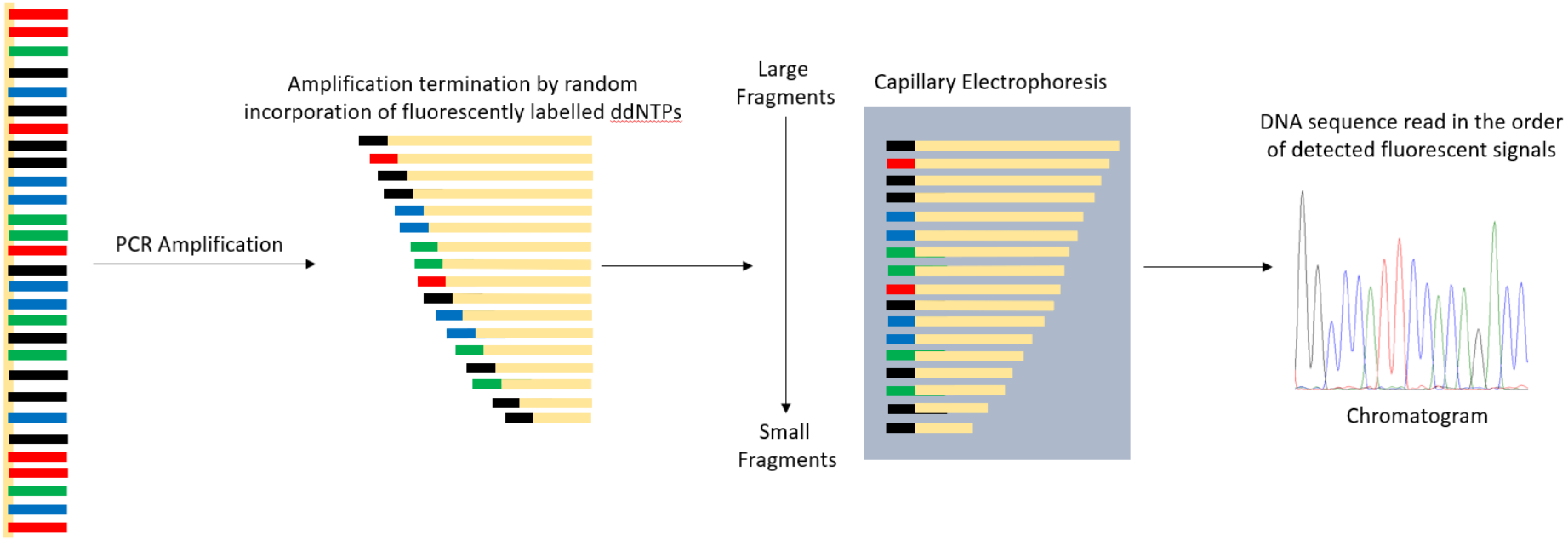


Figure 1.4: Sanger Sequencing Overview.

The sample DNA is amplified with PCR, until a fluorescently labelled dideoxyribonucleotide (ddNTP) is incorporated which stops the elongation of the sequence. The various fragments produced by PCR are separated by capillary electrophoresis and the fluorescent tags are read by the sequencer, outputting the chromatogram.

1.3.2 Segregation

One of the tests used to determine whether a candidate variant is associated with the phenotype in affected people or is instead a chance identification of a neutral polymorphism, is genetic variant segregation in families. This is only possible when additional family members, affected or unaffected, consent to support the research and give a DNA sample. In contrast to genome wide association studies (GWAS), which test for association of genotypes with phenotypes in affected and apparently unrelated individuals versus well matched controls (Uffelmann et al., 2021), in this type of segregation study specific variants are examined for co-segregation with the disease phenotype in a manner consistent with Mendelian inheritance (Wright et al., 2011). The variants were identified either by examining genes already associated with AI, or by comparing the variants found in an individual's exome to the WT, instead of comparing the affected individuals directly, as would be the case with GWAS.

In the context of genetic studies of inherited diseases segregation describes the practice of examining whether a genetic variant that is present in the genotype of people affected by a disease phenotype is also present in the genotype of people that are not affected by the disease (Wright et al., 2011). The variant naturally has to be present in all affected people, but that presence might show correlation without being causative, as the variant examined might be subject to genetic linkage with the variant that is actually causative for the phenotype, which might be in the same gene or a neighbouring gene of the chromosome. Genetic linkage refers to adjacent genomic regions being inherited together on the same chromosome and being less likely to be separated by chromosomal recombination. To exclude this possibility in addition to the affected family members, the unaffected have to be examined as well. Assuming the disease diagnosis is correct all the unaffected people should not carry the variant, if the disease is inherited in an autosomal dominant (AD) mode, or they should be heterozygous carriers, if the disease is inherited in an autosomal recessive (AR) mode. Diseases that are caused by a combination of defects on different genes, also called multigenic diseases might explain the lack of correlation of variant and phenotype in the case that the variant contributes only partial on the phenotype. Additionally, even in monogenic diseases, exceptions can occur that will complicate the segregation process of the variants for example with partial, or incomplete, penetrance of the disease phenotype in AD diseases. Incomplete penetrance leads to mild or even no symptoms for the disease, a phenomenon that has been reported for example for *ENAM* mutations (Seymen, Lee, Koruyucu, et al., 2014) and *FAM83H* mutations (Bai et al., 2022) with AI presenting with variable phenotypes even among members of the same family. In family studies, the segregation is mainly conducted among family members. Similarly to the above explanation, for monogenic diseases affected members are expected to carry the variant, whereas unaffected family members are expected to carry the wild type of the gene, or in the case of AR inherited phenotype to only be heterozygous carriers of the variant. Unrelated families with the same phenotype that are shown to carry the same variant can also be included in the study, after considering the possibility of partial penetrance, as discussed earlier. Obtaining the family medical history can help with segregation, but it is often obtained from the family members, along with the pedigree, so both might be inaccurate and lead to misleading results.

1.3.3 Next Generation Sequencing Technologies

In more recent studies the advances in sequencing technologies have made it possible to increase the amount of data that can be included in a study by high-throughput sequencing and data analysis. Next Generation Sequencing (NGS) technologies allow us to massively sequence in parallel samples in a high-throughput fashion, with increased efficiency/speed and scalability and decreased cost compared to older sequencing methods, such as Sanger sequencing (Reuter et al., 2015). Such technologies can use the entirety of the genome for the sequencing, the technique being called Whole Genome Sequencing (WGS) or specific parts of the genome, such as the exome, with Whole Exome Sequencing (WES) (Hodges et al., 2007). The exome refers to only the protein coding regions, i.e.: the exons of genes, which account for ~1 % of the genome but carry the majority of variants identified in Mendelian diseases (Ng et al., 2009). WES has been shown to be efficient in a clinical setting, to identify and diagnose the genetic cause of Mendelian genetic diseases (Yang et al., 2013).

Both WGS and WES take a non-targeted shotgun approach to large scale sequencing. Instead of using PCR amplification with primer pairs that target specific regions of the genome like Sanger sequencing would, these technologies fragment the genome and construct PCR amplified libraries of the fragments. These are then mapped against the human reference genome, to assemble the genome or exome of the sample. For WGS, the libraries retain all fragments, whereas for WES, only the exome and a limited number of flanking nucleotides, typically less than 200, at each intron-exon boundary, are captured. The number of fragments from a specific region determine the read depth of the sequencing for this region and although the fragmentation is random, a large starting quantity of sample can better ensure that a sufficient number of reads will be available for the majority of the sequenced regions.

Initially, different platforms were developed to perform WES by different companies, with different approaches to capture the DNA targets. The first approach, by Roche's NimbleGen, was array-based capture of the fragmented DNA. The fragmented ends of the DNA are repaired so that blunt ends are produced, and universal linkers can be attached to them. These fragments are then introduced in a microarray with oligos attached, that are specific to capture the appropriate fragments. The fragments hybridise with the oligos and are retained on the micro-array, whereas any fragments that are not hybridised are removed (Albert et al., 2007). After elution, a library of exon fragments can be constructed for sequencing. The alternative to the array-based protocols, are the solution-based protocols that were developed later. Platforms like Agilent's SureSelect Exon kit and Illumina's TruSeq Exome Enrichment kit were designed to work with oligonucleotide probes, suspended in a solution, that would selectively hybridise to the targeted regions of the genome. The solution-based platforms quickly outperformed their array-based predecessors, due to lower input quantities of DNA template being required and also due to the greater number of probes that could be used in a solution compared to the limited number of probes that had to fit on a micro-array (Warr et al., 2015).

Even when using NGS technologies, Sanger sequencing is still used to validate the NGS results and also the families of affected people are still recruited for family studies and segregation by Sanger sequencing. This is necessary, as the high-throughput nature of NGS means that, despite the generally deep coverage of sequencing for most regions, there will be regions that have lower coverage and that can lead to misreading a nucleotide. WES has been shown to have high sensitivity and provide good coverage for the majority of the exome (Ng et al., 2009) but even if improbable, the millions of bases read by NGS increases the probability of error compared to Sanger sequencing of small regions. Sanger sequencing to specifically confirm the validity of a potentially pathogenic variant helps to eliminate false positives.

Furthermore, by extending the Sanger sequencing to include family members not analysed by NGS, Sanger sequencing can both confirm variants and test segregation of the variant with disease within the family, making sequencing more cost effective overall.

1.3.4 Pathogenicity prediction

NGS routinely generates information on over 80000 SNVs and over 12000 indels per sample sequenced (Belkadi et al., 2015) and therefore, it is impossible to manually examine all of the variants identified. To prioritise NGS results, the obviously non-pathogenic variants need to be differentiated from the variants of unknown significance (VUS) and excluded from the downstream analysis. To assist in filtering, specialised software is used. This can identify and remove the variants that are common in the population and can estimate and score the likely effects that the remaining variants will have on the protein expression level of the gene and functionality of the encoded protein.

1.3.4.1 Frequency Filtering

For filtering the common variants, information from publicly available databases of minor allele frequency (MAF) of variants in human populations is used. Databases such as gnomAD (https://gnomad.broadinstitute.org/) and dbSNP (Sherry et al., 2001) can be used to identify variants that have a MAF over 1 % in the population and can be considered a single nucleotide polymorphism (SNP) and so are typically unlikely to be pathogenic. Exceptions to this can be found in some populations, such as the variants causative for α-thalassemia, on the α-globin gene (*HBA*), and for sickle cell anaemia on the β-globin gene (*HBB*) which are found in high MAF in African, Mediterranean and Indian populations as an evolutionary adaptation to confer resistance to the effects of malaria (Kwiatkowski, 2005), or the Phedel508 allele of CFTR, associated with cystic fibrosis, that has a MAF of 3.15 % in the Caucasian population (Shi et al., 2020). These databases can also provide the MAF for each geographically distinct population separately, so if the geographic origin of an affected person is known, the population can be examined for the variant. In the case where a variant has a very low global MAF, but it is very frequent among members of a specific population, without that population having an increased rate of affliction by the disease of interest, then it can be assumed that the variant is not causative for the disease. Additionally, to be more specific when studying the genetic basis of a rare disease, the upper threshold of MAF that a variant can have to be causative for it, can be estimated using the alleleFrequencyApp (http://cardiodb.org/allelefrequencyapp/)(Whiffin et al., 2017).

An inherited genetic disease is considered rare when its prevalence in the population is less than 1 : 2000 (definition according to the European Commission, https://ec.europa.eu/info/research-and-innovation/research-area/health-research-and-innovation/rare-diseases_en) and this formula provides a threshold for the filtering of variants that can be considered as too common in a population to be causative for a rare disease. The prevalence of AI has not been calculated in many parts of the world (Section 1.2.2) but, based on the number of families recruited locally to the Leeds AI genetics study, it is estimated that

the true frequency of AI is closer to the higher estimate of 1 : 14000 and no more than 1 : 2000, so AI can be considered as a rare inherited disease.

After filtering for rare variants, a number of strategies and software packages are used to prioritise likely pathogenic variants, and to flag known pathogenic variants, for further analysis. To estimate the pathogenicity of rare variants, various programs are used that calculate the effect of a variant on the structure and function of the protein. Each of these programs has been optimised to estimate the effect of a specific type of mutation on the gene or the protein. In general, the prediction tools can be categorised as: nucleotide sequence based, protein structure based and supervised learning programs.

1.3.4.2 Nucleotide Sequence Based Pathogenicity Prediction

Nucleotide sequence-based pathogenicity prediction utilises information about the evolutionary conservation of a specific site in the nucleotide sequence among species that are closely related to human and then assesses the pathogenic effect that the specific variant could have. In this category are programs that estimate pathogenicity for single nucleotide variants (SNVs) and short deletion-insertions (delins), such as the Sorting Intolerant From Tolerant algorithm (SIFT) (Sim et al., 2012), PROtein Variation Effect ANalyzer (PROVEAN) (Choi et al., 2012) and Mutationtaster2 (Schwarz et al., 2014).

1.3.4.3 Protein Structure Based Pathogenicity Prediction

Protein structure based pathogenicity prediction examines the effects on the amino acid sequence, the structure of the protein product and the potential changes in interactions with small molecules and ions. Widely used tools from this category are MutPred2 (http://mutpred.mutdb.org/index.html, Pejaver et al., 2020) and PolyPhen-2 (Adzhubei et al., 2010).

More recently developed programs like CADD (Combined Annotation Dependent Depletion, Rentzsch et al., 2019) combine both nucleotide and protein based pathogenicity prediction and consider both the nucleotide sequence change and the change on the protein and then score the probability of pathogenicity for the variants identified.

1.3.4.4 Supervised Learning Pathogenicity Prediction

Assessing the potential effect on splicing for each intronic variant requires complex algorithms. The specialised tools use supervised learning, where a known dataset is used to train a program to predict pathogenicity in other contexts. An example of pathogenicity prediction of splicing alteration is the Human Splicing Finder v3.1 (http://umd.be/HSF3/HSF.shtml)(Desmet et al., 2009) and Splice-AI (Jaganathan et al., 2019).

1.3.4.5 Resolving the Pathogenicity Prediction Results

The variants highlighted by the previously mentioned tools as potentially pathogenic are then investigated further to see if there is any correlation with the disease phenotype.

Specifically, from the list of VUS indicated as potentially pathogenic ,the variants on genes that are already associated with AI are prioritised, to determine whether there is an allele or an allele combination that could explain the observed AI phenotype. The phenotype can also give an indication of which pathway the causative gene might be involved in, since mutations in genes involved in different stages of amelogenesis cause different phenotypes (Sections 1.2.4 and 1.2.5,). In the absence of a clear candidate gene, unsolved cases presenting with a similar phenotype can be grouped and compared to examine if there are variants in genes that were identified in multiple families. This will potentially highlight new genes that, when mutated, will cause the phenotype observed. Segregation of the newly observed variants and the AI phenotype will be used to confirm the correlation of the presence of the variants across multiple cases with AI.

1.3.5 Protein Structure

1.3.5.1 Use of the Protein Structure in a Clinical Setting

In the context of clinical diagnosis, the knowledge of the tertiary structure of a protein can help with estimating the effect that a variant will have on its functionality. For enzymes, mutations in the active centre of the protein will affect its ability to catalyse reactions it participates in and, as such, these are easy to identify as potentially pathogenic. However, the effects of variants on the structural parts of an enzyme or any variants in the sequence of a structural protein, such as the enamel matrix proteins, are harder to categorise as potentially pathogenic without examining the changes the variant will cause on the tertiary protein structure. Assessment can be performed by directly observing the protein structure in its natural and mutated forms, by means of X-ray diffraction imaging or Nuclear Magnetic Resonance (NMR) spectroscopy, which visualise the protein directly, or alternatively by *in silico* simulations of the structure, see section 1.2.5.3 below.

By analysing the structure of a protein, researchers can also get information on how it interacts with small inorganic molecules, as well as with other proteins, and can better understand how a mutated peptide can lose its normal functionality and also disrupt the function of other proteins it interacts with.

1.3.5.2 Proteins with an experimentally observed structure

Of the genes associated with AI (Section 1.2.3), parts of the protein structures of only a few are available in online databases, such as the protein database (PDB, https://www.rcsb.org/ Berman et al., 2000), as being experimentally observed. Most proteins that are involved in amelogenesis and tooth formation are active only at the stages before the eruption of the tooth, which increases the difficulty in isolating them and studying their structure. Of the genes associated with non-syndromic AI that are discussed in this study, only five have an experimentally observed protein structure. These proteins are shown in Table 1.5, along with the method used to observe them.

Table 1.5: Protein structures of AI associated genes.

The PDB-ID and the reference that described the structure are given, along with the experimental method used.

Protein	PDB-ID	Method	Reference	Notes
DLX3	4XRS	X-ray Diffraction	Jolma et al., 2015	As part of a heterodimeric complex with MEIS1
FAM20A	5WRR	X-ray Diffraction	Cui et al., 2017	
ITGB6	5FFG	X-ray Diffraction	Dong et al., 2017	As part of the $\alpha_v\beta_6$ integrin structure
KLK4	4KGA	X-ray Diffraction	Riley et al., 2016	
MMP20	2JSD	NMR Solution	Arendt et al., 2007	Structure of the active site

1.3.5.3 Homology threading

In many cases the simulation of the tertiary structure is more practical than the direct observation through an experimental setup, which can be impossible to achieve. The protein's amino acid sequence is known from the coding sequence of the respective gene. From that amino acid sequence, the secondary structure, including alpha helices, beta coils and loops that they form, can be calculated. The information of how the local segments of a protein are organised can be used to compare them to similar segments in other proteins that have an experimentally observed structure. These can be used as a template to simulate the tertiary structure of the protein of interest. These simulations are made by considering the homology among the structures and by finding the best fit for how the protein will fold in the cellular environment. Commonly used tools to calculate the best fit for a protein structure by homology searching are SWISS-MODEL (Waterhouse et al., 2018) and I-TASSER (Yang et al., 2014). Although the accuracy of these programs is shown to be adequate, the structures that result from the simulations are only *in silico* predictions and lack the experimental evidence to support them. Additionally, the accuracy of the predicted structure increases when the templates used are obtained from proteins from the same family as the protein of interest and decreases if the protein family is not well studied and the template selected with the best "fit" is of an unrelated protein, even if the secondary structure is similar.

Recently the I-TASSER suite has been expanded, so it can use non-homologous protein structures to predict the structure of a protein of interest, with the C-I-TASSER protocol (Zheng et al., 2021). This employs the contact maps of the residues of the protein sequence to predict the interactions among them and simulate the folding of the protein without the need to use a homologous template as a reference (Zheng et al., 2021).

1.3.5.4 Molecular Dynamics Simulations

The most accurate simulations that can be performed on the estimation of the atomistic details of the tertiary structure of a protein are the molecular dynamics (MD) simulations (Karplus and Kuriyan, 2005). MD provide details of the atomic motions of each individual residue and its interactions with the other neighbouring residues. The folding of the protein, is then simulated, based on the contacts among the residues and the change in free energy that is calculated for each potential conformation of the structure and for each folding alternative. The simulated structure is also examined for potential interactions with molecules of water, ions and small organic molecules in its microenvironment, that might affect the conformation of the protein. Variants of the peptide that are introduced in the simulations can show the change in free energy compared to the wild type. The distance in Ångström (Å) that each molecule or each individual residue of the mutated peptide will have, compared to the wild type peptide, can also be calculated. Having measured the distance between the superimposed molecules of the corresponding positions, the root mean square deviation (RMSD) can be plotted to then illustrate the potential increase or decrease in these distances and the stability of the protein (Arnittali et al., 2019).

1.4 Molecular Evolution

1.4.1 Mutations driving molecular evolution

Molecular evolution is the field of observing and recording the changes in the sequence and structure of cellular molecules, such as DNA, RNA and proteins over time, as well as the rate of that change and the effects on the functional product in the different species. Following the observation and description of the DNA double helix by Watson and Crick in 1953 (Watson and Crick, 1953) and the increased improvement of sequencing technologies and molecular techniques the field of molecular evolution experienced a dramatic development.

Mutations can occur randomly in the genome of a cell and they can be integrated and retained or removed from populations depending on a set of factors, such as the selective pressure acting on the specific region of DNA, and the effective population size (N_e) (Lynch, 2007). Selective pressure is discussed later in section 1.4.3. The effective population size (N_e) refers to the minimum number of reproducing individuals in a population that a species must have so that through random mating the dynamics of the allele frequency remain equivalent to the dynamics of the total population. The N_e can be calculated with the formula:

$$N_e = (4 * N_m * N_f) / (N_m + N_f)$$

Where N_m is the number of reproducing males and N_f the numbers of reproducing females in the population (Lynch, 2007). The mathematical framework used to infer and understand the allele frequency data is based on the models described by Ronald Fisher (Fisher, 1930) and Sewall Wright (Wright, 1931), and is referred to as the Wright-Fisher model (Tataru et al., 2017). Fisher's principle describes that for the majority of species, and by extension populations, that reproduce through sexual reproduction the sex ratio nears a 1 : 1 ratio (Fisher, 1930), while Wright's discoveries of the inbreeding coefficient and genetic drift describe the fate of gene variants, i.e.: alleles. The inbreeding coefficient describes the probability that two alleles in an individual are identical and are inherited by the same common ancestor of the parents. Genetic drift expresses the changes of the frequency of existing alleles in a population caused by randomness, as opposed to as a result of selective pressure (Wright, 1931). Genetic drift has been used as the null hypothesis to examine whether models of molecular evolution are accurate or they are indistinguishable from random chance (Orr, 1998; Ackermann and Cheverud, 2004; Smith, 2011; Lynch et al., 2016).

In a population with a set number of members that are reproducing sexually and in which no gene offers an adaptational advantage, the allele frequencies depend on the inheritance of the allele to the next generations through the gametes. Fluctuations of the frequency can be explained by a declining population (decreasing N_e), an unequal sex ratio within the population or a population bottleneck because of a natural disaster (Nei and Tajima, 1981).

Motoo Kimura based his neutral theory of molecular evolution on the aforementioned Wright-Fisher models, claiming that the effect of genetic drift on neutral mutations is driving the majority of evolutionary changes and that this effect occurs at the molecular level (Kimura, 1983). Neutral mutations are the mutations that in a given population do not lead to a fitness advantage or disadvantage.

Mutations that are slightly advantageous or slightly disadvantageous, and thus visible to selection, are considered nearly neutral mutations (Ohta and Gillespie, 1996). Nearly neutral mutations can have a chance to become fixed in a population by affecting the reproductive

success of the individuals that carry them. While neutral mutations can only become fixed in a population randomly by genetic drift, the nearly neutral mutations are also affected by selective pressure that will determine if they become retained or removed from the genome (Kimura and Ohta, 1974). However, different genomes, and even different regions of a genome, show different mutational rates, leading to differing diversification rates. Other than the N_e, the body size and life expectancy of species also contribute to these differences, by affecting the metabolic rate and germline generation time respectively. These factors are also positively correlated with the age of reproductive maturity and length of gestation. Smaller body mass has been shown to correlate with higher rate of both synonymous and nonsynonymous substitutions in nuclear DNA (Welch et al., 2008). Higher metabolic rates are linked with increased production of potentially toxic metabolic by-products that can cause damage to the DNA, which with the increased frequency of repair will lead to increased rate of mutation (Bromham, 2011). The accumulation of mutations will slowly alter the information coded from a gene and eventually lead to a shift in its function and divergence from the other genes in the same family.

1.4.2 Sequence Homology

1.4.2.1 Comparative Genomics

Part of the field of molecular evolution is comparative genomics, which observes and describes the differences in the genomes across species. The aim of many comparative genomics studies is to analyse shared regions of conservation across genomes, the changes in gene clusters and protein domains and other syntenic regions of the genomes, in a phylogenetic context, to describe the evolution, conservation, diversification and function of genomes and the underlying evolutionary mechanisms, so that the evolution of these species can be better understood (Wei et al., 2002). The species studied can be closely or more distantly related and the comparative evolutionary studies that are conducted are based on examining the homology among genes, gene families and genomic regions.

At the heart of comparative genomics is the concept of homology. Homology was first defined by Owen, first in anatomy lectures in 1843 and subsequently published in 1848, who wanting to differentiate between homology and analogy described three types of homology, general, serial and special homology: features that are part of the same organ regardless of their form or function, organs that are repeated in the an organism, such as vertebrae, and organs that in different organisms perform the same function, such as the fins of aquatic animals. The homologous features or organs were thought to follow an archetype that was the root of their homology and the criteria to decide whether two features were homologous were their position, development and composition (Owen, 1848). The concept of homology was implemented by Darwin, as he suggested that homologous structures originated from a common ancestor, instead of following an archetypal ideal form. Consequently, organs that perform the same function and possibly fulfil Owen's criteria are not homologous if they do not originate from a common ancestor but are analogous. Following this definition, the wings of birds and of bats are analogous in their function as wings but are homologous as the front limbs originating from the front limbs of the last common tetrapod ancestor of the respective clades.

With the advancements in genetics and evolutionary biology, as well as the identification of DNA as the carrier of genetic information the definition of homology evolved to include the similarities of features at the molecular level. Homology between genes and proteins is characterised by their respective nucleotide or amino acid sequence and defined as significant

sequence similarity resulting by means of shared ancestry. Shared ancestry, however, can be obtained by gene duplications, speciation events or horizontal gene transfer (Koonin et al., 2002), leading to the formation of paralogs, orthologs and xenologs respectively (Figure 1.5). In the cases when this shared ancestry only affects part of the sequence, or specific domains of a protein, it can be characterised as partial homology (Fitch, 2000).

Considering the effect of partial homology and knowing that the evolutionary history of the majority of gene families is complicated, the aforementioned criterion of shared ancestry to define homology needed to be further refined. A gene can be homologous to other genes belonging in the same gene family, as they have been traditionally adefined until now, but also contain sequences homologous to genes from other gene families.

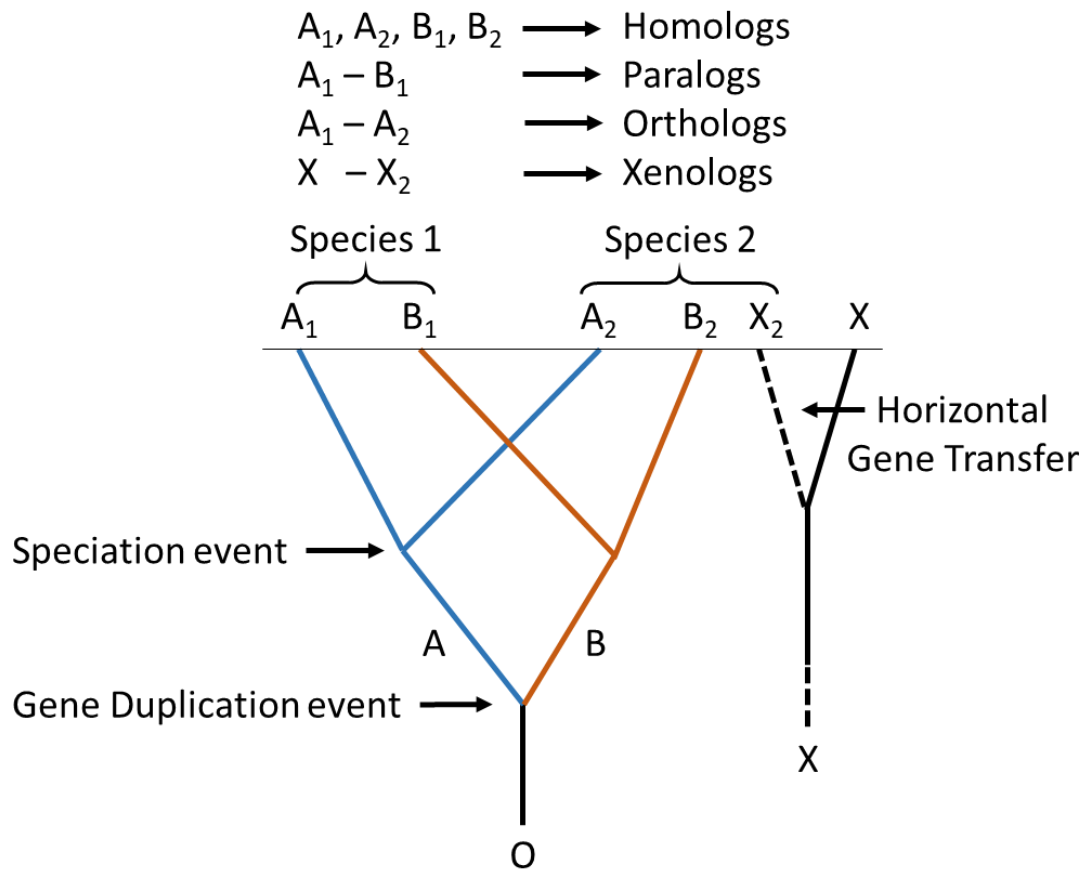


Figure 1.5: Gene homology among species.

The gene sequences that originated from an ancestral sequence can diverge after a gene duplication event occurring within a species' genome, producing paralog genes. The genes also diverge after speciation events, with each species inheriting a copy of the ancestral gene, producing ortholog genes. Gene homology is complicated further with the introduction of xenologs, genes that have been introduced to the genome of a species via horizontal gene transfer.

The identification of homologous genes across species presents a number of challenges, due to gene duplication, differential retention or loss of sequences, and the variable rates of change of these sequences that lead to the divergence of homologs. Application of multiple sequence alignment (MSA) helps to identify which regions of a sequence are homologous by comparing the nucleotide or protein sequences of related species.

1.4.2.2 Multiple Sequence Alignments as Tools for Comparative Genomics

To demonstrate the similarity of genes the coding DNA sequences can be compared by aligning them, constructing Multiple Sequence Alignments (MSAs) and highlighting the positional homology among the sequences. Various aligners have been developed with the more popular being MUSCLE (Edgar, 2004) and MAFFT (Katoh and Toh, 2008) algorithms. These can be used as standalone programs or as part of a suite of tools for comparative genomics, such as MEGA7 (Kumar et al., 2016) which can also perform the phylogenetic analysis and tree reconstruction. An alternative to MEGA7 is AQUA (Muller et al., 2010) which can perform the alignment with both MUSCLE and MAFFT and has integrated Rascal (Thompson et al., 2003) for alignment correction and norMD (Thompson et al., 2001)to score the quality of the resulting alignments, to identify the best fitting MSA for the data. Rascal and norMD can also be used independently on any MSAs provided.

Protein sequences are preferred for MSAs of protein coding regions; however, the DNA protein coding sequence is used instead to increase the amount of information obtained. The protein coding sequence is translated with the appropriate genetic code to create the corresponding protein sequence, which is then aligned, and the translation is reversed to the align the codons of the nucleotide sequence. This is preferred as the protein alignment is obtained, but the information of the codons encoding the amino acid sequence is retained and can be assessed.

1.4.2.3 Genomic duplication and structural variation

The products of the gene duplication, the paralog genes discussed previously, don't always retain the same function as the original gene, becoming the subject of functional divergence. The most common fate for paralog genes is a loss of function for one of the duplicates, in a process called nonfunctionalisation, which results in a gene becoming a pseudogene (Shakhnovich and Koonin, 2006). Alternatively, one copy of the duplicated gene may retain the original function, while the second after accumulating mutations will undergo a shift in function and acquire a novel function (neofunctionalisation). On the occasion that both copies of a duplicated gene may alter their function so that they work in a complementary manner (subfunctionalisation) (Conrad and Antonarakis, 2007). Duplicated genes that remain functional have been shown to follow the neofunctionalisation fate more often, compared to subfunctionalisation, as it has been observed that one copy of the gene will be under stronger purifying selection to retain the ancestral function compared to the second copy, which is also reported to show a lower expression pattern (Sandve et al., 2018).

The structural variation detected within genomes is not limited to single nucleotide polymorphisms or other mutational processes (e.g.: motif level insertion or deletion) localised to a short range within a sequence but can extend to more than 1 kb of the gene sequence

creating a structural variation (SV). SVs include phenomena like balanced translocations and inversions, or imbalanced genomic variants such as deletions, insertions and duplications of parts of a genomic sequence, also called copy number variants (CNV). These also contribute to the phenotypic variation and have occasionally been associated with disease phenotypes, such as a genomic duplication on the human chromosome 17p that has been associated with the Charcot-Marie-tooth disease type 1A (CMT1A) (Lupski et al., 1991), one of the first CNVs to be shown as causative for a specific disease. The variety in type and size of SVs has historically obstructed their accurate identification in genomes, but the advancements in high-throughput sequencing technologies are expected to present a solution to these problem and integrate the identification and functional characterisation of SVs into the study of pathophysiological processes (Ho et al., 2020). CNVs and other SVs have also been shown to cause AI (Seymen, Lee, Tran Le, et al., 2014; Hentschel et al., 2016) proving that the SV analysis is imperative from both an evolutionary and a genetics perspective.

1.4.3 Selective pressure

1.4.3.1 Defining Selective Pressure

Natural selection is the process through which a population alters its genetic diversity as a response to adaptation, guided by its impact on reproductive success. Protein coding regions are comprised of 3 letter codons - the degenerate code that encodes the 20 amino acids building blocks of proteins. Synonymous mutations are mutations that change the nucleotide sequence coding for a protein, but they do not alter the protein sequence due to the new variant coding for the same amino acid residue as the ancestral variant. Conversely, non-synonymous mutations change the protein sequence encoded. Synonymous mutations at synonymous sites (dS) are expected to reflect underlying rate of change in a sequence – accumulating according to the rate of genetic drift (this assumption is dealt with in more detail later in this section). Nonsynonymous mutations at non-synonymous sites (dN) change the protein sequence encoded and are visible to natural selection. Therefore, the ratio of dN/dS, also called ω (Yang, 2007), is used as a measure of selective pressure variation on a given protein coding region. New mutations that are detrimental are removed by purifying selection, but this process is heavily influenced by Ne (more efficient in larger Ne than smaller Ne (Cvijović et al., 2018)). In the case that a mutation conveys a fitness advantage for the organism it can become fixed in the population and the efficiency at which it becomes fixed is linked to the strength of the selection coefficient and the effective population size. According to this definition when $\omega < 1$, (dS more common than dN) purifying/negative selection acting on the sequence, and $\omega > 1$ indicates positive selection. Lastly when $\omega = 1$ the protein coding region is evolving neutrally. Purifying selection, genetic drift or neutral evolution and positive selection are the mechanisms by which natural selection occurs. By examining the patterns of substitution in aligned positions of homologs, the selective pressure variation can be estimated and any changes in the constraints acting on these genes.

1.4.3.2 Estimating the Selective Pressure

Estimating selective pressure variation can be achieved at the population level and at the species comparative level. At the population level, there are multiple approaches to estimate the selective pressure acting on a sequence, these approaches can be classified into three

groups: i) methods that measure the frequency of derived alleles and the effects of genetic hitchhiking (Kim and Stephan, 2002; Nielsen et al., 2005), ii) methods that use the length and structure of haplotypes (Sabeti et al., 2002; Voight et al., 2006) and iii) methods that are based on genetic differentiation, or Linkage Disequilibrium (LD), between populations (Lewontin and Krakauer, 1973; Beaumont and Balding, 2004; Innan and Kim, 2008; Bonhomme et al., 2010).

One of the earliest frequency-based approaches developed was Tajima's D test (Tajima, 1989) which is in effect a test of deviation from neutrality. Using a set of pairwise sequence alignments sampled from a population and measuring the mean observed nucleotide diversity and the expected heterozygosity in polymorphic sites in the population sampled, using the following formula for diploid DNA:

$$E[\pi] = \theta = E\left[\frac{S}{\sum_{i=1}^{n-1} 1/i}\right] = 4 * N_e * \mu$$

where E represents the expectation for equilibrium in nucleotide diversity in the population, S is the number of segregating sites (polymorphisms), n the number of samples, which is also expected to be equal or greater than 3, N_e the effective population size and μ the rate of mutation at a genomic locus. For haploid DNA the formula simplifies to $\theta = 2 * N_e * \mu$.

θ is the population mutation parameter and as shown in the formula above it is a measure of both the Ne and μ of a population, becoming a fundamental parameter in population genetics and demographics (Knudsen and Miyamoto, 2009). The expectation and thus the null hypothesis, is that if the sequences are evolving neutrally the number of polymorphisms between the homologs and the number of pairwise differences between the pairs samples are at an equilibrium and equal to θ. The difference between these two values is calculated as Δθ and Tajima's D is defined as:

$$D = \Delta\theta / \sqrt{\hat{V}(\Delta\theta)}$$

When the value of Tajima's D equals zero (D = 0) then the observed variation in the population is similar to the expected variation, which can be interpreted as a lack of evidence for selection. In terms of population dynamics D = 0 indicates that the population is evolving in an equilibrium between mutations and genetic drift (Payseur and Cutter, 2006). A negative D value (D < 0) indicates an excess of rare alleles in the population, compared to the expected due to drift, resulting in a lower level of average heterozygosity compared to the number of segregating sites among the members of the population. This commonly occurs when a population expands rapidly after a recent selective sweep (Johri et al., 2022). A positive D value (D > 0) on the other hand shows there is a lack of rare alleles in a population, indicating that there has been a rapid decrease in population size resulting in a higher level of average heterozygosity than the number of segregating sites in the population (Peery et al., 2012). Alternatively, a positive D value can be due to the act of balancing selection, which maintains rare alleles in a population at a rate higher than expected from drift (Hedrick, 2007).

As Tajima's D is a statistical test, the statistical significance of its results needs to be calculated. Tajima identified a similarity between the distribution of the test statistic in the data available and a beta distribution with mean zero and variance one (Tajima, 1989), although several problems have been observed with adopting this distribution. Specifically, D is not a continuous variable, especially with low θ values, D's minimum and maximum values depend on θ and lastly the mean and the variance of the distribution are largely dependent on the sample size (Fu and Li, 1993). Significance has also been assigned empirically to D values that surpass +/- 2, although this method does not represent the statistically critical value of a

significance test (Biswas and Akey, 2006). More recently Tajima's D was calculated with sliding windows across genomic regions, with regions for which the value of D deviates from the distribution of the majority of the windows being considered significant (Korneliussen et al., 2013). Significant D values lead to the rejection of the null hypothesis, i.e.: the sequences are not evolving neutrally.

Another major limitation of this method is that it expects that the sequences used are from a random sampling of the population (Tajima, 1989). Furthermore, neutral sites that are linked to sites under selective pressure may not be registered as neutral but follow the selection of their neighbouring site. Naturally, past population events such as population bottlenecks can affect Tajima's D introducing a bias in the calculations.

A common phylogenetic based method for assessing selective pressure variation in populations within species, versus sequences between species, is the McDonald and Kreitman (MK) test (McDonald and Kreitman, 1991). Using a combination of within species polymorphism and between species divergence to determine shifts in selective pressure. The assumption here being that the evolutionary rates of different sites can be modelled independently of one another. The null hypothesis is that under neutral evolution the ratio of dN to dS of a gene or genomic region will equal the ratio of nonsynonymous to synonymous polymorphism or P_n/P_s between species, or $dN/dS = P_n/P_s$. An adaptation of this formula can be used to calculate the proportion of sites that are driven to fixation by positive selection, using the following equation: $a = 1- (dS * P_n)/(dN * P_s)$ (Smith and Eyre-Walker, 2002). A limitation that arises from this approach is the failure to reject the null hypothesis when slightly deleterious polymorphisms are included in the dataset (Charlesworth and Eyre-Walker, 2008; Messer and Petrov, 2013). The linkage effects of these polymorphisms affect the results as they violate a key assumption of the test (independence of sites) (Messer and Petrov, 2013). Additionally, as the selective pressure is estimated at the population level, the lack of available polymorphism data at the same level for many species is also limiting the use of the test.

The tools for assessing selective pressure variation between species differ to those for population level data. Here the approaches developed can combine codon-based approaches with phylogeny and indeed with population-based data and include popular packages such as PAML (codeml) (Yang, 2007) and HyPhy (Kosakovsky Pond et al., 2020). Codeml from the PAML package offers the testing of alternative hypothesis of selective pressure using lineage-specific (or branch specific) and site-specific codon-based models in a maximum likelihood framework. It has been shown that codeml models are robust even with variable GC content (Gharib and Robinson-Rechavi, 2013) which is a common feature of mammalian genomes (Romiguier et al., 2010). The comparison of nested models using Likelihood Ratio Tests (LRTs) allows one to assess the significance of fit of a given model to the data, thereby allowing you to assess the fit of a hypothesis as compared to its appropriate null (Anisimova and Gascuel, 2006). LRT generally follows a chi squared distribution, so the degrees of freedom (df) between the models need to be considered. In LRT the df corresponds to the additional parameters in the more complex model and can be used with the chi-squared table to assess the statistical significance of the LRT results. The LRT is expressed as the difference of the log-likelihoods of the two hypotheses, according to the formula:

$$\lambda_{LR} = -2\,[l(\theta_0) - l(\theta_1)]$$

where λ_{LR} is the LRT value, $l(\theta_0)$ is the log-likelihood of the null hypothesis and $l(\theta_1)$ is the log-likelihood of the alternative hypothesis, where $0 < \lambda_{LR} < 1$. Low LRT value indicates that the

observed result is more likely to occur under the alternative hypothesis, so the null hypothesis can be rejected, while high LRT indicates that the result is likely to occur under the null hypothesis, so the null cannot be rejected. Codeml has been adapted to analyse the selective pressure variation across multiple genomes, using the VESPA pipeline (Webb et al., 2017), which also includes the LRT analysis. HyPhy was originally designed for multi-gene datasets with the option to implement it through the command line or via a graphical user interface and it offers both lineage-specific and site-specific models for testing variation in selective pressure analysis (Kosakovsky Pond et al., 2020).

1.4.3.3 Examples and limitations of selective pressure variation

Selective pressure variation has been assessed in a lineage-specific (sometimes referred to as branch specific) and site-specific manner in the literature, however, the approaches outlined above, specifically Hyphy (Kosakovsky Pond et al., 2020) and codeml (Yang, 2007) have a range of limitations to consider. Poor-quality alignments or errors in the genome assembly and annotation can produce elevated values for nucleotide changes which in turn can produce erroneous false positive predictions of positive selection. Additionally, a major assumption of these frameworks for studying selective pressure variation is that Ds reflects the neutral rate of mutation i.e.: that there is no selection on silent sites. However, we know that selection on silent sites is observed in cases where the specific codon is selected for (codon usage biases). For these dN/dS approaches, the result of selection also acting on silent sites is that dN/dS can become elevated due to decrease in dS in a region rather than increase in dN and producing false positive predictions of positive selection. This has been shown for the *BRCA1* where selection is acting on silent sites to improve exon splice site recognition (Hurst and Pál, 2001). Codon usage biases and first to third codon mutational relationships are widespread in animal genomes (Eyre-Walker, 1991; Galtier et al., 2018). Large scale consortia such as the Zoonomia project (https://zoonomiaproject.org/) are aiming to provide high quality genome assemblies for 131 mammalian species, all but nine of which were previously uncharacterised, to investigate their common and specialised traits (Genereux et al., 2020). Additionally, the genomes assembled by Zoonomia are also examined for signatures of natural selection at a single-base-pair resolution.

Selective pressure variation has been studied at the level of systems, e.g.: the innate immune system in mammals (Areal et al., 2011; Quintana-Murci and Clark, 2013). Signatures of species-specific positive selection were found among human and mouse genes, clustering them by their function and revealing patterns that correlate phenotype with genotype (Webb et al., 2015; Hyland et al., 2021). Webb et al calculated the nucleotide diversity and Tajima's D in 1 kb windows, including the DNA regions flanking the gene. They incorporated filters to detect alignment errors and recombination events, to reduce the false positive rates. Webb et al report that purifying selection acting on silent sites increases the rate of false positives, which is in agreement with earlier reports from Hurst and Pál (2001).

Signatures of positive selection leading to a potential shift in protein function of the TIR domain-containing adapter inducing IFN-β (TRIF) protein and 11 other genes involved in TRIF dependent signalling were detected across 43 mammalian genomes (Hyland et al., 2021). TRIF is a protein utilised by the innate immune system, promoting the antiviral and proinflammatory responses, that was previously reported as an example of species specific positive selection, among other innate immune system genes that are under positive selection (Webb et al.,

2015). The study by Hyland et al confirmed the previous findings of species-specific positive selection, detecting multiple differences between human and mouse sequences of TRIF related proteins. However, it is also reported that the positive selection results from the two studies don't overlap completely, as the addition of new species to the study altered the MSAs used for the codeml analysis (Hyland et al., 2021).

Similarly, a study to investigate the genome wide evolutionary pressures guiding the morphological and physiological traits of the Weddell seal (*Leptonychotes weddellii*) used both site specific and branch specific selective pressure analysis. This study revealed genes with sites under positive selection that could be interpreted as adaptive evolution to the ecological niche of the animal, while the genes show signs of species specific positive selection (Noh et al., 2022). Despite the phenotypic convergence in the marine mammals examined in this study, no evidence of molecular convergence is reported (Noh et al., 2022).

Genome wide studies have also found patterns of positive selection acting on genes for six eutherian mammals (human, chimpanzee, macaque, mouse, rat and dog) (Kosiol et al., 2008). Over 400 genes showed evidence of positive selection and another 144 showed lineage or clade specific positive selection. These genes contribute to various pathways which were strongly enriched for positive selection, indicating the co-evolution of the genes (Kosiol et al., 2008).

Regarding the formation of teeth, signatures of positive selection have been linked with the formation of the diverse dentition patterns in mammals after assessing 236 tooth-associated genes (Machado et al., 2016). Of the 236 genes evidence of positive selection was found in 31, with more recently evolved genes having faster evolutionary rates than older genes. The difference in evolutionary rates and selective pressure between recent genes that are mammal-specific and older genes that are common amongst vertebrates indicate that the younger genes are linked to the diversification of mammalian dentition patterns (Machado et al., 2016). Further evidence of positive selection was found in tooth related genes, also linking them with the dietary adaptations of mammals (Mu, Tian, et al., 2021).

1.4.4 Mammal Phylogeny

The mammal clade includes marsupials and *Eutheria*, with a divergence of circa 93 million years ago (Tarver et al., 2016). These groups expanded rapidly, via adaptive radiation to fill the niches that were made available after the mass extinction event of the end of the Cretaceous era (Lillegraven et al., 1987). The rapid expansion of the clade and extreme morphological variation among the species, has complicated attempts to describe the phylogenetic relationships amongst the major groups. Particularly challenging has been placement of the root of the placental mammal phylogeny (Teeling and Hedges, 2013). Three different hypotheses emerged to place the most basal branch of the mammalian tree, the first of which suggests that the *Afrotheria* (group containing elephants and manatees) is the earliest diverging clade (Murphy et al., 2001), while the second proposes that the *Xenarthra* (group containing armadillos and sloths) are the earliest diverging clade (Nishihara et al., 2007). The third and most recent hypothesis suggests that *Afrotheria* and *Xenarthra* form the Atlantogenata group, which is a sister group to the rest of the placental mammals (Hallström et al., 2007; Morgan et al., 2013; Tarver et al., 2016). The lack of high-quality genomes has contributed to the difficulty of placing the root of the clade (Teeling and Hedges, 2013). It has also been shown that heterogeneous models are needed to describe the evolutionary history

of these data and that the application of homogeneous models places the root in erroneous positions (Morgan et al., 2013). Recent studies using models for heterogeneous data and a data driven approach show that the earliest branching eutherian mammal was the ancestor to the afrotheria and xenarthra sister groups, i.e. the Atlantogenata, is the best fit to a range of molecular data (Moran et al., 2015; Tarver et al., 2016).

The controversy of the mammal phylogeny is not limited to the root of the tree. Clades that underwent rapid divergence gained increased complexity and although most of the taxa can be safely grouped within the respective clades, the relations among them have not been fully resolved. One such example is the large group of *Laurasiatheria* (Hallström and Janke, 2008), where the major branches have been resolved, but the placement of individual species remains challenging (Hawkins et al., 2019). The importance of a well characterised phylogeny is critical for the lineage-specific selective pressure analysis, discussed earlier, as it can lead to a misrepresentation of the relationships between species. The mammal phylogeny used in this study is adapted from Morgan et al (2013) and is presented in Figure 1.6.

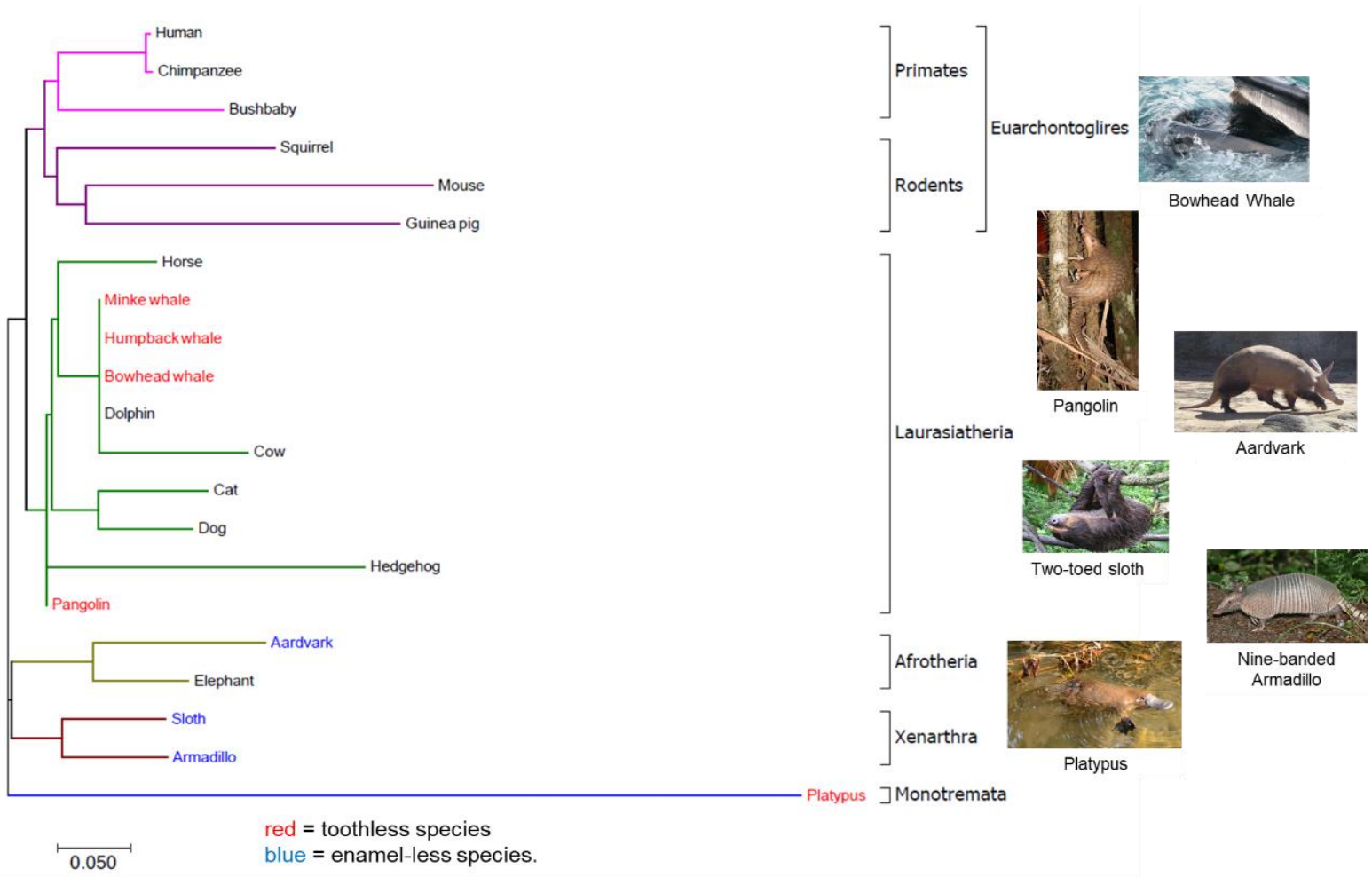


Figure 1.6: The mammalian phylogeny.

The reference species tree, adapted from Morgan et al, 2013. The toothless mammals are highlighted in red and the enamel-less are in blue. The length of the branches corresponds to the distance in nucleotide substitutions among the species.

1.4.5 Convergent evolution

To adapt to environmental selective pressures, species often follow similar strategies, leading to developing similar phenotypes/traits, this is described as independent evolution. Well known examples include the development of wings from the adaptation of the front limbs in birds and bats, or the many cases among marsupial and eutherian mammal species. Specifically, similarities among carnivorous marsupials and placental mammals of the order Carnivora have been demonstrated with an example being the extinct Tasmanian wolf, *Thylacinus cynocephalus*, and the common wolf, *Canis lupus*, among others (Wroe and Milne, 2007). Both the marsupial sugar glider, *Petaurus breviceps*, and the placental rodent flying squirrel (Tribe Pteromyini) have developed the ability to glide through the air with a skin membrane extending in both cases from the front to the hind legs (Figure 1.7). Independent evolution of the same phenotype due to behavioural similarities is also observed among marsupial mole of the genus *Notoryctes* and the placental moles, of the family Talpidae, as both have adapted to burrowing, as they are both blind, with no external ears and limbs adapted for digging (Figure 1.7).

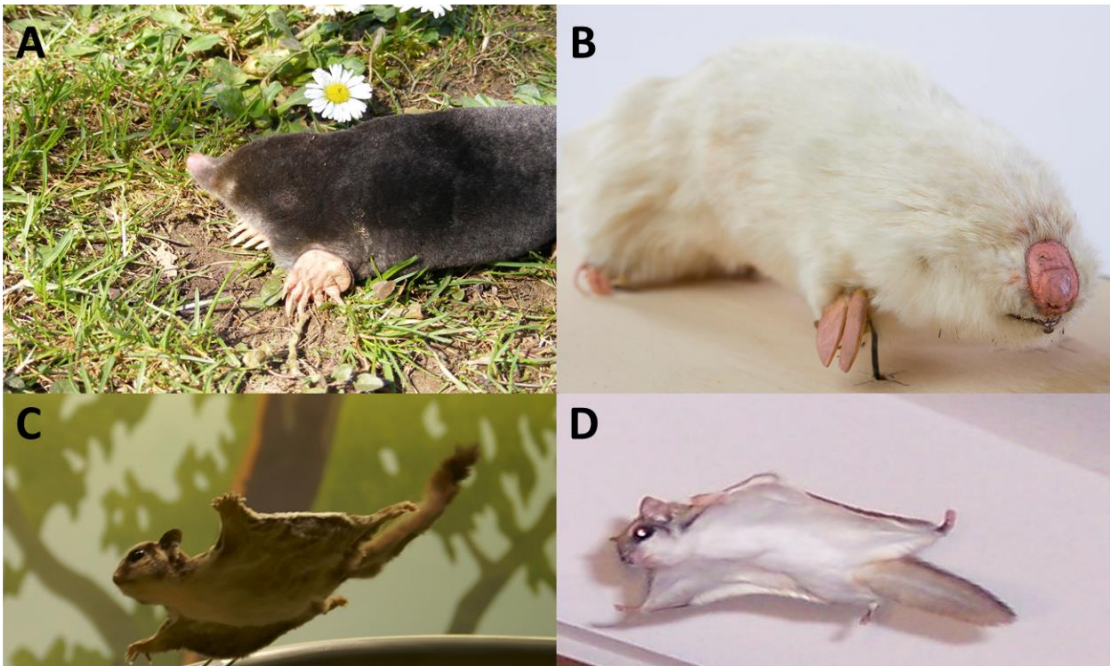


Figure 1.7: Examples of convergent evolution among marsupial and placental mammals.
(a) European mole, *Talpa europaea*, (b) Marsupial mole of the genus *Notoryctes*, (c) Marsupial sugar glider, *Petaurus breviceps,* (d) Southern flying squirrel, *Glaucomys volans.* Photo credits: (a) I,Stanislaw Szydlo, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6433426, (b) Heath Warwick - Museums Victoria, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=113225896, (c) David J. Stang, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=63144901, (d) Bluedustmite, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=2132835.

Furthermore, independent evolution extends beyond phenotypic similarities to the underlying peptide sequence changes, e.g.: the emergence of echolocation in bats and dolphins, where we observe signatures of adaptation in the *prestin* gene, responsible for receiving and processing the ultrasound signals (Liu et al., 2010). In contrast whales, also echolocate under water but they developed alternative adaptations and were shown to have decreased support for sequence convergence compared to the other two groups (Liu et al., 2010). Additionally, many genes involved in echolocation in mammals have been shown to have signatures of convergent evolution across mammal species using genome-wide studies (Parker et al., 2013). Similarly phenotypic traits of marine mammals that evolved to adapt to their aquatic environment were shown to be the product of convergent evolution at the molecular level, as species emerging from different mammalian clades: cetaceans, sirenians and pinnipeds, were adapting to similar environmental challenges (Foote et al., 2015). The morphological similarities of visual sensory organs in species would lead to the expectation that the organs are homologous or originating from the same tissue and their formation was directed by homologous proteins. However, different tissues were recruited to develop the eye structures indicating that the eye was formed independently multiple times across the ages (Fernald, 1997).

1.4.6 Loss as a means of adaptation and phenotypic change

In addition to the emergence of new traits as response to adaptation, phenotypic change can also occur due to the loss of molecular markers, a common phenomenon during the emergence of the mammals. The adaptive radiation that occurred during mammal evolution was complemented by a variety of novel traits and phenotypic innovations (Close et al., 2015) but also by the loss of molecular regions that were no longer selected for, that being a loss of gene function or of their respective regulatory elements in the different mammalian lineages. Loss or modification of regulatory elements has also been linked with phenotypic changes during animal evolution (Lowe et al., 2011), as genetic adaptation to environmental changes.

On a larger scale, in the animal kingdom gene loss has been associated with the development of phenotypic novelty, despite the initial loss of diversity from the loss of function of the implicated genes (Guijarro-Clarke et al., 2020; Murray, 2020), a process termed reductive evolution (Guijarro-Clarke et al., 2020). Outside of the mammal clade, loss of limbs occurred multiple times in reptiles with signalling pathways and regulatory elements associated with limb formation showing shared divergence amongst reptile species (Roscito et al., 2022).

Focusing again on mammals, a specific example of adaptation after reduction of diversity is the loss of teeth and enamel in species that have highly specialised diets, such as the baleen whales, anteaters, armadillos and other species. This illustrates that the loss of teeth happened independently in the mammalian clade in at least five lineages and the loss of enamel in at least two lineages independently (Davit-Béal et al., 2009). It could be assumed that the genes involved in cooperation to form mammal teeth and the enamel organ will have similar evolutionary histories and will be under similar selective pressure in species with similar tooth/enamel status.

1.4.7 Premature Termination Codons, Functional Translational Readthrough and Recoding

Traits that developed due to loss of function mutations develop from the inactivation of the genes involved in the ancestral phenotype. This inactivation leads to a truncated and non-functional protein product and can be achieved in several ways by disrupting the coding sequence of a gene: (1) introducing premature termination codons (PTCs), (2) transposable elements, (3) insertions of a sequence or deletion of a region, (4) missing a regulatory region, an intron, or an exon (Figure 1.8).

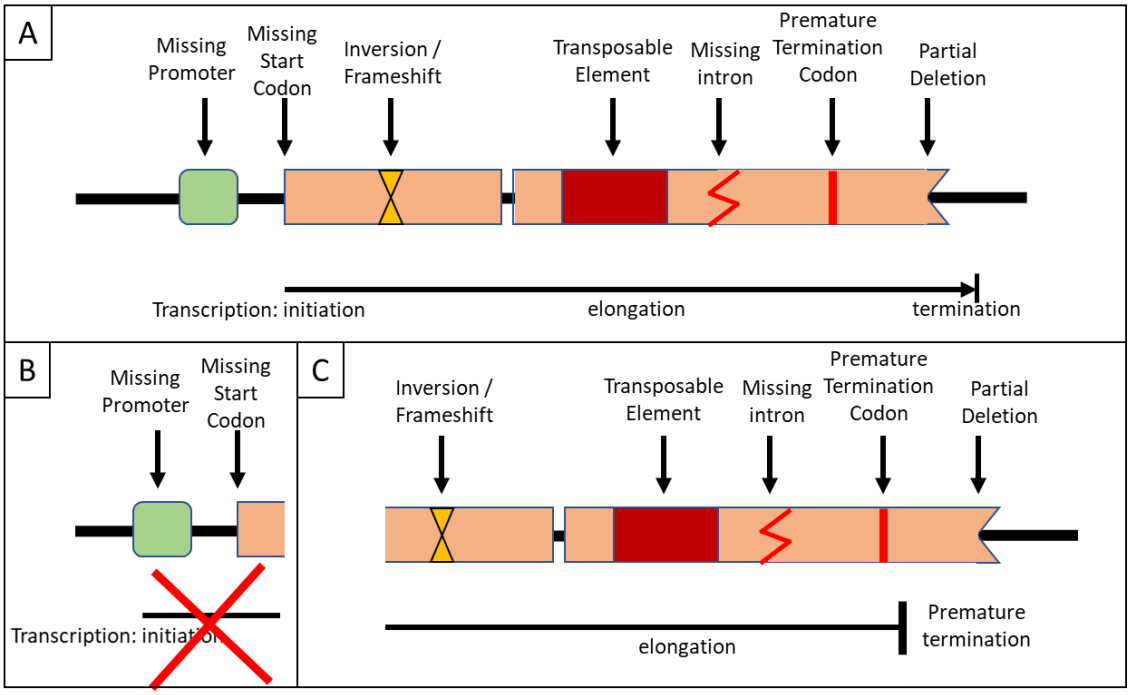


Figure 1.8: Modes of pseudogenization of a coding sequence.
(a) The coding region is shown with the possible modes that a mutation can disrupt it and lead to the pseudogenization of the gene. (b) Effect of the loss of promoter or start codon on the transcription. (c) Effect of the other types of mutation on the gene body and the transcription.

Pseudogenisation, the change of an active gene to a pseudogene which by the accumulation of disruptions in the gene sequence or its respective regulatory elements can no longer be translated, is the generally accepted means of terminating the translational activity. However, although rare there are instances that the premature termination codon aspect of the pseudogenisation model is invalidated. Some of the modes of pseudogenisation presented in Figure 1.8 are too disruptive to be able to be rectified and produce a protein product, such as missing promoter elements, large deletions and insertions and the accumulation of a large number of termination codons. However, for the rest of the pseudogenisation modes, it is possible to bypass the premature termination codon thus allowing translation to continue after a stop codon, albeit with a greatly reduced translational efficiency and an irregular protein product. The two distinct mechanisms by which cells can bypass internal stop codons are by (i) Translational Recoding (TR) (Gesteland et al., 1992), and (ii) Functional Translational Readthrough (FTR) (Doronina and Brown, 2006), both mechanisms which are shared by prokaryotic and eukaryotic cells.

Translational recoding uses mRNA elements to allow the ribosomes to alter the meaning of codons, decode mRNA in alternative reading frames or even skip parts of an mRNA (Dever et al., 2018). A prominent example of translational recoding of a stop codon is the incorporation of non-canonical amino acid residues in peptides, such as selenocysteine and pyrrolysine, which are incorporated by recoding the UGA and UAG stop codons respectively (Touat-Hamici et al., 2014; Hoffman et al., 2018). In cases of recoding the mRNA typically contains a 100 nt secondary structure that directs the incorporation of the non-canonical residues, called the selenocysteine insertion sequence element (SECIS) (Touat-Hamici et al., 2014). This element is a shared structure that is indicative of TR and it is expected to be identified in every such case. Additionally, it is expected that multiple in-frame UGA codons will be present so that the selenocysteine recoding can occur, a characteristic absent from genes where internal stop codons were introduced during the pseudogenisation process.

FTR is the process by which the ribosome continues the translation by incorporating a coding tRNA in the place of a stop codon (UAA, UAG or UGA), instead of terminating with the appropriate termination factors: Release Factor (RF) 1 and RF2 in prokaryotes and eRF1 in eukaryotes. It has been described in various organisms and is most associated with viruses where it facilitates compression of a greater amount of genetic information into a smaller genome (Firth et al., 2011; Csibra et al., 2014), but it has also been reported in fungal (*S. cerevisiae*) (Williams and Bowles, 2004), insect (*D. melanogaster*) (Jungreis et al., 2011), and human genomes (Loughran et al., 2014). The peptide produced via stop codon readthrough is elongated and has the potential to be functional by FTR (Schueren and Thoms, 2016).

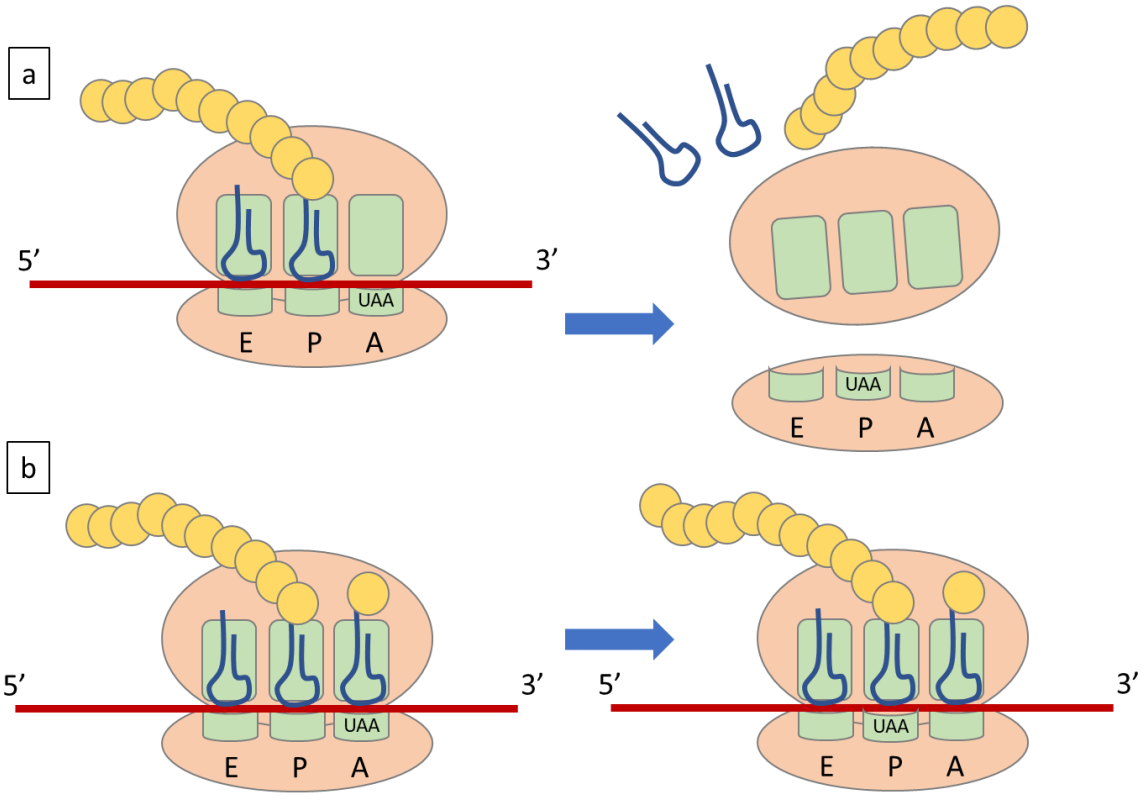


Figure 1.9: The translational mechanism of eukaryotic cells.
Presenting the regular translational termination (a) and stop codon readthrough (b). The amino acid residues are shown as yellow spheres that are added to the elongating peptide. FTR causes an amino acid residue to be incorporated instead of terminating the translation of the mRNA, leading to the continuation of translation.

The occurrence of FTR has been shown to correlate with specific nucleotide motifs adjacent to the stop codon that is being readthrough. In *D. melanogaster* FTR was shown to be guided by a 6-nucleotide motif following the stop codon, the motif being CA(A/G)N(UCG)A (Jungreis et al., 2011), while in S. cerevisiae various combinations of four nucleotides in the six positions downstream of a stop codon can affect the translation termination efficiency, specifically in the +1235 or +1236 positions (Williams et al., 2004). In human genes the tetranucleotide CUAG after a stop codon correlated with a statistically significant FTR efficiency (Loughran et al., 2014; Loughran et al., 2018).

1.4.8 Cloning and Dluc assay

Any *in silico* observations and predictions of FTR need to also be validated with experimental data. To assess whether protein coding sequences with PTCs have stop codon readthrough the standard approach to take is to perform a luciferase assay, with a specialised vector and the region of interest.

A vector in cloning is defined as plasmid, yeast or viral genome that is used to introduce foreign genetic material in a host cell, either to replicate the foreign sequence, or to incorporate it in the host genome or to express the information contained in it. An expression vector is a vector specialised for the last of these functions, as it contains the regulatory elements that will allow for any sequence inserted in it to be expressed in the host cell, also see Figure 1.10.

The luciferase assay uses an expression vector that contains a luciferase gene, most commonly from the firefly *Photinus pyralis* (Smale, 2010). A sequence of interest, usually a regulatory element, such as a promoter or enhancer, is added upstream from the luciferase gene so that the expression of the luciferase protein is directly controlled by the inserted sequence. The luminance of the expressed protein is quantified by a luminometer and compared to an empty vector that is used as a negative control.

To study the effect of a stop codon on the expression of a protein an adaptation of the luciferase assay can be used, where the sequence flanking the stop codon is cloned into a dual luciferase expression vector, see Figure 1.10. This specialised vector has two different luciferase genes, the firefly luciferase as described previously and the luciferase gene from the species *Renilla reniformis*. The two luciferase genes are expressed and measured sequentially, with the renilla luciferase used to normalise the assay. The results are then presented as a ratio of the luminance measured from the two genes (McNabb et al., 2005). The luminance of the expressed product is quantified but also the difference in luminance among different cloned sequences can be shown, as the expression of the different constructs is normalised by the activity of the first luciferase and their effect on the activity of the second can be compared directly.

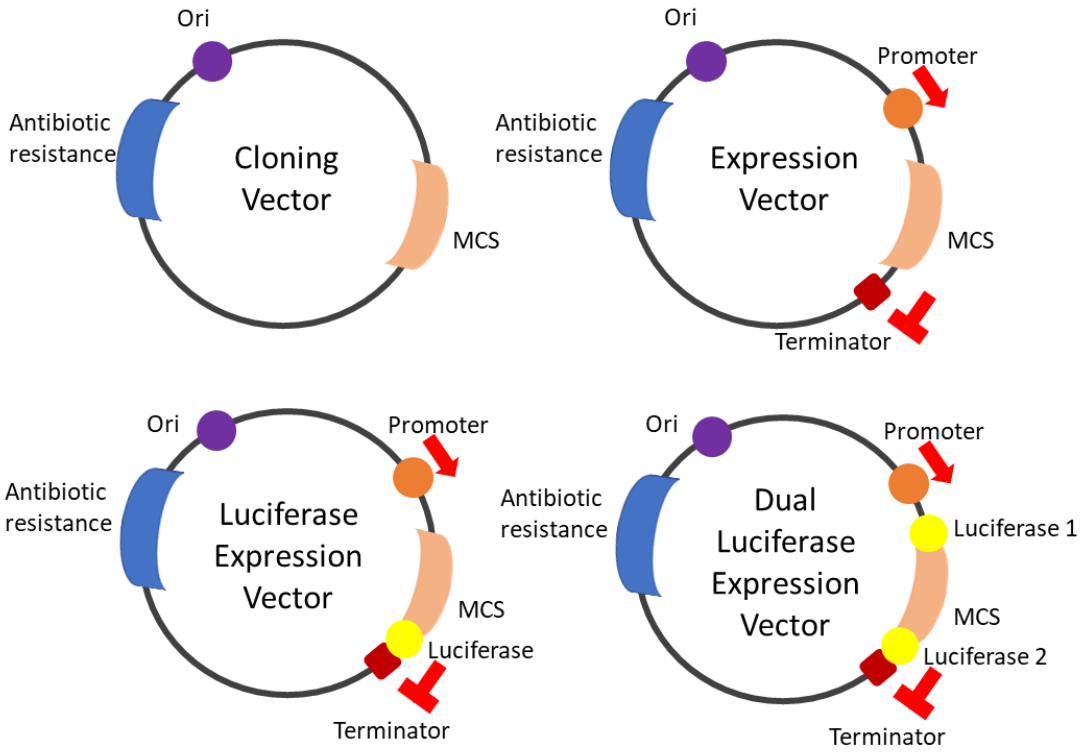


Figure 1.10: Cloning and expression vectors.
All vectors presented here carry an Ori site and an antibiotic resistance gene, but the expression vector also has a promoter and a terminator flanking the MCS so that the inserted sequence can be expressed. The luciferase vectors also show the position of the luciferase gene relative to the cloning site.

1.5 Aims of this Project

This project aims to use high throughput next generation sequencing techniques, to provide molecular diagnoses for AI patients of the Leeds AI cohort, and, in unsolved cases, determine whether the AI phenotype is caused by mutations in unknown AI associated genes or by missing mutations in known genes and attempt to identify them. The analysis of the evolutionary history of the genes that have been associated with AI can be beneficial to our aim, as it will provide information on the evolution of the genes that are being studied. Discerning patterns of natural selection that are unique to this group of genes is a novel approach that is proposed here, that will help to distinguish genes that should be associated with AI among the candidate genes indicated by WES.

The objectives of this project can be summarised as:

- To investigate the genetic basis of AI in people recruited in the Leeds AI cohort, that are affected by AI in any of its described forms
- To examine the selective pressure acting on genes associated with AI
- To identify common/shared patterns in the evolution of the genes involved in amelogenesis

Chapter 2 - Identifying the genetic causes of AI

2.1 Introduction

2.1.1 Amelogenesis imperfecta (AI) and gene discovery

The formation of teeth, as discussed in section 1.1, is a complex process involving the action of many proteins in the different phases of amelogenesis. Defects in the proteins that participate in the formation of the enamel layer, either structurally or in a catalytic capacity, can lead to the rare autosomal inherited disease characterised as amelogenesis imperfecta (AI). In addition, defects in protein components of the hemidesmosomes, thought to be involved in cell adhesion to other cells or surfaces, can also give rise to AI, as can defect in genes involved in ion transport, among others. The heterogeneity of the underlying genetic basis of AI means that a mutation causative for AI can be found in a wide range of genes. The AI phenotype can be inherited by any mode of inheritance (MOI). Different genes, encoding different proteins can be causative when their function is fully disrupted, leading to AR MOI, or even lowering the amount of functional protein to half can lead to the disease phenotype, leading to an AD MOI (Table 1.3).

As seen in section 1.2.7 and summarised in Table 1.3, a defect in one of the EMPs will lead to a hypoplastic AI phenotype, sometimes mixed with hypomaturation AI, and defects on the proteins involved in the formation and function of hemidesmosomes will also lead to hypoplastic AI. Mutations of enamel matrix proteases will lead to a hypomaturation AI phenotype, as the cleavage of the EMPs and maturation of the enamel layer will be incomplete, while defects on proteins involved with the transfer of mineral ions to the enamel matrix will lead to hypomineralised AI phenotype. Occasionally mixed or unclear phenotypes are observed, which can be caused by incomplete penetrance of the phenotype, as has been reported for families carrying ENAM or FAM83H variants (Seymen, Lee, Koruyucu, et al., 2014; Bai et al., 2022). Additionally, as the condition of the teeth deteriorates with age and use, the classification of the AI phenotype is often inconclusive, as the AI symptoms may be masked by the consequence of external forces acting on them.

By associating the phenotypes with specific genetic pathways new candidate genes and variants associated with unsolved AI cases can be identified. Unrelated patients can be grouped by the phenotype, on the basis of the genetic pathway that is expected to have been disrupted, and research can focus only on the genes involved in the functions relative to that specific pathway. Additionally, genes associated with a disease phenotype can be examined for their role in tooth development and help expand the knowledge of which genes are essential to the tooth formation and to amelogenesis, by investigating how the identified variants on these genes are linked to the observed effects.

The more comprehensive and complete the knowledge of the genetic causes of AI is, the better the people affected by it can be informed of their condition and receive counselling and treatment that is personalised to them. Personalised treatment can target more efficiently the clinical features caused by a specific gene and avoid unnecessary strain to the patients from treatment that doesn't benefit them. Consequently, to identify more genes that can potentially carry a variant causative for AI, there is a need for researchers to investigate

and identify the genetic causes of the disease in as many affected people as possible, to allow us to account for even the most uncommon of variants in highly conserved genes. To that end, the Leeds Dental Genetics group is collecting DNA and/or teeth samples from paediatric dentistry clinics in Leeds and other parts of the UK, as well as from collaborators around the world. At present the AI cohort archived in Leeds includes over 400 AI families.

2.1.2 Family studies

In this project, genomic DNA from unsolved and previously unscreened AI patients in the Leeds cohort were subject to exome sequencing to identify the variants causing disease in each case. The selection of a particular sample for WES was informed by a number of factors, most important of which was the patient phenotype and the number of family members, including affected and unaffected, recruited for the study. Samples which had previously been screened by exome sequencing by others were excluded from this analysis. The selected samples were quantified and libraries prepared and sequenced by the University of Leeds Next Generation Sequencing (NGS) facility. The processes of sample preparation and analysis of sequencing results are described in detail in section 2.2.10.

There are four distinct results expected from the family studies, which offer different amounts of information. These are:

a) identification of variants predicted to be pathogenic within or affecting a gene for which pathogenic variants are already known to cause AI. This information is helpful for dentists and patients as it provides a definite answer to the cause of the disease, helping with counselling and treatment options
b) identification of variants that are within a gene for which variants have not been previously associated with AI. This is similarly useful to dentists and patients, but also enriching the knowledge of the biology guiding amelogenesis and the disease phenotypes
c) inconclusive findings, where there are multiple candidates that require supportive findings in other families to be able to discern the gene(s) causing AI, which as a group overlaps with
d) there are no findings, meaning that the cause of the disease is not identifiable from WES analysis, which will lead us to look for the cause of the disease outside of the exome, e.g.: genome variants that were not in the scope of the WES analysis. If a candidate variant is identified within a gene that is not associated with an AI phenotype, further investigation into its genetics as well as the roles and functions of the gene's protein product are warranted.

2.1.3 Pathogenicity prediction

The presence of the same variant, or a different variant on the same gene, in other families is examined, as that would support the hypothesis that the variant is causative for AI. For the variants identified, their prevalence is examined, by looking the variant up in gnomAD and dbSNP, to find its minor allele frequency (maf) in the general population. As mentioned earlier,

in section 1.2.4.1, AI is a rare disease with an estimated prevalence of less than 1 : 2000 in the population, so a variant with a high maf is unlikely to be pathogenic and causative for the disease as this would mean that a large portion of the population would present with AI. The pathogenicity of the variants is calculated, using pathogenicity prediction tools such as CADD and SIFT, a low pathogenicity score meaning that a variant is less likely to be pathogenic, which allows us to focus on variants with a score that passes a threshold, i.e.: CADD > 15, with CADD = 10 being variants that are among the 10 % most damaging (Rentzsch et al., 2019), see section 2.2.10.3 for further detail. Proteins that have been reported to contribute to the development of teeth through the study of animal models, or genes encoding proteins that interact with any of the proteins previously associated with AI, are also prioritised for investigation.

In cases where there is no clear candidate variant(s) that can be investigated as causative for the phenotype, the family is categorised as unsolved and is put on hold until more family members affected by AI can be recruited or until new information gets published that will report findings in animal models, or in new transcriptome analysis or a genome analysis that will be associated with an AI phenotype.

2.1.4 Examination of the AI laboratory phenotype

In families that the variants causative for the disease are identified the resulting phenotype can be investigated in more depth. The AI phenotype strongly correlates to the genes involved and to the pathway that has been disrupted, so any observations can be compared to the phenotype reported in the literature to identify any similarities or differences. Any teeth that are available from the recruited affected people can be examined for structural findings. The microstructure of the enamel of the teeth can be observed by either a Scanning Electron Microscope (SEM) or by micro-Computerised X-ray Tomography (µCT). The SEM is used to study the surface of the tooth and the enamel organ, enabling the observation of the enamel rods and any abnormalities in their shape or form, as discussed in the next paragraph, while µCT is used to examine the mineral density and thickness of enamel.

2.1.5 Microstructure analysis

Normal enamel is formed in a highly organised structure, that consists of prisms, also called enamel rods, and interprismatic enamel that exists in between the prisms. Both forms are comprised of HA crystals and are both produced by the ameloblasts Tomes' processes, as seen in section 1.2.5, (Nanci, 2017). The complex structure of interwoven prisms can be observed with SEM microscopy and by comparing the WT to a tooth from an individual affected by AI it is simple to show how the prismatic form of the enamel is affected and disrupted by the disease.

An example of the complex architecture that is formed by prismatic enamel is shown in Figure 2.1, with a mutated phenotype shown next to it (Lacruz et al., 2012). Lacruz et al used transgenic mice to express a mutated AMTN peptide, which caused the phenotype observed in Figure 2.1. The enamel abnormalities caused by variants in *AMTN* were confirmed by Smith

et al., (2016) on teeth donated by people affected by AI and again by Smith et al., (2019) on teeth from people affected by AI caused by variants on *LAMB3*. Similar abnormalities were reported in teeth of people presenting with AI caused by KLK4 variants, with enamel retaining its prismatic form, but with the prisms being disorganised and not having the same orientation.

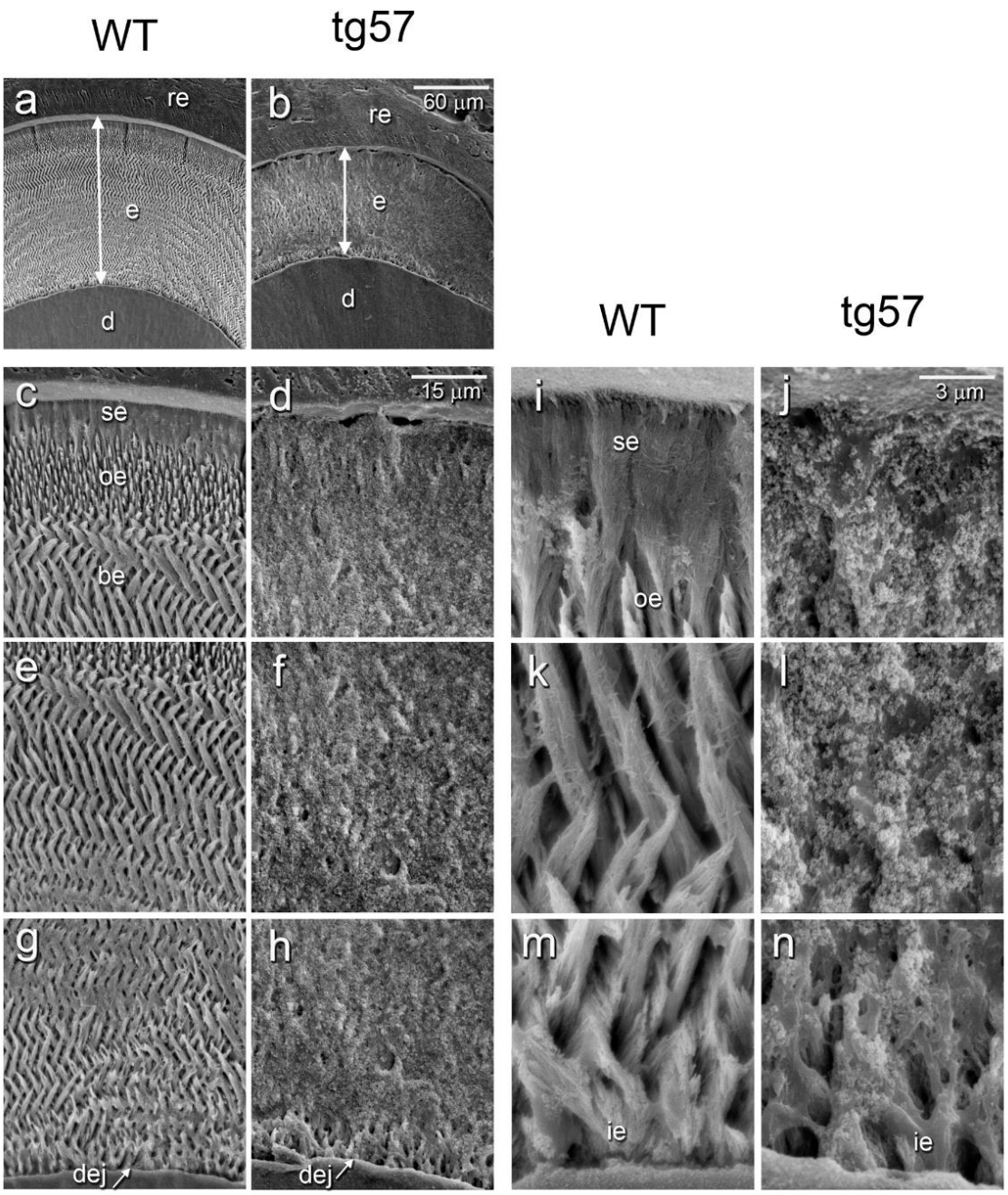


Figure 2.1: Representative SEM pictures of normal and disease enamel prisms.
The panels in the columns indicated as WT are from sections of normal teeth and the panels in the columns labelled tg57 are from sections of the teeth obtained from the transgenic mice. Abbreviations: be: bulk enamel; d: dentine; dej: dentine-enamel junction; e: enamel; ie: inner enamel; oe: outer enamel; re: embedding resin; se: surface enamel; ep: enamel prism. Image partially reproduced from Lacruz et al., 2012, CC-BY-SA 4.0.

2.1.6 Protein Structure Analysis

As described in section 1.2.5, the structure of a protein can be utilised in studying the clinical symptoms of a disease. The changes that a variant causes in the structure have an effect on the functionality of the protein, which can be estimated by simulating the final folding of the protein. In this chapter the protein structure of *RELT* and *MMP20* and the effects that variants associated with AI have on them, are discussed in sections 2.3.5 and 2.3.6 respectively.

As seen in Table 1.5 there is an experimentally observed tertiary structure for MMP20, but not for RELT, leading to different methods being preferable for studying the variations in each protein. For RELT the lack of a known structure means that a structure needs to be constructed, leading to the homology searching approach, to find a protein that is similar to RELT and base the new structure on it. There is an NMR structure for MMP20, however only the active site of the protein is available (see Table 1.4). This allows for the much more detailed study of the changes, at the atomic level, by using MD simulations with the AMBER suite of biomolecular simulation programs (Case et al., 2018). The various modules of AMBER can simulate the interactions among the amino-acid residues, their interactions with small inorganic molecules, such as mineral ions, and the stability of the protein structure.

2.1.7 Founder effect

The founder effect refers to the phenomenon of reduced genetic variation observed in a population, making that population genotypically different to the general population (Provine, 2004). The effect can arise in a population that either is a new population comprised of only a few members originating from a larger parent population but are isolated from it, or a population that has suffered through a catastrophic event leaving only some members remaining from the parent population, which is also called a population bottleneck. Variants that have a low maf in the parent population will have an exaggerated maf in the new population if some of the founder members carry them. If different families with the same ethnic origin, that according to the family history are not related to each other, present with the same AI phenotype and carry the same AI associated variant, they can be examined with microsatellite markers to find if there was a founder effect that led to an increased maf of the specific variant in it. The way to identify the presence of the founder effect in a population is by analysing the alleles shared by the members of the population and a common method to examine the alleles is by utilising microsatellite regions existing close to the gene of interest.

Microsatellites, also called Short Tandem Repeats (STRs), are a category of repetitive DNA motif, comprised of 1 - 10 nucleotides that are repeated from 5 to 50 times. They have been shown to have significantly higher mutation rates compared to other DNA regions (Brinkmann et al., 1998), with a measured 10-fold increase in mutation rates in cell generations compared to SNPs (Gemayel et al., 2012). The high level of variability shown by STRs makes them ideal for use as genetic markers in linkage and association studies. Different individuals will have differences in the number of repeats in a STR, so it follows that members of the population who carry STRs with the same number of repeats in each of them are highly likely to be related

and carry the same chromosome that has been inherited from a common ancestor (Vieira et al., 2016). Naturally, recombination events and genetic drift can cause dissimilarities in STRs among more distant relatives, but by examining a large number of STRs extending out from the region of interest, this possibility can be taken account of. Additionally, STRs that are close to a gene often will often move together in linkage disequilibrium and so are a good indication that individuals that share the same copy of an STR share the same allele of the neighbouring gene.

2.1.8 Aims of this Chapter

The objective of this chapter is to attempt to provide a molecular diagnosis for the genetic basis of the AI phenotype for some of the families of the Leeds AI cohort. By examining the exome of affected members of these families, candidate variants will be identified that could be linked to the AI phenotype. The segregation analysis will help to determine whether the candidate genes segregate together with the phenotype among affected and unaffected family members, while pathogenicity prediction algorithms and structural analysis of the proteins encoded by the genes will improve our understanding of the pathogenicity of each of these variants.

Additionally, in the case of seemingly unrelated families that share the same gene variant the presence of a founder effect will be investigated, using microsatellite genomic markers, that will reveal any common ancestry shared among those families. This provides information on the frequency of the alleles, that were examined with the microsatellite markers, in the population, along with a view on whether the specific variant is at a mutational hotspot or if its frequency is due to a founder of the population. Finally, the microstructure of the enamel on teeth donated by the affected people will be examined, to compare the effect that pathogenic variants on different genes have on the structure of enamel.

Specific aims for this chapter are:

a) To investigate and identify the genetic basis of AI in the Leeds cohort of AI patients using WES and Sanger sequencing.
b) To assess if there is evidence for a founder effect in families that share the same variant, to look for the presence of a founder effect.
c) To examine the microstructure of teeth, where available, from patients with mutations in the same gene or functional subtype.

2.2 Materials and Methods

2.2.1 Patients

Families with members diagnosed by a dentist as affected with AI (OMIM # PS104500) included in this study are residents of the UK, Costa Rica, Oman or Pakistan. Samples were collected with informed consent in the country of residence of each individual, after appropriate local ethical approval. For the individuals recruited in the UK, ethical approval was given by the Yorkshire and The Humber - South Yorkshire Research Ethics Committee Leeds (ref: 13/YH/0028).

2.2.2 Patient samples

Saliva samples were collected from patients in dental clinics by trained clinical staff. The samples that were included in this study are presented on Table 2.1. DNA was obtained from the saliva samples using Oragene® DNA collection kits (DNAgenotek Inc., Ottawa, Canada). Teeth, obtained through natural exfoliation or by extraction for clinical reasons, were also donated by some affected individuals for enamel microstructure analyses. These samples are stored in the Human Tissue Act compliant Skeletal Tissues Research Tissue Bank (School of Dentistry, University of Leeds).

Table 2. 1: Patient samples included in this study.

The number of family members recruited is reported here, along with the reported or inferred mode of inheritance, the phenotype reported by the clinician that recruited the family when available and whether teeth were also donated.

Family	Members Recruited	Mode of Inheritance	Reported Phenotype	Teeth donated
162	6	AR	-	Yes
222	4	AR	Hypomaturation	-
224	3	AD	ENAM like	-
248	2	AR	Hypomaturation	-
266	4	AD	-	-
269	3	AD	-	Yes
273	3	AD	Hypoplastic	-
279	5	AD	-	-
282	4	AD	-	-
283	4	X-linked	Hypoplastic	-
287	1	Sporadic	-	-
291	5	AR	-	-
292	5	Sporadic	-	-
293	9	X-linked	Hypoplastic	Yes
296	1	AR	-	-
297	1	AD	-	-
300	4	AD	-	Yes
311	1	AR	-	-
317	3	AR	Hypoplastic pitted	Yes
318	4	AD	Pitted	-
332	1	Sporadic	-	-
334	1	Sporadic	-	-
335	2	Sporadic	-	-
336	2	Sporadic	-	-
337	5	AR	Hypoplastic	Yes
344	8	AD	-	Yes
345	10	AD	Hypoplastic	-
346	4	AD	Hypocalcified	-
348	5	X-linked	Hypoplastic	-
349	10	AD	-	-
350	1	Sporadic	-	-
351	3	X-linked	Hypoplastic	-
355	3	AD	-	-

2.2.3 DNA extraction

DNA extraction from the Oragene® saliva collection kits was performed according to the manufacturer's instructions. In detail, the Oragene tubes were incubated at 50 °C for 3 h, being mixed every 1 h, by inverting the tube, to ensure homogenous heating of the sample. After thorough mixing, 500 µl was transferred to a 1.5 ml microtube and 20 µl of PT-L2P reagent was added. The tube contents were gently mixed by inversion and incubated on ice for 10 min. Samples were then centrifuged at room temperature for 10 min, at 15,000 x g. The supernatant was transferred to a new microtube, and the pellet was discarded. 600 µl of ethanol was added and the sample was mixed by inversion and left at room temperature for 10 min. The sample was then centrifuged for 10 min as previously described, the supernatant was discarded, and the pellet was washed through the addition of 500 µl 70 % ethanol. The sample was again centrifuged for 5 min as previously described then the supernatant was removed and the pellet was left to air dry. The pellet was redissolved in 50 µl of TE by incubation at room temperature overnight to ensure complete rehydration of the sample.

2.2.4 DNA Quantification

The dsDNA concentration of the sample was quantified using either the ND-2000 Nanodrop™ (Thermo Fisher Scientific, CA, USA), or a Qubit Fluorometer with the Qubit dsDNA Broad Range assay kit (Invitrogen). The Nanodrop was preferred when confirming the quantity in samples that had been quantified before and were known to have abundant DNA, whereas Qubit was preferred when quantifying newly extracted samples, because of its higher sensitivity to detecting low quantities of DNA. That is achieved by the fluorescent dye that binds to the DNA and amplifies the signal emitted, so very low quantities can be detected more accurately compared to other methods.

2.2.5 Amplification of DNA with Polymerase Chain Reaction (PCR)

Polymerase chain reaction (PCR) was used to amplify SNPs identified by WES and the regions flanking them, to confirm they were not NGS artifacts and exclude sample mix-up. It was also used to amplify the microsatellite loci for the microsatellite genotyping. All PCR reactions were performed using the following parameters: initial denaturation step at 94 °C for 120 s, followed by 30 cycles of: 94 °C for 30 s, annealing temperature for 30 s, 72 °C for 30 s and followed by a final extension step at 72 °C for 300 s. The extension step (72 °C for 30 s) would be adjusted to account for the length of the product, with 30 s added for every 500 bp of expected product. The annealing temperature varied depending on the primers used and was generally 5 °C lower than the Tm of the least stable of the two primers. To calculate a more accurate annealing temperature (T_a) for a given primer pair with a known product this formula can be used:

$$T_a = 0.3 \times (T_m \text{ of primer}) + 0.7 \times (T_m \text{ of product}) - 14.9$$

with T_a being the annealing temperature, T_m of primer referring to the least stable primer of the pair and T_m of product the melting temperature for the product (Rychlik et al 1990). The melting temperature of DNA is the temperature in which 50 % of the DNA is in its normal double stranded helix form and 50 % in a single stranded, coil form. The T_m of a DNA sequence can be roughly calculated by the formula:

$$T_m = (wA+xT) * 2 + (yG+zC) * 4$$

where, w, x, y and z are the numbers of A, T, G and C respectively.

PCR reactions contained: 25 ng genomic DNA, 1.25 µM of each primer, 1.25 mM each dNTP (Invitrogen), 1.5 mM $MgCl_2$ (Promega), 1x PCR buffer (Invitrogen) and 1 u Taq DNA Polymerase (0.2 µl of 5 u/µl stock, Invitrogen) and sdH_2O up to 12.5 µl total volume. All PCR primers were designed using AutoPrimer3 (https://github.com/gantzgraf/autoprimer3) and are listed in Appendix A.1.

2.2.6 Sex Determination PCR

To confirm the sex of a DNA sample, primers amplifying a section of the amelogenin genes were used. In humans, there are two homologous copies of the amelogenin gene, located on the X and the Y chromosomes and named *AMELX* and *AMELY* respectively. *AMELX* and *AMELY* have diverged to have different sequences. As a result, particular primers will amplify a product of 977 bp for *AMELX* and 790 bp for *AMELY*, a difference easily noticeable on an agarose gel as two separate bands if the sample is from a male, but only one band if female. The primer sequences are: Forward primer: 5'-CTGATGGTTGGCCTCAAGCCTGTG-3' and Reverse primer: 5'-TAAAGAGATTCATTAACTTGACTG-3' (Eng et al., 1994).

2.2.7 Microsatellite Analysis

Primer sequences for known poly-CA microsatellite and other short terminal repeat (STR) markers were obtained from the UCSC genome browser. The primer pairs were obtained from Merck Life Science UK Limited (Dorset, UK) with each forward primer carrying a 5'-HEX tag. The microsatellite markers were selected based on the distance from the gene of interest, the variability of the marker and the recombination rate of the genomic region of the locus, with at least two on each side of the test region tested in order to establish a haplotype across the genetic interval. PCR was carried out as described previously. Amplified DNA was subsequently diluted 2x – 5x with HiDi Formamide (Applied Biosystems) and 1 µl of the dilution added to 8.5 µl HiDi Formamide and 1 µl GeneScan 500 ROX size standard (Applied Biosystems). The PCR products were resolved on an ABI3130xl sequencer (Life Technologies) using a 36 cm array, POP7 polymer and 3730 buffer with the FragmentAnalysis36_pop7_1 module. Amplified DNA was sized relative to GeneScan 500 ROX (Life Technologies). The results were analysed using Genemapper v4.0 (Life Technologies) with manual confirmation. Ambiguities were judged individually and manually resolved.

2.2.8 Agarose gel electrophoresis

The visualisation of the PCR products from section 2.2.5 was carried out by agarose gel electrophoresis according to the protocol outlined in Sambrook and Russell (Sambrook and Russell, 2000). Agarose gels were formed by dissolving molecular grade agarose in 1x TAE, to achieve a final concentration of 1.5% (w/v) agarose. All gels are of this standard concentration, unless noted otherwise. A gel made with higher agarose concentration would be preferred when there is a need for a higher resolution of the DNA bands, whereas for smaller size of product and when the high resolution of the bands is not necessary the agarose concentration can be reduced. Agarose is used for the electrophoresis gels because of its ease of use, lack of toxicity and ability to separate a broad range of product sizes. To enable the visualisation of DNA bands, Midori Green Advance dye (Nippon Genetics, Japan) was added to a final concentration of 0.0047% (v/v). Prior to gel loading, samples were mixed with a loading buffer to a 1x final buffer concentration. Loading buffer stock was 6x concentrated and consisted of: 20% (w/v) Ficoll 400, 0.2% (w/v) bromophenol blue, 0.2% (w/v) xylene cyanol, 6x TAE. Electrophoresis was performed at a potential difference of 100 V for 1 hour. Sample migration was visualized under UV light, using a BioRad Gel Doc molecular imager and displayed using the ImageLab software (Bio-Rad, Hemel Hempstead, UK).

2.2.9 Sanger sequencing

Prior to sequencing, PCR products were treated with ExoSAP-IT® (Applied Biosystems), by adding 1:2.5 (v/v) ratio of ExoSAP-IT® to PCR product and incubating at 37 ^{o}C for 15 minutes and then at 80 ^{o}C for 15 minutes to deactivate enzymes. The sequencing reactions were carried out using the BigDye® Terminator v3.1 kit (Applied Biosystems). Each sequencing reaction contained 1x sequencing buffer (Applied Biosystems, supplied as 5x), 1.6 µM primer, 1 µl BigDye® Terminator, 1 µl ExoSAP-IT treated PCR product and sdH_2O, up to 10 µl total volume.

The sequencing reaction consisted of an initial denaturation step at 96 ^{o}C for 60 s followed by 25 cycles of: 96 ^{o}C for 10 s, 50 ^{o}C for 5 s and 60 ^{o}C for 240 s. All temperatures are ramped at 1 ^{o}C / second.

Precipitation of DNA was carried out by adding 5 µl 125 mM EDTA (final concentration: 8 mM EDTA) and 60 µl ethanol (final concentration: 70% ethanol), then the mix was centrifuged at 3000 x g at 20 ^{o}C for 30 minutes. The plate was upended on tissue to remove the supernatant, 60 µl 70% ethanol was added and the plate was again centrifuged at 800 g at 4 ^{o}C for 15 minutes. The plate was upended on tissue again and then air dried at 56 ^{o}C for 5 minutes to completely remove ethanol, before redissolving the precipitate in 10 µl Hi-Di Formamide (Applied Biosystems). Sequencing was performed on an ABI3130xl sequencer (Applied Biosystems) using a 36 cm array, POP7 polymer and 3730 buffer with the FragmentAnalysis36_pop7_1 module. The results were analysed using either SeqScape v2.5 or Sequencing Analysis v5.2 (Applied Biosystems).

2.2.10 Whole exome sequencing and data analysis

2.2.10.1 Sample preparation and Sequencing

The samples were quantified prior to WES by the ND-2000 Nanodrop™ (Thermo Fisher Scientific, CA, USA), and at least 800 ng of DNA were transported to the Leeds NGS facility to be processed. All whole exome sequencing (WES) was performed at the Leeds NGS facility on an Illumina HiSeq 3000 Sequencer. Exome preparations were captured using the SureSelectXT Human All Exon V6 capture library (Agilent, Santa Clara, CA, USA). The preparation of the samples was conducted by the Leeds NGS facility. In brief, these preparations included the fragmentation of the DNA by sonication in fragments of 150 – 200 bp, purification of the DNA fragments, repair of the ends of the fragments, hybridisation with the probes of the SureSelectXT kit and adding the index to the fragments, prior to sequencing (Chen et al., 2015).

2.2.10.2 Data Analysis

The WES adapters were trimmed from the sequences and the quality of the exomes was examined by TrimGalore (https://github.com/FelixKrueger/TrimGalore), which incorporates the Cutadapt tool for trimming sequences (Martin 2011) and FastQC for quality control of raw sequences obtained from high throughput sequencing (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

The output files were analysed using in house PerlScripts provided by Dr James A. Poulter, University of Leeds, UK. The sequences were trimmed and aligned to the hg19 human genome using TrimGalore software (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), to identify and remove adapters and bad quality ends from the reads. The reads were sorted and realigned locally around any indels present by using GATK v3.5 (McKenna et al., 2010) and PCR duplicates were marked using Picard tools v2.5.0 (http://broadinstitute.github.io/picard/). The marked PCR duplicates were not removed but were ignored for the exome depth calculations described below. The resulting SNP and indel variants were hard filtered, that is filtering with a specific threshold value chosen and discarding variants that fail that threshold. The filters used were QD, FS, MQ, MappingQualityRankSum and ReadPosRankSum (De Summa et al., 2017), as shown in Table 2.2.

Table 2.2: Parameter values of the thresholds used for hard filtering of SNPs and Indels with GATK.

Filter	Threshold SNP	Threshold INDEL	Details
QD	< 2.0	< 20	Variant confidence normalised by unfiltered depth of variant samples
FS	> 60.0	> 200.0	p value of number of reads calling the allele at the variant site on either DNA strand
MQ	< 40.0	-	Count of reads that have Mapping Quality (MAPQ) = 0 across all samples
MQRankSum	< -20.0	-	Count of reads that have MAPQ = 0 for each sample
ReadPosRankSum	-	< -20.0	Bias in the variant site within the reads supporting them, between the reference and alternate alleles

The list of variants called for each sample was obtained in a combined variant call format (vcf) file, using the HaplotypeCaller function of the GATK suite. Pedigrees were used to infer the mode of inheritance (MOI) of AI, either autosomal dominant, autosomal recessive or X-linked, where possible. All variants with minor allele frequency (maf) of 1% or higher in gnomAD v2.1.1 (https://gnomad.broadinstitute.org/) for families with recessive MOI or 0.1% or higher for families with dominant MOI were excluded from the following steps of the analysis. The maf filtering was performed with respect to the ethnicity and the geographic origin of the sample if that was known, otherwise the average maf for the human population was used. Where MOI was not known, samples were filtered for potential dominant inheritance then again with different filtering criteria for recessive inheritance.

The lists of variants were filtered to select only those with a minimum read depth of 5 at the position of each variant and were prioritised depending on the pathogenicity prediction score given by the mutation prediction software packages described in the section 2.2.10.3. Copy Number Variation (CNV) was also calculated for each sample, using the ExomeDepth R script (https://CRAN.R-project.org/package=ExomeDepth, Plagnol et al., 2012) to identify any potential pathogenic deletions or amplifications. The annotated variant lists were manually filtered to exclude all intronic variants apart from splice site variants, which were defined as intronic variants within 20 nt distance of the splice site. Additionally, in families with more than one affected family member sequenced, only variants that were shared by all affected members were retained. The commands used to execute the pipeline are presented in Appendix B.

2.2.10.3 Variant Pathogenicity Prediction

The predicted effect of the identified variants was scored in terms of either pathogenicity or deleteriousness, using the following software, with higher priority given to variants classified as pathogenic by a greater number of software packages. CADD (Combined Annotation Dependent Depletion, http://cadd.gs.washington.edu, Rentzsch et al., 2019) is used to score the effect of single nucleotide variants, insertions or deletions in the human genome, while being able to account for multinucleotide variants as well. Variants with a CADD score that signified that the variant is within the 10% most deleterious of all possible variants in the human genome were examined in detail, whereas variants with a lower CADD score were considered unlikely to be pathogenic and were excluded from subsequent analyses. CADD scores are calculated on a compressed scale with the 90 % most common variants scoring values 0 – 10 and the next 9 % get the values from 10 – 20 and the top 1 % most deleterious get values over 20 (Kircher et al., 2014; Rentzsch et al., 2019). The CADD score threshold used was decided to be CADD = 15, based on prior experience. PROVEAN (PROtein Variation Effect ANalyzer, v1.1: http://provean.jcvi.org/genome_submit_2.php) predicts if a protein sequence variation has a pathogenic effect on the function of the protein, by analysing the effect of the amino acid sequence change in the context of the surrounding sequences, and characterising variants as deleterious or neutral (Choi et al., 2012). Human Splicing Finder (v3.1, http://www.umd.be/HSF3/index.html) calculates the effect of intronic and exonic variants on pre-mRNA splicing and suggests whether a variant could lead to the

introduction of a new splice site, the deactivation of an existing one or if it will have no effect (Desmet et al., 2009).

The phenotype resulting from knocking out a given gene in a mouse model was also examined via the Mouse Genome Informatics (MGI, http://www.informatics.jax.org) database. All mammalian teeth have the same characteristics, however, in rodents the intense use of their incisors leads to increased incisor abrasion which is compensated by the continuous eruption of their incisors, as described in section 1.2.2. As a result, mouse models are preferred for amelogenesis studies, as they allow us to study teeth that continue the process of amelogenesis post eruption, along with providing us with parts undergoing all steps of enamel formation. The protein-protein interactions of candidates were examined via STRING, the protein-protein association networks database (https://string-db.org, Szklarczyk et al., 2017).

2.2.11 Structural Analysis of Teeth

In rare cases, teeth (mostly exfoliated primary teeth) were available for study from study participants. These teeth were first studied macroscopically, including assessment by a dentist and photography by the Medical and Dental Illustration department (University of Leeds, Leeds, UK). Teeth were then studied microscopically to examine the microstructure of the enamel and the dentino-enamel junction, to provide information about the mineralisation status of enamel, as mature and properly mineralised enamel is clearly distinct, radiographically and by microscopy, from dentine. All teeth were analysed alongside a matched control tooth from either an unaffected family member or a control obtained from the Human Tissue Act Compliant Skeletal Tissues Research Tissue Bank (School of Dentistry, University of Leeds; National Research Ethics Service Leeds East Research Ethics Committee ref: 07/H1306/95+5). These were obtained with written informed consent from patients attending clinics at Leeds Dental Institute. Control teeth matched the affected teeth by tooth type and age of the donor (Table 2.3).

Table 2. 3: Tooth samples included in this study

Family	Tooth type	Age	Sex
AI-317	Lower 7 Molar	Adult	Female
AI-337	Incisor	Juvenile	Female

2.2.11.1 X-ray Micro-Computerised Tomography (µCT)

Teeth were analysed by high resolution µCT using a Skyscan 1172 (Bruker, Billerica, MA, USA) X-ray tomographer. It was operated at 100 kV, with a source current of 100 µA. The optimal operating conditions were determined by the experience of using the scanner in previous experiments. Two aluminium sheets were used as filters to reduce beam hardening artefacts. Three controls, consisting of hydroxyapatite mineral / mineral suspensions of known densities (0.25, 0.75 (Bruker) and 2.9 g/cm^3 (Himed, NY, USA)), were included in every scan and were used to calibrate the mineral density in the resulting images. The µCT images were reconstructed using Skyscan Recon software (Bruker). Calibrated false colour maps of mineral density were generated from these images, using ImageJ2, from the FIJI package (Schindelin et al 2012) and the interactive 3D surface plot plugin (https://imagej.nih.gov/ij/plugins/surface-plot-3d.html). Videos of the imaged teeth were constructed using the CTVox software (Bruker). All imaging with the µCT was performed by Dr Steven J. Brookes at the Department of Oral Biology in the School of Dentistry of the University of Leeds.

2.2.11.2 Preparation of Samples for Scanning Electron Microscope Imaging

The tooth was embedded in thermoplastic before being fixed to a glass block that fits the Accutom-5 cutter (Struers, Ballerup, Denmark). Using a peripheral diamond cutting disc, cooled by the machine with minimal dH_2O, sections of each tooth were cut across the bucco-lingual axis. The settings used were: 0.1 mm.s^{-1} feed, cut-off wheel WHE25 (250 µm) and blade speed 4000 rpm. The cut edge was subsequently manually polished by grinding against 600 and 2000 grade carborundum paper (3M, Maplewood, MN, USA), to remove any scratches from the Accutom blade, followed by further polishing with a nail buffer. The progress of the polishing was continuously examined under a microscope until the sample surface looked smooth like glass. The sections were then etched in 30% phosphoric acid, by immersion for exactly 20 s, and then washed with excess dH_2O for 2 hrs, in a protocol established in the lab through multiple trials (Poulter, Brookes, et al., 2014; Brookes et al., 2014; Smith et al., 2016). The sections were placed on filter paper and put in a vacuum chamber to be dried overnight under vacuum. The next day, the samples were mounted on adhesive-topped aluminium stubs. These aluminium stubs were of two sizes: 15 mm diameter if a small tooth sample was to be fitted, or 32 mm diameter for larger samples. The mounted sample was then sputter coated with gold, using an auto sputter coater (Agar Scientific, Elektron Technology, Stansted, UK).

2.2.11.3 Scanning Electron Microscopy

Microstructural analysis was conducted using a Hitachi S-3400N scanning electron microscope (Hitachi, Tokyo, Japan), fitted with a 123 eV Nano XFlash® Detector 5010 (Bruker). All images were taken with the same settings; with an accelerating voltage of 15 kV, current of 60 kV and using secondary electron detection. The magnification used was 300x for the wide field photos and 1000x and 2500x for the detailed views at the points of interest.

2.2.12 Protein Structure Analysis

The tertiary structure of the proteins was predicted by homology modelling when it had not been experimentally observed. The amino-acid sequence was used to estimate the secondary structure of the protein and then predict the tertiary structure, using the I-TASSER and C-I-TASSER suites (Yang et al., 2014; Zheng et al., 2021) to predict the folding of the peptide based on comparison to homologous and non-homologous known structures respectively. Site Directed Mutator (SDM) is used to predict the effect that each mutation can have on the stability of the protein structure, which can also indicate its effect on the functionality of the altered protein (Pandurangan et al., 2017). SDM can calculate the free energy of the folding of a protein structure (ΔG), as well as the changes in free energy (ΔΔG) caused by each mutation compared to the WT and also to estimate the Occluded Surface Packing density (OSP) of each amino acid residue that is neighbouring the mutated residue. The ΔG that is calculated for each potential folding of the tertiary structure of the protein shows which folding possibility is the more stable, as it has lower ΔG value, and so the most likely to be true (Gruebele, 2002). When examining protein variants, the ΔΔG indicates if the change makes the protein more stable, with a negative ΔΔG value, or less stable, with a positive ΔΔG value. OSP density shows how available each surface level residue is for interactions with other residues or mineral ions and the changes in OSP due to a mutation can show the effect that the mutation can have on the functionality of the protein (Pattabiraman et al., 1995).

In the cases that a protein structure is already available, e.g.: the structure of the active site of MMP20, the protein structure model was obtained from PDB. The effect of variants was simulated with the AMBER suite of biomolecular simulation programs using the xLeap module and AMBERTools18 (Case et al., 2018), also using the ff14SB parameters for the protein force fields (Maier et al., 2015). The analysis of the trajectories of the protein were performed with the CCPTRAJ module of AMBER18. Protein structures were visualised using UCSF Chimera (Pettersen et al., 2004) and the trajectories from the MD simulations were visualised with the Visual Molecular Dynamics (VMD) program (Humphrey et al., 1996). The MD simulations and subsequent analysis of the structure's trajectories and stability were conducted in collaboration with Dr Sarah A. Harris, School of Physics and Astronomy, University of Leeds.

2.3 Results

Two separate rounds of exome sequencing were performed at the Leeds NGS facility, designated as the WES-2018-Batch and WES-2019-Batch, with 17 and 16 samples respectively in each batch. The samples were processed in two separate batches due to the limit of the sequencer to concurrently sequence a large number of samples as this would have a negative effect on the coverage depth of the sequencing. Each run of the sequencer produces a specific number of reads which are shared equally among the samples comprising one lane of the run. By increasing the number of samples, the share of reads that corresponds to each sample is

reduced. The number of samples was decided to maximise the number of families examined, while keeping the coverage depth of the majority of the exome at a sufficient level and was based on prior experience with the performance of WES with sequencing from 10 to 18 samples. The samples were chosen from affected individuals recruited to the Leeds AI cohort. Priority for WES was given to families where a large number of informative people had been sampled as this simplifies the subsequent segregation process for the confirmation of findings. The second criterion used was the mode of inheritance. Families with unclear or AR mode of inheritance were selected for the first batch, WES-2018-batch, see Figure 2.2 for the pedigrees of the selected families. Families with AD, or suspected x-linked inheritance were preferred in the selection for the second batch, WES-2019-Batch, the family pedigrees are shown in Figure 2.5. WES results were analysed as discussed previously. The gene variants considered to be the most likely to be causative for the observed disease phenotype are presented in Tables 2.2 and 2.3. Other variants that were identified as potentially pathogenic, by having a CADD score of over 15 and low maf in the databases, are included on the electronic Appendix. These variants were not examined further in families that a variant in a gene already associated with AI was found, while for some families it was not possible to prioritise any of these variants due to the extremely high number of candidates.

In the case where a variant in a gene associated with AI was identified, the disease phenotype conformed to previously published descriptions and the variant segregated together with the disease phenotype in the family, it was concluded that the variant was causative for the AI phenotype and the family was considered solved. The phenotype of the families that were included in the WES analysis was also compared to the phenotype associated in the literature with each of the genes that the variants listed on the Tables 2.2 and 2.3 are on, so that variants which could cause a completely different phenotype can be filtered out and the rest prioritised for subsequent analysis. In cases where multiple variants in the same gene were identified for multiple families, further studies were carried out to prepare these data for publication. These are described in detail in sections 2.3.5 and 2.3.6 for families carrying genetic variants in *RELT* and *MMP20* respectively.

2.3.1 Families Sequenced in WES-2018-Batch

For WES-2018-Batch families that were considered as following an AR mode of inheritance were prioritised for sequencing, with other families with sporadic or undetermined MOI included as well. In many cases segregation of the phenotype among the family members was not possible but the candidate variants were still validated with Sanger sequencing. The pedigrees of the families included in WES-2018-Batch are shown in Figure 2.2, and representative phenotypes of members of these families affected by AI are shown in Figure 2.3.

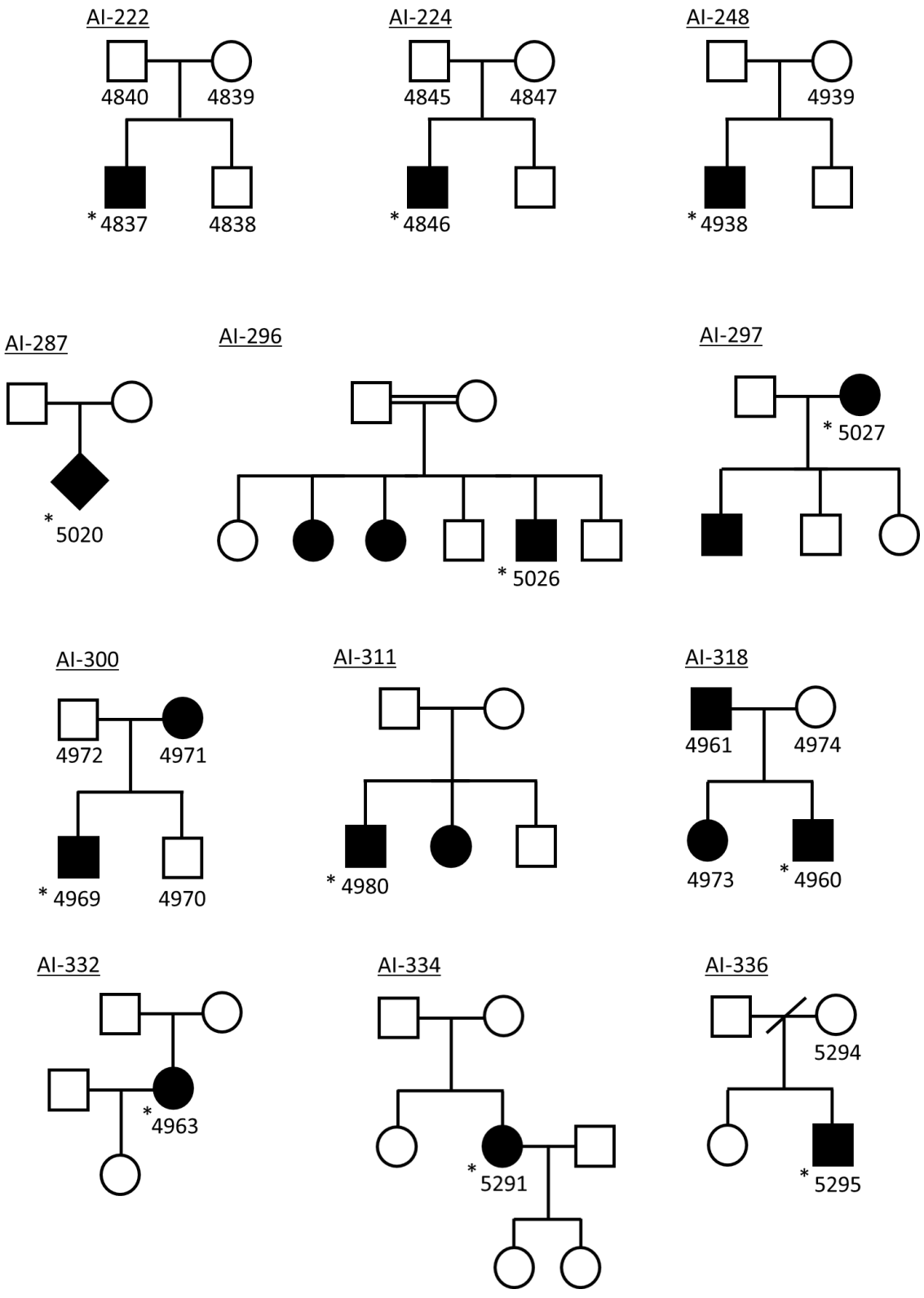


Figure 2.2: Pedigrees of the families included in WES-2018-batch.

The family members recruited for this study have 4-digit codes assigned to them. The proband is indicated with an asterisk (*). Individuals presenting with an AI phenotype are shown as filled in squares or circles, with square indicating a male and circle a female.

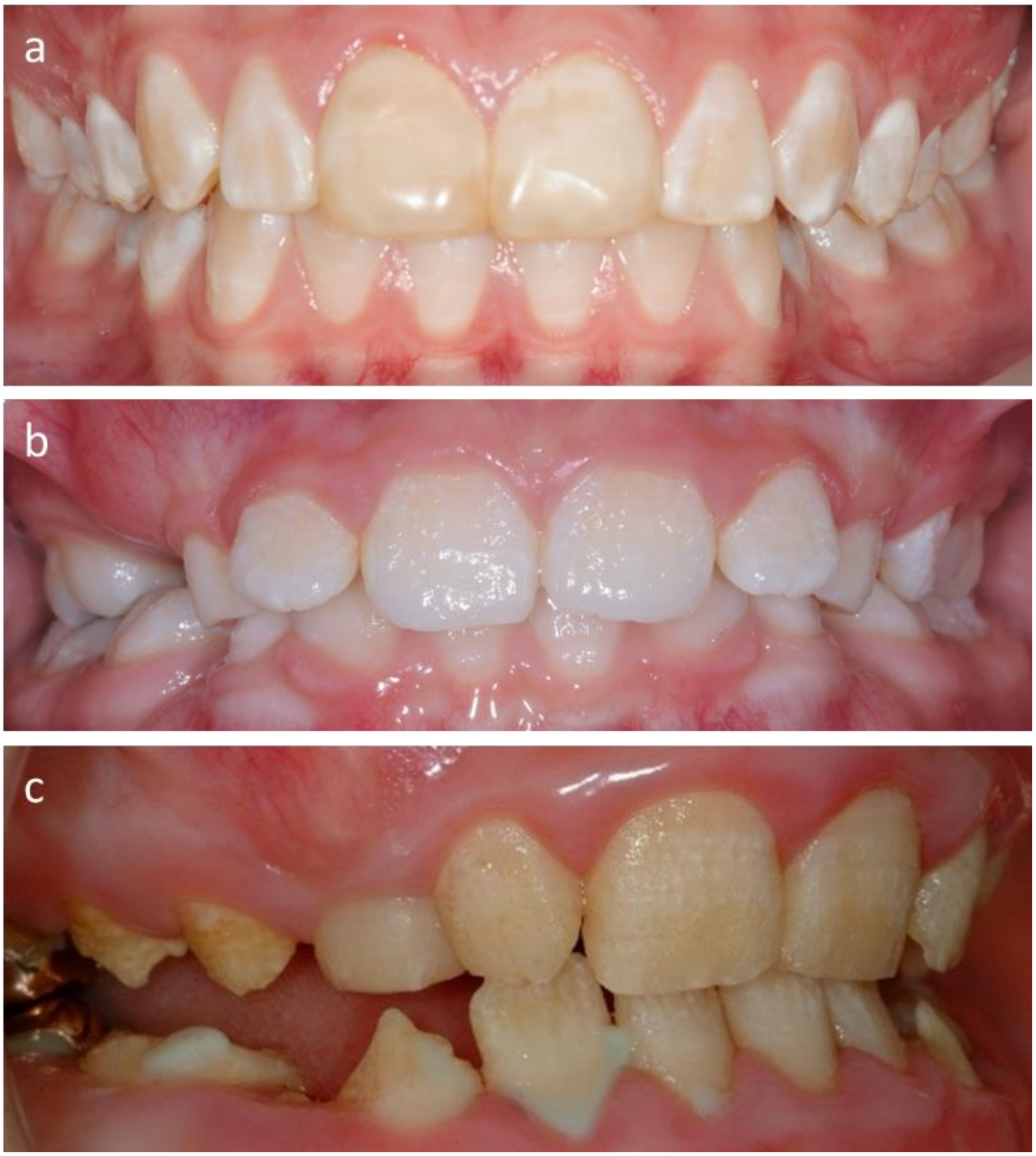


Figure 2.3: Example dental photographs of families included in WES-2018-Batch.
These photographs highlight the limitations of classification by clinical phenotype alone, especially given that post-eruption changes will have changed the tooth appearances. (a) Dental photo of 5295, the proband of Family AI-336, presenting with Hypomaturation AI. (b) Dental photo of 4973, the proband of Family AI-318, presenting with rough pitted Hypoplastic AI. (c) Dental photo of 4980, the proband of AI-311, presenting with an AI phenotype that could not be determined, described as a mixed hypoplastic / hypomaturation phenotype.

2.3.2 WES-2018-Batch Key Findings

Table 2.4: WES results from the 2018 batch of samples.

In the case of variants that have been previously published as causative for AI the reference is provided, whereas variants marked as 'new variant' have not been previously associated with AI. The most likely candidate variant is shown here for each family, other candidates with CADD > 15 are presented in the electronic Appendix. Variants marked 'rejected' did not segregate with the disease phenotype among the family members.

WES-2018-batch							
Family	Candidate Gene	Notes	Coding	Protein	CADD	maf (gnomAD)	dbSNP
162	*RELT*	rejected	c.800C>T	p.(A267V)	8.3	0.0014	rs148007314
222	-	-	-	-	-	-	-
224	*ENAM*	(Hart et al., 2003)	c.1259_1260insAG	p.(P422Vfs*27)	23.3	0.000178	rs587776588
248	*LAMB3*	new variant	c.2666C>T	p.(R889Q)	20.5	0.000021	rs139896242
287	-	too many candidates	-	-	-	-	-
291	*RELT*	new variant	c.164C>T	p.(T55I)	27.6	N/A	-
296	*WDR72*	(El-Sayed et al., 2011)	c.2686G>A	p.(R896*)	36	0.000007	rs557128345
297	*AMBN*	new variant	c.76G>A	p.(A26T)	26	N/A	-
300	*COL17A1*	new variant	c.2812+2T>C	p.(?)	23.4	N/A	-
311	*AMBN*	new variant	c.209C>G	p.(S70*)	36	0.000114	rs146148316
317	*RELT*	new variant	c.164C>T	p.(T55I)		N/A	-
318	*COL17A1*	new variant	c.3162G>T	p.(Y1054*)	38	N/A	
332	*AMBN*	new variant	c.209C>G	p.(S70*)	36	0.000114	rs146148316
334	*AMBN*	new variant	c.209C>G	p.(S70*)	36	0.000114	rs146148316
	AMBN	new variant	c.295T>C	p.(Y99H)	25.8	0.000114	rs148944860
335	*COL17A1*	new variant	c.4304G>A	p.(A1435V)	26.7	0.002653	rs146841330
336	*ITGB6*	rejected	c.2170C>G	p.V724L	10	N/A	rs146397669
337	*RELT*	new variant	c.1264C>T	p.(R422W)	34	0.001633	rs139368769

As shown in Table 2.4 above, for most of the families included in the WES-2018-Batch, a variant was identified in an AI associated gene that has a low maf, < 0.01 and high score from the pathogenicity prediction software, i.e.: CADD > 15. Following the identification of these variants the phenotype observed in the affected members of each family was compared to the phenotype reported in the literature for the variants on the various genes that are associated with an AI phenotype. As a suitable candidate variant was found on a gene that has already been associated with AI, those variants were prioritised for segregation and further analysis. Other possible pathogenic variants that were identified by the WES analysis, with maf < 0.01 and CADD > 15, are presented on the electronic Appendix.

After filtering for maf and pathogenicity prediction score there were no candidate gene variants remaining for AI-222 or AI-336. In Family AI-336 more closely an *ITGB6* variant was found, that had a low CADD score and was removed from the final list of candidates. The proband of AI-336, sample 5295 (Figure 2.2), presents with hypomaturation AI, also see Figure 2.3a, which is not consistent with the hypoplastic/hypomineralised AI reported in the literature for AI associated with *ITGB6* (Poulter, Brookes, et al., 2014). Additionally, the low pathogenicity score and the fact that sample 5295 was heterozygous for the variant while *ITGB6* is associated with AR inheritance (Poulter, Brookes, et al., 2014; Wang et al., 2014) indicated that it was not a good candidate to explain the AI phenotype in the family. The WES data for 5295 were re-examined for a second *ITGB6* variant that could explain the inheritance type as a compound heterozygote, but none were found. For AI-222 no candidate variants were found, even after lowering the filtering thresholds used. Although surprising this result indicates that the cause of the AI phenotype is likely not a single nucleotide variant on a gene's coding sequence, but possibly a mutation in the intronic regions of the genome or in any of the regulatory elements that are not included in the regions sequenced by WES. A remedy to this limitation of WES is to perform WGS on the affected members of the family and to examine whether there are regions with increased mutational frequency that can be associated with the phenotype.

On the contrary, in AI-162 and AI-287 too many candidates remained even after filtering with stricter thresholds for the pathogenicity prediction programs and by stricter filtering for maf, maf < 0.001. To narrow down the candidates it would be suggested to recruit additional available members of the families to sequence, so that the variants common among the family are excluded and only variants that segregate with the phenotype are considered. In Family AI-162 a *RELT* variant was found with low CADD score (CADD = 8,3) which will be discussed in section 2.3.5 below.

For families AI-224 and AI-296 variants that are on AI associated genes and that have already been published as causative for AI, also see Table 2.4, were identified. After confirming that the variants segregate with the phenotype among family members and that the phenotypes of the families correspond to the phenotypes reported in the literature the families were considered as solved.

In families that a variant was identified in a gene already associated with AI, that variant was prioritised and proceeded to segregation. For families: AI-300, AI-318 and AI-335 heterozygous variants on *COL17A1* were found, with the phenotype of the affected members

of the families presenting with rough pitted enamel and signs of hypoplasia, see Figure 2.3b. In the literature variants on *COL17A1* are causative for hypoplastic AI with pitted enamel (McGrath et al., 1996; Tasanen et al., 2000; Prasad et al., 2015), which is consistent with the findings in these three families. Additionally, AI-300 and AI-318 seem to follow an AD MOI, also see Figure 2.2, which is expected of COL17A1 variants, while the family pedigree and medical history for AI-335 could not be determined due to insufficient information. After confirming the variant segregates with the phenotype for the available family members, the families were considered as solved.

A heterozygous *LAMB3* variant was identified in Family AI-248. LAMB3 variants have been associated with Hypoplastic AI, while the proband of AI-248 presents with a mixed hypoplastic / hypomaturation AI phenotype, see Figure 2.4 below. The variant was validated with Sanger sequencing and segregates with the phenotype.

Figure 2.4: Dental photo of 4938, the proband of AI-248.
The proband of AI-248 presents with a mixed hypoplastic / hypomaturation phenotype with severe enamel attrition.

In families AI-297, AI-311, AI-332 and AI-334 *AMBN* variants were found, with AI-311, AI-332 and AI-334 sharing the same (c.209C>G) variant, see Table 2.4. In short, the proband of AI-297 carries a heterozygous c.76G>A variant, the probands in AI-311 and AI-332 are homozygous for the c.209C>G variant and the proband of AI-334 is a compound heterozygote for the c.209C>G and c.295T>C variants. Three of the four families present with Hypoplastic AI, with the exception being AI-311, which as can be seen in Figure 2.3c, was considered to have a mixed hypoplastic / hypomaturation phenotype. The identification of the homozygous variant that was also found in the other families reinforced the suggestion that it was the causative variant despite the differences in the observed phenotype. The MOI of three out of the four families is consistent with the literature, as *AMBN* has been associated with AR inheritance (Poulter, Murillo, et al., 2014; Prasad et al., 2015), but the MOI of AI-297 is not. The WES data of the sample 5027, of AI-297, were examined for an accompanying variant that would explain the phenotype as a compound heterozygote, but none were found that would have a low maf and high CADD score. However, a recent study by Lu et al. (2018) has suggested that variants on *AMBN* can be associated with AD inheritance of AI, while causing severe hypoplastic AI, which could explain the hypoplastic phenotype of AI-297 as being caused by the single variant in an AD manner. In all families the variants were validated by Sanger sequencing and segregation was conducted that confirmed that the variants co-segregate with the phenotype when additional family members were available.

Families AI-291, AI-317 and AI-337 all carry homozygous *RELT* variants that have not been previously published. These families are discussed in more detail in section 2.3.5 below.

2.3.3 Families Sequenced in WES-2019-Batch

The pedigrees of the families included in the WES-2019-Batch are presented in Figure 2.5, below.

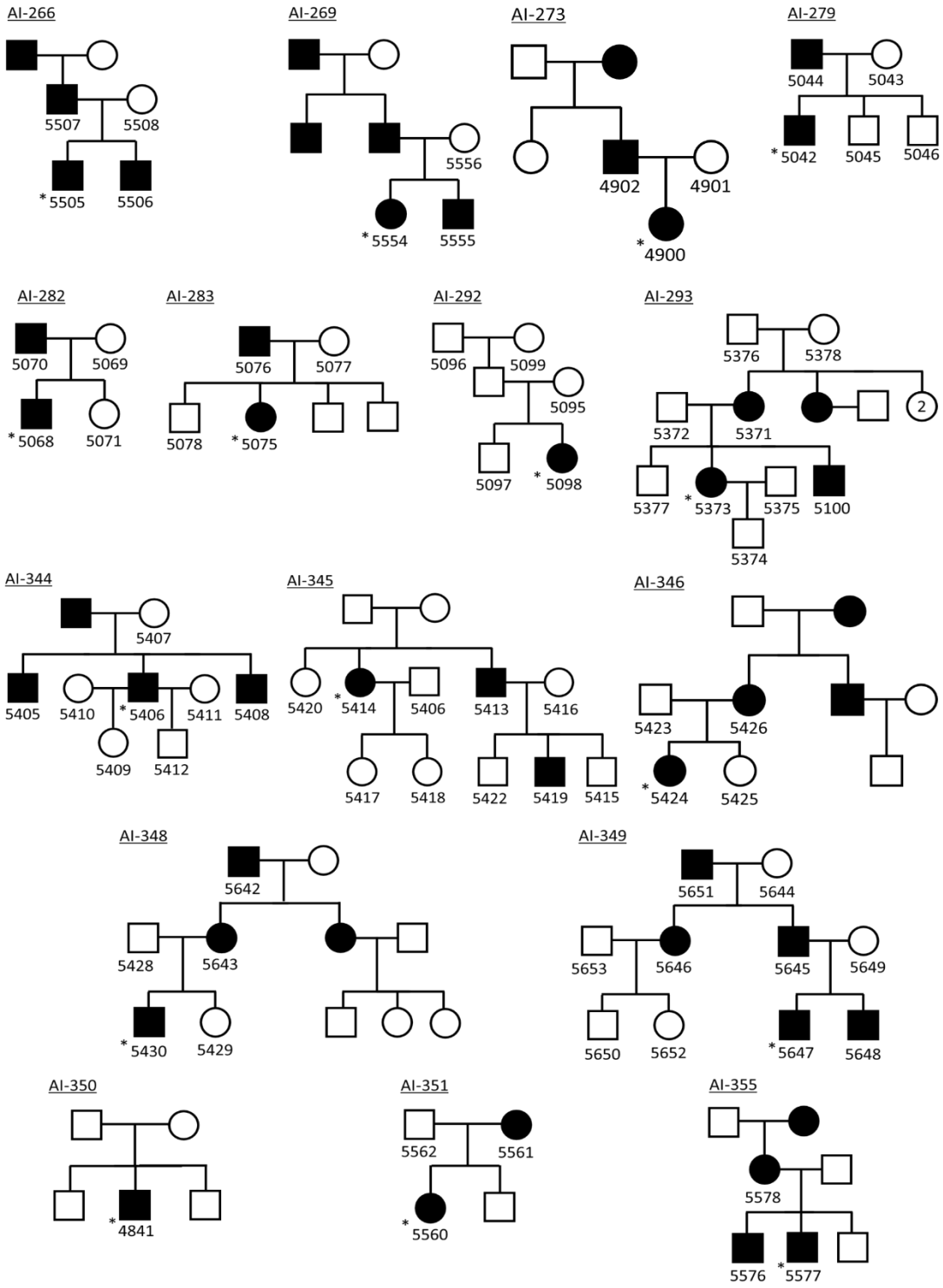


Figure 2.5: Pedigrees of the families included in WES-2019-batch.

The family members recruited for this study have 4-digit codes assigned to them. The proband is indicated with an asterisk (*). Individuals presenting with an AI phenotype are shown as filled in squares or circles, with square indicating a male and circle a female.

Most of these families were determined to be following an AD mode of inheritance, with AI-292 being an obvious outlier. The phenotype of the proband of AI-292, sample 5098, was considered to be presenting with mixed Hypoplastic and Hypomaturation AI, and the lack of an AI phenotype was not clear if it was due to an AR inheritance, incomplete penetrance of a gene the follows AD inheritance, or a *de novo* appearance of a new to the family mutation. The phenotype of 5098 can be seen in the dental photos, shown in Figure 2.6. Other representative examples of Hypomaturation and Hypoplastic AI are shown in Figure 2.7 and 2.8 respectively.

Figure 2.6: Dental photographs of 5098, the proband of AI-292.

(a) View of the upper jaw, (b) view of the lower jaw, (c) front view of the teeth of 5098. The phenotype has the characteristics of Hypomaturation AI.

Figure 2.7: Dental photographs of 5042, the proband of AI-279.
(a) View of the upper jaw, (b) view of the lower jaw, (c) front view of the teeth of 5042. The phenotype has the characteristics of Hypomaturation AI.

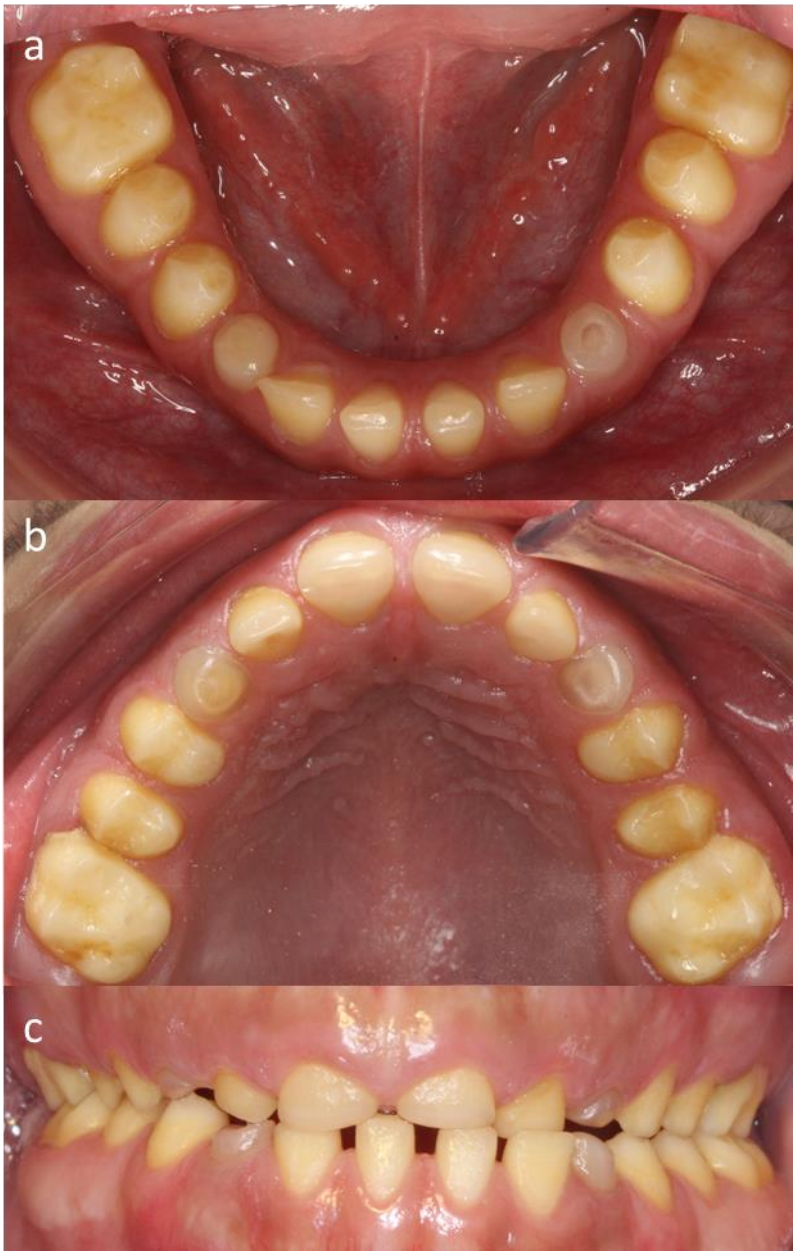


Figure 2.8: Dental photographs of 4841, the proband of AI-350.
(a) View of the upper jaw, (b) view of the lower jaw, (c) front view of the teeth of 4841. The phenotype has the characteristics of Hypoplastic AI.

2.3.4 WES-2019-Batch Key Findings

Table 2.5: WES results from the 2019 batch of samples.

In the case of variants that have been previously published as causative for AI the reference is provided, whereas variants marked as 'new variant' have not been previously associated with AI. The most likely candidate variant is shown here for each family, other candidates with CADD > 15 are presented in the electronic Appendix.

WES 2019 batch							
Family	Candidate Gene	Notes	Coding	Protein	CADD	maf (gnomAD)	dbSNP
266	-	too many candidates	-	-	-	-	-
269	*LAMB3*	(Poulter, El-Sayed, et al., 2014)	c.3394dup	p.(E1132Gfs*28)	35	N/A	rs786201004
273	*ENAM*	(Brookes et al., 2017)	c.92T>G	p.(L31R)	32	N/A	rs1060499539
279	*LAMB3*	new variant	c.2660G>A	p.(R887H)	16.23	0.000049	rs183045589
282	*LAMA3*	new variant	c.1570C>T	p.(R524C)	33	0.000037	rs754718312
283	*AMELX*	new variant	c.100G>A	p.(E34K)	32	N/A	-
292	*ENAM*	new variant	c.263A>G	p.(Q88R)	22.2	0.000267	rs565258194
293	*AMELX*	(Hart et al., 2002)	c.230A>T	p.(H77L)	25.4	N/A	CM022608
344	*COL17A1*	new variant	c.3595G>C	p.(E1199Q)	25.1	N/A	-
345	*FAM83H*	(Urzúa et al., 2015)	c.1669G>T	p.(G557C)	22.7	0.002772	rs312262803
346	*FAM83H*	(Hyun et al., 2009)	c.1354C>T	p.(Q452*)	37	0.000004	CM096324
	LAMB3	new variant	c.898C>T	p.(R300W)	33	0.000238	rs368834085
348	*AMELX*	new variant	c.167A>G	p.(Y56C)	25.1	N/A	-
349	*FAM83H*	(Kim et al., 2008)	c.1192C>T	p.(Q398*)	23.7	N/A	rs137854436
350	*FAM20A*	(Jaureguiberry et al., 2013)	c.727C>T	p.(R243*)	41	0.000024	rs367720325
	FAM20A	(Jaureguiberry et al., 2013)	c.907-908delAG	p.(S303Cfs)	35	0.000037	rs750880244
351	*AMELX*	(Prasad et al., 2015)	c.155C>T	p.(P52L)	27.4	N/A	rs387906487
355	*FAM83H*	(Kim et al., 2008)	c.1192C>T	p.(Q398*)	23.7	N/A	rs137854436

In this second batch in the majority of the families, specifically in AI-269, AI-273, AI-293, AI-345, AI-349, AI-350, AI-351 and AI-355, variants were found in AI associated genes that have previously been published as causative for AI, see Table 2.5. The observed AI phenotype in the affected members of these families were found to be in accordance with the expected phenotypes, as they are reported in the literature.

For Family AI-266 there were too many possibly pathogenic variants identified by WES, even after filtering the results with stricter filters, such as maf < 0.001 and CADD score > 20, similarly to AI-162 and AI-287 that were mentioned previously. Additionally, no dental photos or information about the phenotype of the family were provided, so the genetic basis of the symptoms cannot be narrowed down to specific pathways.

Families AI-279, AI-282 and AI-344 were found to carry heterozygous variants on the cell adhesion proteins, *LAMB3, LAMA3* and *COL17A1* respectively. Both laminins are linked to hypoplastic AI but there is no information about the phenotypes of AI-282, so it was not possible to compare the phenotype of the family to the literature, while for AI-279 the phenotype was determined as hypomaturation AI, shown in Figure 2.7, *COL17A1* is associated with hypoplastic / hypomaturation AI, but there is no information on the phenotype of the affected family members. The variants were validated with Sanger sequencing, using primers designed to amplify with PCR the region containing each respective variant, but the segregation analysis for all three families was inconclusive. For each family the proband and their parents were confirmed to show the expected segregation profile, but some of the other family members did not, for the family pedigrees refer to Figure 2.5. Specifically, in AI-279 and AI-282 the siblings of each proband, 5045 and 5071 respectively, were shown by Sanger sequencing to carry the variant as a heterozygote, while they have been reported as not affected by AI. In Family AI-344 the children of the proband, 5409 and 5412, were also shown to carry the variant as heterozygotes, but had been reported as not affected by AI. A common characteristic of all these four people is that at the time of recruiting and sampling they were of young age and it is possible that they are affected by AI but were not diagnosed as such. Consequently, the rejection of the findings on the basis of the failure to segregate is not possible before a re-evaluation of the youngest family members in each of the three families and a better description of their respective phenotype.

Families AI-283 and AI-348 were shown to carry novel heterozygous missense *AMELX* variants. The variants were validated with Sanger sequencing, using the appropriate primers for PCR amplification, and segregation showed that the variants segregate with the phenotype among the family members. There is no available phenotype information about these two families, which doesn't allow the comparisons to the expected hypoplastic / hypomineralised AI that is associated with variants on *AMELX*. As is the case with *AMELX* variants the female family members are heterozygous for the variant, e.g.: 5075 of AI-283, while the males only have one copy of the X chromosome and show as homozygous in segregation, e.g.: 5430 of AI-348.

The proband of Family 292, 5098, presents with mixed hypoplastic and hypomaturation AI, as mentioned earlier, also see Figure 2.6. A novel *ENAM* variant was identified and was validated by Sanger sequencing and confirmed by segregation that the family members that

are reported as not affected by AI do not carry the variant. The MOI of the family was expected to be AR, due to both parents being unaffected by AI and their child being affected, however, the *ENAM* variant found in 5098 is a heterozygous variant, suggesting an AD MOI. Any attempt to find an accompanying second variant on *ENAM* was unsuccessful. As was mentioned previously, *ENAM* has been associated with both AD and AR inheritance, also see Table 1.3, so a heterozygous variant can be sufficient to cause an AI phenotype. In this case the variant can be explained by a *de novo* mutation in 5098, as none of the parents carry the variant.

In Family AI-346, WES of the proband, sample 5424 as shown in Figure 2.5, revealed variants in both *FAM83H* and *LAMB3* which segregated with the phenotype among the family members. The phenotype of 5424 could not be determined as no dental photos or radiographs of were provided, so neither of the variants could be rejected as not fitting with the phenotype. The *FAM83H* variant has been previously reported as causative for AD hypocalcified AI (Hyun et al., 2009) and as a missense variant it will lead to a truncated peptide and a haploinsufficiency effect. According to Hyun et al this variant is sufficient to cause a severe AI phenotype with soft and uncalcified enamel, but the new *LAMB3* variant that was also identified is also predicted to be pathogenic. Without more information on the phenotype of the affected members of AI-346 it cannot be determined if and how much each of the variants contribute to the AI phenotype and so neither variant can be rejected, nor the possibility that the combination of mutations in both genes is responsible for causing the phenotype, a phenomenon which also called digenic inheritance (Gazzo et al., 2017).

2.3.5 Families Carrying Variants on RELT

The analysis of the results from WES batch-2018 (Section 2.3.1), revealed that members of three families carried homozygous single nucleotide variants in *RELT*, which is a gene that was recently associated with autosomal recessive hypoplastic AI (Kim et al., 2019). RELT is a member of the Tumour Necrosis Factor Receptor superfamily (TNFRs), although its specific function has not yet been determined. *RELT* has been reported to activate the NF-κB pathway in hematopoietic tissues (Sica et al., 2001) and NF-κB has been shown to affect amelogenesis, so the disruption of NF-κB due to mutations on RELT could be a potential explanation of the mechanism that causes the AI phenotype, although further research is required to clarify the role of RELT. Kim et al identified three consanguineous families, presenting with irregular hypoplastic enamel, showing excessive attrition and a generalized AI phenotype. In animal studies with Crispr/*CAS9* transgenic mice *RELT* was found to be expressed in secretory stage ameloblasts and in odontoblasts. Additionally, Kim et al report the medical history of the probands in the three families they recruited, which included frequent infections, febrile convulsions and short stature, and suggest that variants on *RELT* are causative for a syndromic AI phenotype, although they admit that no other systemic symptoms were identified or reported among the family members.

Among the Leeds AI cohort, two of the families, AI-291 and AI-317, were shown to carry the same exonic variant, c.164C>T, p.(T55I), and Family AI-337 carried a different exonic variant, c.1264C>T, p.(R422W). This led to the re-examination of the list of variants found in older WES samples from unsolved families of the AI cohort, with a focus on *RELT* and its

adjacent genomic regions. It was found that members of two further families carried nucleotide variants in *RELT*, one of them, AI-37, sharing the same exonic variant, c.164C>T, p.(T55I), as families AI-291 and AI-317 and the other, AI-162, carrying a heterozygous exonic variant, c.800C>T, p.(A267V).

All families except for AI-162, were biallelic for the variant identified, which is in accordance with the autosomal recessive mode of inheritance reported in the literature on *RELT* (Kim et al., 2018). AI-162 was therefore re-examined by looking into the CNV results and relaxing the hard-filter thresholds of the analysis, to try to find a second variant within *RELT*, the pedigree for AI-162 is presented in Figure 2.9. However, no other variant was identified within *RELT* for sample 4475, the proband from Family AI-162, also see the family pedigree shown in Figure 2.9. Additionally, although families AI-37, AI-291, AI-317, AI-337 present with hypoplastic AI, a phenotype which is consistent with that reported by Kim and colleagues in families with biallelic RELT variants, the phenotype of Family AI-162 could not be clearly determined by the photos provided by the dentist that recruited the family, Figure 2.9a,b. This lack of a distinct hypoplastic phenotype, in addition to the lack of a variant on the second allele in *RELT*, does not correlate with the characteristics of AI caused by *RELT* variants as described by Kim et al and as observed in the other four families. As a result, this variant could not be determined to be causative for AI on its own. Family AI-162 was therefore omitted from any subsequent analysis of the *RELT* variants and patients.

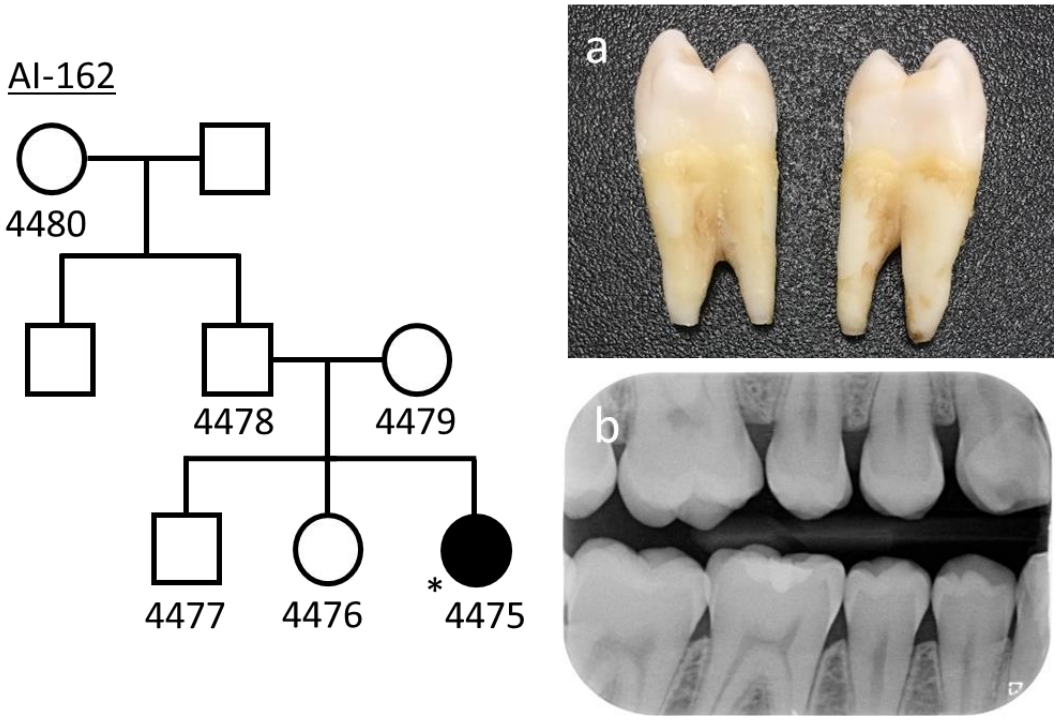


Figure 2.9: Pedigree and photos of teeth from AI-162.
The proband of the family, sample 4475, is indicated by an asterisk (*) and is the only member that has been diagnosed with AI. (a) teeth extracted for clinical reasons from 4475, (b) radiographic (x-ray) images of the dentition of 4475. The radiographic image shows the enamel density and thickness of the enamel layer, showing that there is no clear indication of an abnormal phenotype.

To confirm that the variants identified in samples from the four families segregate among affected and unaffected family members, primers were designed for PCR amplification and subsequent Sanger sequencing. The primers are presented in Appendix A.1. The results of the Sanger sequencing confirmed both the validity of the variants and that the variants segregate as expected amongst the family members.

2.3.1.1 Family AI-37, AI-291 and AI-317

Family AI-37 is a UK Pakistani family with the proband and one sibling presenting with hypomineralised AI, the family pedigree is presented in Figure 2.10. The radiographs of the unerupted permanent teeth demonstrate an apparently normal enamel volume with a normal difference in radiodensity between enamel and dentine and loss of the normal crown contours after eruption. Teeth have variable loss of enamel consistent with post-eruptive changes characterised by irregular surface loss and associated discolouration that progressed over time. The other three siblings and both parents are unaffected and there is no family history of AI reported. The parents of the proband reported that they were thought to be distant relatives, suggesting potential consanguinity. The medical histories of the affected children lack any reference to recurrent infection during infancy and there are no other recognised co-segregating clinical features. The WES analysis of the proband revealed a *RELT* missense variant that segregates with the disease phenotype in all family members in an autosomal recessive manner (Figure 2.11). Affected individuals are homozygous for a variant in *RELT* exon 4: c.164C>T [Refseq: NM_032871.3], p.(T55I) [NP_116260.2]. This variant is not present in gnomAD and is predicted to be pathogenic by both MutationTaster2 (Disease causing, p: 0.951) and CADD v1.3 (score: 27.6).

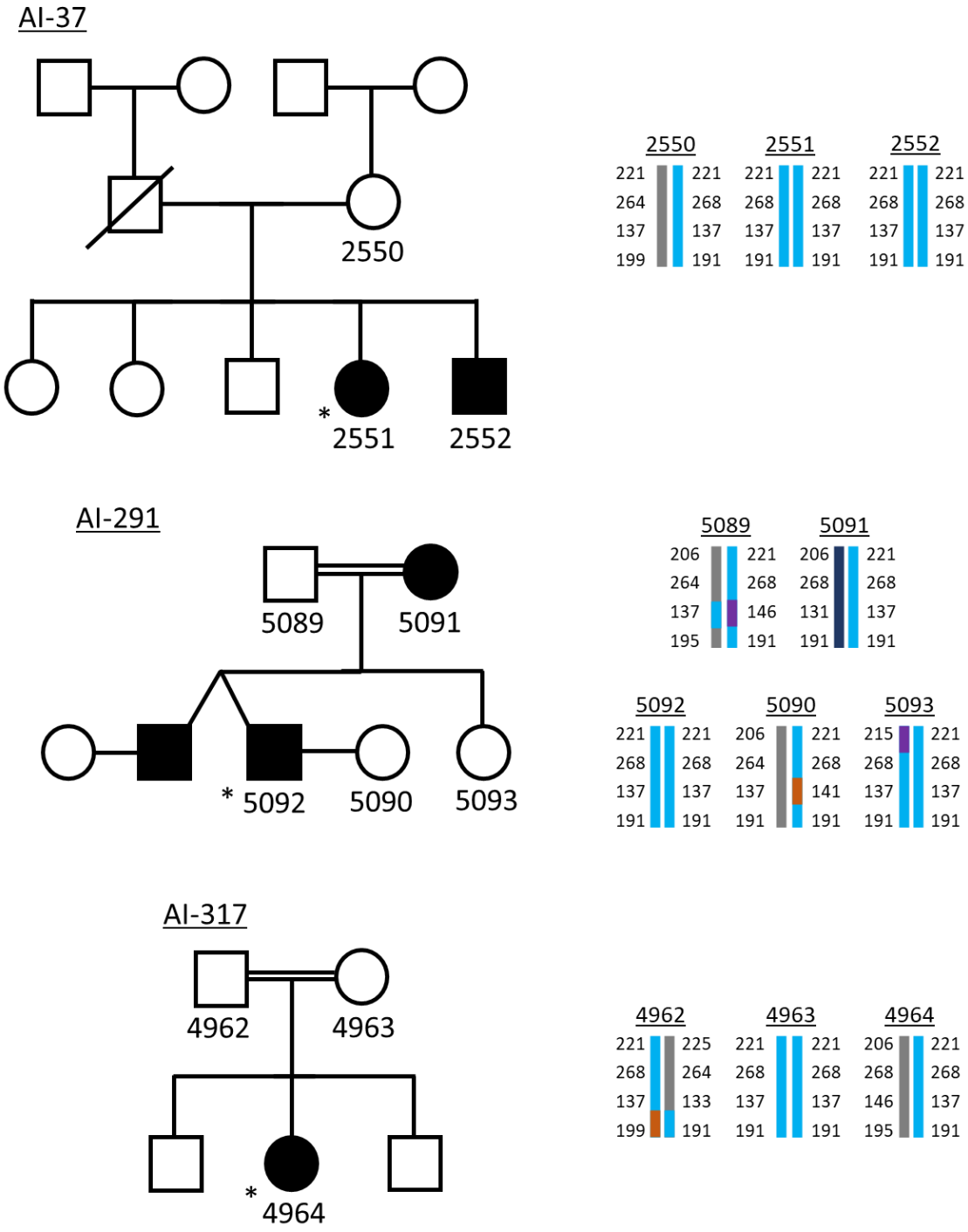


Figure 2.10: Pedigrees and genotyping results for families AI-37, AI-291 and AI-317.
The affected people are shown as filled in circles or squares for female or male family members respectively. The people recruited in this study have been assigned 4-digit codes and the proband is indicated with an asterisk (*). The haplotypes from the microsatellite genotyping of the recruited family members are shown next to the corresponding pedigree. The numbers show the number of repeats found for each microsatellite marker in the order of D11S1314, D11S4184, D11S916, D11S2371 from top to bottom.

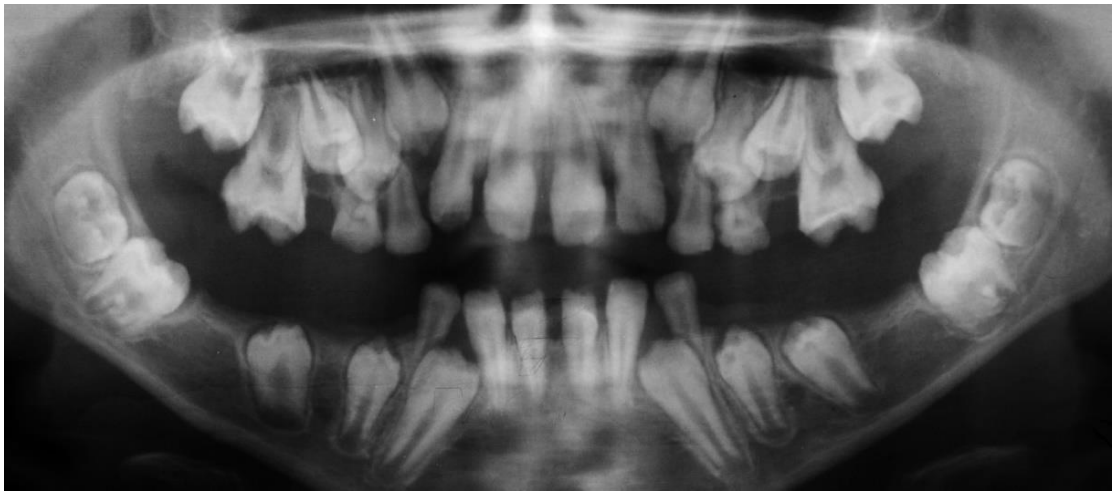


Figure 2. 11: Radiographs of the proband of Family AI-37.

The unerupted permanent teeth show a normal enamel volume with a clear distinction in radiodensity between enamel and dentine. There is visible loss of the normal crown contours after eruption.

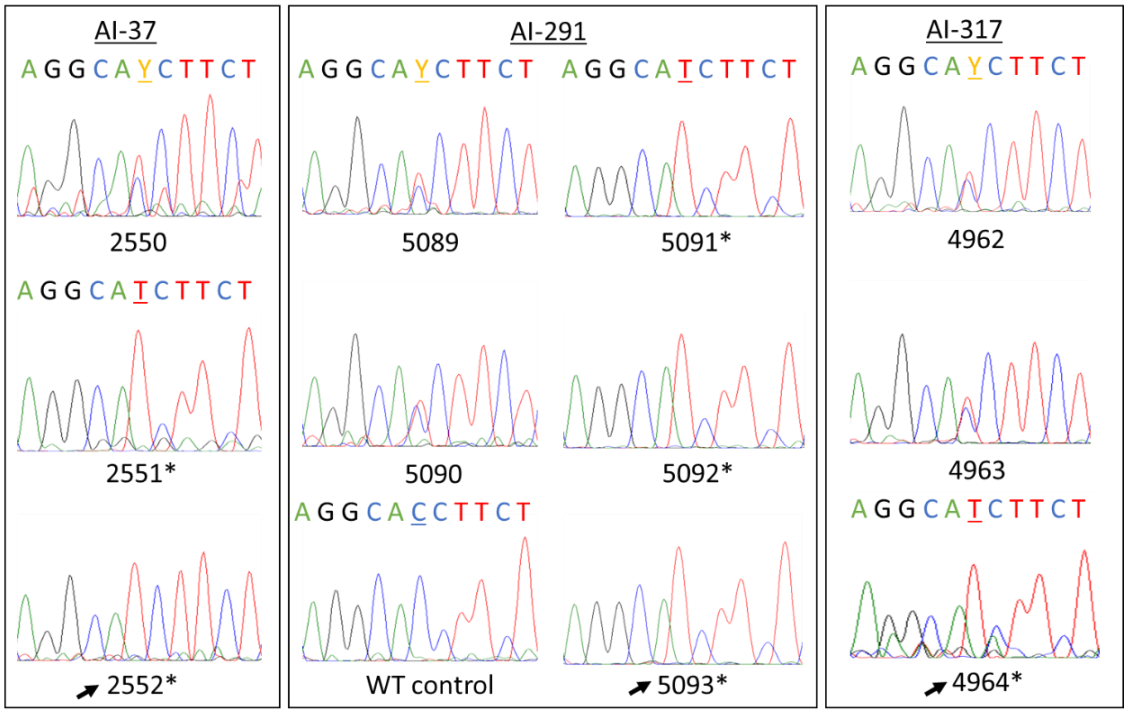


Figure 2.12: Electropherograms of recruited family members of families AI-37, AI-291 and AI-317. The electropherograms are shown centred at the position of the variant c.164C>T [NM_032871.3], p.(T55I) identified in exon 4 of *RELT*. A WT control sample is also included for comparison. The individuals that were reported as phenotypically affected are indicated with an asterisk (*). The proband in each family is indicated by an arrow. The nucleotide sequence is also shown above the electropherograms, with the variable position underlined in each case.

Family AI-291 is a consanguineous UK family, also of Pakistani origin. The proband, her two brothers, mother and maternal grandfather are all reported to be affected, see Figure 2.10 for the family pedigree. The radiographic phenotype was consistent with that of Family AI-37, although no other clinical images were available. No other potential syndromic features were apparent and there were no reports of recurrent infections in infancy or childhood. The WES analysis of the proband identified the same homozygous c.164C>T, p.(T55I) variant that was identified in Family AI-37.

Family AI-317 is also reported to be a consanguineous UK Pakistani family, Figure 2.10. The proband is the only one of three siblings diagnosed with AI, characterised by enamel surface irregularities and evidence of good enamel volumes on radiographs. No other potentially syndromic features were noted and there was no history of recurrent infection in infancy, however, the proband and both other siblings were also diagnosed with isovaleric acidemia (IVA, OMIM # 243500), a condition in which the body is unable to metabolise leucine. IVA is caused by variants in the gene encoding isovaleryl-CoA dehydrogenase (*IVD*, OMIM * 607036) and is inherited in an autosomal dominant manner without any recognised impact on enamel formation. A pathogenic *IVD* variant had previously been identified in this family as part of clinical care. WES analysis of the proband revealed the same homozygous variant in exon 4 of *RELT* as in families AI-37 and AI-291, c.164C>T, p.(T55I), as well as the heterozygous *IVD* variant in exon 3:c.280G>A [Refseq: NM_002225.5], p.(G94S)[NP_002216.3] which was previously mentioned.

As AI in the three families (AI-37, AI-291 and AI-317) results from homozygosity for the same variant and the families have the same ethnic background, the affected family members were examined to investigate whether they share the same haplotype. Genotyping was performed on all available family members from each family, with microsatellite markers flanking *RELT* on chromosome 11q13, across a 1.1 cM / Mb region, in the order: 11 cen, D11S1314, D11S4184, *RELT*, D11S916, D11S2371, 11 qter. It confirmed that they share a common haplotype, implying that the three families are distantly related, Figure 2.10.

2.3.5.1 Family AI-337

The fourth family carrying a *RELT* variant, Family AI-337, is a non-consanguineous Costa Rican family presenting with AI in three children characterised by irregular surface loss of enamel and dental radiographs that indicated good enamel volumes, dental photos of the teeth of the three siblings are shown in Figure 2.13. The parents do not have the same phenotype as the children, although the mother presents with minor enamel regularities on the cusps of the molars and canines. These minor irregularities could potentially be due to a mild AI phenotype, but this could not be confirmed. There is no report of any extra-oral disease phenotypes or recurrent infections in the family medical history. WES analysis of DNA from the proband identified a homozygous *RELT* variant in exon 11: c.1264C>T, p.(R422W), which segregated with the AI phenotype, see Figure 2.14. This variant is present in dbSNP (rs139368769) and also in gnomAD with maf: 0.001633 and in EVS with maf: 0.00077. It is predicted to be damaging by both MutationTaster2 (disease causing, p: 0.683) and CADD v1.3 (score: 34).

Figure 2.13: Dental photos of the three affected siblings of AI-337.
(a) photos of sibling 1, designated 5298 on the pedigree, (b) 5299, (c) 5300. In all views the phenotype is clearly visible, with the irregular surface loss of enamel and discolouration being the most prominent. Examples of both are highlighted with white arrows.

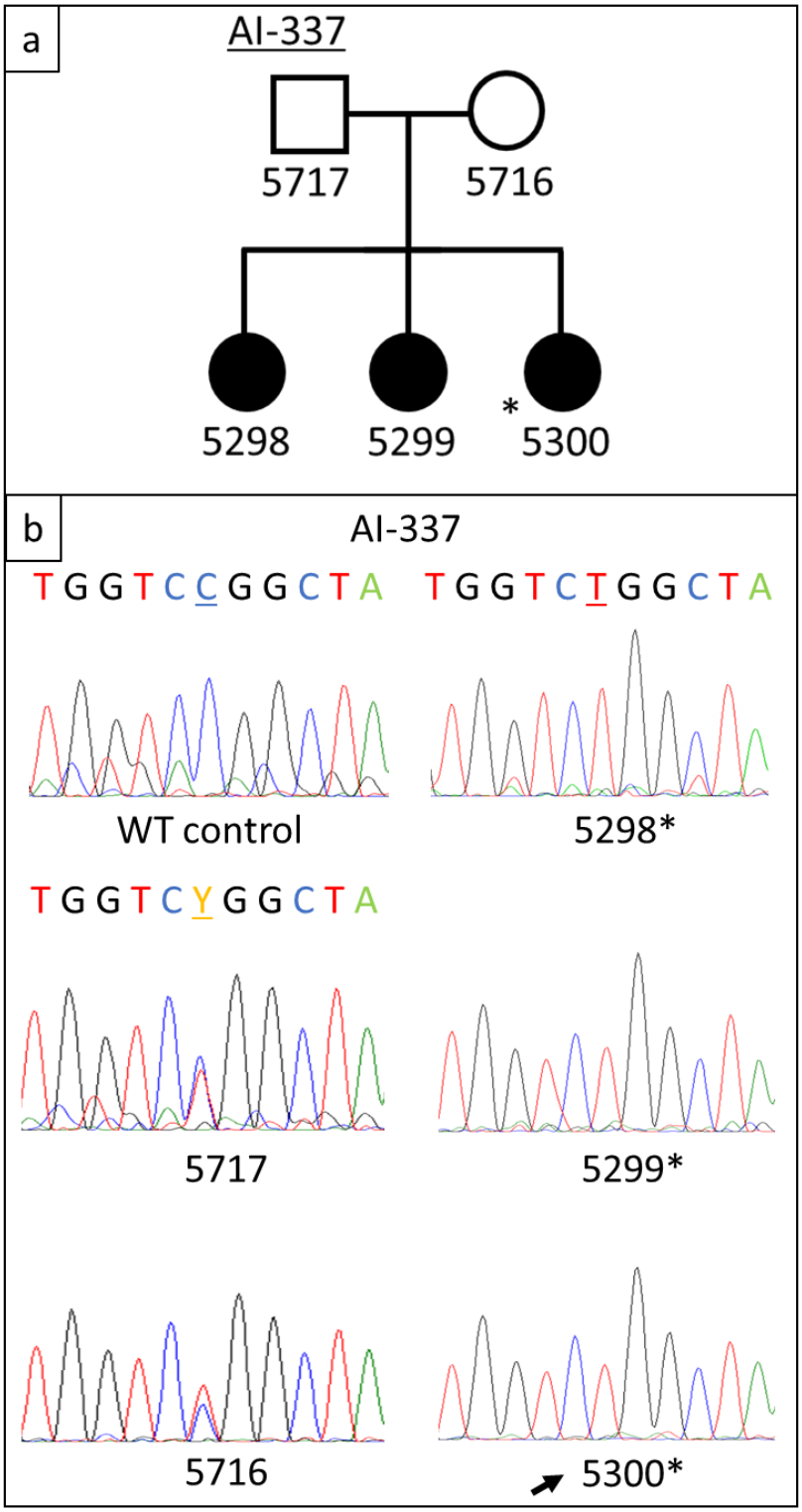


Figure 2.14: Pedigree and electropherograms of AI-337.

(a) The proband of the family is indicated with an asterisk (*) on the pedigree and all affected are shown as filled in. (b) The electropherograms used to validate the variant found and for the segregation within the family, centred at the position of the variant, c.1264C>T, p.(R422W) in exon 11 of *RELT*. The proband is shown with an arrow and all affected members are indicated with and asterisk (*). A WT control is included for comparison and the nucleotide sequences are shown above each electropherogram, with the variable position being underlined.

2.3.5.2 Microstructure of Teeth Affected by RELT Variants

Families AI-317 and AI-337 had donated teeth that had been extracted by a dentist as part of the affected individuals' treatment. These teeth were used to study the microstructure of the enamel affected by the two novel *RELT* variants. Specifically, the microstructure of an adult permanent molar tooth from the proband (IV:2) of Family AI-317, and an exfoliated primary incisor from the proband (II:2) of Family AI-337 were analysed using µCT and then SEM, as described in Section 2.2.11. The teeth affected by AI were compared to age-appropriate controls of unaffected teeth, of the same age and type, obtained from the Leeds Tissue Bank, as shown in Figures 2.15, 2.16 and 2.17. Ideally the samples and the controls would also be matched for the donor's sex, but due to the limitations of the tissue bank and the teeth available to us this was not possible. The enamel density of the teeth shown in Figure 2.15 seems at the normal levels, similarly for the enamel thickness. The wear of the crown in Figure 2.15b, however, impedes an accurate judgement.

Both control teeth display the expected prismatic enamel structure, as shown in Figure 2.16a and 2.16b and 2.17a and 2.17b. Analysis of the affected molar from Family AI-317 revealed two layers of enamel. The outer enamel layer has enamel prisms with abnormalities, while the inner enamel layer is non-prismatic with an abnormal lamellar structure. The affected incisor of the proband (II:2) of Family AI-337 also has two distinctive layers of enamel. The inner layer shows the regular configuration of enamel prisms, which then become disorganised and change orientation to become non-prismatic, layered enamel. No enamel pits were identified and there is no indication in either tooth that the dentine or the dentine-enamel junction are affected. It was not possible to explain how the *RELT* variants can lead to the formation of these distinct layers of enamel, however, they bear similarity to the enamel abnormalities reported as caused by variants on other genes, as described in section 2.1.5, and specifically with the disorganised enamel reported by (Smith et al., 2016), on variants in *AMTN*, shown in Figure 2.18. Although Smith et al do not report distinct layers of disorganised enamel, they observed a formation of enamel that resembles the lamellar structure observed on the teeth presented in Figures 2.16 and 2.17.

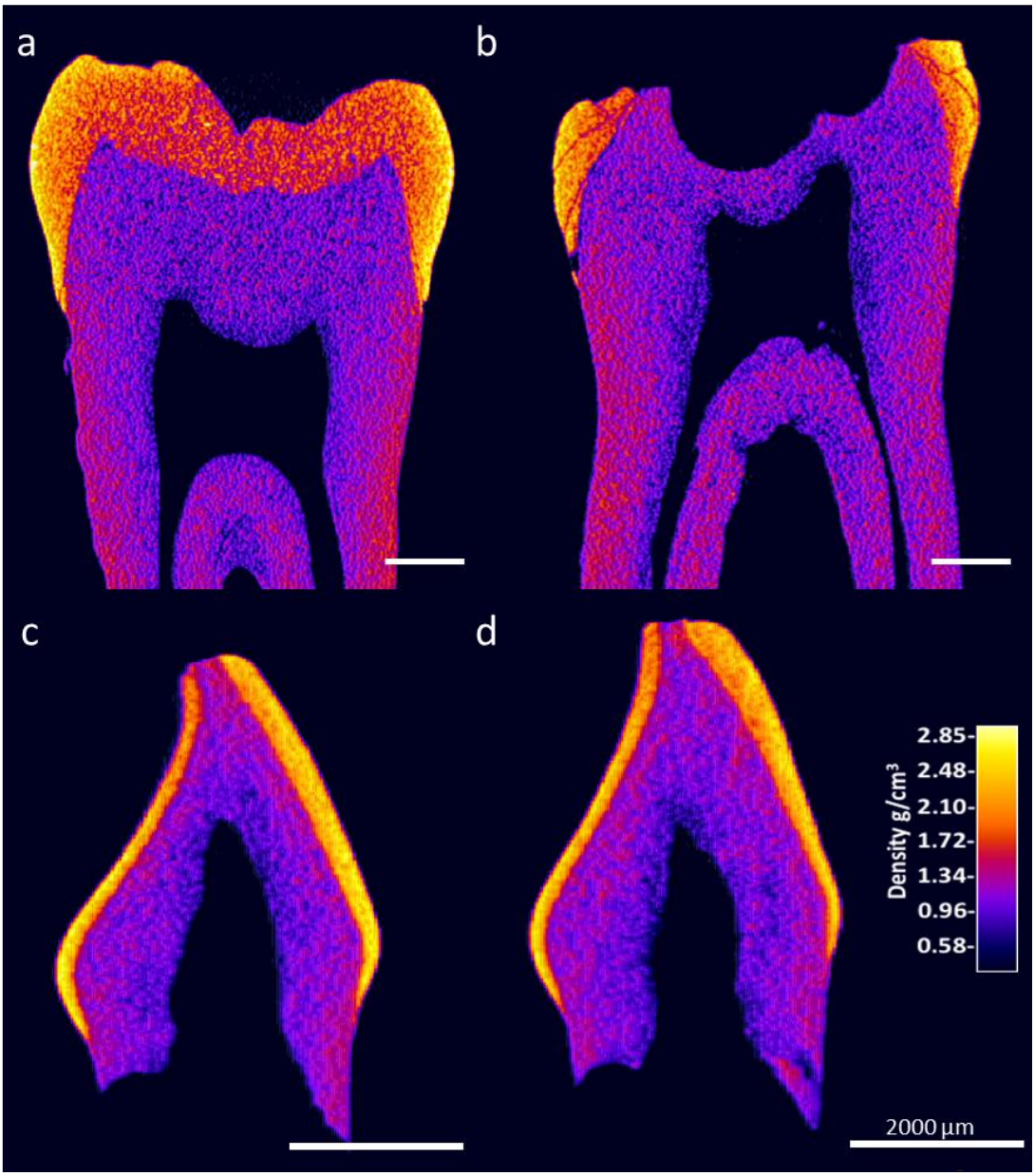


Figure 2.15: Calibrated enamel density heatmaps of microCT scan sections.
(a) Control adult molar, with normal enamel. (b) Affected adult molar from Family AI-291. Enamel density looks normal, but judgement is impeded by the corrosion of the crown. (c) Control deciduous incisor, with normal enamel. (d) Affected deciduous incisor from Family AI-337. The enamel density looks normal, indicative of the observed hypoplastic phenotype. Figure from Nikolopoulos et al., 2020, CC-BY-SA 4.0.

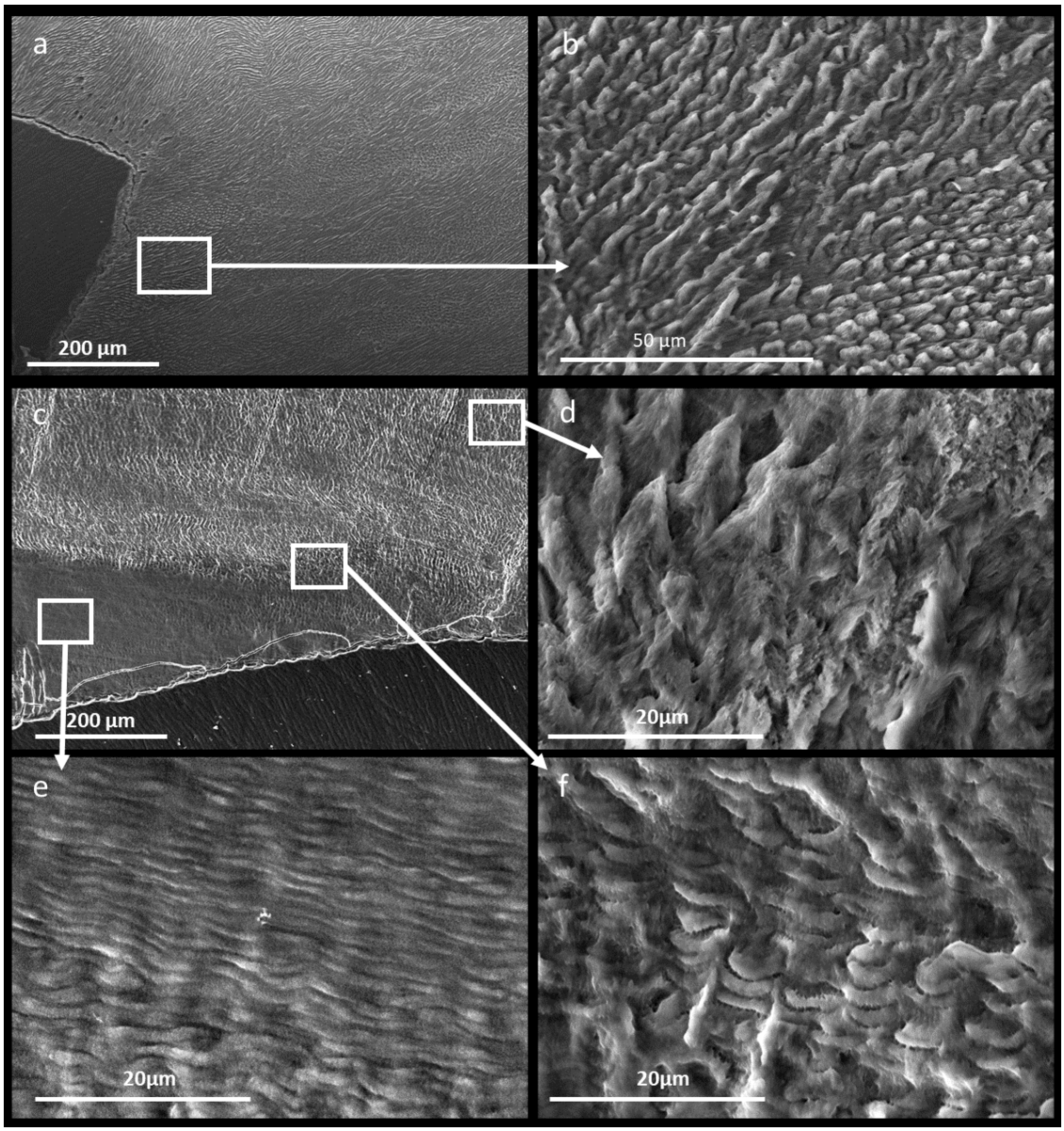


Figure 2.16: SEM photos of sections of adult molars.

(a), (b) Control L7 molar. (c) Section of the affected permanent molar from the proband (IV:2) of Family AI-317, sectioned across the occlusal to basal axis, with the inserts showing: the outer layer of enamel with a few abnormalities (d), the inner layer of stratified enamel (e), and the transitional phase between the layers (f). Both teeth were sectioned across the occlusal to basal axis, then acid etched and gold plated prior to imaging. Figure from Nikolopoulos et al., 2020, CC-BY-SA 4.0.

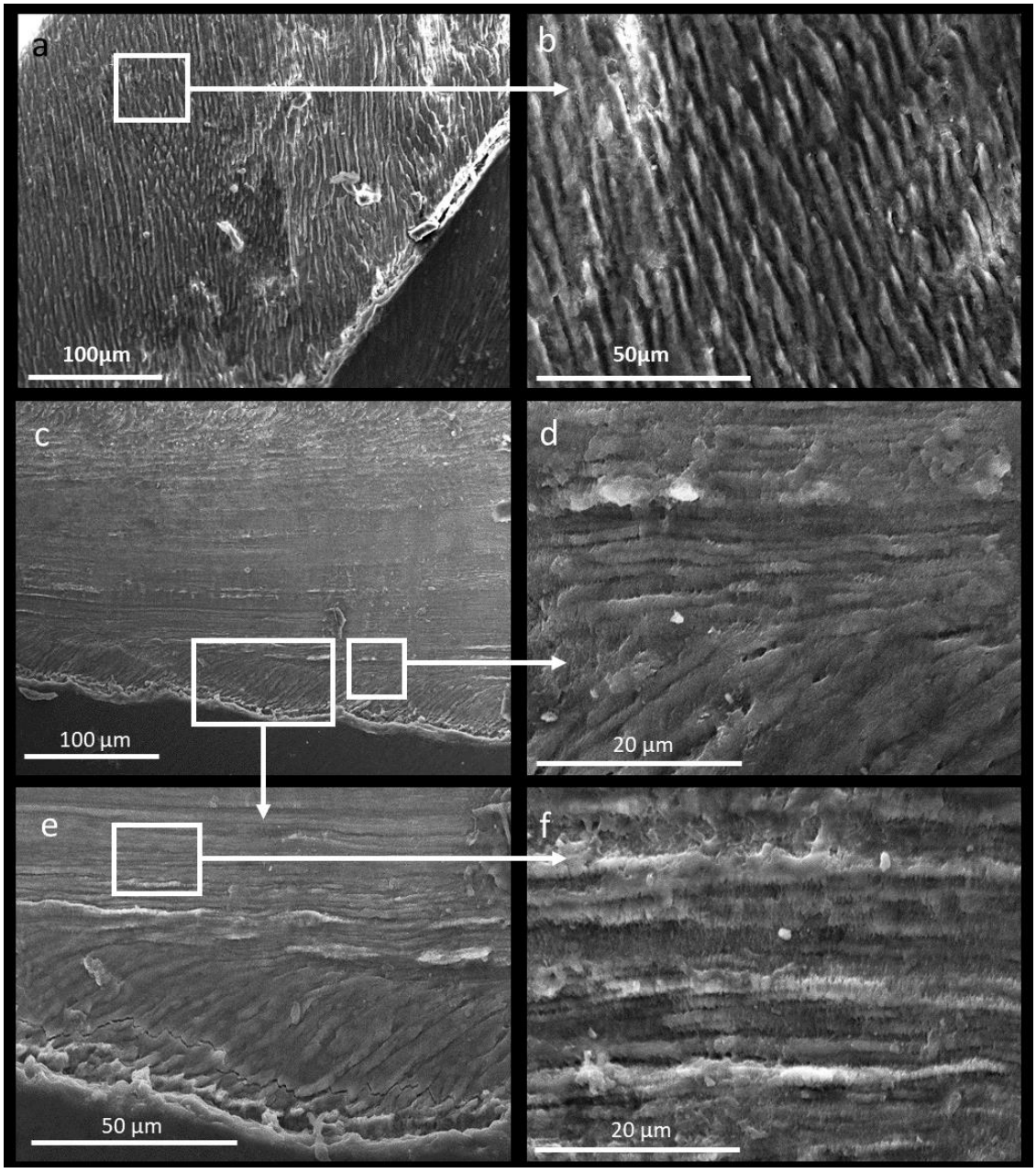


Figure 2.17: SEM photos of sections of deciduous incisors.

(a), (b) Control incisor. (c) Affected incisor from the proband (II:2) of Family AI-337, with the inserts showing the stratified enamel (d, e, f). Both teeth were sectioned across the occlusal to basal axis, then acid etched and gold plated prior to imaging. Figure from Nikolopoulos et al., 2020, CC-BY-SA 4.0.

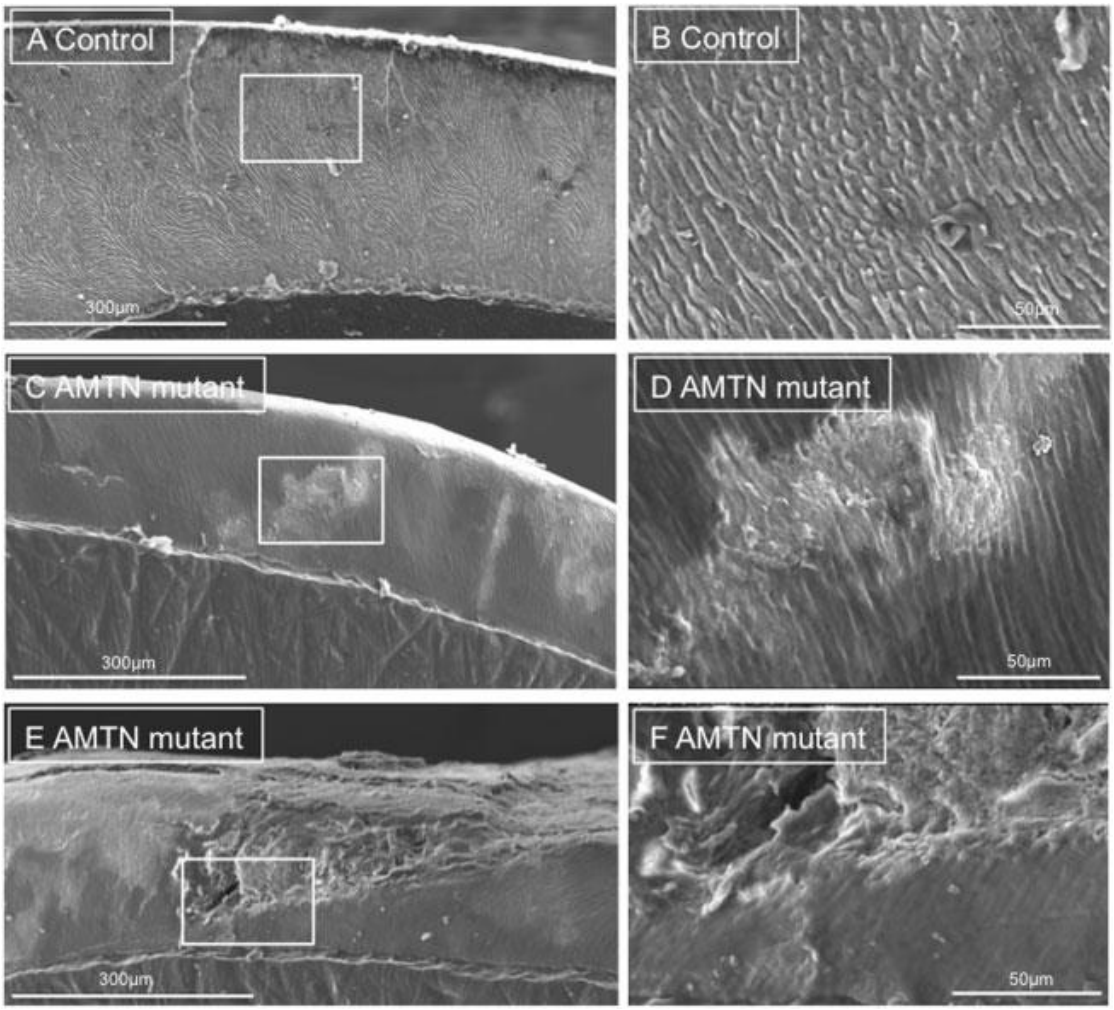


Figure 2.18: SEM of sections of representative exfoliated teeth.
(a), (b) The normal prismatic enamel can be seen on the SEM of the control, (c - f) the disrupted structure of the *AMTN* mutant. The photos on the right side column correspond to the boxed regions of the photos of the left side column. Figure from Smith et al., 2016, CC-BY-SA 4.0

2.3.5.3 Protein Structure Homology Prediction

To better understand the effect that the variants have on the functionality of the protein, the tertiary structure of RELT was predicted using homology models based on structures of other members of the TNFRs that are available in PDB, see section 2.2.12 for details. Domain location analysis identified four main regions: signal peptide (M1 to T25), TNFR domain (T27 to S112), transmembrane domain (Y163 to C183) and RFRV motif (R349 to V352). Secondary structure prediction showed two α-helices and sixteen β-sheets for the protein, shown in Figure 2.19. However, because the structure for RELT predicted is based on homologous structures covering only parts of the protein, instead of being directly observed with experimental support, it cannot be verified whether this is an accurate representation of the protein in its WT form.

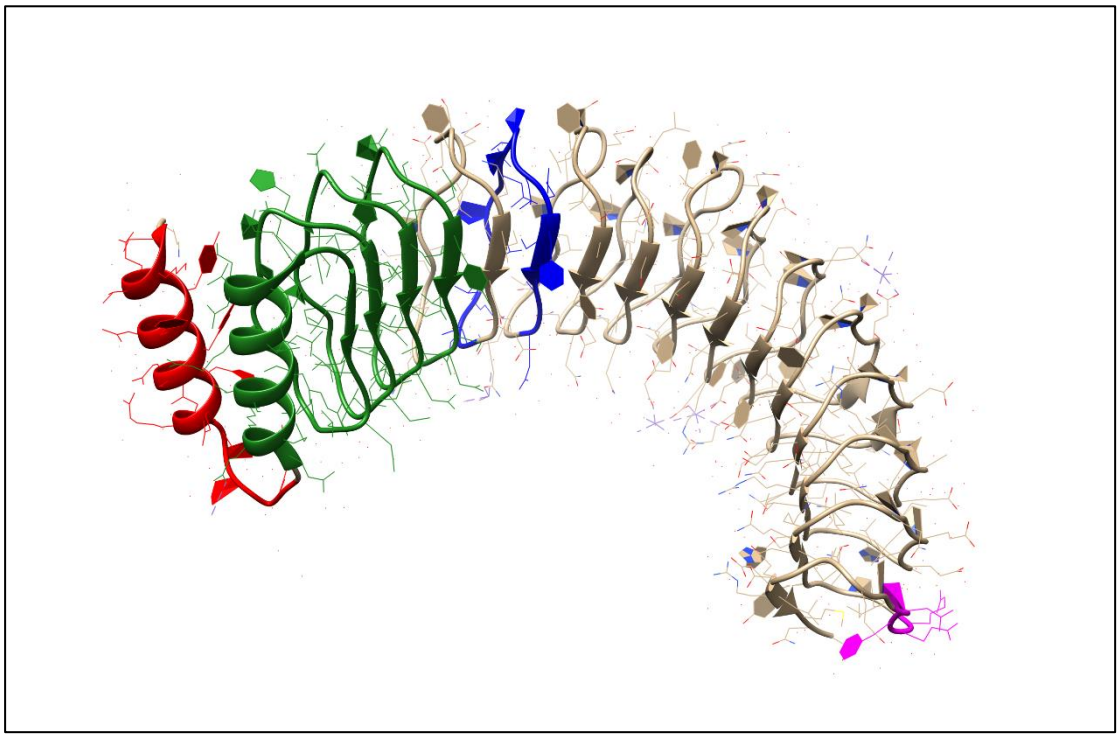


Figure 2.19: Tertiary structure of RELT, predicted by homology searching.
The signal peptide has been coloured red, the TNFR motif green, the transmembrane domain blue and the RFRV motif purple. Only the N-terminus up to the TNFR region, has been predicted with accuracy as there was no homology found between the rest of the peptide and the available structure models.

2.3.6 Families with pathogenic *MMP20* variants and common founder haplotypes

Following up on work conducted by Dr Claire E. L. Smith (University of Leeds) and Dr James A. Poulter (University of Leeds), families presenting with autosomal recessive hypomaturation AI that were examined by WES analysis were shown to contain variants within *MMP20* or its adjacent genomic regions. A total of nine unrelated families carrying homozygous *MMP20* variants and a tenth family with compound heterozygous *MMP20* variants were identified. The position of these variants on the gene is shown in Figure 2.20a. These ten families presented with features consistent with autosomal recessive hypomaturation AI, typical of the phenotype caused by pathogenic variants in *MMP20.* All families also reported the absence of any co-segregating disease. PCR and Sanger sequencing were used to confirm the variants that were previously identified by WES of the affected members of each family and also to show that these *MMP20* variants segregated with the AI phenotype in all available family members, also see Figure 2.20b-i. The DNA extraction, PCR amplification and sequencing for the segregation of the variants in families AI-13, AI-39, AI-52 were performed by Dr James A. Poulter, for families AI-77, AI-79, AI-155 and AI-187 were performed by Dr Claire E.L. Smith and for families AI-218, AI-239 and AI-243 were performed by me, during this project. The subsequent microsatellite genotyping was performed by me for all ten families.

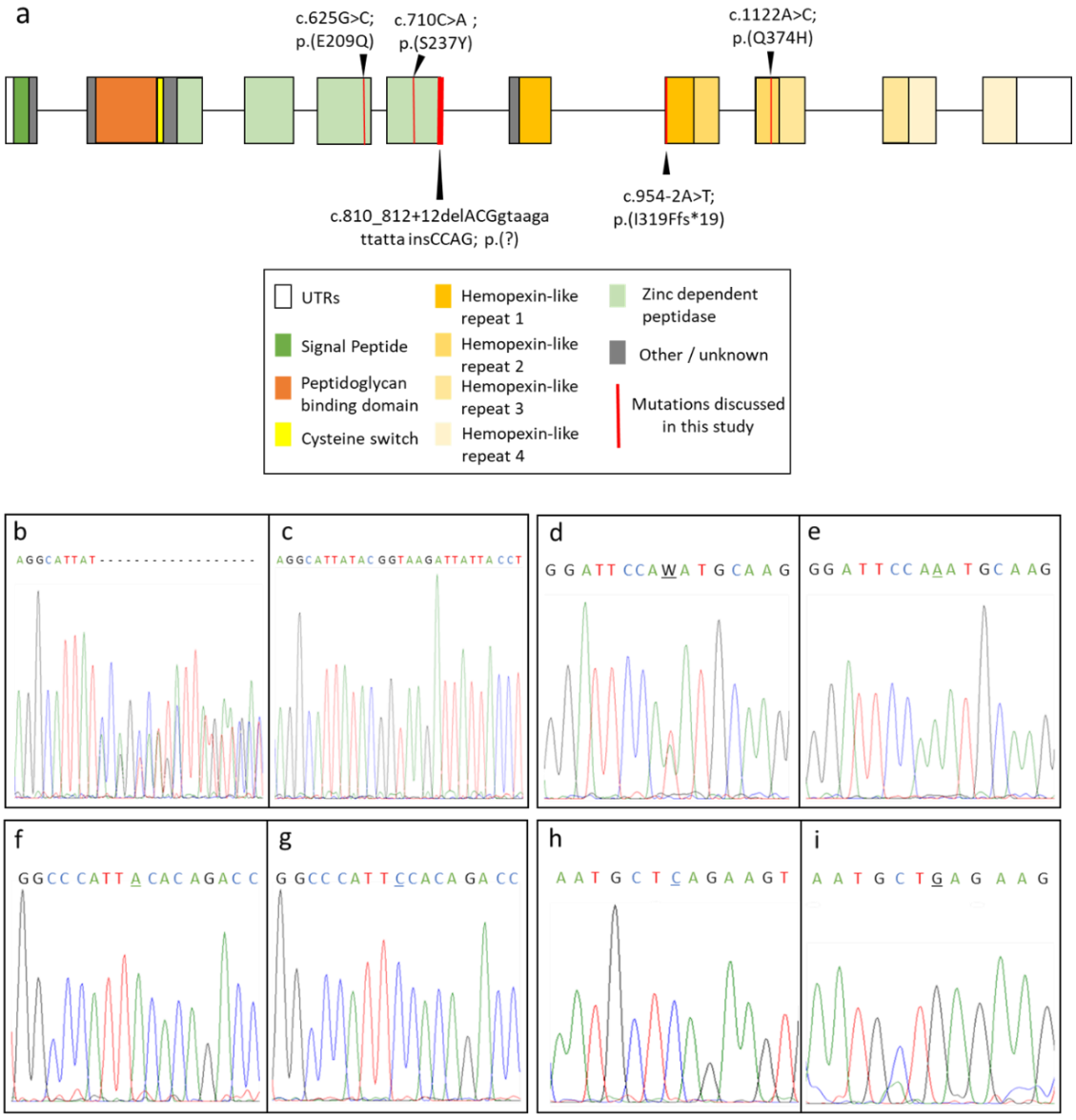


Figure 2.20: Gene diagram of *MMP20* and segregation results of the mutations.

(a) *MMP20* gene diagram, the mutations discussed in this section are marked with a red line. (b – e) segregation of 2102, the proband of AI-13, variant c.809_811+12delinsCCAG (b) and its WT (c), variant c.1122A>C (d) and its WT (e). (f, g) Representative segregation of families AI-39 and AI-52, variant c.710C>A and WT. (h, i) Representative electropherogram from the segregation of AI-77, AI-79, AI-187 and AI-218, showing the c.625G>C variant and WT. The variable position in each electropherogram is underlined.

The families that share variants (Families AI-39 and AI-52 with c.710C>A, Families AI-77, AI-79, AI-187 and AI-218 with c.625G>C and Families AI-155, AI-239 and AI-243 with c.954-2A>T) also originate from the same ethnic backgrounds. To determine whether these families share common founder haplotypes at the *MMP20* locus, see section 2.1.7, five microsatellite markers were genotyped across a 1.5 cM / Mb region of chromosome 11q22 in each family, in the order: 11 cen, D11S940, D11S1339, *MMP20*, D11S4108, D11S4159, D11S4161, 11 qter.

2.3.6.1 Family AI-13

The proband in Family AI-13 was found to be a compound heterozygote for a novel missense mutation c.1122A>C, p.(Gln374His) in exon 8 and a novel deletion-insertion (delins) variant: c.809_811+12delinsCCAG, p(?), spanning the splice donor site of intron 5, see Figure 2.20a. Both are absent from the variant databases. Variant p.(Gln374His) is predicted to be damaging, and Q374 is conserved for the majority of the mammalian clade. Mutation prediction software classifies this variant as pathogenic, with a CADD score of 18.8. It is absent from the gnomAD database, and paired with the second *MMP20* variant found, it has given rise to a hypomaturation phenotype consistent with biallelic *MMP20* disease, also see Figure 2.21. The delins variant is predicted by Human Splicing Finder to disrupt the intron 5 splice donor site, as shown in Figure 2.22, possibly leading to retention of the fifth intron, although this retention was not proven experimentally, due to the difficulty in obtaining mRNA during amelogenesis as discussed previously.

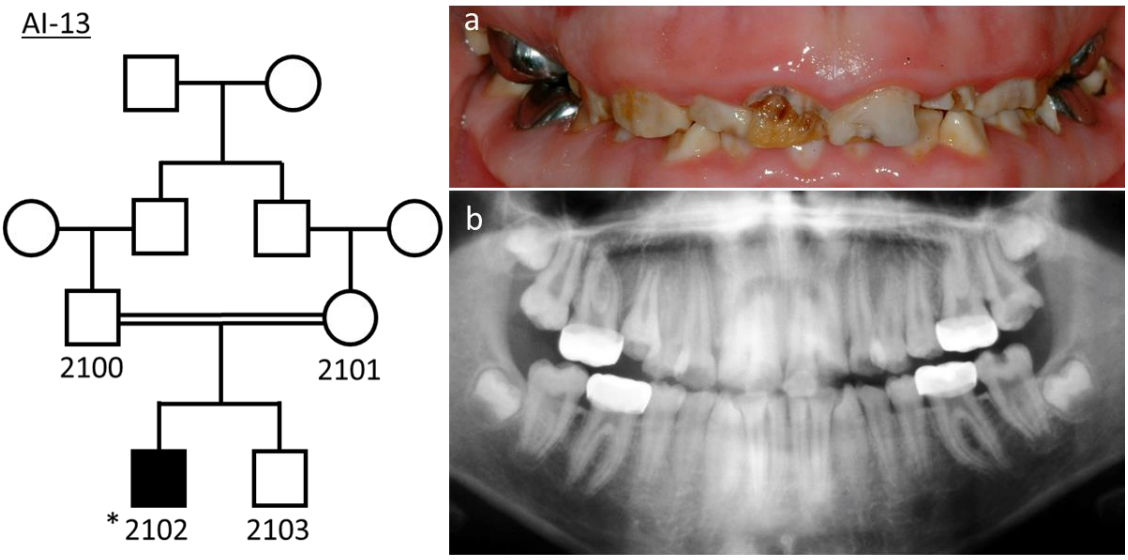


Figure 2.21: Pedigree and dental photos of AI-13.
The proband is indicated with an asterisk (*) and is the only affected member of the family. (a) Dental photos of the proband, showing Hypomaturation AI with post-eruptive changes. (b) Panoramic radiograph of the permanent dentition of the proband.

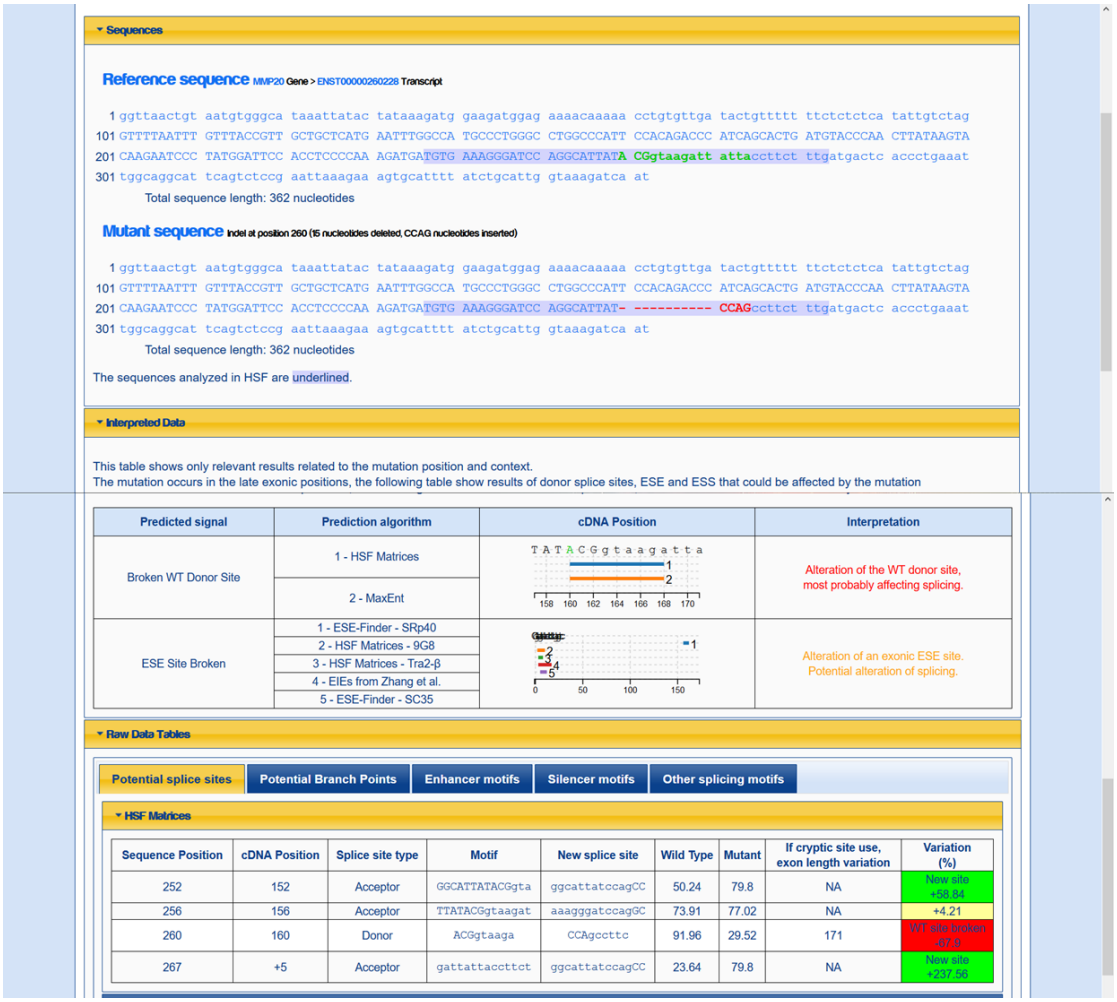


Figure 2.22: Human Splice Finder results for the c.809_811+12delinsCCAG, p(?) variant of *MMP20*. The change in the sequence is shown at the top and the results of the analysis at the bottom, showing the predicted change for the donor site.

2.3.6.2 Families AI-39 and AI-52

In Families AI-39 and AI-52, a novel homozygous missense mutation, c.710C>A, p.(Ser237Tyr) was identified in exon 5. This variant is absent from both gnomAD and dbSNP and is predicted to be damaging by CADD, with a score of 28.4. Both families are of Omani origin and are reportedly unrelated. However, haplotyping with the microsatellite markers mentioned previously suggests that they are closely related; as suggested by the presence of the same allele, identified by the same microsatellite markers on the chromosome, the rare maf of the variant identified and the two families sharing their ethnic origin (Figure 2.23).

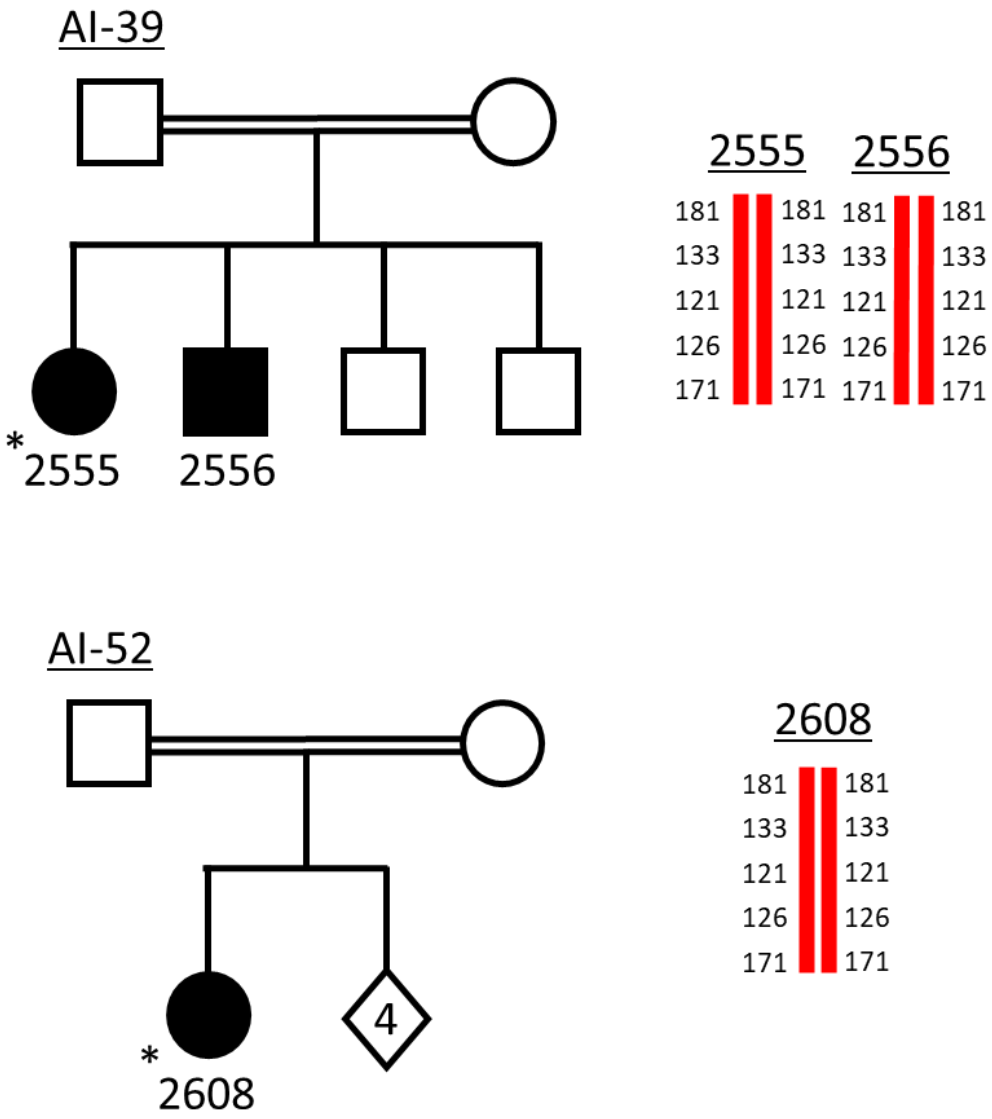


Figure 2.23: Pedigrees and genotypes of families AI-39 and AI-52.

The affected people are shown as filled in circles or squares for female or male family members respectively. The people recruited in this study have been assigned 4 digit codes and the proband is indicated with an asterisk (*). The haplotypes from the microsatellite genotyping of the recruited family members are shown next to the corresponding pedigree. The numbers show the number of repeats found for each microsatellite marker in the order of D11S940, D11S1339, D11S4108, D11S4159, D11S4161 from top to bottom.

2.3.6.3 Families AI-77, AI-79, AI-187 and AI-218

A novel homozygous missense mutation, c.625G>C, p.(Glu209Gln), was identified in exon 4 as the cause of the AI phenotype in Families AI-187, AI-77, AI-79 and AI-218, Figure 2.24. This variant is present in gnomAD as: rs199788797, with maf of 0.000457 in the South Asian population, but has not been previously associated with a disease phenotype. Additionally, the variant is absent from all other populations included in gnomAD. The variant is predicted to be damaging with a CADD score of 27.2. Families AI-187, AI-77, AI-79 and AI-218 are of UK Pakistani origin, which suggests the possibility of a founder mutation. Microsatellite analysis with the markers mentioned previously showed that the proband in Family AI-79 is homozygous for the same haplotype segregating in Family AI-218. However, a second haplotype with a distal recombination can be found in the second affected sibling in Family AI-79, suggesting the affected (unsampled) father carries both haplotypes. Family AI-77 is homozygous for a third haplotype, again identical at the proximal end to that in Family AI-218 but recombinant at the distal end. Family AI-187 in contrast is homozygous for a fourth haplotype identical to the Family AI-218 haplotype at the distal end but proximally recombinant. Again, all four families are homozygous for the marker immediately adjacent to *MMP20*, D11S4108. These observations indicate that it is likely the families originated from the same founding population in the past but had enough time to allow for genetic recombination to alter parts of the chromosome which explains the dissimilarities in some of the markers that are not adjacent to *MMP20* and so can recombine independently from it.

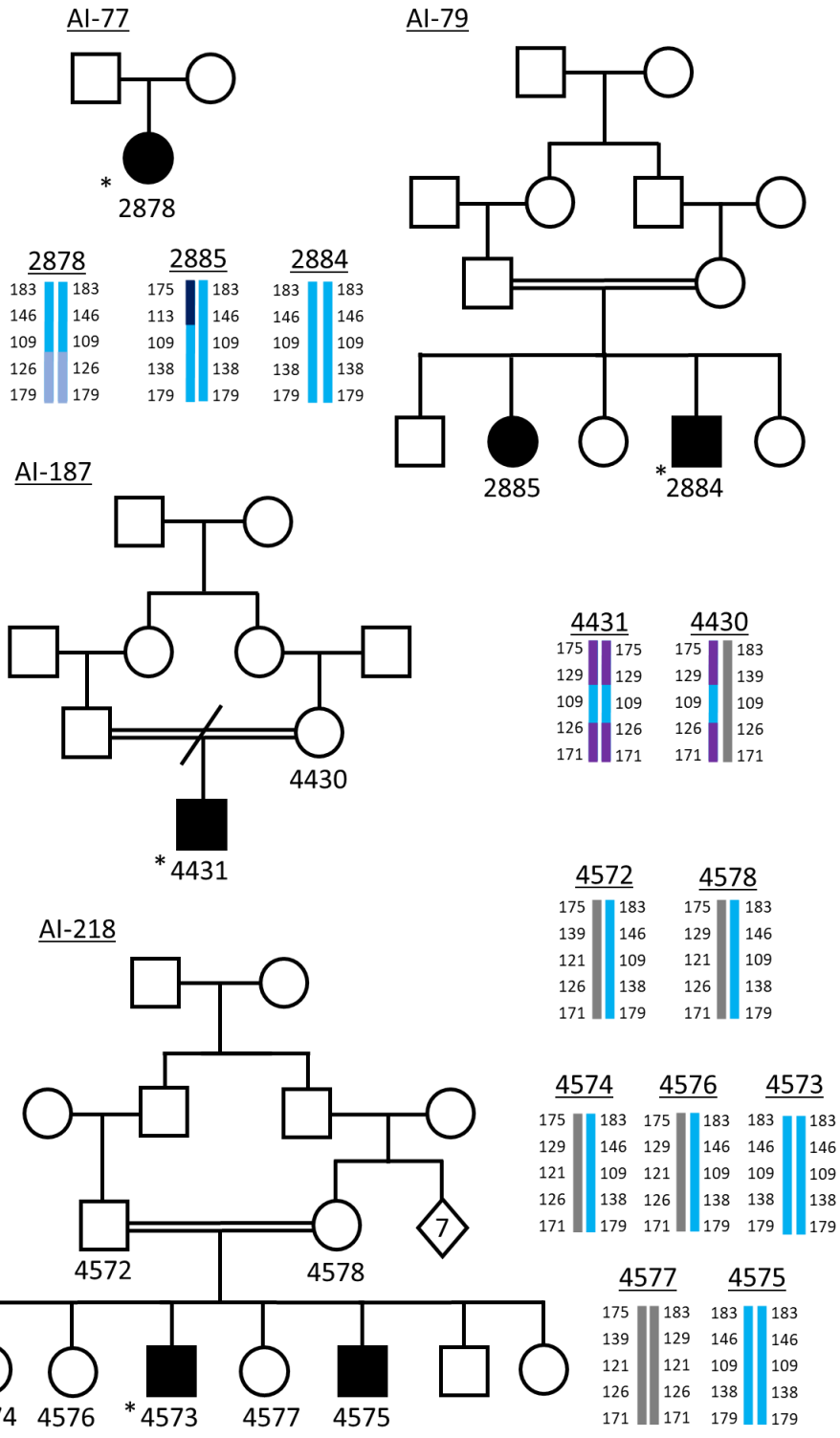


Figure 2.24: Pedigrees and genotypes for families AI-77, AI-79, AI-187 and AI-218.
The affected people are shown as filled in circles or squares for female or male family members respectively. The people recruited in this study have been assigned 4 digit codes and the proband is indicated with an asterisk (*). The haplotypes from the microsatellite genotyping of the recruited family members are shown next to the corresponding pedigree. The numbers show the number of repeats found for each microsatellite marker in the order of D11S940, D11S1339, D11S4108, D11S4159, D11S4161 from top to bottom.

2.3.6.4 Family AI-155, AI-239 and AI-243

In Families AI-155, AI-239 and AI-243, Figure 2.24, a homozygous frameshift variant (NM_004771: c.954-2A>T, NP_004762: p.(Ile319Phefs*19)) was identified in intron 6. Intronic variants have been suspected in previous cases but the proximity of this variant to the exon-intron junction means that there is a very high probability that it is located at either the splice donor sites, leading to a disruption of splicing. This variant has been published previously as the cause of autosomal recessive hypomaturation AI and is expected to lead to retention of the sixth intron (Kim et al., 2005). However, the effect on the MMP20 protein was not confirmed, as its function ends before tooth eruption and is no longer expressed at the erupted tooth or any other tissue of the body, so it would be impossible to isolate the mRNA for confirmation. The construction of a mouse model to examine specific variants was beyond the scope of this project.

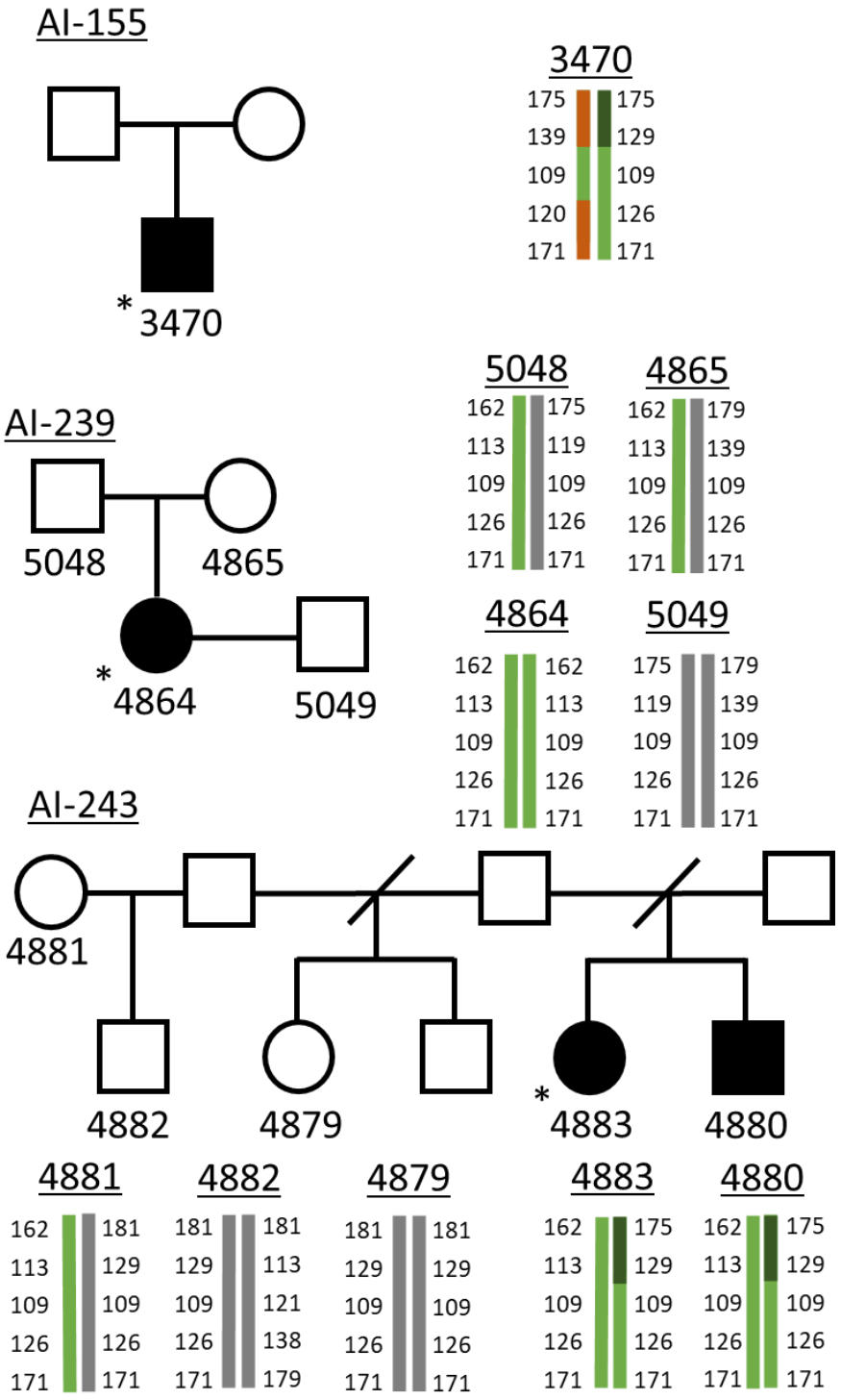


Figure 2.25: Pedigrees and genotypes of families AI-155, AI-239 and AI-243.

The affected people are shown as filled in circles or squares for female or male family members respectively. The people recruited in this study have been assigned 4 digit codes and the proband is indicated with an asterisk (*). The haplotypes from the microsatellite genotyping of the recruited family members are shown next to the corresponding pedigree. The numbers show the number of repeats found for each microsatellite marker in the order of D11S940, D11S1339, D11S4108, D11S4159, D11S4161 from top to bottom.

Families AI-155, AI-239 and AI-243 carry the same c.954-2A>T variant, but do not share the same ethnic origin, the first being a Costa Rican family and the other two being white British families. The proband in Family AI-239 is homozygous for a haplotype also seen in Family AI-243, but a second haplotype in Family AI-243 has a proximal recombination, indicating that there is a more distant relation to the previous allele, while still showing that they are possibly related, or originating from the same population. Family AI-155 could share the recombinant haplotype observed in Family AI-243 together with an unrelated haplotype, but without phase information this cannot be confirmed. Interestingly, all three families are homozygous for D11S4108, which is 100 kb away from *MMP20,* which means that as a locus it can be subject to recombination events that do not include the gene sequence, which could alter it enough to not show correlation with loci closer to the gene that are less likely to recombine independently from it. Lacking information about the variability of that locus in various populations it is not possible to distinguish if that observation is meaningful for the origins of the three families, or a random occurrence.

2.3.6.5 Protein Structure Analysis

As the tertiary structure of the catalytic centre of MMP20 is known, protein modelling can was used to simulate the effect that mutations of the protein sequence can affect the functionality of the protein, in collaboration with Dr Sarah Harris. The catalytic centre of MMP20 is a 160-residue domain containing the zinc dependent peptidase active site. It contains one catalytic and one structural zinc ion, as well as two calcium ions, both structural, as observed in the NMR structure of the MMP20 active site, PDB: 2JSD (Arendt et al., 2007). MD simulations can be performed to assess how missense variants affect protein function of the mutated proteins. In addition, to demonstrate the key role of metal ions in maintaining MMP20 structure, MD simulations were performed in which the metal ions were removed. The protein structures and thermodynamic changes that were observed for the WT with metal ions present, with those obtained for the AI associated mutations: p.(Glu209Gln) and p.(Ser237Tyr), which were described in this study (presented in Figure 2.26a) and p.(Thr130Ile) and p.(Leu189Pro) which have been previously associated with AI (Gasse et al., 2013; Gasse et al., 2017) (Figure 2.26b); and also with the WT in the absence of metal ions were compared. Figure 2.26c shows the root mean squared deviation (RMSD) of each of the simulations from their starting structures. The calculated RMSD values are continuously adjusted and recalculated during the 900 ns of the simulation. As such the final 300 ns, during which the plots have stabilised, are selected as the representative RMSD values for each simulation and so the average values presented in Supp. Table S2.1 are used for the comparison among the variants. An increase in RMSD relative to the WT, implying decreased protein stability, was observed for all variants, apart from p.(Thr130Ile). The changes in three key inter-atomic interactions between the WT and the variants were also analysed, specifically the atomic fluctuations, hydrogen bonding interactions and salt bridges, to provide insight into why these particular variants cause functionally deleterious changes in protein structure. The most significant structural distortions were observed in the simulations performed in the absence of structural zinc and calcium ions.

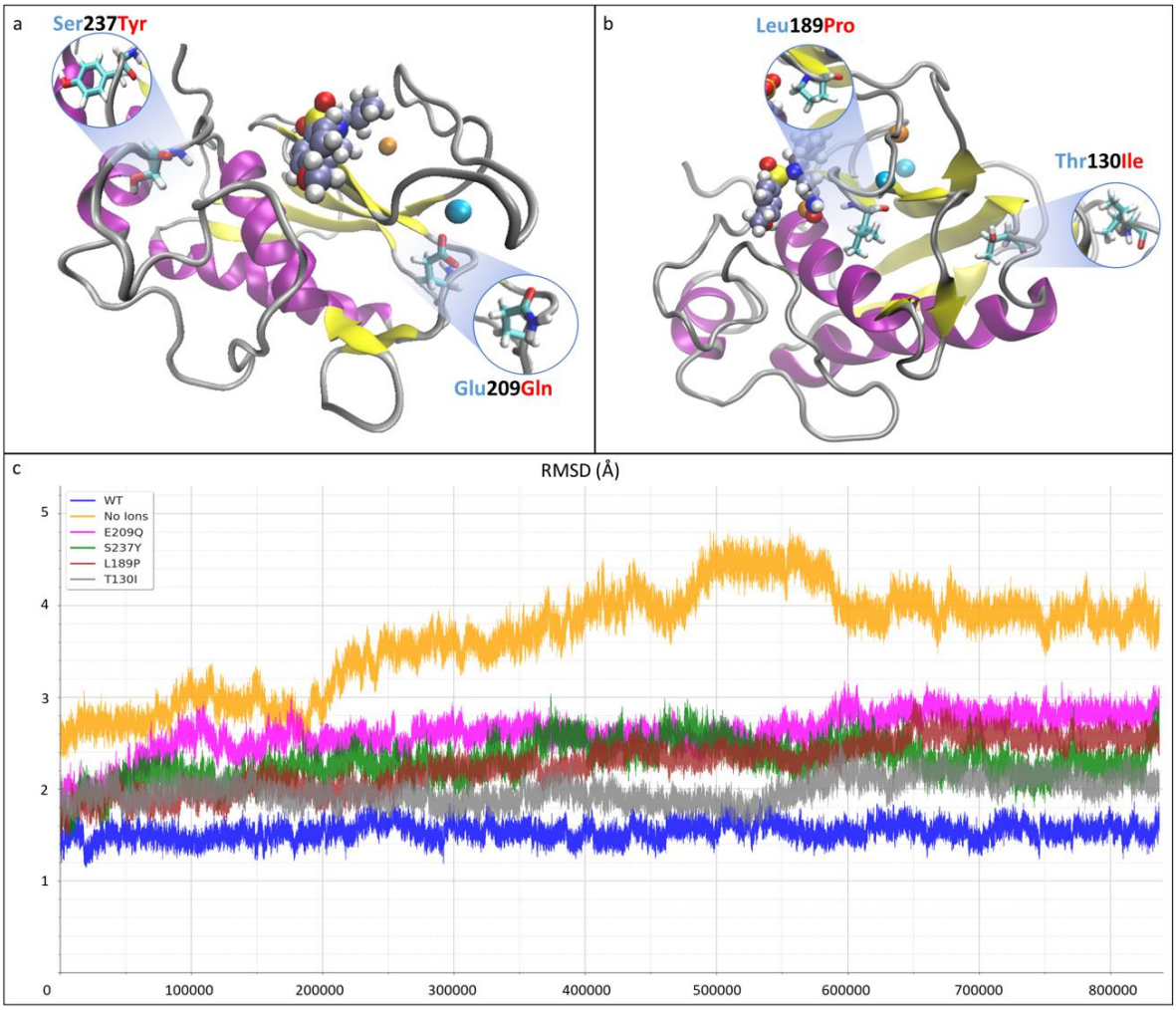


Figure 2.26: The tertiary structure of the catalytic domain of MMP20, based on the PDB:2JSD NMR model.

(a) The WT protein structure of the active site, with the 2 novel AI-causing variants, c.625G>C; p.(E209Q) found in Families AI-77, AI-79, AI-187 and AI-218 and c.710C>A; p.(S237Y) found in Families AI-39 and AI-52, presented in the inlays. (b) The WT protein structure of the active site with inlays showing 2 previously published pathogenic variants in the active site of MMP20, c.389C>T; p.(T130I) and c.566T>C; p.(L189P). (c) The root mean square deviation (RMSD) of each modelled variant during molecular dynamics (MD) simulations of 900ns. An increase of RMSD value corresponds to a loss of stability, with the WT MMP20 structure being the most stable and the MMP20 structure in the absence of structural zinc and calcium ions being the least stable. Figure from Nikolopoulos et al., 2021,

In-silico saturation mutagenesis of the MMP20 active site performed by Rhapsody shows that there are regions of the protein where mutations are significantly more likely to be pathogenic when the residues located there are altered, presented in Figure 2.27. These regions largely correlate with the sites of the known and novel variants and have an increased polyphen-2 score. The Rhapsody analysis was limited to the catalytic domain of MMP20, because it relies on the availability of a tertiary structure and this was unavailable for the other regions of MMP20. The results of the SDM analysis for the four missense variants known to be in the active site of the protein, are presented in Table 2.5, showing the changes of free energy (ΔΔG), residue occlusion (OSP%), residue solvent accessibility (RSA%) and residue depth, in Å, for the WT and each mutant respectively.

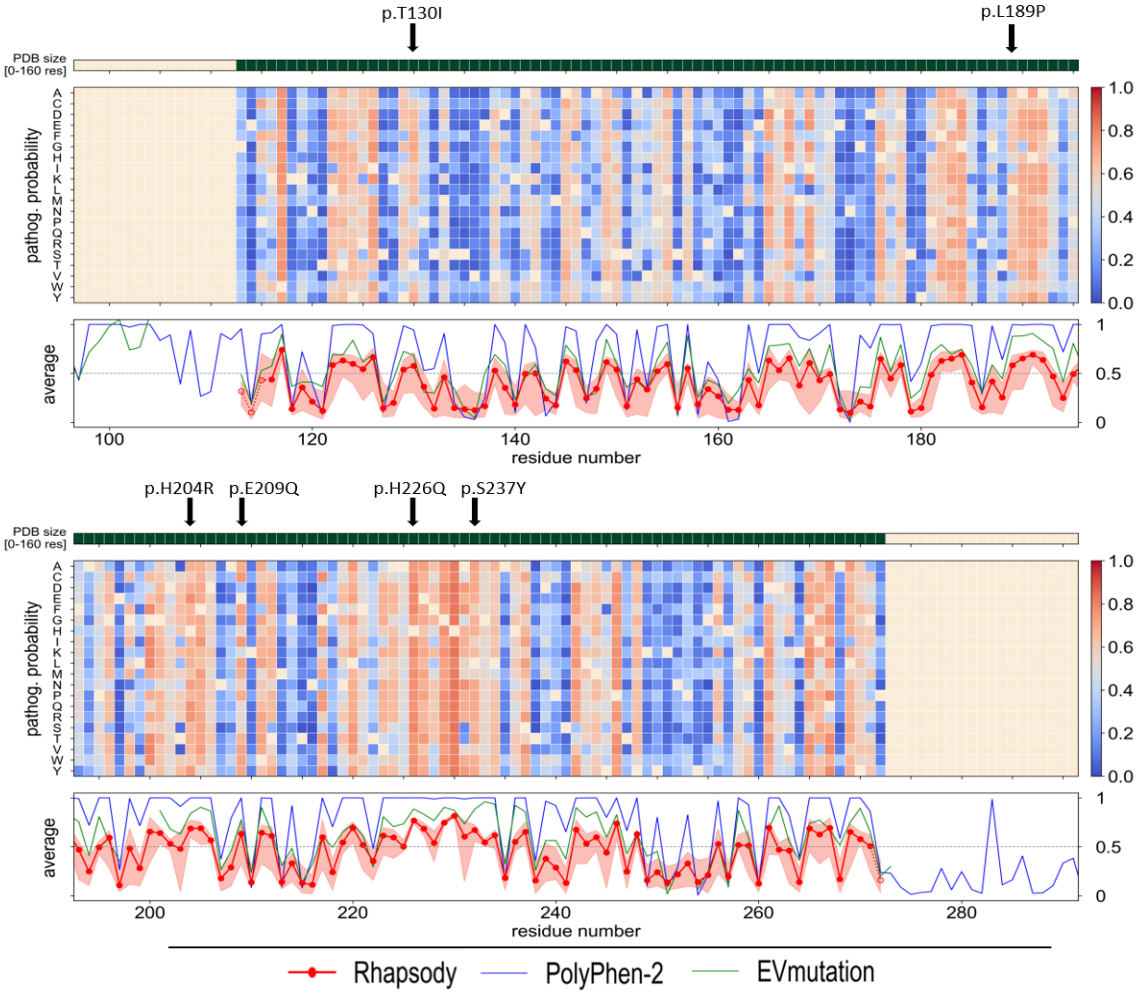


Figure 2.27: Rhapsody score for each possible amino acid change of the active site of MMP20.
Previously published and novel missense variants of the active site of MMP20 are indicated with a black arrow. The pathogenicity probability of each change is shown as a heatmap for each position of the active site, with the pathogenicity prediction score by PolyPhen-2 shown on the graph below, with 1 = high probability for pathogenicity and 0 = low to no probability for pathogenicity.

Table 2.6: Results of the SDM analysis.

SDM predictions of the changes in residue solvent accessibility (RSA%), residue occluded surface packing (OSP%), residue depth in the structure in Å and change in free energy (ΔΔG).

	E209Q	S237Y	T130I	L189P
WT_RSA(%)	5.9	41.2	0.1	7.9
MT_RSA(%)	3.3	31.6	0.8	22.8
MT-WT_RSA(%)	-2.6	-9.6	0.7	14.9
WT_DEPTH (Å)	5.8	3.9	6.4	4.9
MT_DEPTH (Å)	6	4.3	6	4.6
MT-WT_Depth (Å)	0.2	0.4	-0.4	-0.3
WT_OSP(%)	0.45	0.28	0.47	0.48
MT_OSP(%)	0.49	0.28	0.53	0.35
MT-WT_OSP(%)	0.04	0	0.06	-0.13
Predicted ΔΔG	-0.15	0.36	0.7	-1.29

2.4 Discussion

2.4.1 Key findings of the WES analysis

During the course of this project 33 families of the Leeds AI cohort were sequenced by WES and analysed, with potentially pathogenic variants identified in 30 of them, listed in Table 2.4 and 2.5, which were then validated with Sanger sequencing and segregated with the AI phenotype among the family members. Of the 30 families the candidate variant in 25 of them passed the segregation analysis, while two were rejected and three remain as tentative, pending re-evaluation of the clinical phenotype of some of the family members. This included a surprisingly high number of variants found in known AI associated genes, 31 variants of which 19 novel and unpublished, proving that the methodology that was followed has impressive performance in families that are well characterised and with a clearly defined family history.

2.4.2 *RELT* pathogenic variants

Two novel variants were identified in *RELT* for five families, with Sanger sequencing and segregation analysis validating the variants and confirming that they segregated with the phenotype for four of the families. Subsequent microsatellite analyses revealed that three families sharing the same *RELT* variant, AI-37, AI-291 and AI-317, also shared the same haplotype, suggesting that the three families originated from the same founding population, instead of the variant being located at a mutational hotspot. Teeth donated from the affected members of the families showed the characteristics of hypomineralised enamel, with severe attrition, while the analysis of the enamel microstructure showed that the enamel was disorganised and instead of having the highly structured architecture observed in WT teeth. A lamellar structure was observed instead that resembles the structure of enamel that has been previously reported in AI affected teeth. In essence, this study improved our understanding of the genetics of *RELT* with regards to AI, as *RELT* is one of the most recently identified genes to be associated with AI (Kim et al., 2019) and thus one of the least studied. The results presented above contribute to the mutation spectrum of *RELT* and describe the effect that mutations in the gene have on the formation of enamel and the resulting phenotype. Also, in contrast to Kim et al (2019), this study found no evidence that *RELT* mutations have a broader syndromic effect in addition to the AI phenotype, a finding that highlights the need of further research on *RELT* variants.

2.4.3 *MMP20* pathogenic variants

This study enriched the mutation spectrum of *MMP20* and provided new information on the phenotype caused by variants on *MMP20*. Ten families that were found to carry homozygous or compound heterozygous variants in *MMP20* were examined here. Four of the five variants found had not been previously reported as causative for AI. The families that shared the same variant were also examined for the presence of a founder effect on them, with the results indicating the for two of the three groups there was a common population from which they originated, for AI-39 and AI-52 that common founder population was recent,

as the microsatellite markers examined showed an identical haplotype shared among them, while for the group of AI-77, AI-79, AI-187 and AI-218 the founding population was further back in the family history, as for some of the microsatellite positions genetic recombination had time to change the marker. Despite these differences the presence of a founder effect is clear for these variants in contrast to variant c.954-2A>T found in families AI-155, AI-239 and AI-243, which was shown to be a frequently found variant in unrelated populations, possibly on a mutational hotspot.

Additionally, the effect of some of the missense mutations on the active site of the protein was examined by MD simulations of the protein structure. Four of the mutations on the active site were examined for their effect on the stability of the protein structure, the interactions among the residues and the change in solvent accessibility caused by the mutations. The effect of the mutations was compared to the WT protein and a protein that does not have any metal ions, which are integral to the structure of MMP20. All variants were found to disrupt the protein structure at a significant level, although not as severely as the no-ions model (Figure 2.25). Studies on the protein structure of a gene product that causes an AI phenotype have not been previously reported in the literature and can become an indispensable tool in assessing the pathogenic effect of gene variants.

2.4.4 Future Work

Despite the significant findings in most of the families reported here, a limitation of WES is also shown, in that it can only detect nucleotide polymorphisms in the exonic sequences that are captured, without examining the majority of the regulatory elements or any intronic sequences, limiting the variants that can be found. Additionally, via WES it is more likely to identify variants in genes that are already associated with AI than to identify as causative a variant in a new gene, due to the need to observe variants in a novel gene in multiple families before it can be definitively linked with an AI phenotype. In part because of these limitations, for some samples there was no one gene variant that could be identified as the best candidate to be causative for the phenotype. When a candidate variant was shown to not segregate with the phenotype in all family members it was rejected as causative for the phenotype and the family was designated as unsolved. Importantly, most of the families investigated here can be considered solved, after findings in genes already known to be associated with AI, creating the need for a rapid testing approach that would target these known genes, so that research efforts can instead be focused on the families where the cause of AI is not clear. Despite not identifying a new gene that can be associated with an AI phenotype this project helped to reduce the number of unsolved families to a set of well-defined exomes.

Going forward, all families that are considered unsolved will be grouped together based on the observed AI phenotype and the WES data obtained from them will be examined for genes with an elevated mutational burden. Additionally, grouping the families will make it easier to search for genes common in multiple families, which would indicate that they might be associated with the AI phenotype.

Chapter 3 – Evolutionary Analyses identify signatures of positive selection on putative pseudogenes involved in tooth / enamel formation in toothless / enamel-less mammals

3.1 Introduction

3.1.1 Loss of teeth in mammals

The enamel organ has been considered as one of the most conserved organs in mammals and the genes that are involved in its formation are well characterized. Genes that when disrupted are associated with non-syndromic AI in humans were selected as the gene set of interest. Given the conserved nature of this organ it is proposed that the genes underpinning enamel formation and associated with enamel disorder will be under similar evolutionary constraints across mammal species that have retained functional enamel. This proposal is tested in this chapter, by sampling independent sets of closely related species across the mammal phylogeny that have retained or lost either teeth or enamel, and whether they exhibit similar or different selective pressures in each of these lineages is examined. According to the "pseudogenisation model" for tooth or enamel loss in mammals, it is expected that orthologs in lineages without teeth or enamel would be evolving at a more rapid rate and without the selective constraints that would act to conserve the functional genes, with genetic drift allowing them to accumulate mutations. This could potentially lead to establishing new mutations, with the possibility of pseudogenising the gene, or in rare cases they might lead to a shift in function for the genes.

The basic stages in the development of teeth and the generalised morphological characteristics of tooth formation have remained consistent since the origin of teeth in jawed vertebrates, also called gnathostomes, and are considered heavily conserved at the molecular level. Whilst the loss of specific tooth types is not uncommon, the vast majority of mammals have retained a form of dentition. Examples of loss of specific tooth types include: the absence of canines in herbivores such as the tapinocephalid lizards, where the dental occlusion of neighbouring teeth leads to a mammalian-like dentition (Whitney and Sidor, 2019); the substitution of some tooth types with others such as observed in the Xenarthran clade (Vizcaíno, 2009); and the loss of tooth complexity, such as the simplified teeth observed in sloths which is thought to have resulted from adaptation to food niche (Hautier et al., 2016). Indeed, the independent loss of teeth or loss of the enamel organ can be observed at least 5 times during mammal evolution, in the Marsupialia branch (e.g.:platypus), the xenarthra branch (e.g.: sloths, anteaters and armadillos), the afrotheria branch (e.g.: aardvark), the cetacean branch (e.g.: baleen whales) and the pholidota branch (e.g.: pangolins). For example, anteaters that feed using their tongues or baleen whales that feed on plankton and algae have undergone adaptive parallel evolution. The examples of enamel-less and toothless mammal species are shown in Table 3.1 whilst their phylogenetic relationships to one another are shown in Figure 1.6. Genes that when pseudogenised lead to the inability to form teeth or enamel, that are involved in tooth formation and amelogenesis, correlate with being causative for AI when mutated. Subsequently, the question arose whether these genes can be characterised by a pattern in their evolutionary history or show a unique selective pressure

motif that would allow researchers to identify and relate them to possible candidates to causing AI when mutated even before finding their variants in the genome of people affected by AI. To examine this possibility the genes that are associated with AI and the genes that are suggested to have led to the loss of teeth and enamel in other mammalian species need to be examined for any shared characteristics.

Table 3.1: Mammalian species without enamel or without teeth.

The species presented in this table include all species used here and shown in Figure 1.6 and some additional species that are mentioned but not analysed in this study.

Mammals without enamel	Mammals without teeth
Aardvark	Anteater
Armadillo	Baleen whales
Pygmy sperm whale	Pangolin
Sloth	Platypus

The fact that these are independent occurrences of loss of teeth or enamel indicates that there was no uniform pressure acting on the genes that were pseudogenised, but different kinds of selective pressure would act differently on individual cases of tooth or enamel loss. These separate occurrences will be examined to find any similarities among the toothless and enamel-less species and also to compare them to the toothed mammalian species. The possibility that due to convergent evolution there is a discernible footprint of selective pressure acting on the genes that have been made functionally redundant will be explored as well as whether there is a trend towards a shift in function for genes that would code for enamel in toothless and enamel-less species.

3.1.2 The pseudogenisation model

The mechanism that has been shown to lead to the convergent loss of teeth and enamel in species is the inactivation of the relevant genes by pseudogenisation. As described in section 1.4.7, pseudogenisation is the evolutionary change of an active gene to inactive by the accumulation of disruptions in its coding sequence. These disruptions can be in the form of the introduction of premature termination codons (PTCs) in the coding sequence, as mutations that cause the inactivation of the promoter region, of the start codon or the exon-intron junction and mutations that cause a frameshift in the coding sequence, such as insertions and deletions (Figure 1.6). Regardless of the underlying mechanism, disruption of the coding sequence leads to a truncated and otherwise non-functional protein product often leading to functional implications downstream of the peptide.

Returning to the central premise of independent loss of enamel and teeth, there are a set of pseudogenised genes (e.g.: *ACP4, AMBN, AMTN, ENAM, MMP20* and *ODAPH*) proposed as underpinning the degeneration and potential loss of tooth or enamel in aardvarks, armadillos, sloths and the baleen whales (Deméré et al., 2008; Meredith et al., 2011; Gasse et al., 2012; Delsuc et al., 2015; Springer et al., 2016; Huang et al., 2017; Sharma et al., 2018; Mu, Huang, et al., 2021). In more detail, the protein expression of *ACP4* has been described to be disrupted in toothless and enamel-less species, (Mu, Huang, et al., 2021), specifically baleen whales, the pygmy sperm whale, aardvark and armadillo, all of which show signs of pseudogenisation of ACP4. Mu et al report that two single nucleotide deletions in exons 4 and 5 are shared among all living baleen whales and are responsible for the inactivation of ACP4 by altering the reading frame, while other mutations introduce premature termination codons, splice site effects and start codon alteration, uniquely in other whale species (Mu, Huang, et al., 2021). Sharma et al., (2018) also report the pseudogenization of *ACP4* in minke whale, aardvark and armadillo by incorporation of nonsense mutations in the gene body. Similarly, *AMTN* is believed to be pseudogenised in sloths and armadillos (Gasse et al., 2012). In their study Gasse et al (2012) used the SLAC program to estimate the selective pressure acting on the codons of AMTN, with the results indicating 51 codons under purifying selection, 2 codons under possible positive selection and 156 variable sites under neutral evolution. *In silico* searches and selective pressure analysis reveal that *AMBN* has been inactivated in armadillos, sloths and aardvarks by accumulation of nonsense mutations (Delsuc et al., 2015) and in baleen whales, of the clade Mysticete, by either single base pair indels or nonsense mutations (Deméré et al., 2008). *ENAM* has been reported to be pseudogenised in baleen whales (Deméré et al., 2008). Among

the 13 Mysticete species examined various single base indels were responsible for inactivating the gene, by introducing a deleterious frameshift in the sequence. Demere et al also report that, according to the parsimony reconstruction of the ancestral sequences, they estimate that the inactivation of both *AMBN* and *ENAM* occurred after the loss of mineralised dentition in the Mysticete evolution, they did not, however, examine the selective pressure that is possibly acting on these genes even though they say that a relaxation of selective constraints should be expected. *ENAM* has additionally been found to be pseudogenised in an extinct toothless member of the sirenia clade, Steller's sea cow (*Hydrodamalis gigas*) (Springer et al., 2015), by a transversion mutation (A -> C) in the acceptor splice site of intron 2. Springer et al sampled bone fragments from six specimens of individuals of the species that had been preserved in a museum in two rounds of experiments. Initially, they performed DNA extraction and sequencing to obtain a set of provisional sequences, which were then used to design more accurate probes for the second round of sequencing. Branch specific codeml analyses conducted by Springer et al, showed that sites (codons) of *ENAM* evolved under positive selection for the sirenia but were not statistically significant for *H. gigas.* Additionally, the sites reported to be under positive selection are unique for the sirenia, which lead Springer et al to deduce that these residues may have mediated the evolution of the unique dietary adaptations of the Stellar sea-cow and the sirenia.

Transposable elements (TEs) are mobile genetic elements that can be integrated in parts of the genome and be the cause of pseudogenization of a gene. TEs have different characteristics that categorize them in different classes (Saleh et al., 2019), also see Figure 3.1 for a summary of the TE types and are typically inserted in introns of genes or intergenic regions, as the insertion in exons is usually deleterious and can often affect the expression of the host gene, as new stop codons can be formed, splicing can be affected by creation or loss of exons and the methylation of the DNA region can be affected (Chenais, 2015).

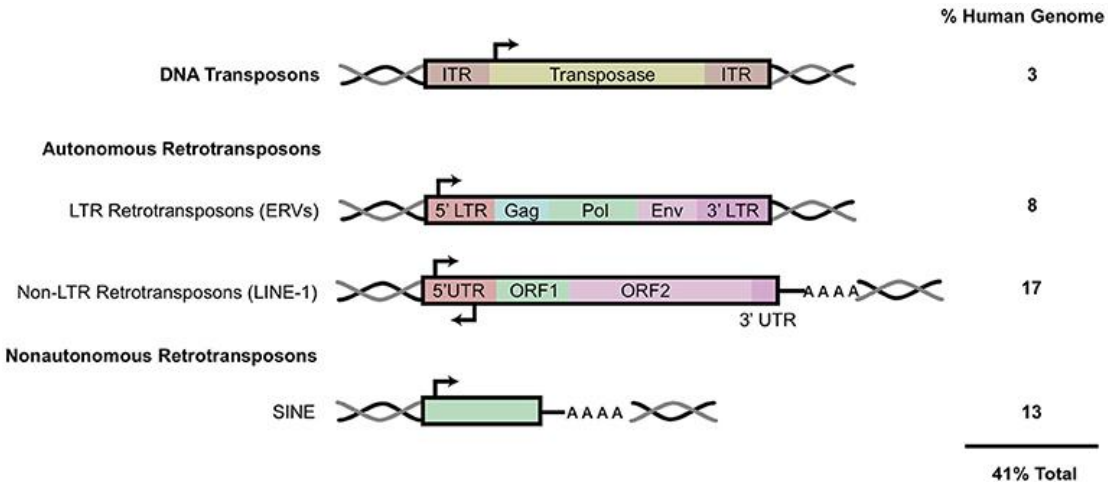


Figure 3.1: Types of transposable elements and their frequency in the human genome.
Image from Saleh et al (2019), CC BY-SA 4.0.

An example of this deleterious effect is observed by Meredith et al., (2011) with the insertion of a CHR-2 SINE retroposon in exon 2 of the *MMP20* gene of the Mysticete species. This SINE insertion introduces stop codons in all possible reading frames and is considered the causal factor for the pseudogenization of the MMP20 gene in these species (Meredith et al., 2011). The lack of other stop codons or frameshifts in *MMP20* reinforces the assumption for the effect that the identified transposon has on the functionality of the protein. The dN/dS analysis conducted with the codeml program showed that all toothless whale species with pseudogenised *MMP20* had elevated ω values (ω = 1.84) in the *MMP20* protein coding region, indicating positive selection and possibly a functional shift for the protein, which is contrary to the branches with intact *MMP20* that showed signs of purifying selection (ω = 0.19) (Meredith et al., 2011).

In *AMBN* the gene sequence is shown to generally be under purifying selection in most mammals, with some codons (5 out of 447 codons) shown to be under positive selection (Delsuc et al., 2015) in the catarrhine primates, while the gene sequence has been pseudogenised in armadillos, sloths and aardvarks, while also having an elevated ω value in sloths and aardvarks. However, they detected higher than expected ω values in some other toothed species as well, i.e.: elephant (*Loxodonta africana*), the elephant shrew (*Rhynchocyon cirnei*) and the aye-aye (*Daubentonia madagascariensis*), with the high ω values correlating with the presence of additional exons in AMBN compared to the ancestral sequence, which Delsuc et al interpret as the exon duplications being adaptive and encouraged by positive selection.

ODAPH has been found to be inactivated in pangolins, as well as in minke and bowhead whales (Springer et al., 2016). The selective pressure analysis was performed by estimation of the dN/dS ratio with the codeml program and was inconclusive in detecting positive selection. Springer et al., (2016) attribute the high ω ratio they estimated to either positive selection or possible mistakes in the alignment of the sequences or the selection of the models used for the codeml analysis. Springer et al., (2016) examined the sequences and selective pressure acting on four genes: *WDR72, SLC24A4, FAM83H* and *ODAPH* and found no evidence of statistically significant positive selection in the sequences of toothed and toothless mammals. Rather they identify that ODAPH is pseudogenised only in the toothless species, whereas *WDR72, SLC24A4, FAM83H* are not pseudogenised suggesting that they have broader function not restricted to the teeth, but that *ODAPH* is tooth specific. The analysis of the sequence and function of *FAM83H* showed that there are multiple sites on the C-terminus of the protein that are heavily conserved across the mammalian clade, but also some sites that are under significant positive selection (Huang et al., 2017), which are suggested by Huang et al to have played an adaptive role during evolution. The dN/dS analysis was performed using a combination of both codeml and SLAC, identified the sites under positive selection, however, as these methods use a different approach on creating the phylogeny that used for the analysis (see section 1.4.3) it was unsurprising that the specific sites identified as under positive selection did not overlap between the two methods, both methods locating them at the C-terminal end of the protein.

As can be seen from the previous studies, genes that have been associated with AI are consistently found to be inactivated or under positive selection in toothless and enamel-less

species. Studies examining the selective pressure variation of each gene among multiple mammalian species find that even genes that are under purifying selection for the majority of the branches are found to have sites under significant positive selection in toothless and enamel-less species, as is the case with studies on *MMP20* (Gasse et al., 2017) and *FAM83H* (Huang et al., 2017) among others.

However, these studies examined the evolutionary history of each gene in isolation, without considering that these genes are involved in forming the enamel organ, which is the characteristic of this study. The methodology used in the literature is similar to the one used for this project, i.e.: using codeml to estimate the selective pressure acting on the genes of interest for toothed and toothless/enamel-less mammalian species, however the published studies focus solely on a gene of interest without examining the other tooth or enamel related genes. This study will attempt to identify any similarities shared by the genes that are associated with an AI phenotype across the toothless and enamel-less mammal species, compared to the toothed mammals. The selected sets of species are composed of three toothless and three enamel-less mammals, along with 13 toothed representatives of each major branch (superorder) of the mammalian clade (class Mammalia), that demonstrate the independent loss of teeth or enamel (see Figure 1.6). The selection of species, toothless, enamel-less or even toothed, was often restricted by the availability of high-quality genomic sequences.

3.1.3 Aims of this chapter

This chapter aims to investigate the investigate the natural selection acting on the genes that have been associated with AI and attempt to identify a signature of selective pressure that could distinguish the genes of this group from the other genes of the genome. Previous studies that show that some of the genes involved in amelogenesis have undergone pseudogenisation in some toothless or enamel-less mammalian species, however, this study will attempt to examine whether the pseudogenisation model is the most suitable to explain the loos of teeth or enamel in those species. This is also the first study that provides a comprehensive look at all the genes associated with AI and across all the major mammalian clades.

The specific aims for this chapter are:

a) Identify homologs of genes associated with AI in humans across a range of carefully selected mammals - those with and without teeth or enamel.
b) Assess selective pressure variation in AI associated genes across mammals without teeth or enamel, to determine if there is a common change in selective pressure variation associated with AI genes in mammals without teeth or enamel.
c) Examine whether the patterns of change present in the multiple sequence alignments of homologous AI associated genes support the pseudogenisation model for the convergent loss of teeth or enamel in mammals.

3.2 Materials and Methods

An overview of the steps of the methodology used for the selective pressure analysis, starting from selecting the mammalian species that were included and leading to the selective pressure results analysis are mapped out in the workflow diagram (Figure 3.2) and each stage is detailed below.

3.2.1 Dataset assembly

The dataset of coding nucleotide sequences for 22 genes across 13 toothed species and 6 toothless/enamel-less species was assembled from Ensembl, version 90 (Aken et al., 2017) and Genbank (Benson et al., 2013) (last accessed December 2019, see Table 3.2 for summary and Appendix C.1 for each ID), by manual text based searching of the species and then selecting the longest among the transcripts that are annotated as functional, if alternative transcripts were available. The genes selected for this analysis were the 20 genes that are associated with a non-syndromic AI phenotype and are essential for amelogenesis, along with *RHO* and *TUBA4A* which do not belong in that group.

As the 20 non-syndromic AI genes can be described as enamel specific genes, *RHO* and *TUBA4A* were included to provide a comparison baseline of the enamel specific genes to an eye specific gene and a housekeeping gene respectively. Other genes involved in amelogenesis, or tooth formation could not be included as controls, as even without being associated with an AI phenotype until now, it is not possible to know that there is no link between them and AI and, also, any potential interactions with the AI genes.

RHO is an eye specific gene and was included to examine whether any patterns found in AI and tooth associated genes are specific for genes associated with enamel or are relevant for a more general group of genes. Mutations in *RHO* have been associated with causing Retinitis Pigmentosa 4 (OMIM # 613731), Retinitis Punctata Albescens (OMIM # 136880) and Congenital Stationary Night Blindness (OMIM # 610445), but *RHO* has not been associated with any tooth phenotype or linked with any tooth specific genes and so it is considered unlikely to be inherently biased towards having a similar evolutionary history as the tooth specific genes that are the focus of this study, or to have been affected by any selective pressure acting on them. *TUBA4A* is a housekeeping gene that is thought to have been greatly conserved, due to its essential function for every cell of all organisms, for example it has been reported as phylogenetically uninformative and under purifying selection in the protozoan class Litostomatea (Rajter and Vďačný, 2018). Similarly to RHO, it has not been linked with any of the tooth specific genes and as such is considered unlikely to have been affected by any selective pressure acting on them.

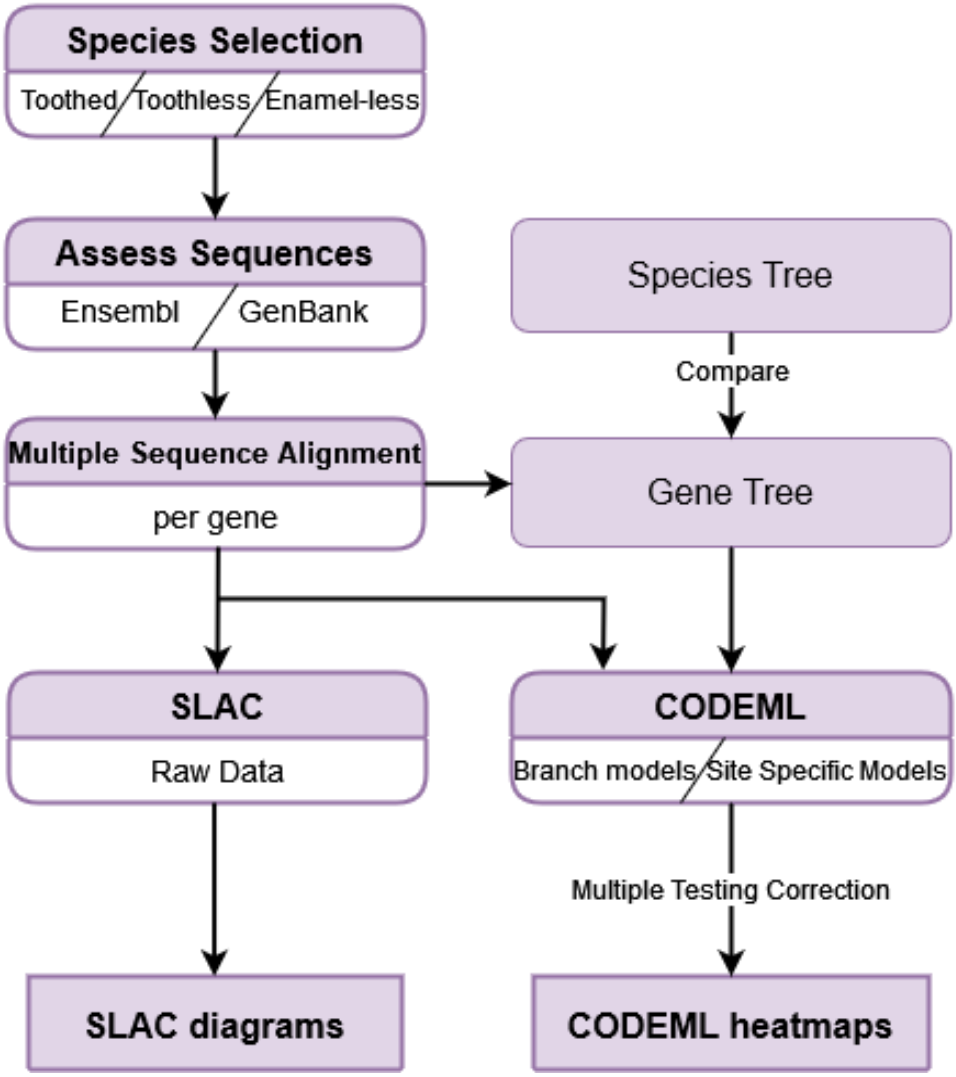


Figure 3.2: Methodology used for selective pressure analysis.

Flowchart of the methodology used for the selective pressure analysis, starting from the selection of the species and composing the dataset to constructing the MSAs and the phylogenetic trees and analysing the results.

Table 3.2: List of toothed and toothless / enamel-less species and the genes included in this study.

Table of the species and the genes used for this study, with the species name in the top row and the genes in the first column. The sequences marked as N/A were not available in either GenBank or Ensembl, those shaded green were available and those shaded blue were available but were found to contain internal stop codons. The names of the toothless species are highlighted in red and the enamel-less in blue.

	Human	Chimpanzee	Bushbaby	Squirrel	Mouse	Guinea Pig	Horse	Minke whale	Dolphin	Cow	Cat	Dog	Hedgehog	Pangolin	Aardvark	Elephant	Sloth	Armadillo	Platypus
ACP4	ENST00000270593													N/A			N/A		
AMBN	ENST00000322937																		
AMELX	ENST00000380712																		
AMTN	ENST00000339336													N/A					N/A
COL17A1	ENST00000353479																N/A		
DLX3	ENST00000434704								N/A										N/A
ENAM	ENST00000396073																N/A		
FAM20A	ENST00000592554																		
FAM83H	ENST00000388913																N/A		
GPR68	ENST00000531499								N/A								N/A		
ITGB6	ENST00000283249																		
KLK4	ENST00000324041													N/A	N/A		N/A	N/A	N/A
LAMA3	ENST00000313654													N/A					
LAMB3	ENST00000391911																		
MMP20	ENST00000260228																		N/A
ODAPH	ENST00000435974							N/A									N/A		N/A
RELT	ENST00000064780																		N/A
RHO	ENST00000296271																N/A		
SLC10A7	ENST00000507030																		
SLC24A4	ENST00000532405																		
SP6	ENST00000536300																N/A		N/A
TUBA4A	ENST00000248437																		
WDR72	ENST00000396328																		

3.2.2 Multiple Sequence Alignment and Quality check

The coding DNA sequences were initially aligned in MEGA 7 (Kumar et al., 2016) using MUSCLE (Edgar, 2004) with the option to align the codons instead of individual nucleotides and with the default settings for codon alignment. The terminal stop codons were removed from each sequence. The alignments quality was also assessed with norMD (Thompson et al., 2001) using the AQUA protocol (Muller et al., 2010). Positions that appear at the same place in multiple MSAs are considered more reliable and can help to indicate the quality of a MSA, norMD finds and scores such positions in different MSAs or in different iterations of the same MSA, with alignments getting a norMD score > 0.6 being regarded as reliable MSAs, with the higher value indicating the better alignment. AQUA examines alignments made by either MUSCLE or MAFFT (Katoh and Toh, 2008), performs norMD scoring and then employs RASCAL, an alignment correction tool (Thompson et al., 2003) to improve any low scoring alignments. The refined MSAs are scored with norMD again and the best scoring MSA is marked as the best to use for the phylogenetic analyses. AQUA was conducted using the default parameters.

3.2.3 Tree construction and visualisation

The published mammal species tree of (Morgan et al., 2013) was pruned to include the species in this study (Figure 1.6). Phylogenetic trees were constructed using the Maximum Likelihood (ML) method from the Multiple Sequence Alignments (MSAs) of Section 3.3.1 as input in MEGA 7, using the default parameters. The best fitting substitution models for the sequences were calculated using the ModelFinder function of IQ-Tree (Kalyaanamoorthy et al., 2017). The GTR+G model was selected as the best fitting of the models available in MEGA 7. The resulting trees were saved in the newick format (.nwk) for use in subsequent steps. The optimal topology of the resulting trees was found by the Approximately Unbiased (AU) test (Shimodaira, 2002) which calculates a likelihood value for each candidate tree and outputs the tree with the largest likelihood value, which is expected to better represent the true evolutionary history of the gene. IQ-Tree was used to conduct the AU test, using the ModelFinder (Kalyaanamoorthy et al., 2017) and Tree reconstruction modules (Nguyen et al., 2015).

The Robinson-Foulds (RF) method (Robinson and Foulds, 1981) was selected to compare the phylogenetic trees that were constructed from the gene sequences to the reference tree. This test calculates the distance between the topologies of a pair of trees by partitioning the tree branches of the first tree and labelling them and then reconstructing the second tree with the same branches, splitting, and merging them if necessary. If a branch remains unaltered after the second tree is reconstructed it keeps the label that was assigned to it. At the end of the reconstruction of the second tree, the end nodes that have kept their labels are counted and together with the nodes that differ between the trees give the RF distance for the two trees. The CLANN software package (Creevey and McInerney, 2009) was used to perform these calculations and find the RF distances.

3.2.4 Assessing selective pressure variation: models and statistical analysis

Codeml is part of the PAML package (Yang, 2007) and is used to estimate the ω ratio of a gene based on a given MSA of the coding DNA sequence (CDS) and a phylogenetic tree. The codeml models used in this study are presented in Table 3.3 and in summary are designed to examine the selective pressure acting on a gene under different parameters, e.g.: allow a uniform selective pressure among sites, model M0, allow the selective pressure to vary among the sites without allowing for positive selection, model M1a, whether there is positive selection on the sites of the MSA, models M2a and M8, so that by comparing the estimated ω ratios for each model, the best fitting model can be found to describe the data more accurately. Additionally, whether the gene of interest has signs of positive selection in specified foreground lineage compared to the other lineages in the dataset, models Clade Model A and C. Each model has its respective null hypothesis, as specified in Table 3.3. An example of a control file with the parameters used for the codeml models is shown in Appendix D.1. Codeml estimates the maximum log likelihood of each model, therefore for those models where omega ratios are estimated from the data it is important to consider various initial omega vales across the likelihood plane to reduce the risk of reporting from local maxima. Different values were used for the initial ω ratio in repeats of the models; these were initial-ω= 0, 1.3, 2 and 10. The varied starting ω values help with reducing this risk as now the starting point for the estimates will be at different locales and with different starting ω local highs and lows can be avoided, approaching more accurately the global maximum value instead of getting stuck in a local maximum.

Table 3.3: Codeml models used.

The models used for codeml with their respective null hypothesis models are shown, along with the parameters specific for each model.

Model	Null Hypothesis Model	Type of Analysis	Model code	NSsites	ω
M0	-	Site specific	0		Estimated <1
M1a	M0	Site specific	0	1	Estimated ≤ 1
M2a	M1a	Site specific	0	2	Estimated
M7	-	Site specific	0	7	Estimated
M8	M7 or M8a	Site specific	0	8	Estimated
M8a	-	Site specific	0	8	Fixed=1
Clade Model A	A_null	Branch specific	2	2	Estimated
A_null	-	Branch specific	2	2	Fixed=1
Clade Model C	M2a_rel	Branch specific	3	2	Estimated
M2a_rel	-	Branch specific	0	22	Estimated

Statistical analysis of the fit of each model to the data is performed by Likelihood Ratio Test (LRT) which allows us to compare the difference in log likelihood between models and their appropriate respective nulls taking the number of free parameters estimated into account as follows :

$$LRT = 2 * \Delta L \text{ (Model - nullModel)} = 2 * (\ln L1 - \ln L0)$$

where lnL0 is the log likelihood of the null model and lnL1 is the log likelihood of the alternative model, with the alternative being the more parameter rich model compared to the null. If the value of 2ΔL is greater than the chi-square value then the more parameter rich model is statistically significant and the null hypothesis can be rejected. The difference of the number of parameters of the two models is equal to the degrees of freedom for the chi-squared distribution. To find the model that is the best fitting for the data the LRT was performed for each pair of models described on Table 3.3.

Due to the many models used, multiple testing correction was conducted to reduce the false positive (Type I) errors, by using the Benjamini–Hochberg approach to identify the False Discovery Rate (FDR), with the formula:

$$(i/m) * Q$$

where 'i' is the individual p-value's rank, 'm' is the total number of tests and 'Q' is the expected FDR (Benjamini and Hochberg, 1995) and to have a 95 % p-value confidence, Q = 5 % for a result to be statistically significant.

3.2.5 Single-Likelihood Ancestor Counting - SLAC

SLAC (Kosakovsky Pond and Frost, 2005) is part of the HyPhy package on the datamonkey.org server (http://www.datamonkey.org/slac, Kosakovsky Pond et al., 2020) and uses the ML tree construction method and a counting approach to estimate the omega ratio on a per-site basis for any coding nucleotide alignment. SLAC performs a site specific selective pressure analysis and provides an ω value for each codon site of the MSA, while assuming a constant selective pressure for all species in the dataset. Specifically, branch lengths and nucleotide substitution parameters are optimised under the MG94xREV model and then ML is used to find the best fitting tree that represents the most likely ancestral sequence at each node of the phylogeny. Using the inferred ancestral sequences SLAC employs an adaptation of the Suzuki – Gojobori counting method (Suzuki and Gojobori, 1999), counting any synonymous and non-synonymous changes observed in each position of the ancestral sequences and then assigning statistical significance to each position, following a binomial distribution. Due to this counting-based approach, Kosakovsky Pond and Frost, (2005) warn that, when examining data sets with high divergence levels, SLAC analysis may not be accurate. The same MSA file that was produced from MEGA 7 for codeml was also used as input for SLAC for each gene and SLAC analysis was performed using the default parameters.

3.3 Results

3.3.1 Phylogenetic Trees – Species Trees and Gene Trees

Multiple sequence alignments (MSAs) for the 22 genes examined, varied greatly in length, ranging from 561 nucleotides for *ODAPH* to 10707 nucleotides for *LAMA3* (Figure 3.3). The distribution of the gene sequence length within each alignment is presented as a boxplot for each gene in Figure 3.4, with the detailed sequence lengths presented in Appendix E.2. The norMD quality scores of the MSAs for each gene, obtained by AQUA are presented on Table 3.4, the majority of the MSAs produced obtained a score > 0.6 indicating the high quality of the dataset. The MSA with the higher norMD score was selected in every case to be used for the subsequent phylogenetic analyses and in the case that more than one MSAs had the highest score MUSCLE MSAs were preferred, as they exhibited generally more consistent high quality.

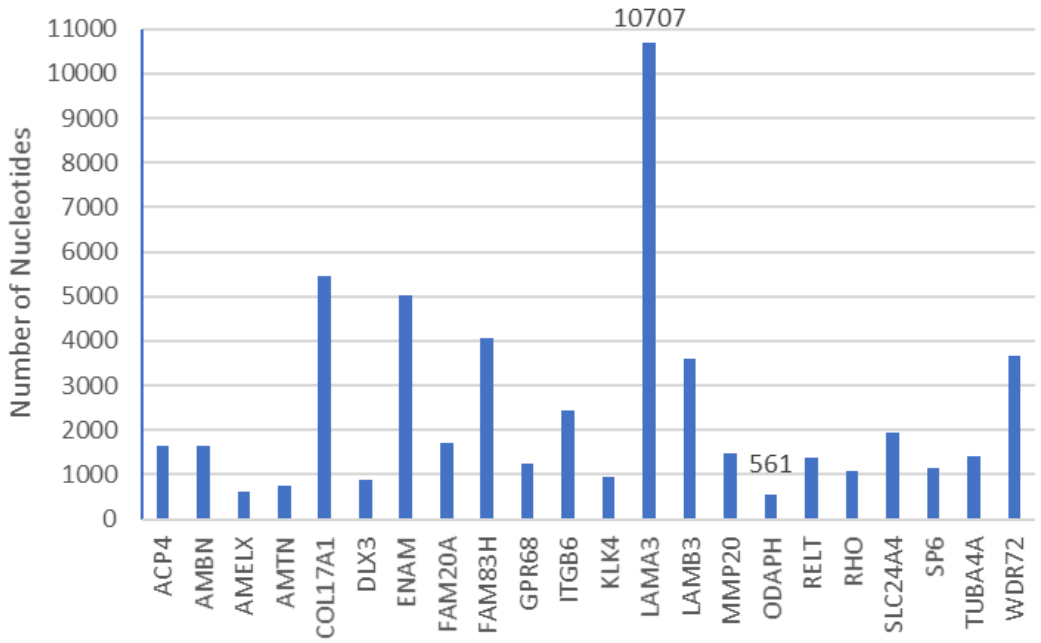


Figure 3.3: Length of genes included in the MSAs.

The x axis shows the length of the sequences included in the MSAs, in nucleotides, while the longest (*LAMA3* with 10707 nucleotides) and the shortest (*ODAPH* with 561 nucleotides) genes are marked on the graph.

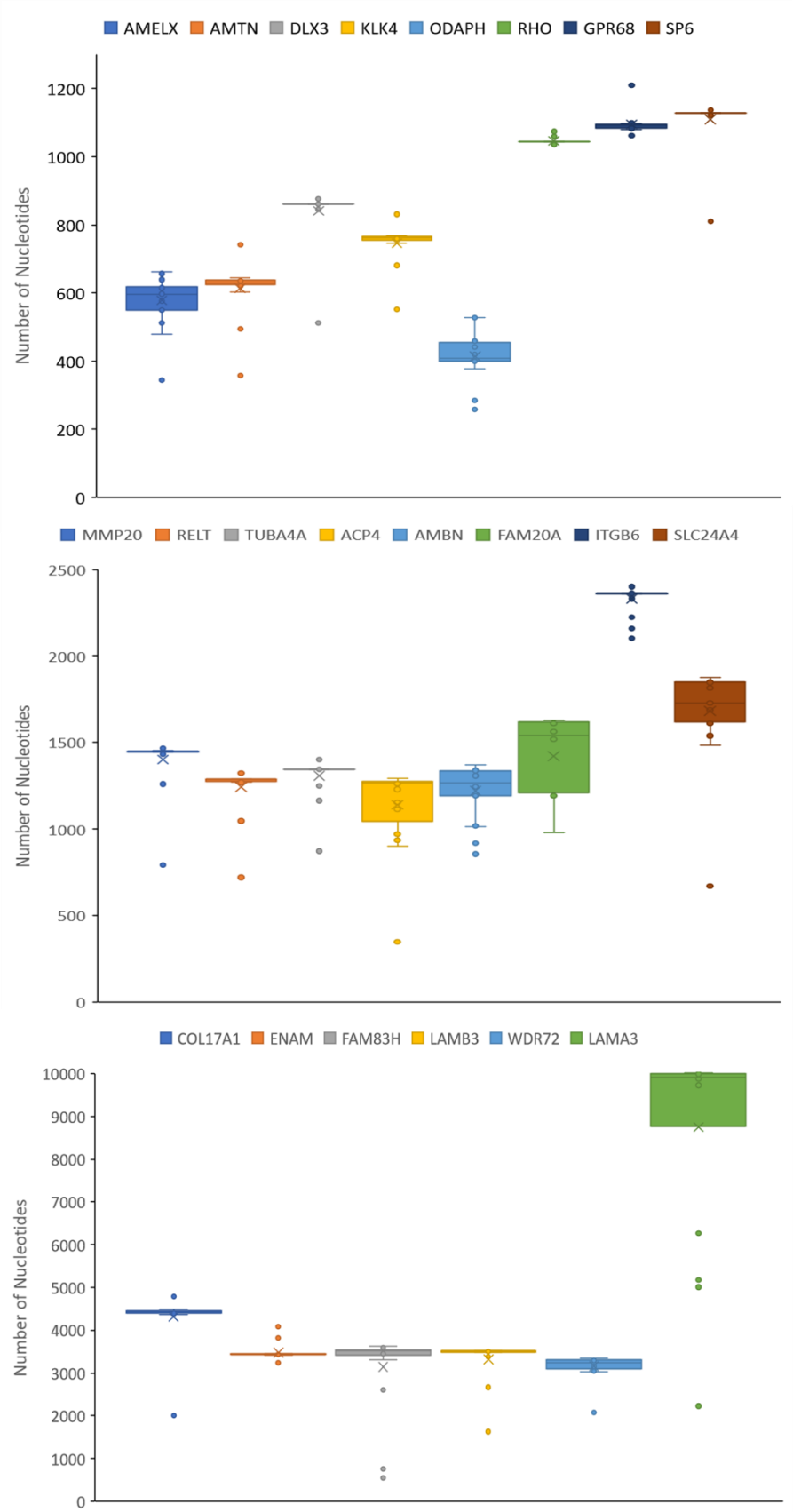


Figure 3.4: Distribution of the length of gene sequences that were used in this study.
For the purpose of this plot, in each gene the species for which the sequences are missing were marked as having length equal to zero.

Table 3.4: NorMD scores for the MSAs for each gene examined.

A score > 0.6 indicates a reliable MSA and a higher score shows a better quality alignment. The alignments were obtained using either the MAFFT or MUSCLE aligner and then refined by RASCAL. The MSA chosen for the subsequent analyses is indicated with an asterisk (*) next to the respective norMD score, if all MSAs had the same score the MUSCLE_RASCAL MSA was preferred.

Gene	MAFFT	MAFFT_RASCAL	MUSCLE	MUSCLE_RASCAL
ACP4	1	1	1	1*
AMBN	0.911	0.909	1.064*	1.043
AMELX	1	1	1	1*
AMTN	1	1	1	1*
COL17A1	1.741	1.813	1.467	1.542
DLX3	1	1	1	1*
ENAM	2.23	2.288*	1.971	1.964
FAM20A	1	1	1	1*
FAM83H	0.463	0.502	0.727	1.363*
GPR68	1	1	1	1*
ITGB6	1	1	1	1*
KLK4	2.416	2.392	2.43	2.481*
LAMA3	1	1.117*	1	1
LAMB3	1.426	1.43	1.743*	0.907
MMP20	1	1	1	1*
RHO	1	1	1	1*
SLC24A4	1	1	1	1*
SP6	1	1	1	1*
TUBA4A	1	1	1	1*
WDR72	1	1	1	1*

Table 3.5: AU test results of the gene trees.

The pAU value represents the bootstrap values from 10000 resamplings of the tree. The '+' sign denotes that the tree is within the 95 % confidence for statistical significance.

Gene	pAU	95 % confidence
ACP4	1.0000	+
AMBN	0.7841	+
AMELX	0.9817	+
AMTN	0.9867	+
COL17A1	0.9995	+
DLX3	0.8239	+
ENAM	0.9988	+
FAM20A	0.9994	+
FAM83H	0.9998	+
GPR68	0.9132	+
ITGB6	1.0000	+
KLK4	0.7888	+
LAMA3	0.9997	+
LAMB3	0.9999	+
MMP20	0.9990	+
RHO	0.9468	+
SLC24A4	0.9982	+
SP6	0.9125	+
TUBA4A	0.9952	+
WDR72	0.7165	+

The approximately unbiased (AU) test was conducted with IQ-Tree, using the ModelFinder module to find the best fitting model for the MSAs and the tree reconstruction to construct the best fitting ML phylogenetic tree. The p-values of the AU test are presented on Table 3.5 and correspond to bootstrap values derived from 10000 resamplings of the trees.

A representative MSA of a nucleotide sequence is shown in Figure 3.5 and the corresponding protein alignment is shown in Figure 3.6. The rest of the alignments can be found in the electronic Appendix. The nucleotide MSAs are presented to show changes at the nucleotide level that might lead to a silent mutation at the protein level which would not be seen at the protein MSA but is useful when examining the levels of variation among the DNA sequences. The protein MSAs are more convenient to compare the conservation levels of the amino acid residues of the peptides among the species.

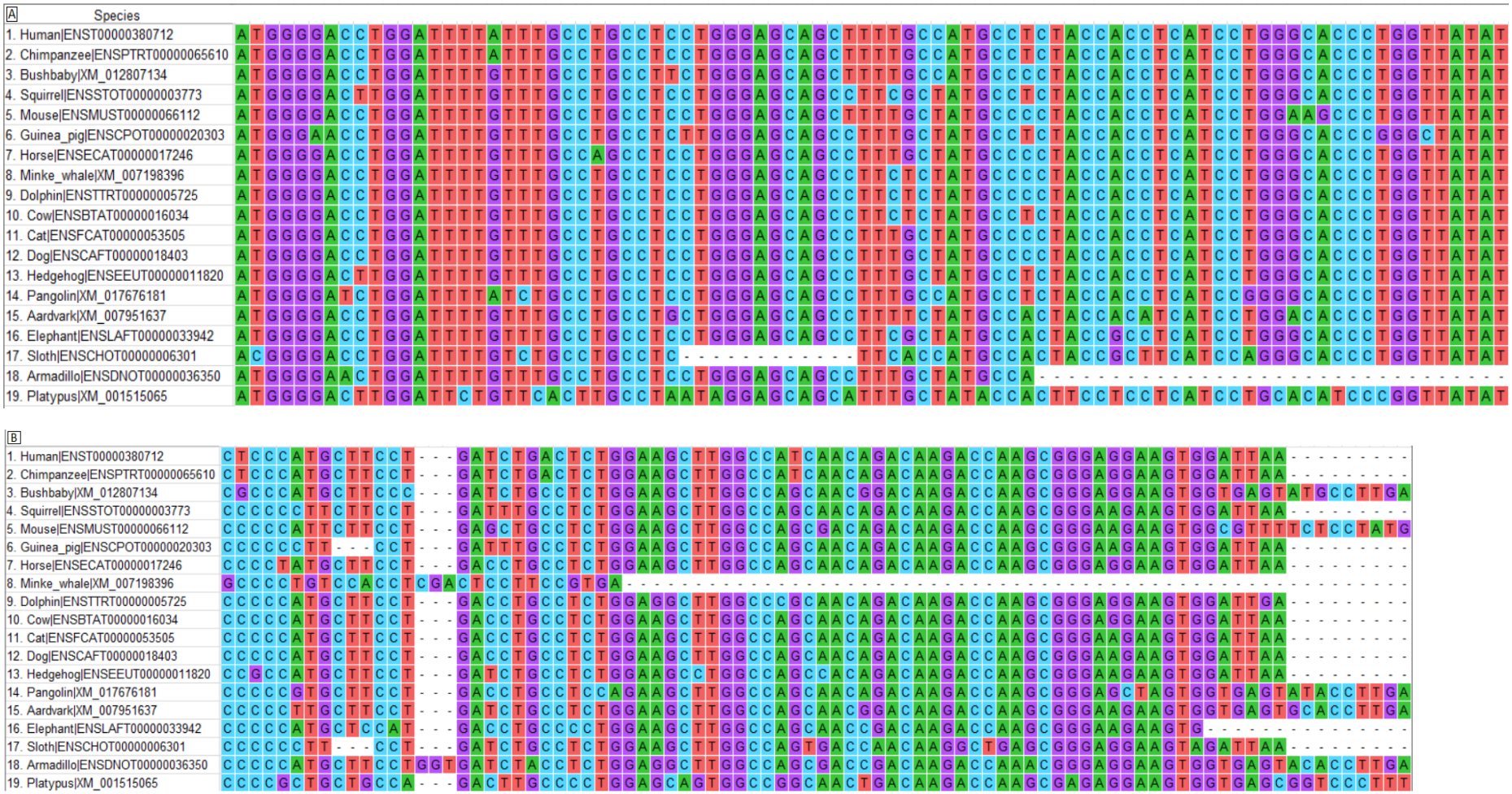


Figure 3.5: Representative MSA of *AMELX*.

MSA of *AMELX* at the nucleotide level, showing the degree of sequence conservation observed in the MSAs of all genes, both at the start of the sequence (a) and at the end (b).

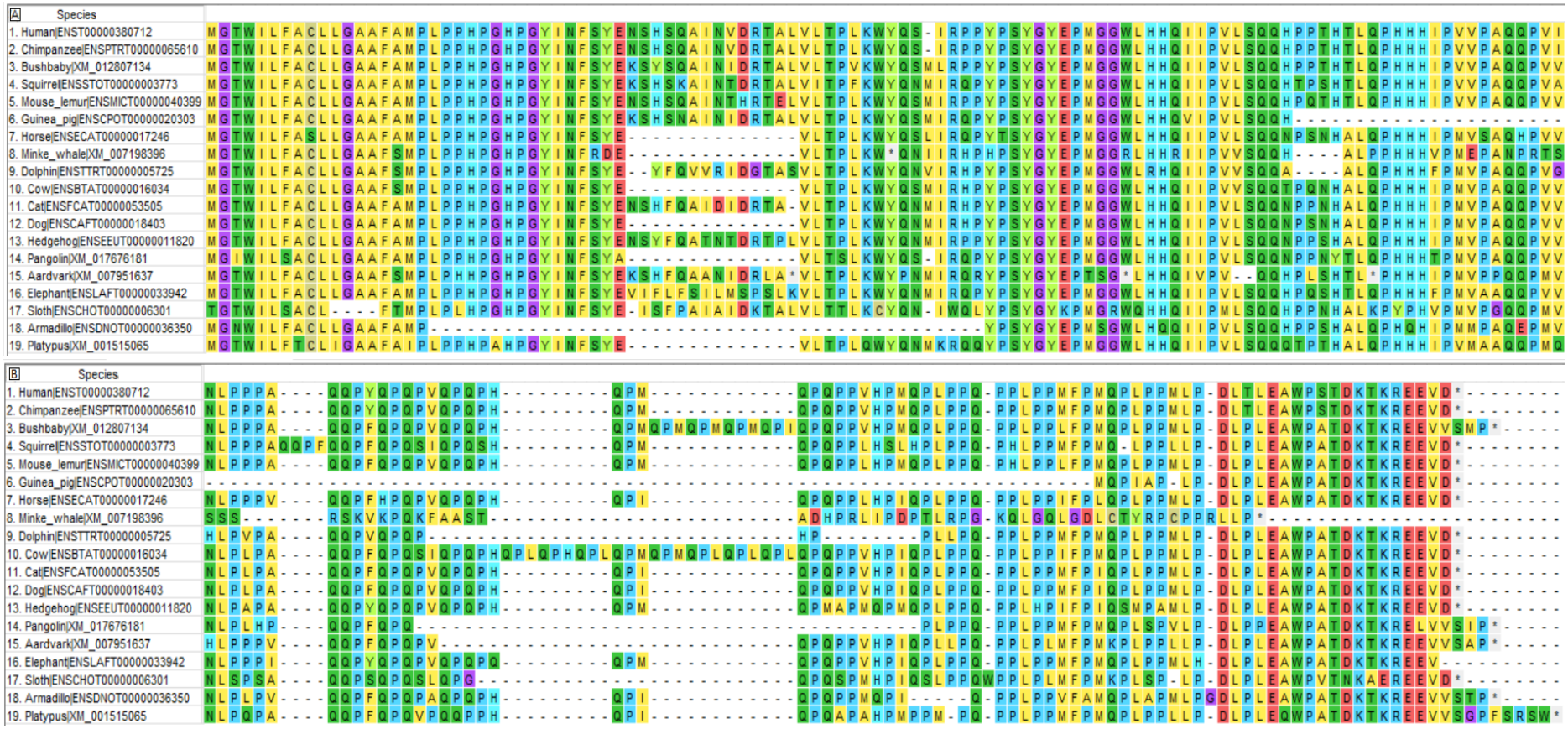


Figure 3.6: Representative protein alignment of AMELX.

The corresponding protein sequence alignment to the nucleotide MSA of Figure 3.5. The degree of sequence conservation that is presented here is representative for the MSAs of all genes, both at the start of the sequence (a) and at the end (b).

The adapted reference species tree from the literature (Morgan et al., 2013; Tarver et al., 2016) displaying the taxa included in this study is shown in Figure 1.6. For each gene family the ML trees were reconstructed (Appendix F) and an example is shown in Figure 3.7 for *ACP4* and *GPR68*.

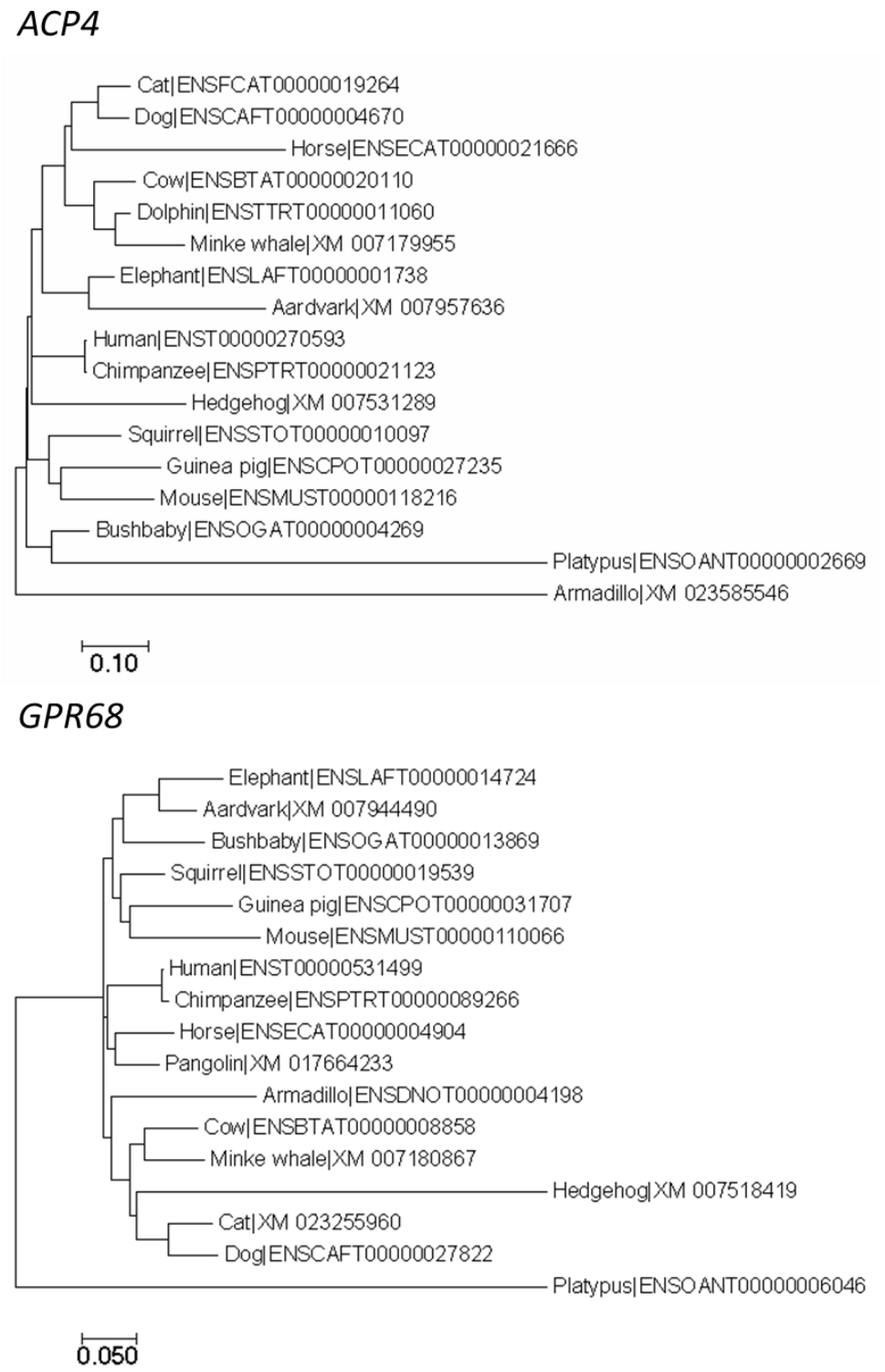


Figure 3.7: Representative gene trees.

Phylogenetic trees for *ACP4* and *GPR68*, constructed with the GTR+G model and the ML method. The length of the branches on the tree represent the number of nucleotide differences amongst the sequences.

By comparing Figure 1.6 and 3.7, the gene trees generally follow the same grouping for related species as the reference tree, with only a few exceptions, such as Bushbaby in the *GPR68* gene tree. There are however gene trees that have a completely different topology, like *ACP4* in the example of Figure 3.7b, in which the Euarchontoglires branch is not retrieved as a monophyletic branch. Some species are not represented in the gene trees and are simply marked as N/A on Table 3.2. As would be expected, the different rates of evolution among different species means that the length of the branches of the gene trees varies greatly. Relaxation of functional constraints and indeed pseudogenisation can also increase the branch lengths observed, as has been shown in other instances of loss of function of genes by pseudogenisation (Feng et al., 2014).

Using a RF distance calculation, the nodes of each gene tree to the species tree were compared, by taking into account the structure of the trees and transforming each gene tree to the reference tree. The more steps need to be taken to transform each tree the greater the value of the distance between them, these distances are summarised in Figure 3.8. Gene trees that have a value of 0 on the graph have an identical topology to the reference species tree (excluding branch lengths), with the higher values showing an increasing dissimilarity of the respective gene tree to the species tree.

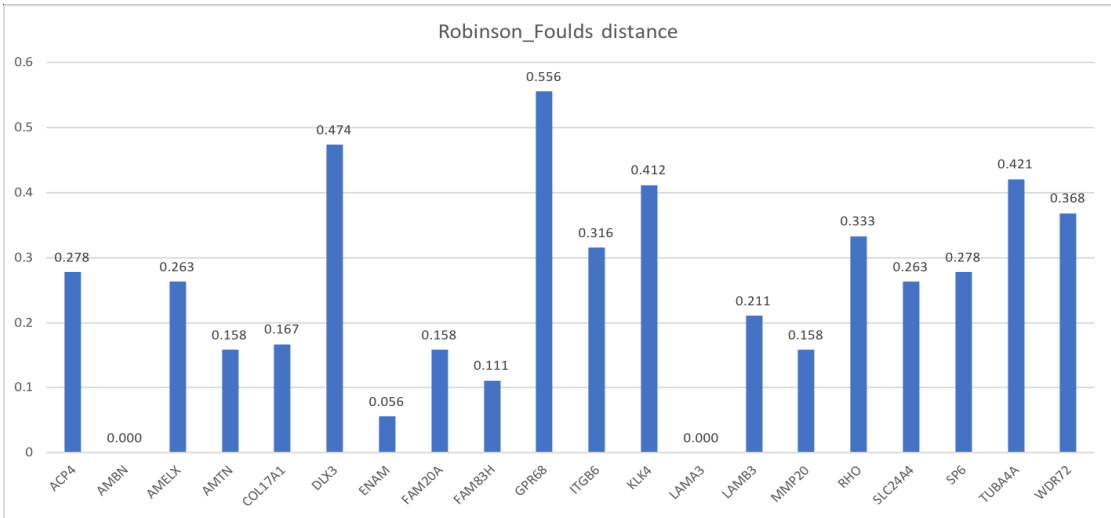


Figure 3.8: Dissimilarity of gene trees, according to the Robinson-Foulds distance.
The RF distance value of each gene tree is calculated by comparing to the reference tree of Morgan et al 2013. A value of 0 indicates no differences and the value proportionately to the dissimilarity.

3.3.2 Selective pressure analysis results – codeml

To determine if there was significant selective pressure variation across the sites in the MSAs the fit of the site-specific models M0, M1a, M2a, M7, M8 and M8a to the genes of interest was assessed. The estimates that were the results of the models are presented in Appendix D.2 and the detected positive selection along with its statistical significance are summarised in Figure 3.9.

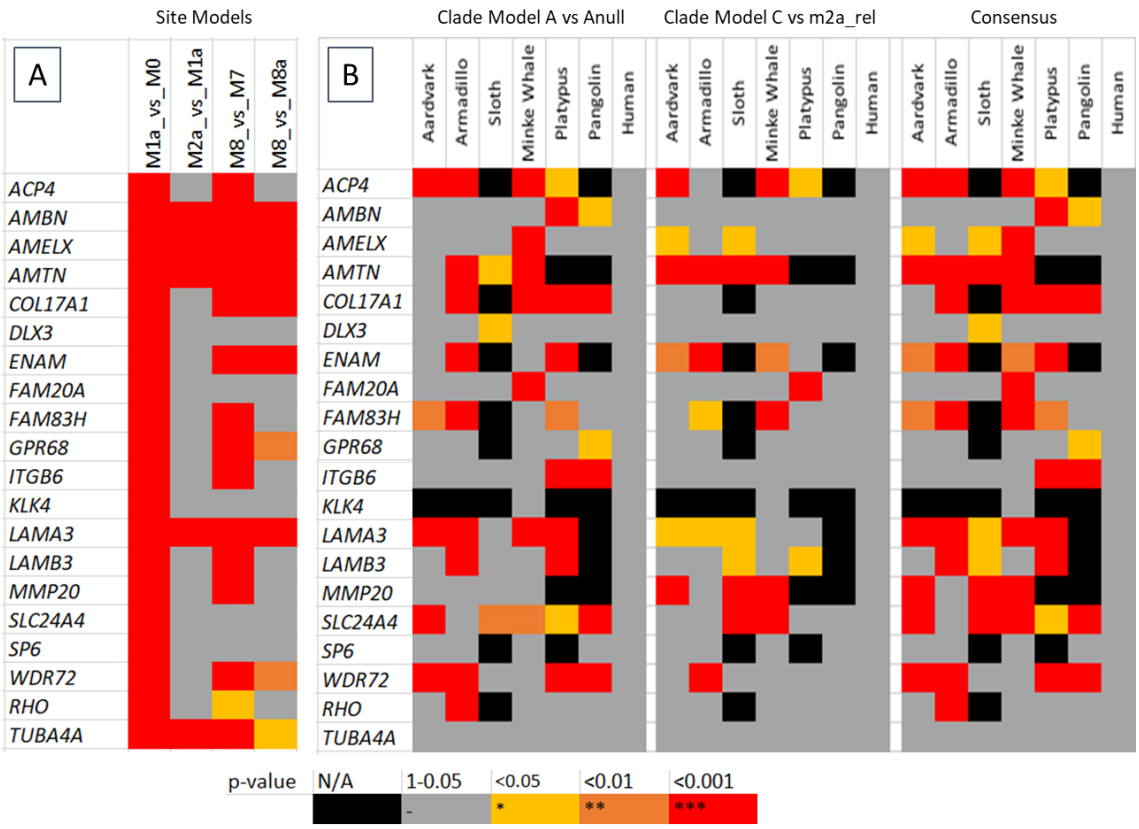


Figure 3.9: Heatmap of positive selection present in the genes.

(a) the heatmap according to the site models, each column is a different codeml model used, named in the format model_vs_null. Model M1a indicates the presence of selective pressure. Positive selection is accepted if either models M2a or M8 show significance. For M8 it has to be significant against both null models, so both M8_vs_M7 and M8_vs_M8a, to be accepted as positive selection. (b) Branch specific convergence models, i.e. “Clade models A and C” and their consensus, each column shows the species used as foreground for each test. Positive selection is accepted if either of the models shows statistical significance. Both heatmaps are colour coded to show the statistical significance of the LRTs according to the p-value of the results.

The branch specific models, Clade model A and Clade model C, were employed to address whether there were specific lineages in the MSAs where there were signatures of selective pressure variation that were unique to that lineage. These specific lineages of interest are referred to as foreground lineages and are presented in the columns of Figure 3.9b. A total of 7 foreground lineages were assessed, across 22 genes and 2 clade models for each of them, totalling 308 models. The lineages labelled as foreground were the 3 enamel-less (Aardvark, Armadillo, Sloth), the 3 toothless (Minke whale, Platypus, Pangolin) and a representative of the toothed species (Human). The selective pressure at play in these lineages was in direct comparison of the foreground species to the remaining 18 mammals in the dataset. That is also a direct comparison of the toothless / enamel-less species that are used as foreground, showing that the different species have different profiles of selective pressure variation, without these profiles being linked to the phylogenetic relations among the species. Additionally, Human, which was included as a foreground as a control representative of the toothed species, shows that in toothed species that still have functional copies of the genes there is no branch specific selective pressure. Additionally, in the clade models the two genes that are not linked to teeth or enamel, *RHO* and *TUBA4A*, are among the genes that show the least amount of branch specific positive selection acting on them, with the housekeeping *TUBA4A* showing no statistically significant evidence of positive selection among the branches.

3.3.3 SLAC analysis results

The results of the SLAC analysis on the examined genes are presented in Figures 3.10 and in the Appendix G, with ω values close to 0 being considered as neutral evolution, negative values given to amino acid residues indicate conserved evolution and positive values indicate positive selection. Only sites with p-value < 0.05 are accepted as under positive or purifying selection. The two genes that are not enamel specific, *TUBA4A* and *RHO,* were also included, with the same purpose as during the codeml analysis, to be compared to the tooth or enamel specific genes. Each gene presents a different selective pressure variation profile, but all tend to show most of the sites being under neutral evolution, many sites being under purifying selection and only a limited number of sites being under positive selection. In these figures, the X axis is the position of the site on the protein alignment and the Y axis represents the dN/dS value of the site, with 0 being neutral evolution, > 0 being positive selection and < 0 representing purifying selection.

The SLAC results (Appendix G) of the genes do not show any correlation with the results from the codeml analysis, emphasising that the results obtained by the two approaches are not comparable, but can be used complementarily, as SLAC estimates the ancestral sequences and calculates the dN/dS based on this predicted sequence. This discrepancy between approaches is expected, as reported by studies employing both programs in the literature (Kulmuni et al., 2013; Huang et al., 2017), due to the difference in the approach each program takes to calculate the dN/dS ratios.

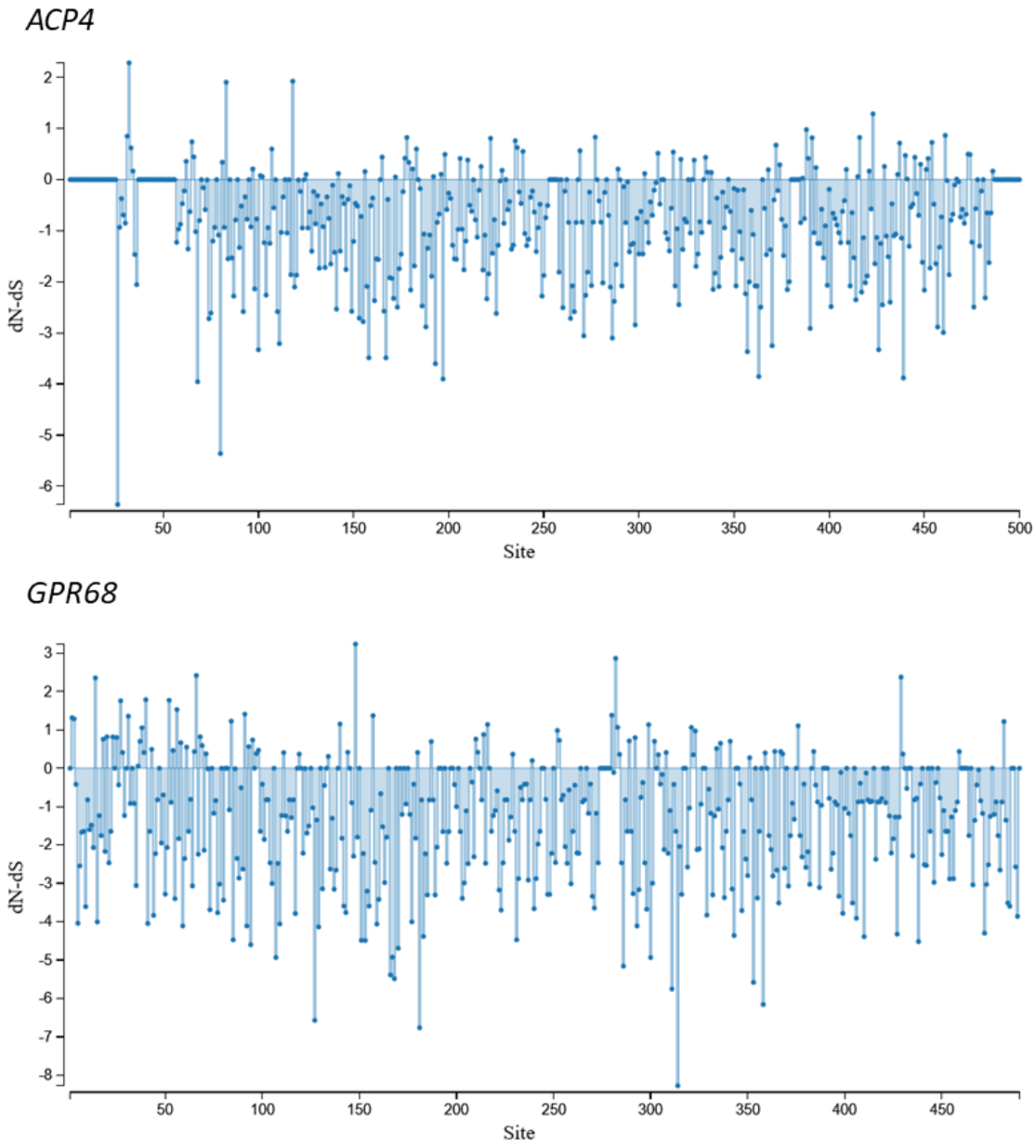


Figure 3.10: Representative SLAC results.

The SLAC results for *ACP4* and *GPR68*, positive values indicate positive selection and negative values indicate purifying selection.

3.4 Discussion

In this chapter, the evolutionary history of the genes associated with an AI phenotype was examined in an attempt to identify patterns of natural selection that are shared among them, via selective pressure analysis. Codeml examines the variation in selective pressure across individual nucleotide sites and/or on specified branches of a phylogeny for a set of homologous protein coding sequences, at the nucleotide level. However, in both cases codeml outputs the estimated ω ratio for the entire gene and does not provide the values per site, so that an overview of the selective pressure acting on the gene is obtained, but not which specific regions of the gene might be more (or less) conserved than others. SLAC was selected to complement the codeml site specific analysis , as it is also used to calculate the selective pressure acting on each individual site in the nucleotide alignment. SLAC also estimates the dN/dS substitution rate per site of the sequence, treating each codon as an individual site, while examining the nonsynonymous and synonymous substitution rates of the nucleotides comprising that codon. Outputting the values of dN/dS per codon in an alignment, it allows us to plot these values in more detail and better visualise the variation in selective pressure across the sequence. This can be invaluable when trying to identify regions of the sequence that show increased positive selection or conserved in a greater degree than the rest of the sequence.

The sequences used to construct the MSAs and to then perform the codeml and SLAC analysis were selected from the longest canonical transcripts, with the least number of ambiguities, amongst the sequences available on the online databases. Typically, Ensembl was shown to contain better quality sequences and was preferred, with missing sequences, or species that Ensembl does not include, being retrieved from GenBank. All terminal stop codons were removed from the alignments of the sequences, as the aligner would need to translate the nucleotide sequence to the corresponding protein sequence to align based on the codons, instead of each nucleotide in isolation. The internal stop codons in sequences suspected as pseudogenised were changed to ‘NNN’ to be interpreted as ambiguity from the aligner for the same reason. The sequences of suspected pseudogenes were examined to locate internal stop codons, also called premature termination codons (PTCs). PTCs were found exclusively in toothless and enamel-less species, an example being Aardvark in the protein MSA of Figure 3.6, suggesting a causative relation to the pseudogenisation process. An interesting observation that arose when masking these PTCs is that after the PTCs the rest of the sequence would be unexpectedly conserved and similar to the coding sequences of the toothed species, often to the end of the sequence. This finding will be discussed in detail in Chapter 4.

Having aligned the sequences to get the MSAs for each gene, the respective gene trees can be constructed. Another factor to consider is whether these gene trees will be used for the selective pressure analysis with codeml, or the reference species tree of Figure 1.6 from Morgan (Morgan et al., 2013; Tarver et al., 2016) should be used instead, after it is pruned so it only contains the species included in this study, instead of the entire mammal clade. Gene trees, as shown in Figure 3.7, are not necessarily identical to the species tree, so they don’t always reflect the true phylogeny of the species. There is a number of reasons for this dissimilarity, with the common reasons being that gene families’ evolution can be subject to

incomplete lineage sorting, duplications, deletions, recombination and gene hybridization or horizontal gene transfer (Maddison, 1997; Swenson and El-Mabrouk, 2012). In the organisms examined for this project, i.e.: mammalian species, horizontal gene transfer is not likely and can be rejected, but the other causes of tree dissimilarities cannot be disregarded. Except for the biological reasons for the difference, gene trees might also be limited by a small sample size, that might not be enough to resolve all nodes on the tree. To avoid all these possible inconsistencies between gene tree and species tree and after observing how dissimilar some of the gene trees are compared to the accepted mammal phylogeny (Morgan et al., 2013; Tarver et al., 2016), the pruned species tree was used instead of the gene trees.

Some housekeeping genes are known to be undergoing positive selection, such as the genes of the histone protein family (Ponte et al., 2017), which poses an interesting question; whether housekeeping genes can be used as examples of heavily conserved genes to be compared against in studies of positive selection. Ideally, it is preferable to have genes that can be used as negative 'controls' for selective pressure, but given that each gene has its own independent evolutionary history they cannot be used as a control for a gene from a different gene family.

Despite not being able to use them as negative controls, two genes that were not enamel or tooth specific were included among the AI genes, *RHO* and *TUBA4A*, to examine whether they would follow any patterns identified from the enamel or tooth specific genes. As can be seen in Figure 3.9, *TUBA4A* is positively selected in the mammal clade but when specific foreground species were examined, i.e. human and the toothless/enamel-less species, *TUBA4A* does not show any sign of positive selection, showing that there is no difference in the rate of evolution of *TUBA4A* among the mammal species with or without enamel/teeth and the positive selection observed from the site specific models is common among the mammalian clade. This is consistent with the findings on the histone proteins of Ponte et al., (2017) as it shows that *TUBA4A* is among the housekeeping genes that are under positive selection in mammals. *RHO* is not shown to be under positive selection by the site-specific models, whereas the branch specific models show it to be under positive selective pressure in Armadillo. This observation can be explained by the *RHO* in armadillos being part of the ongoing adaptation of the animals to their environment, while the gene is under purifying selection in the rest of the mammalian clade as it needs to be preserved.

The branch specific models, summarised in Figure 3.9b, show that toothless/enamel-less species have a different profile of selective pressure variation compared to the toothed mammals, which is consistent with the findings of large scale studies on the mammalian clade for individual genes, such as Huang et al., (2017) and Gasse et al., (2017) that showed that tooth specific genes can be under purifying selection for the majority of the species and under positive selection in toothless or enamel-less species. Among the genes included in the analysis, many are considered as pseudogenes, or non-functional, in these species, so the selective pressure constraints are expected to be relaxed so they can change by genetic drift to either be fully pseudogenised and have their coding sequence completely disrupted, or to shift in function with positive selection aiding to incorporate new mutations in their sequence. The shift in function is dependent on the coding region still producing a peptide, even if it is

not fully functional for its original purpose, on which new mutations that get established with the aid of positive selection can shift its function.

From the genes examined, three categories of genes came up from the codeml analysis, the first being genes that have sites under positive selection as shown by the site-specific models, genes that are under positive selection in the majority of the toothless / enamel-less species and genes that are not under significant positive selection (Figure 3.9). Of the first category *AMBN*, *AMELX, GPR68* and *TUBA4A* show signs of positive selection that is shared across all mammals and may be a sign that the genes are still adapting to improve their fitness for their function. Together with them *AMTN*, *COL17A1*, *ENAM*, *LAMA3* and *WDR72* show the same signs of positive selective pressure acting on them for the entirety of the mammalian clade, but also show significant positive selection when the toothless / enamel-less species are used as foreground for the branch specific models, but not when human is used as the foreground, which could mean that the adaptation is ongoing for these genes in the mammalian clade, but they evolve at a faster rate, a potential indication that these genes are headed towards a functional shift in the toothless / enamel-less species. Similarly, the genes in the second category, specifically *ACP4, FAM83H, LAMB3, MMP20* and *SLC24A4* are not shown to be under significant selective pressure from the site-specific models but are shown to be so for the majority of the toothless / enamel-less species. As mentioned previously, this is consistent with the literature (Huang et al., 2017; Gasse et al., 2017) and these genes are the ones more often reported to be pseudogenised in the toothless / enamel-less species (Section 3.1), indicating that the positive selection detected is part of the process of pseudogenisation, or of a shift in function. The last category of genes, including *DLX3, FAM20A, ITGB6, SP6* and *RHO* are not under statistically significant positive selection from the site-specific models or for the majority of the species examined with the branch specific models, indicating that there are selective constraints acting on them, to preserve their functionality. Some exceptions can be noticed, such as minke whale for *FAM20A*, platypus and pangolin for *ITGB6* and armadillo for *RHO*, for which the positive selective pressure can be attributed to ongoing adaptation of the gene, that is specific to the branch in which these species are found. For *KLK4* the coding sequence of all toothless / enamel-less species, other than Minke whale, was missing from the databases, preventing the branch specific analysis and potentially affecting the site-specific analysis, so the results of codeml for *KLK4* cannot be properly interpreted.

The results of the SLAC analysis, as reported in section 3.3.3 and the examples of Figure 3.10, show that for the genes examined most of the sites (codons) are estimated to undergo neutral evolution, many sites are under purifying selection and only a few sites are under positive selection. For the sites under purifying selection, it is often observed that their ω value can reach extreme negatives, pinpointing the position of amino acid residues that are essential for the function of the protein. The codeml analysis estimates whether a sequence is under positive selection by calculating the ω values for each site of the nucleotide sequence and averaging the values to estimate the effect on the gene. A direct comparison of the sites indicated as under positive selection by codeml and by SLAC is not appropriate, as codeml is used to estimate the total selective pressure acting on a gene with the site models and the branch specific selective pressure with the clade models and SLAC is used to estimate the pressure acting on the individual codons. So, in this study the two programs were used as

complementary tools to have a more thorough description of the selective pressure variation of the genes of interest. However, in studies that compare the two programs they concluded that there is little overlap of sites indicated as positively selected by both programs, as reported by Huang et al., (2017) for FAM83H as well as other reports in the literature that compared the results of the two programs (Kulmuni et al., 2013), or similar to us didn't compare them but used them complementary (Duarte et al., 2021) to avoid this complication. This is due to codeml estimating the best fitting ω value for the entire dataset and then forcing all positively selected sites to have the same ω value for its analysis, while SLAC estimates the ω value of each site independently, site by site, allowing the ω to vary among them (Kosakovsky Pond and Frost, 2005). Additionally, SLAC is based on reconstructing the ancestral sequence of the MSA used as input and conducting the selective pressure analysis on this, instead of using the given sequences.

It was expected that by investigating the selective pressure acting on the genes associated with non-syndromic AI we would be able to discern a unique pattern of natural selection that could aid in the identification of novel genes to be associated with AI. Ultimately, the results of the site-specific analysis, with either codeml or SLAC, did not reveal any such patterns. Furthermore, the lineage specific analysis showed that in the majority of the toothless or enamel-less species most of the genes show signatures of positive selection, an observation that indicates an evolutionary drive to a potential shift of function for these genes, as an alternative to the pseudogenisation model for their inactivation.

Chapter 4 - Premature termination codons in genes and potential stop codon readthrough

4.1 Introduction

4.1.1 Premature termination codons in coding sequence of genes

As shown in section 1.4.7 and in Figure 1.8, The 'pseudogenization model' of gene inactivation posits that functional genes accumulate mutations that generate premature termination codons (PTCs) that lead to a truncated, non-functional peptide. The standard translation mechanism in eukaryotic cells (Figure 1.9), dictates that the presence of a termination codon in a coding sequence indicates the position that the translation ends. Unlike the termination codons expected to be found at the end of the coding sequence of a gene, PTCs are found inside exonic regions of the coding sequence and lead to incorrect termination of translation and the production of truncated proteins that have lost part or all function. In pseudogenised genes, after the PTC the sequence is no longer under selective pressure to be conserved, as the product of the sequence is no longer translated into a functional protein. As a consequence, any mutations that occur due to genetic drift (Section 1.4.3.1), will not be removed to preserve the ancestral sequence, but will remain and slowly alter the sequence to lose any resemblance to its homologs. Additional stop codons will also be introduced, establishing the pseudogenisation of the former coding sequence and leading to even further divergence of the sequence (Figure 4.1).

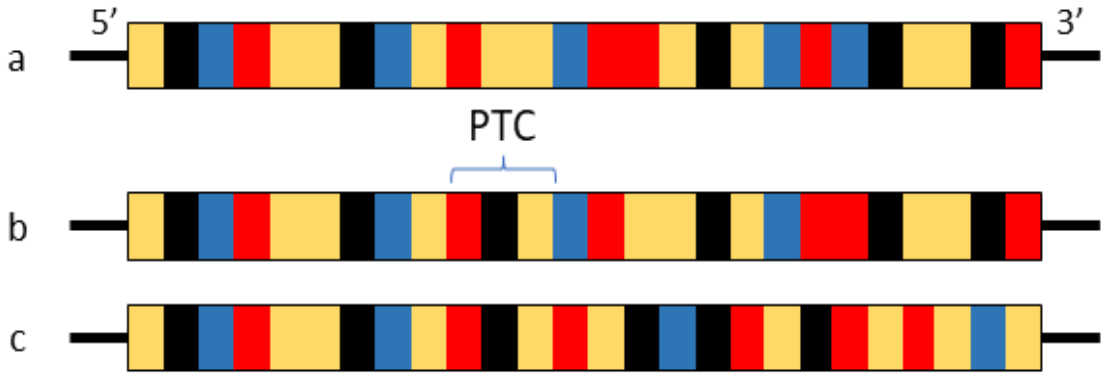


Figure 4.1: Fate of a coding sequence after inactivation by PTC.
(a) The ancestral state of a functional sequence. (b) A PTC is inserted but the downstream sequence remains conserved. (c) The same PTC is inserted but the downstream sequence diverges completely leading to a loss of any resemblance to the ancestral sequence.

The current consensus is that inactivation of tooth or enamel specific genes in toothless or enamel-less mammals, by PTCs are causal for the enamel-less or toothless phenotype (Section 3.1). To reiterate, examples of inactivation of genes by PTC are *ACP4* in aardvark, armadillos and the minke whale (Sharma et al., 2018), *AMTN* in armadillos and sloths (Gasse et al., 2012), *AMBN* in aardvark, armadillo and sloths, *ODAPH* in pangolin and baleen whales (Springer et al., 2016), amongst other genes. The findings presented in the MSAs of the previous section, Chapter 3, identified PTCs that would be sufficient to cause the inactivation of the genes that they were located in, but in some of them the sequence after the PTC was still conserved as if it was still functional and under selective pressure (Figure 3.6).

4.1.2 Mechanisms that circumvent termination of translation

The retention/conservation of the sequence after the PTC, although unexpected, is not unheard of, as it can be explained by two different but similar mechanisms that have been observed in both prokaryotic and eukaryotic cells. These mechanisms are translational recoding and stop codon readthrough or functional translational readthrough (FTR).

Translational recoding uses mRNA elements to allow the ribosomes to alter the meaning of codons, decode mRNA in alternative reading frames or even skip parts of an mRNA (Dever et al., 2018). With regards to the PTCs identified here, translational recoding could induce the recoding of the stop-to-coding codons, preventing the termination of translation. A prominent example of translational recoding of a stop codon is the incorporation of non-canonical amino acid residues in peptides, such as selenocysteine and pyrrolysine, which are incorporated by recoding the UGA and UAG stop codons respectively (Touat-Hamici et al., 2014; Hoffman et al., 2018). As described in Section 1.4.7, when translational recoding occurs the coding sequence contains a multitude of stop codons, as they are needed to incorporate the non-canonical amino acid residue in the peptide being synthesized (Mariotti et al., 2012). Consequently, this study focuses on the second mechanism for continuing translation after a stop codon, FTR.

FTR occurs when translation doesn't terminate upon reaching a termination codon but continues as if the stop codon is a coding one (Figure 1.9b). The process by which this happens is that near-cognate tRNAs (nc-tRNAs), which have anticodons complementary to two of the three nucleotides of a stop codon, introduce a regular amino acid residue into the position (Figure 1.9). The single nucleotide mismatch has been observed to occur in either the 1st or 3rd position of the anticodon (Roy et al., 2015). The nc-tRNAs compete with the translation termination factors at the tRNA binding site of the ribosome, Figure 1.9, and allow the translation of the protein to continue, albeit in a greatly reduced capacity. The efficiency of FTR in a genome is estimated to be about 0.001 – 0.1 % of total transcripts in a cell, which increases to 1 - 6 % of total transcripts with induced FTR (Brooks et al., 2006; Roy et al., 2015; Loughran et al., 2018). Some general principles of this process have been established, e.g.: different stop codons allow for different readthrough efficiency, with UGA being the easier to be readthrough and UAA almost never allowing FTR (Manuvakhova et al., 2000). In addition, different nucleotides downstream of the stop also have different degrees of accommodating FTR, the order being: C>T>G>A (Jungreis et al., 2011).

One of the first examples of FTR reported was in the Tobacco Mosaic Virus (TMV) (Pelham, 1978), where readthrough of UAG was shown to be regulated by the 6 nt following the stop

codon (Skuzeski et al., 1991). In other words, FTR is informed by the sequence context in which the stop codon is found. In the Murine Leukaemia Virus (MuLV) the *gag* and *pol* ORFs are separated by an in frame stop codon and readthrough is induced by a downstream RNA secondary structure, called a pseudoknot (Staple and Butcher, 2005), at a 5% efficiency, showing FTR can be structurally dependent (Csibra et al., 2014). Similarly, in many other plant and animal viruses, there are structural elements at the 3' end that stimulate readthrough in a structure dependent manner (Firth et al., 2011; Rodnina et al., 2020).

The first mammalian gene candidates for FTR were: *OPRL1, OPRK1, ACP2, MAPK10, SACM1L*, all with highly conserved UAG stop codon followed by CUAG (Jungreis et al., 2011), but without any functional similarity or indication of interaction. The same tetranucleotide motif (CUAG) was also found after the stop codon of 23 further human genes, and follow up experiments on these candidates using *in vitro* constructs showed that FTR also occurred here but with varying degrees of efficiency (Loughran et al., 2018). The motif present in all of these human genes was UGA_CUAG, which is significantly depleted in the human genome (Loughran et al., 2018), indicating a selective pressure to reduce the instances of accidental readthrough. The CUAG footprint was proposed as a key motif in identifying context dependent readthrough from primary sequence data. Jungreis et al (2016) report that they identified over 300 instances of FTR in mosquitos of the genus *Anopheles* and 51 in *D. melanogaster*, estimating that FTR happens for ~600 mosquito stop codons and ~900 in the fruitfly genome suggesting this is reasonably commonplace.

4.1.3 Chemical induction of stop codon readthrough

Readthrough has also been chemically induced, via use of antibiotics, specifically gentamicin and other aminoglycoside antibiotics, with varying degrees of efficiency (Brooks et al., 2006; Dabrowski et al., 2015; Sabbavarapu et al., 2018). This research was conducted with the aim of using drugs to induce readthrough to rescue the phenotype in diseases associated with peptides being disrupted by PTCs. Similarly, in studies on JEB, discussed in section 1.2.7, attempting to rescue the expression of *LAMB3* and the formation of laminin 332, the use of gentamicin was found to normalise the morphology, proliferation rates, cell matrix adhesion and hypermotility of JEB cells (Lincoln et al., 2018). The use of these antibiotics has been shown to affect both the PTCs and the natural stop codons at the end of the coding sequence, causing concerns about cytotoxicity (Lincoln et al., 2018). Additionally, although experimental evidence shows that they increase the readthrough efficiency, as of yet there are no reports of significant improvement in the clinical phenotype of the patients that participated in these studies (Dabrowski et al., 2018).

4.1.4 Studying stop codon readthrough

To reliably assess the presence and efficiency of FTR the dual luciferase assay is the most common approach taken. Luciferase assays are a type of reporter assay, that have been used for a multitude of reasons from assessing gene expression to studying the effect of promoter and enhancer elements, to comparing the expression levels and localisation of proteins (McNabb et al., 2005; Smale, 2010; Nair and Baier, 2018) among others. The principle underlying the method being that by adding a regulatory element of interest (e.g. a promoter), upstream from a luciferase gene in an expression vector (a plasmid), and transfected in an

appropriate cell line, one can observe and quantify the efficiency of that regulatory element based on the intensity of the read out from the luciferase component of the construct (Nair and Baier, 2018). The dual luciferase assay builds on this approach by placing the region of interest between the coding sequences for two different luciferase genes, so we can assess how our insert affects the translation of its downstream sequence (Figure 4.2). The inserted sequence will be expressed consecutively after the sequence of the first luciferase gene and before the second. In the case that the insert contains a PTC it is expected that only the first luciferase gene will be expressed, and the translation will stop before the second gene, but if there is FTR, both luciferase genes will be expressed. This setup allows the study of other non-regulatory elements that can be inserted between the luciferase genes, whereas The dual luciferase assay has been previously applied to identifying and assessing instances of FTR in human genes (Loughran et al., 2014; Loughran et al., 2018) and is the approach of choice in this chapter.

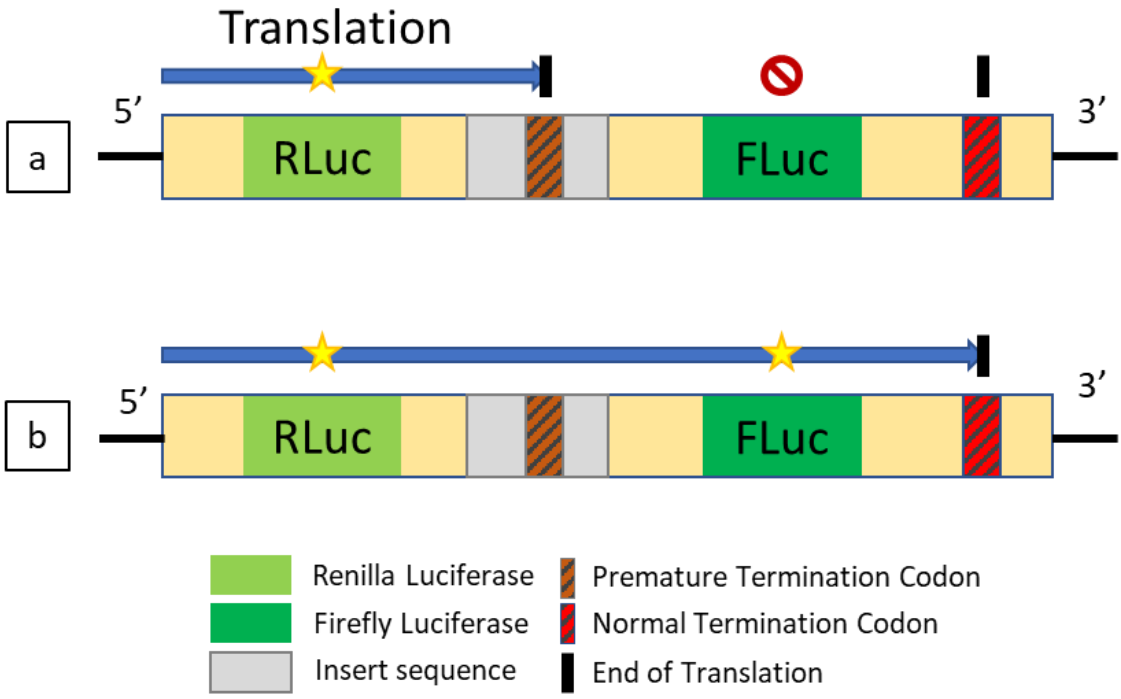


Figure 4.2: Dual luciferase construct with a region of interest that contains a PTC.

(a) On the mRNA level, translation stops at the PTC and the *FLuc* is not expressed, whereas in (b) the PTC is being readthrough so both reporter genes are expressed, and their luminance can be quantified.

In summary, from examining the Multiple Sequence Alignments (MSAs) generated for the AI associated genes as presented in Chapter 3, unexpected levels of sequence conservation were observed after PTCs in six genes, *ACP4, AMBN, AMELX, ENAM, MMP20, FAM20A* and *SLC24A4* in four toothless or enamel-less species (aardvark, armadillo, minke whale and pangolin). In this chapter, we explore the patterns of conservation and divergence in these sequences along with concepts of transcriptional readthrough to understand the patterns we are observing in our data to determine whether these genes should be considered as deactivated/pseudogenised and nonfunctional or whether they are indeed producing peptides via FRT.

4.1.5 Readthrough or genome sequencing error?

However, before committing to the investigation of possible signatures of FTR in the genes identified, the possibility that the PTCs are an artefact of a low-quality genome needs to be considered. The sequences used in this study were of the best available quality, but the need to include toothless and enamel-less species narrowed down the options for available species. As a result, the regions containing the PTCs need to be independently examined and the sequence validated with Sanger sequencing before the investigation on FTR can proceed.

4.1.6 Aims of this chapter

This chapter seeks to clarify the nature of the internal stop codons that were identified in the MSAs of AI associated genes in toothless or enamel-less species. This is a novel approach to the pseudogenisation of these genes in toothless or enamel-less species, as it is proposed that the sequence of these genes is conserved due to stop codon readthrough, which produces functional gene products.

Specifically, the aims of this Chapter are to:

a) Determine whether the stop codons we identified in the alignments are real or artefacts of the genome sequencing and annotation by resequencing regions of specific genes
b) Investigate if the high level of sequence conservation after the stop codons found in our alignments contain the context dependent signatures associated with stop codon readthrough, and
c) Examine if these genes have dual luciferase readouts consistent with FRT, i.e.: do they produce a translated protein that extends beyond the internal stop codons for these specific genes in toothless/enamel-less mammals.

4.2 Materials and Methods

4.2.1 PCR and Sanger Sequencing

In order to determine the validity of the PTCs identified in the MSAs, the sequences in the regions containing the PTC in each gene were validated by Sanger sequencing. The PCR and Sanger sequencing reactions for the minke whale sequences were performed as described in section 2.2, albeit in the lab of Professor Per Palsboll and his group at the Groningen Institute of Evolutionary Life Sciences, University of Groningen, The Netherlands. PCR cleanup was conducted using Sera-Mag beads, made in-house, at a ratio of 0.9x beads : PCR product, per reaction, when necessary due to amplification of non-specific secondary bands in the PCR products. The sequence of the PCR products was then read with Sanger sequencing (Section 2.2.9). The primers used for the reactions can be found in Table 4.1 below. The DNA samples used for the validation were selected from the cohort of minke whale samples, maintained at the University of Groningen by the group of Prof Palsboll.

Table 4.1: List of primers used for the validation of the minke whale sequences.
Each primer strand is given in a 5' – 3' orientation.

Gene	Direction	Sequence 5'-3'
ACP4	Forward	CCTAAGGGGAGCATGAAGAG
	Reverse	AAAGCCAGTCCTTCCAGAAA
ENAM	Forward	GGGAGGTCCAAGAAGTCAAA
	Reverse	GGGCAAATTCTGGCTTTGGG
FAM20A	Forward	CCAGCAACGTGTGTTTCTTC
	Reverse	TCAGCCTTCACGAGTCCCTA
MMP20-A	Forward	AAGGAGCTGCAGGCTTTCTTC
	Reverse	GGTTGGTACCATCAGGCTAT
MMP20-B	Forward	ATGAGGAAACGGAGGCTTAG
	Reverse	GTGGCAAGAGCATAGTAGGT

To examine the PTCs found in the other toothless or enamel-less species, stool samples were donated from Chester Zoo for anteater, aardvark and platypus, to be used for DNA extraction and downstream applications, as described for minke whale in the previous paragraph. The certificate of donation can be found in Appendix H. The E.Z.N.A.® Stool DNA Kit (Omega Bio-tek Inc., GA, USA) was selected for the DNA extraction from the stool samples.

4.2.2 Cloning and dual luciferase assay

The dual luciferase assay protocol of Loughran et al (2018) was adapted to determine if there is evidence for stop codon readthrough in six genes (*ACP4, AMBN, AMELX, ENAM, MMP20, FAM20A* and *SLC24A4*) in four species of interest (aardvark, armadillo, minke whale and pangolin). The Stop & Glo Dual Luciferase Reporter Assay (Promega) was obtained to be used according to the manufacturer's instructions.

The expression vector constructed by Loughran et al (2018) was selected for the cloning of the regions of interest, procured by Addgene, plasmid: pSGDluc, catalogue No: 119760 and its sequence was validated with Sanger sequencing upon receipt. The plasmid map, as provided by Addgene, was visualised with SnapGene (https://www.snapgene.com/snapgene-viewer/) and is shown in Figure 4.3.

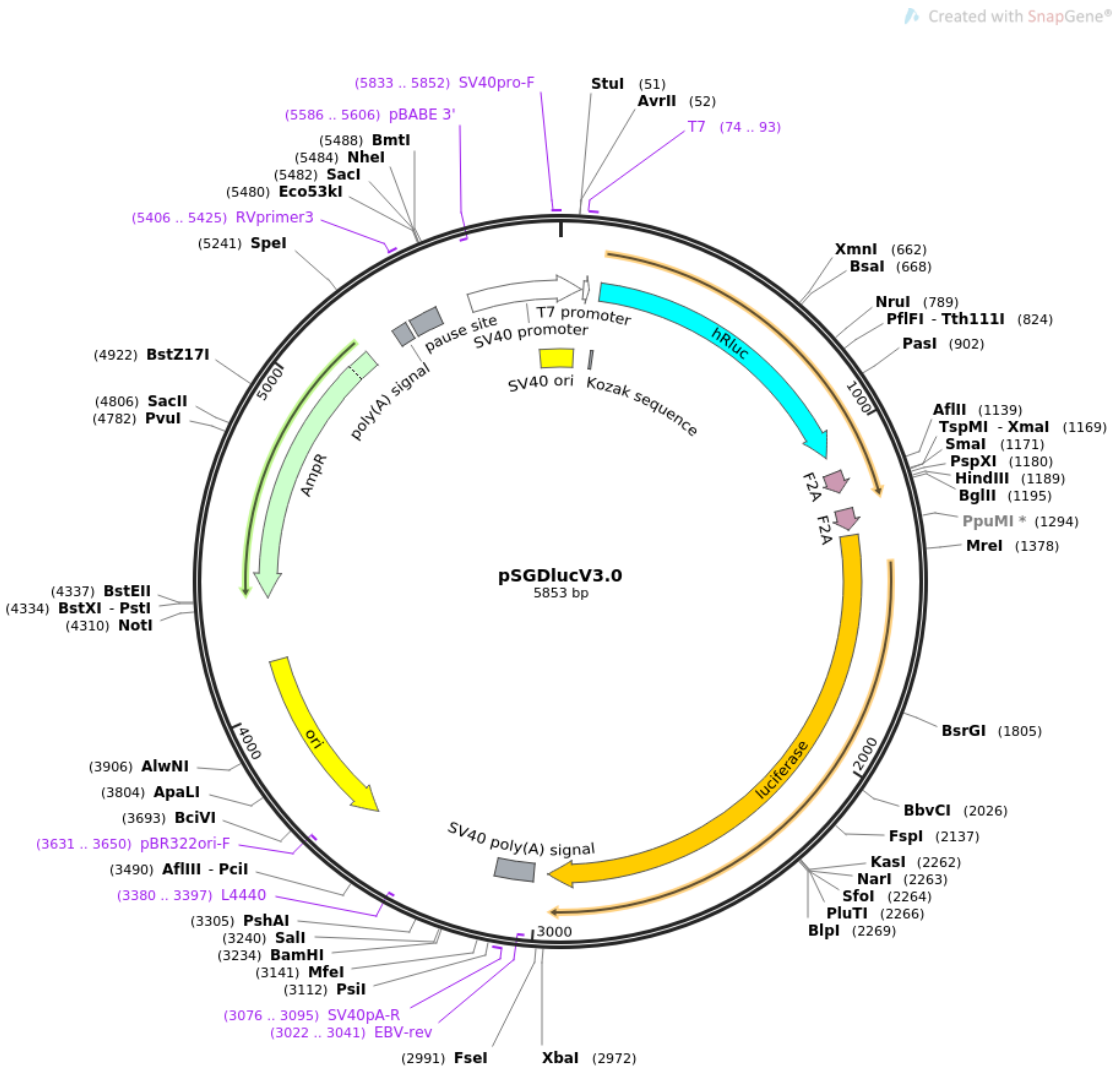


Figure 4.3: Plasmid Map of pSGDluc (cat no: 119760, Addgene).

The restriction enzyme recognition sites are marked on the map, as well as the two luciferase genes and the T7 promoter that is used to initiate their transcription. The sequence of interest is inserted between the F2A regions.

Bacterial transformation was performed in chemically competent DH5α *E. coli* cells (Merck KGaA, Darmstadt, Germany). After thawing the cells on ice, 5 µl of chilled ligation product was added to 50 µl cells, according to the instructions provided by the manufacturer. These were very gently mixed and incubated on ice for 30 min. The cells were then heat-shocked in a water bath at 42 ^{o}C for exactly 45 secs, followed by immediate incubation on ice for 5 min. Subsequently, 200 µl SOC medium (Super Optimal Broth with Catabolite repression medium, Merck) was added to the cells which were incubated in a shaking incubator at 37 ^{o}C, at 220 rpm for 1 h. Cells were subsequently plated and incubated at 37 ^{o}C overnight, or for 16 h.

Luria Broth (LB) agar consisted of an autoclaved mixture of 7.5 g of agar and 10 g of LB broth base in 500 ml sdH_2O with appropriate antibiotic selection (ampicillin, 1:1000) added.

Mini-prep was performed using the QIAprep Spin Miniprep Kit (Qiagen), according to the manufacturer's instructions. In detail, from a single colony, a new liquid culture consisting of 5 ml LB and appropriate selective antibiotic was inoculated and grown overnight at 37 ^{o}C, 220 rpm. The following day, the culture was centrifuged to pellet the cells at 3000g for 10 mins at 4 ^{o}C. The supernatant was discarded and the pellet resuspended in 250 µl buffer P1 and transferred to a microtube. Subsequently 250 µl buffer P2 was added and mixed by inversion to initiate cell lysis. After a maximum of 5 mins, 350 µl buffer N3 was added and mixed immediately by inversion. The sample was centrifuged for 10 min, 13000 rpm or ~ 18000 x g. Following centrifugation, 750 µl of the supernatant was applied to a QIAprep. This was centrifuged for 1 min at 13000 rpm. Washed with 70% ethanol and finally the plasmid was eluted from the column with the elution buffer (50mM Tris HCl pH 8.5) and quantified by Nanodrop™. The extracted plasmids were sequenced by Sanger sequencing on an ABI3130xl sequencer (Applied Biosystems) as described in section 2.2.9.

Restriction endonuclease enzyme digestions were carried out according to the instructions provided by the manufacturer for each enzyme respectively. The reaction for *BglII* was prepared with 10 µl Buffer O (Thermo Fisher Scientific, CA, USA), ~100 ng plasmid pSGDluc, 0.9 µl *BglII* (10 u/µl) (Thermo Fisher Scientific) and sdH_2O up to 100 µl, incubated in a heatblock at 37 ^{o}C for 1 h and inactivated by adding 125 mM EDTA pH 8.0, to a final concentration of 20 mM. The reaction for *HindIII* (Thermo Fisher Scientific) was prepared similarly but with 10 µl Buffer R (Thermo Fisher Scientific) and the enzyme was inactivated by heat inactivation, by incubation at 80 ^{o}C for 20 min. The reaction for *PspXI* (New England Biolabs, Ipswich, MA, USA) was also prepared similarly, with the supplied NEBuffer (New England Biolabs) used instead. *PspXI* cannot be heat inactivated, so the restriction digest product was purified by gel extraction. The result of each digestion was visualised by electrophoresis with an 1% agarose gel, at 50 V for 2 h, along 10 kb ladder and undigested plasmid as a negative control.

Double digestion was set up with the following concentrations, 10 µl Buffer R, 2.5 µl *BglII*, 1.25 µl *HindIII*, ~400 ng plasmid pSGDluc and sdH_2O up to a total volume of 100 µl. The reaction was incubated in a heatblock, at 37 ^{o}C for 1 h, according to the enzyme manufacturer's instructions. Then the reaction was terminated by inactivation of the restriction enzymes and visualised using agarose gel electrophoresis, as described previously. The quantity of ~400 ng of plasmid was determined to be sufficient to make sure that there would be sufficient DNA to be visible on the agarose gel during the electrophoresis.

After each step of inactivating the restriction enzymes the product of the digestion was purified by gel extraction, using the QiaQuick Gel extraction Kit (Qiagen, cat no: 28704), according to the manufacturer's instructions, and was quantified again by nanodrop. Next the

product was treated with Shrimp Alkaline Phosphatase (SAP, USB/Affymetrix, cat no: 70092Y) to remove the phosphate from the digested ends and prevent the plasmid from re-ligating without the insert. After the phosphatase treatment the plasmids were stored in -80 ^{o}C to be used for ligation.

The inserts that were selected to be used for ligation were single stranded oligos, ordered from Sigma-Aldrich in pairs of reverse complement sequences, to be annealed in the lab. The inserts comprised of the genomic sequence containing the PTC and the flanking sequences, the two codons preceding the PTC and the four codons following it, with ends compatible to the restriction recognition site of each enzyme, flanked with an additional adapter sequence at the ends to allow the restriction enzymes to work, as shown on the Table 4.2 below.

Table 4.2: List of primers designed to examine the readthrough potential of the PTC identified in minke whale.

The name of the primer consisted of the name of the gene and the strand. The strand corresponds to S: sense of WT, AS: antisense of WT, UGG-S: sense of positive control, UGG-AS: sense of negative control, Stop-S: sense of negative control, Stop-AS: antisense of negative control. The 5' and 3' adapters consist of the restriction enzyme recognition site, in capital letters, and a short adapter. The inserts contain the PTC, underlined in the sense strands, and the flanking nucleotides of the two codons upstream of the PTC and four codons downstream.

Primer Name	Strand	5' adapter	5'-3' insert	3' adapter
ACP4	S	gtactACTCGAGC	aaggcctagctgtctgggggt	AGATCTgag
	AS	ctcAGATCT	acccccagacagctaggcctt	GCTCGAGTagtac
	UGG-S	gtactACTCGAGC	aaggcctggctgtctgggggt	AGATCTgag
	UGG-AS	ctcAGATCT	acccccagacagccaggcctt	GCTCGAGTagtac
	Stop	gtactACTCGAGC	aaggcctagtaatctgggggt	AGATCTgag
	Stop-AS	ctcAGATCT	acccccagattactaggcctt	GCTCGAGTagtac
ENAM-1	S	gtactACTCGAGC	ttttactgaaatcaacaaatt	AGATCTgag
	AS	ctcAGATCT	aatttgttgatttcagtaaaa	GCTCGAGTagtac
	UGG-S	gtactACTCGAGC	ttttactggaatcaacaaatt	AGATCTgag
	UGG-AS	ctcAGATCT	aatttgttgattccagtaaaa	GCTCGAGTagtac
	Stop	gtactACTCGAGC	ttttactgataacaacaaatt	AGATCTgag
	Stop-AS	ctcAGATCT	aatttgttgttatcagtaaaa	GCTCGAGTagtac
ENAM-2	S	gtactACTCGAGC	aataaatgaaactgtaaactg	AGATCTgag
	AS	ctcAGATCT	cagtttacagtttcatttatt	GCTCGAGTagtac
	UGG-S	gtactACTCGAGC	aataaatggaactgtaaactg	AGATCTgag
	UGG-AS	ctcAGATCT	cagtttacagttccatttatt	GCTCGAGTagtac
	Stop	gtactACTCGAGC	aataaatgataatgtaaactg	AGATCTgag
	Stop-AS	ctcAGATCT	cagtttacattatcatttatt	GCTCGAGTagtac
MMP20-1	S	gtactACTCGAGC	ccaggttaacccaaatggaaa	AGATCTgag
	AS	ctcAGATCT	tttccatttgggttaacctgg	GCTCGAGTagtac
	UGG-S	gtactACTCGAGC	ccaggttggcccaaatggaaa	AGATCTgag
	UGG-AS	ctcAGATCT	tttccatttgggccaacctgg	GCTCGAGTagtac
	Stop	gtactACTCGAGC	ccaggttaatagaaatggaaa	AGATCTgag
	Stop-AS	ctcAGATCT	tttccatttctattaacctgg	GCTCGAGTagtac
MMP20-2	S	gtactACTCGAGC	tccagctaagcctttgatgct	AGATCTgag
	AS	ctcAGATCT	agcatcaaaggcttagctgga	GCTCGAGTagtac
	UGG-S	gtactACTCGAGC	tccagctgggcctttgatgct	AGATCTgag
	UGG-AS	ctcAGATCT	agcatcaaaggcccagctgga	GCTCGAGTagtac
	Stop	gtactACTCGAGC	tccagctaatagtttgatgct	AGATCTgag
	Stop-AS	ctcAGATCT	agcatcaaactattagctgga	GCTCGAGTagtac
AQP4	S-WT	gtactACTCGAGC	tcagtatgactagaagatcgc	AGATCTgag
	AS-WT	ctcAGATCT	gcgatcttctagtcatactga	GCTCGAGTagtac
	UGG-S	gtactACTCGAGC	tcagtatggctagaagatcgc	AGATCTgag
	UGG-AS	ctcAGATCT	gcgatcttctagccatactga	GCTCGAGTagtac
	Stop-S	gtactACTCGAGC	tcagtatgataagaagatcgc	AGATCTgag
	Stop-AS	ctcAGATCT	gcgatcttcttatcatactga	GCTCGAGTagtac

The annealing reaction was performed by adding ~2 ug of each complementary strand of oligos in the annealing buffer: 10 mM Tris, pH 7.5-8.0, 50 mM NaCL, 1 mM EDTA, 50 µl reaction volume and incubate at 95 ^{o}C for 5 min and then let on the bench to cool down to room temperature over at least 30 min.

The ligation reaction was set up using the T4 DNA ligase (Promega), with the reaction mix including: 1 µl T4 DNA ligase 10x Buffer, 1 u T4 DNA ligase (3 u/µl), 100 ng vector DNA, variable insert DNA depending on the insert : vector ratio used, ranging from 1:1 ratio up to 5:1 ratio, and nuclease free sdH_2O up to 10 µl. The reaction was incubated at RT for 3 h, or at 4^{o}C overnight, or at 15 ^{o}C for 4 – 18 h. The ligation reaction was terminated by incubating at 65 ^{o}C for 10 min. Constructs were stored at 4 ^{o}C if they were to be used immediately, or at -80 ^{o}C for long term storage. To confirm that the ligation was successful and that the plasmid contained the intended sequence, the constructs were validated with Sanger sequencing as described previously.

The next phase of the protocol continues with the transfection of human HEK293T cells, following the recommendations of Loughran et al (2018), on 24-well plates, incubating the cells to express the luciferase of the expression vector. The luminance of the constructs is quantified with a luminometer. This part of the experiments was prepared but not conducted.

4.3 Results

4.3.1 Observation and confirmation of premature termination codons

The observation from Chapter 3 is that there were 33 PTCs present in four species in the dataset (Table 4.3). A representative example of a MSA that contains PTCs is shown in Figure 4.4, showing both the nucleotide and the peptide alignment of the aardvark sequence of *AMELX*. To determine whether these PTCs were genuine or whether they were the result of sequencing error, an attempt was made to independently confirm the sequences in the species of interest, in collaboration with Prof Palsboll and his group at the University of Groningen, a group with access to an extensive collection of samples from cetacean species and the expertise in studying them.

The presence of the PTCs in the minke whale genome was examined by Dr Martine Berube, University of Groningen, for at least three individuals for each of the four genes (*ACP4, ENAM, MMP20* and *FAM20A*). These experiments were conducted by our collaborators due to time and material constraints and as the project was nearing its completion. The MSAs of the Sanger sequencing of the regions containing these PTCs are shown in Figure 4.5. The presence of the PTCs in *ACP4*, *ENAM* and *MMP20* were confirmed in this way, as expected, the PTC in *FAM20A* however was not, as the primers used produced an amplified PCR product that did not correspond to the expected nucleotide sequence. Specifically, in the case of minke whale's *FAM20A*, during the time needed to perform the Sanger sequencing to validate the PTCs, the minke whale entry of *FAM20A* in GenBank, which we used to obtain the sequence, was updated and the region with the PTCs was amended to be highly similar to the other mammalian species, with the PTCs removed.

As the PTCs in *ACP4*, *ENAM* and *MMP20* were validated in minke whale, the investigation of potential FTR can proceed.

Regarding the PTCs in other species, stool samples from three individuals per species were obtained from Chester Zoo, but the experiments necessary to confirm or reject the PTCs couldn't be performed, due to time constraints caused by delays in the collection and transport of the samples, as well as by the impact of the covid lock-down.

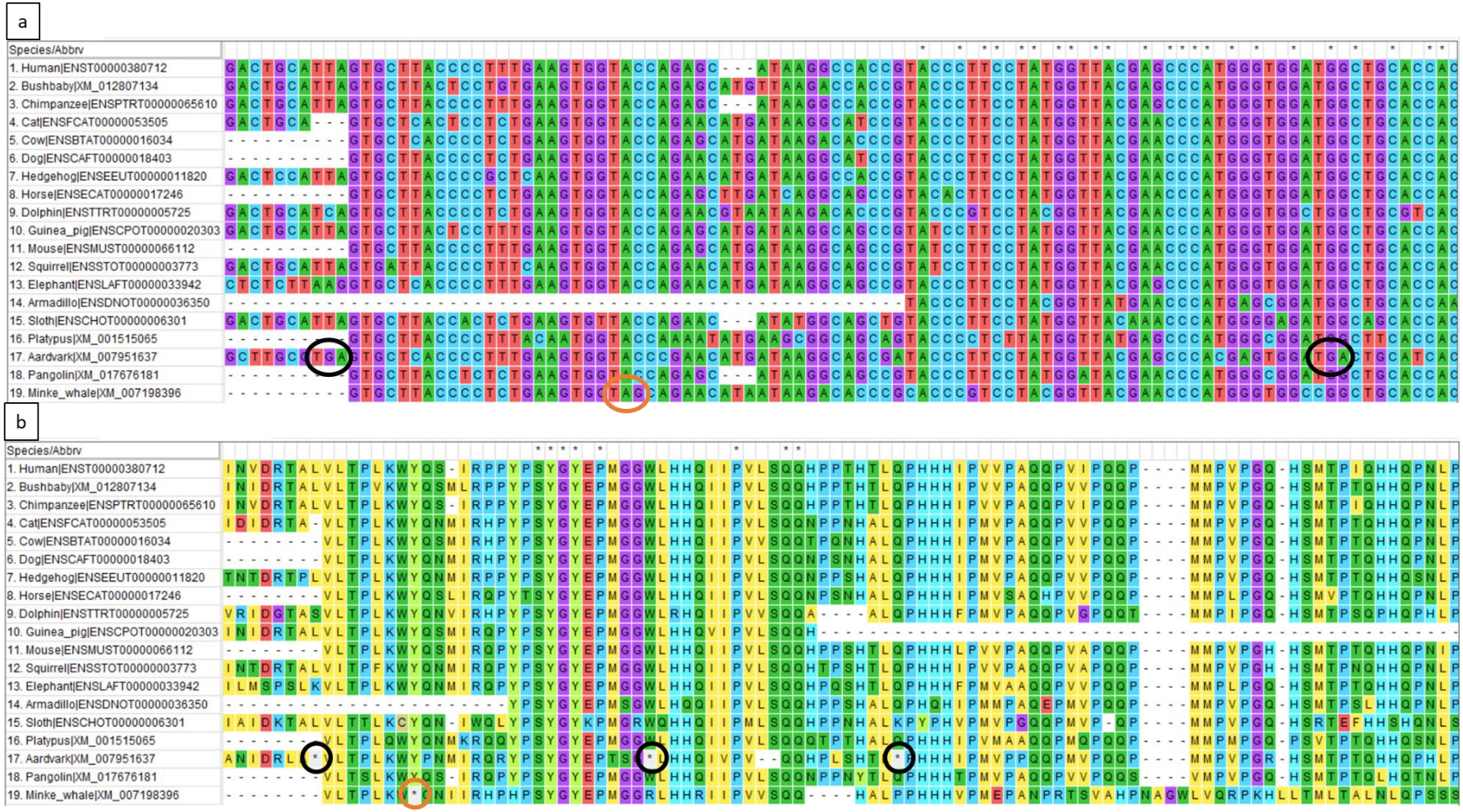


Figure 4.4: MSA of AMELX including the stop codons in the sequences of Aardvark and Minke whale.

By examining the Aardvark sequence, we can see that it is conserved after the PTCs (highlighted with black circles) by a degree comparable to the other toothed species, unlike the Minke whale sequence, which is not conserved after the PTC (highlighted with an orange circle), a divergence clearly visible on the protein MSA.

Table 4.3: List of premature termination codons observed in the MSAs of toothless and enamel-less species, and the sequence surrounding the PTC.

The list is sorted by species, with the position of the PTC in the protein sequence indicated in the 2nd column. The PTCs are shown, along with the nucleotide sequences flanking them, 18 nt either side of the PTC, sorted by gene within each species.

	stop site	18nt before	stop codon	18nt after	
Armadillo	115*	tgcagcacagacttcgag	tag	accctggagcgcgcccag	*ACP4*
	326*	ggtgggcacgtgcccccc	tag	gccgcctgcctgggcttc	
	1658*	cctccatttgaantcctt	taa	aaagggaccaatccacaa	*ENAM*
Aardvark	246*	gcctccctggatgtcccg	tga	actcttactcagctctcg	*ACP4*
	289*	gccaacttctctcgggcc	tag	tgcctgggcctacccctc	
	363*	accatctccctcttctac	tga	aacgactcttctcgcccg	
	150*	tcctctttcctatggata	tga	ccaagagaacatgatgct	*AMBN*
	49*	aatattgacaggcttgcg	tga	gtgctcacccctttgaag	*AMELX*
	76*	tacgagcccacgagtgga	tga	ctgcatcaccaaatcgtc	
	98*	cccctgagccacaccctg	tag	cctcatcaccacatcccc	
	364*	ccctatcaacaaccacca	tag	caagtcccacagagggta	*ENAM*
	276*	acattgtattctcccctt	tga	atcttaagactttggaaa	*MMP20*
	359*	tccaatgtggatgcagct	taa	gaagtggctgagaggggc	
	376*	ttcaaaggtccccagtac	tag	ctaacaaaaggattccaa	
	87*	gcagtacacaagccaccg	tag	tatgacagactgacaaac	*SLC24A4*
Pangolin	113*	ataagacagttgggaagt	taa	cagagattaaacatgctt	*AMBN*
	525*	atgatcccagacattgag	taa	acaatgatgccagcaaac	
	881*	agagatccaactggctgc	tag	agaaactctcaagactat	*ENAM*
	920*	caaagcaaaaactcttat	tag	ccaagaggagattccaga	
	1059*	taccttagcagcaattca	tga	gatgagaaagatgattct	
	1522*	tgtctcaacagtgatctt	tga	ggagacaggaacaatgtt	
	130*	ttcatcataatatccaaa	tag	acatcttccgtgacttct	*MMP20*
	188*	tgtccattcgatgggcct	tga	aggactcaagctcatgca	
	213*	ttcagcaatgctgagaag	tag	actacgggaatgaatggt	
Minke whale	270*	ggggcagcagagaaggcc	tag	ctgtctgggggaatcctg	*ACP4*
	57*	cttacccctctgaagtgg	tag	cagaacataataagacac	*AMELX*
	577*	cgtaatgggccttttac	tga	aatcaacaaattcaaagg	*ENAM*
	892*	gactatggacttaataaa	tga	aactgtaaactgcctcac	
	377*	ctggcgggcaaagaggag	tga	gtgggcccctggaccaag	*FAM20A*
	393*	gcctggctcggggagccc	tga	cgctgcctcagtcctcct	
	429*	ccactgcgtacccagcac	tag	atgctgggggggacagaa	
	115*	tatcgcctcttcccaggt	taa	cccaaatggaaaaaaaat	*MMP20*
	308*	gacctctgtgactccagc	taa	gcctttgatgctgtgaca	

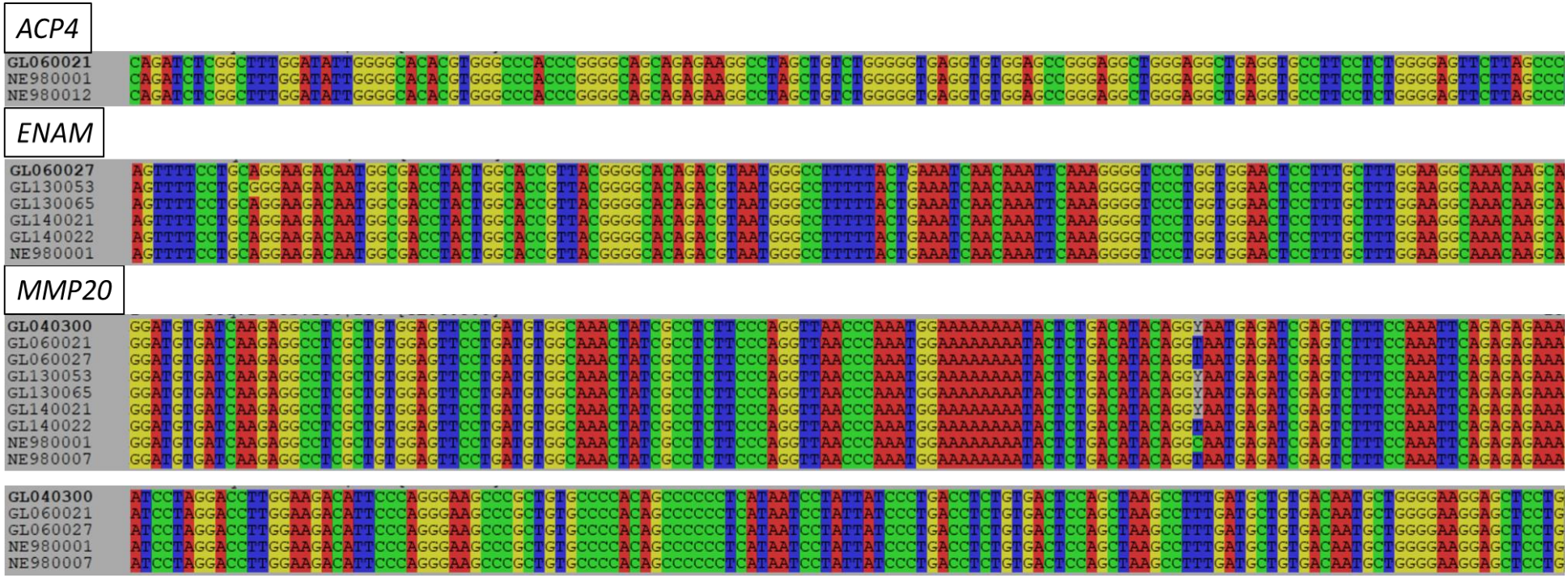


Figure 4.5: MSAs to confirm the presence of the PTCs found in minke whale genes.

At least three individuals were sequenced for each gene, the unique identifiers of the individual given to the left of the sequence. The stop codons, that were validated here, are indicated on the sequences.

4.3.2 Searching for a footprint

As described in section 4.1.2 of this chapter, there are well defined motifs that occur in close proximity to PTCs and facilitate readthrough, i.e. the CUAG motif found after stop codons in human genes (Dabrowski et al., 2015; Loughran et al., 2018). To determine whether there is such a context dependent motif present in the alignments presented in Chapter 3, that may affect or even induce readthrough, the regions flanking the PTC were analysed at the nucleotide level using the WebLogo tool (http://weblogo.berkeley.edu/logo.cgi), to determine if there were any common patterns shared amongst PTCs with the same stop codon (Figure 4.6).

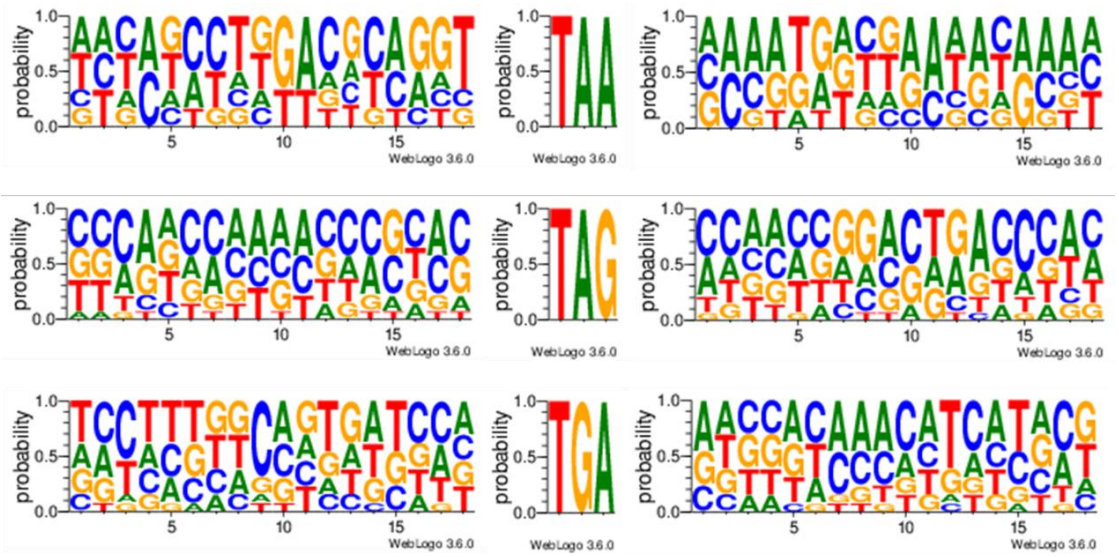


Figure 4.6: WebLogo of the sequence surrounding PTCs having the same stop codon.
The x axis shows the position of the nucleotide on the sequence and the y axis shows the probability that a specific nucleotide will be found in a sequence of the set examined. No motif is clearly discernible in the sequences flanking any of the three stop codons.

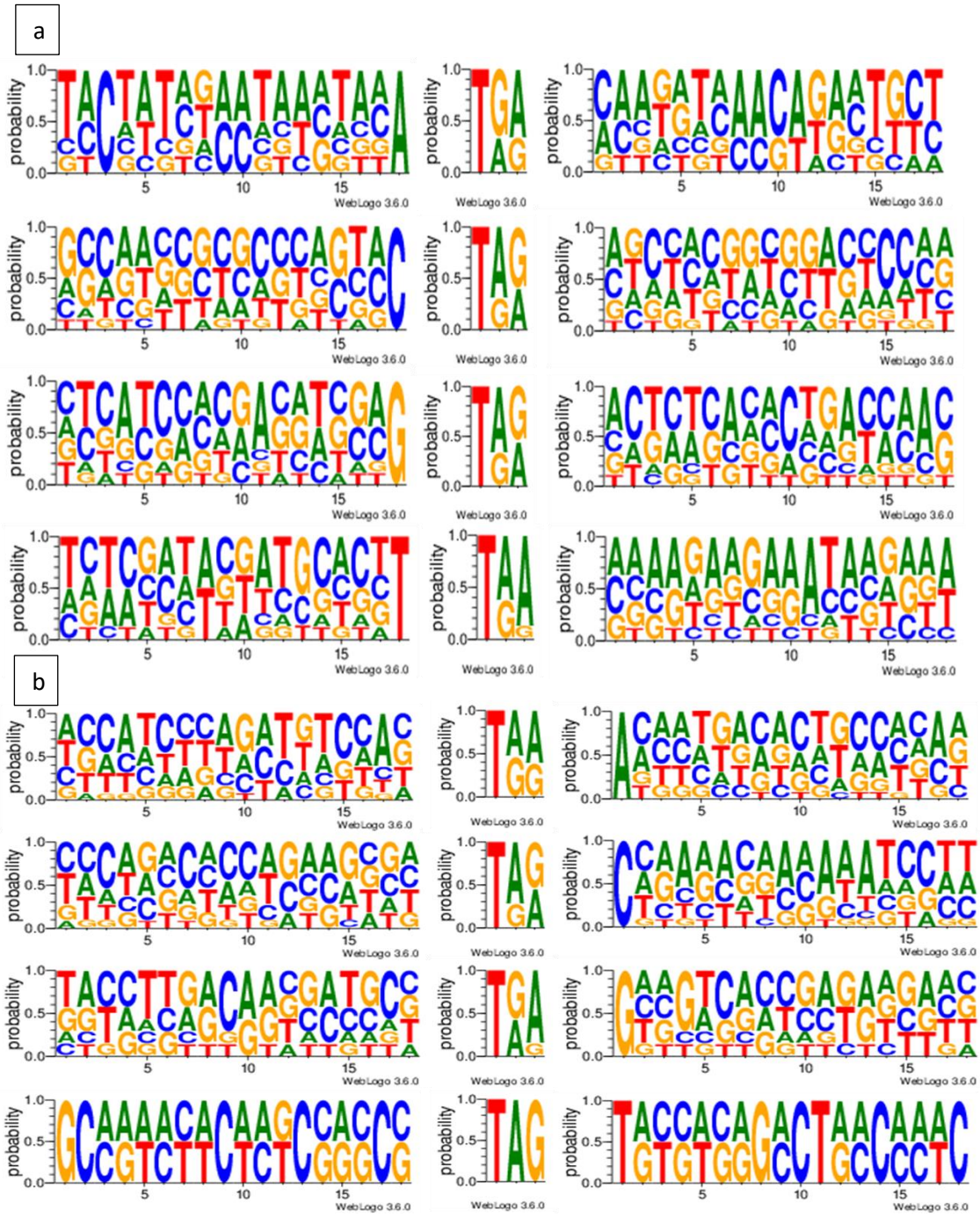


Figure 4.7: WebLogo of the sequence surrounding PTCs, with alternative sorting methods.
(a) The sequences are sorted by the -1 position nucleotide and (b) sorted by the +4 position nucleotide. No patterns are clearly discernible here, as assessed by an optical evaluation.

The weblogos shown are the result of aligning the 18 nt flanking a PTC from either side, irrespective of the gene the PTC was found in or the species, sorted only by the stop codon of the PTC. Alternative sorting methods were examined, such as sorting the sequences by the nucleotide at the -1, with regards to the PTC, position (Figure 4.7a) and sorting by the +4 position (Figure 4.7b). Due to the small sample size there is no statistical significance to our observations. Interestingly, only in Aardvark do we see PTCs followed immediately by a T in the +4 position, but this could also be an artefact of the small sample size. This study will be aided by the dual luciferase expression assays, as they will narrow down the sequences to be used to only the ones with detectable levels of FTR.

4.3.3 Cloning

In order to proceed with the dual luciferase assays, the regions of interest (Table 4.2 and Appendix I) need to be inserted in the chosen expression vector by cloning. The plasmids were successfully cut with the selected restriction enzymes, initially with *HindIII* and *BglII* and then with *BglII* and *PspXI*. The inserts were cloned into the cut plasmid and then used to transform chemically competent *DH5α* cells. The first attempt at transformation, which used the *HindIII* and *BglII* restriction enzymes did not produce any transformed colonies. The second attempt with *BglII* and *PspXI* did, so these colonies were used for all downstream applications. The transformed colonies were grown in liquid culture overnight and then used both for mini-prep plasmid-DNA extraction and to make 25 % glycerol stocks of each colony for long term storage. After DNA extraction the sequences that were inserted in the plasmids were verified with Sanger sequencing, primers used are shown on Table 4.4, to confirm that these were the intended inserts. The Sanger sequencing results are presented in Figure 4.8. Unfortunately, due to time limitations this part of the project stopped at this step, to be concluded at a later time.

Table 4.4: Primers used to validate the pSGDluc plasmid before and after cloning.

Both primer pairs were used to validate the plasmid sequence after receiving it from the manufacturer. Primer p3F was also used to validate the inserts after cloning. 'Distance from restriction site' indicates the distance of the forward primer from the cloning site.

Primer p3	~200 bp from restriction site		
	Sequence (5'->3')	Tm	GC%
Forward primer	CACCGAGTTCGTGAAGGTGA	59.97	55
Reverse primer	CCGCCAGCTTAAGAAGGTCA	60.04	55
Product length	200		
Primer p8	~50 bp from restriction sites		
Forward primer	TGACCTTCTTAAGCTGGCGG	60.04	55
Reverse primer	ATGGTGGCAGATCCGAAAGG	60.11	55
Product length	185		

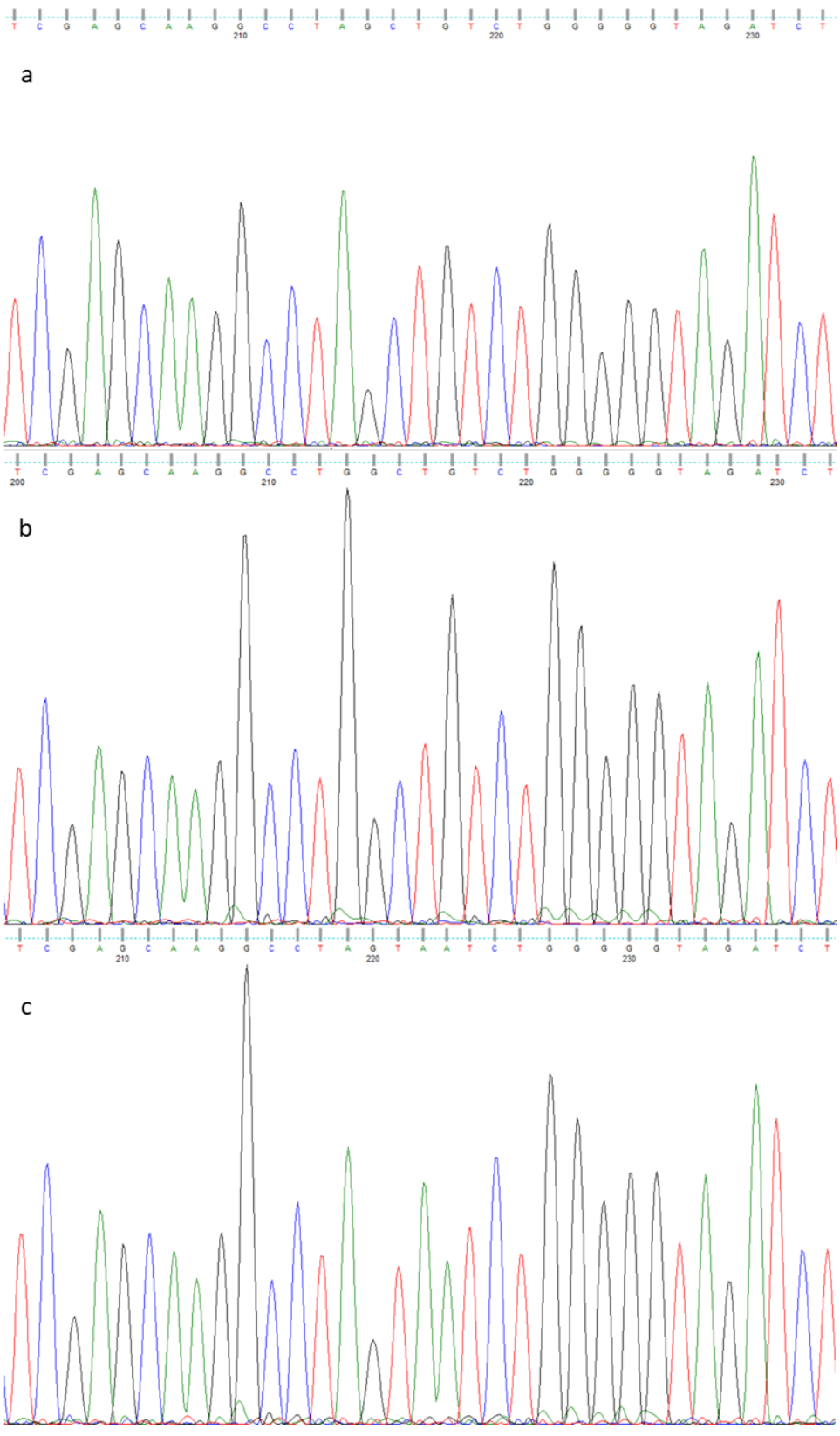


Figure 4.8: Electropherograms validating the cloned sequence for ACP4.
(a) ACP4-WT, (b) ACP4-UGG and (c) ACP4-STOP. The height of the peaks corresponds to the strength of the sequencing signal, but the entire sequence is of high quality as there is no background noise or other artefacts.

4.4 Discussion

4.4.1 Premature termination codons found and the pseudogenization model

The PTCs identified from the MSAs of poorly sequenced species (low coverage numbers for the genome sequence) needed to be confirmed by Sanger sequencing before proceeding. As was shown by the case of *FAM20A* in minke whales annotated sequences from online databases can contain errors, which are not limited to the regions containing the PTCs (Section 4.3.1). This update of the database reinforced the need to confirm the validity of the sequences studied, even for species with high quality genomes. Thanks to the help of Prof Palsboll the PTCs identified in minke whale were validated for three of the four genes (*ACP4*, *ENAM* and *MMP20*) with the *FAM20A* sequence was amended in the meantime, as mentioned earlier. The PTCs in other species could not be confirmed but the genomes available on GenBank and Ensembl were queried regularly to monitor for any changes or updates on the sequences.

The accepted hypothesis in the literature is that genes become inactivated by accumulation of internal stop codons or other mutations that disrupt the reading frame of a coding sequence, via the process of pseudogenisation (Meredith et al., 2011; Gasse et al., 2012; Kawasaki et al., 2014; Springer et al., 2016; Mu, Huang, et al., 2021). The unusually high levels of sequence conservation observed in the genes that should be pseudogenised led to the proposal of an alternative hypothesis, that the pseudogenisation model is not the only mechanism that affects genes that would contribute to a loss of the respective phenotype. As seen in Figure 4.4, the levels of sequence conservation after the PTC were comparable to the sequence conservation in toothed species and indicative of a gene that is under constraints due to producing a functional product. As it is irrefutable that these species lack teeth or enamel, if tooth specific genes produce a functional protein, then it must be used for different purposes than its original function. Proteins undergoing a shift in function have been reported before (Philippe et al., 2003; Abhiman and Sonnhammer, 2005), not however for those encoded by genes involved in tooth formation or amelogenesis. With this alternative hypothesis it is proposed that the identified PTCs are not being used as translation termination codons but are subject to alternative uses that circumvent the termination. The mechanism to achieve this is not yet determined, with the most common processes being translational recoding and functional translational readthrough (FTR).

4.4.2 Why is translational recoding rejected?

Although the mechanism for introducing alternative amino acid residues in peptides is present in mammals it has some unique characteristics that are absent from the genes and sequences identified in this study. Translational recoding is typically driven by downstream mRNA elements (Mariotti et al., 2012; Touat-Hamici et al., 2014) which have not been reported for the genes in this study. The underlying mechanism that drives the incorporation of these alternative residues is only now starting to be described and understood, however it is clear that the proteins affected are from conserved protein families that are common among related species (Doronina and Brown, 2006; Rajput et al., 2019). As Mariotti et al (2012) report, mammalian selenoproteomes are shared and originate from the mammalian ancestral proteome, which would suggest that all species that originate from that ancestor should share the selenoprotein. In a clade as closely related as the mammals, it would follow that evidence of translational recoding would be found in most species and not only in toothless or enamel-

less species as was the case here. Naturally, a shared ancestry of such traits is not always observed and as can be seen from traits that are characterized by loss of function, such as the loss of teeth or enamel, where novel mutations have been shown to be responsible for the loss of the trait. Contrary to these, however, the translational recoding mechanism is not dependent on single nucleotide mutations but requires a multitude of changes to the genomic sequence to arise *de novo* and also be functional. Due to the absence of these characteristics from the identified sequences the hypothesis of translational recoding being responsible for the conservation of the sequence after the PTC was rejected.

4.4.3 Footprints downstream of the stops

To identify patterns in the sequences flanking the PTCs the method chosen was to use weblogo, following the example of many published studies (C. Lee et al., 2008; Döring et al., 2017; Ullah et al., 2018), although the number of sequences that can be examined for the toothless and enamel-less species are far fewer than what is used for the published examples.

The weblogo results did not provide a clearly discernible motif in the sequences before or after the PTCs. This result can be interpreted in many ways, most likely of which is that the small sample size of species and genes examined were not sufficient to differentiate a motif from the random nucleotides surrounding PTCs that are not under FTR. Alternatively, the list of flanking sequences that are included to make the weblogo needs to be modified, as there is a possibility that any motif is species specific and is masked by the inclusion of sequences from other species. There is a limitation on this as well, as the number of PTCs identified per species is also very small which will likely impact our ability to confidently recognise a motif. The results of the dual luciferase expression assays will aid this investigation, as it will help identify specific examples of FTR, allowing the focused research on solely these sequences.

4.4.4 Cloning and dual luciferase assay

Initially each enzyme was examined individually at first to confirm that it can digest the plasmid in a single digestion reaction, before attempting the double digestion. The digested plasmid was used for ligation and the protocol described in Loughran et al (2018) was followed. The first attempt using *HindIII* and *BglII* for the cloning produced no colonies after the bacterial transformation. After re-examining the protocol, a few problems were identified, specifically that adaptor sequences flanking the restriction enzyme recognition sites should be added to the inserts and that the two enzymes selected, *HindIII* and *BglII*, had their restriction recognition sites too close to each other, which prevented them from both binding on the plasmid at the same for the double digestion (Figure 4.3). Overhang adaptor nucleotides were added to the inserts after consulting the literature and the manufacturer of the restriction enzymes (Thermo Fisher Scientific) and the new pair of *BglII* and *PspXI* were selected to allow for some extra space between the enzyme restriction recognition sites. The *AQP4* sequence that was demonstrated to be being readthrough by Loughran et al (2018) was also included in this study as a positive control for FTR. A negative control was also included in the form of a double stop codon insert (Table 4.2), the second stop codon used was TAA as it has been shown to be the least accommodating to readthrough (Manuvakhova et al., 2000). The dual luciferase assay has been used successfully (Grentzmann et al., 1998; McNabb et al., 2005; Loughran et al., 2018; Nair and Baier, 2018) to detect and quantify readthrough and it

is expected that on the occasion of FTR there will be a clear distinction in luminance between the samples and the negative control. This part of the study was not completed due to time constraints, but as the constructs have been created and validated there is the possibility to conclude it at a later time.

Chapter 5 - General Discussion

Amelogenesis imperfecta (AI) is a Mendelian inherited disease that affects the tooth enamel (Smith, 1998; Vogel et al., 2012; Smith, Poulter, et al., 2017). Genetic studies that investigate the cause of the disease at a nucleotide level have long used whole exome sequencing (WES) to find genetic variants causative for the AI phenotype (Poulter, Brookes, et al., 2014; Lu et al., 2018). Evolutionary studies can inform the pathogenic effect of the genes when disrupted (Gasse et al., 2012; Delsuc et al., 2015; Springer et al., 2016; Smith et al., 2016; Huang et al., 2017; Mu, Huang, et al., 2021). The aims of this study were to further investigate the genetic basis of non-syndromic AI, and to assess the selective pressure acting on the genes with variants that cause non-syndromic AI to improve the identification of novel genes and variants that can cause AI.

5.1 Findings of WES and family studies on 33 families with AI

5.1.1 Key Variants identified

The phenotype of AI is dependent on the mutated gene involved, as genes with different functions perturb different stages of amelogenesis (Gadhia et al., 2012; Smith, Poulter, et al., 2017). The variety of underlying genetic causes for AI means that for many affected people the genetic cause for their phenotype has not been identified. This thesis focused on non-syndromic AI and attempted to identify novel variants that cause AI, in any of its described phenotypes. In Chapter 2, people recruited to the Leeds AI cohort presenting with different types of AI were sampled and their exomes were sequenced by WES. Affected members of 33 families were included in this study, with potentially pathogenic variants identified for 30 of the families. As described previously (Section 1.2.2) segregation analysis in the context of genetic studies refers to the detection of a genetic variant in the genome of people affected and unaffected by a disease, with (Møller et al., 2011; Eichler, 2019).

Through segregation analysis 25 of the 30 variants identified segregate with the disease phenotype in the families, an unusually high level of success. There were two variants that did not segregate and were discarded as not causative for AI and three variants that could not be validated with Sanger sequencing, due to high GC content in the genomic regions surrounding each variant making the PCR amplification of the region difficult. Of the twenty genes that have been associated with non-syndromic AI in the literature (Smith, Poulter, et al., 2017; Kim et al., 2019; Smith et al., 2020) candidate causative variants in nine of these genes were identified (Table 2.4 and 2.5), thereby enriching our knowledge on how specific mutations affect amelogenesis and cause an AI phenotype. A selection of these variants were submitted to the online database that documents pathogenic variants causative for non-syndromic AI, LOVD (dna2.leeds.ac.uk/LOVD/genes), following their publication (Nikolopoulos et al., 2020; Nikolopoulos et al., 2021), as the database includes only published data.

Novel and published *MMP20* variants were identified in ten of the families in the Leeds AI cohort, presenting with similar phenotype but originating from various locations in the UK and abroad. Furthermore, three unrelated families recruited from various locations in the UK that shared the same novel *RELT* variant were also investigated for a shared haplotype. As all families included in the microsatellite analysis were self-reported to be unrelated to each

other it is unexpected that there are shared haplotypes among them, as that indicates that all families within one of these groups originated from the same founding population.

These findings helped with the clinical consultation offered to the recruited families, as the genetic basis of their disease was clarified, and more personalized advice could be provided. The identification of novel variants in known AI genes contributed to the enrichment of the mutation spectra of the genes on which the variants were found, while also offering novel insights in the effect that disrupting these genes has on amelogenesis. Specifically for *RELT*, the results presented in this study are only the second such study of *RELT* variants that has been conducted with regards to AI, adding to the published knowledge novel findings such as detailed descriptions of the phenotype and the microstructure of enamel.

The findings of this study of patients with *RELT* mutations and of patients with *MMP20* mutations have been published in two peer reviewed papers (Nikolopoulos et al., 2020; Nikolopoulos et al., 2021).

5.1.2 Founder effect in families and microsatellite analysis of haplotypes

Due to mutational hotspots (Hart et al., 2000; El-Sayed et al., 2010) or shared ancestry (Hart et al., 2004; Kim et al., 2005) variants are often found on the same gene and in the same genomic region among different families. Here, microsatellite analysis was used to determine the presence of a founder effect among families that shared the same pathogenic variants. Evidence of a founder effect was identified for two of the three groups with *MMP20* mutations and also for the group of families sharing a *RELT* mutation. This discovery helps with the genetic consultation that can be provided to other members of these populations, as it shows that there is an elevated frequency of the variants within the respective populations, and thus more care should be taken by the members of these populations to avoid intermarriage.

5.1.3 Structural abnormalities in enamel caused by AI

In rare cases exfoliated teeth or teeth that had to be extracted for clinical reasons were available to examine. Published studies of enamel in AI affected teeth have reported abnormal enamel structure, as the enamel can have reduced mineral content or reduced thickness, both conditions visible with a micro-computerised X-ray tomographer (µCT) or have enamel with abnormal morphology, which has lost the typical ordered structure of normal enamel, a change that can be seen with a scanning electron microscope (SEM) (El-Sayed et al., 2011; Poulter, Murillo, et al., 2014; Smith, Kirkham, et al., 2017).

Teeth were made available for study from two families for which mutations had been identified in *RELT*. The microstructure of the enamel of these teeth was examined both by µCT and SEM, to examine the mineral density, thickness, and enamel morphology of the teeth. The enamel density and thickness that was observed is within the normal levels (Figure 2.14), which is consistent with the published literature for hypoplastic AI (Collins et al., 1999; Poulter, Murillo, et al., 2014; Lu et al., 2018). An abnormal enamel structure was, however, observed with the SEM, as instead of the enamel rods of normal enamel, disordered layers of deposited enamel were discovered (Figures 2.16 and 2.17). This lamellar structure has not been reported in any other published study, but it is similar to the structures reported by Smith et al (2016), on teeth of patients affected by hypomineralised AI caused by mutations in *AMTN*. The better

characterization of the abnormal enamel microstructure improves our understanding of the disease phenotype and will present new approaches to treating the clinical features of the specific microstructural features of AI that characterise different phenotypes.

5.1.4 Molecular dynamics simulations of MMP20

Molecular dynamics simulations of the MMP20 active site revealed that the pathogenic mutations identified in this study greatly reduce the stability of the tertiary structure of the protein. AMBER (Maier et al., 2015; Case et al., 2018) was used to conduct the molecular dynamics simulations. This methodology had not been used to model the effects of mutations in AI genes on their encoded proteins' structures before this study, but has been used previously to give insight into genome research (Biagini et al., 2018; Tam et al., 2020) and also to predict the pathogenic effect of missense mutations (Poli et al., 2022) and to further investigate the effect of known pathogenic variants to the function of a protein (Parveen et al., 2019).

The limitation of the molecular dynamics simulations is that it requires an experimentally observed tertiary structure of a protein to conduct the simulation, which limits the options of AI associated genes as the structure of the protein for the majority of them has yet to be determined. The AlphaFold system attempts to remedy this limitation by computationally predicting the folding of proteins with the use of neural network-based model and a machine learning approach (Jumper et al., 2021). The structures predicted with AlphaFold are deposited on an online database (AlphaFold DB, https://alphafold.com/) and are freely available (Varadi et al., 2022). However, despite the accuracy demonstrated by AlphaFold in predicting tertiary structures, to precisely calculate the minute effects that single residue changes will have on the stability of a protein a high resolution experimentally observed structure is necessary to be used as the starting point of the molecular dynamics simulations.

The findings presented here, of the greatly reduced stability of the mutated protein need to be examined further in the context of the effect of pathogenic and non-pathogenic variants on the stability of other proteins of genes involved in amelogenesis, to establish whether this result is an indication of the pathogenicity of the variant or a change that should be expected from any variant, regardless of its pathogenicity. This method offers an alternative approach to predictions of the pathogenicity of candidate variants and although it is limited by the need for an experimentally observed protein structure, the advancements in technology and the further study of the other AI associated genes will render these simulations critical in future studies.

5.2 Selective pressure analysis of genes associated with AI

Chapter 3 focused on comparing the evolution of the genes that have been associated with an AI phenotype when disrupted, across a range of 19 mammals including those that have independently lost tooth and / or enamel to determine if there is a shared pattern that can be identified in the evolution of these genes in the mammal clade. Some housekeeping genes were also included as controls: *RHO* and *TUBA4A*, to contrast the levels of natural selection acting on genes outside of the group of interest. Following the methodology described in published studies, for the evolutionary analysis of genes and their pathogenic variants, the

selective pressure acting on the protein coding sequences of these genes was examined to identify shared patterns of natural selection that would help to identify other genes that co-evolved with this set and could affect amelogenesis when disrupted (Meredith et al., 2011; Kawasaki et al., 2014; Delsuc et al., 2015; Smith et al., 2016; Huang et al., 2017; Webb et al., 2017; Mu, Huang, et al., 2021). A pruned version of the most up-to-date phylogeny of the mammal clade was used (Morgan et al., 2013; Tarver et al., 2016), and all genes associated with non-syndromic AI were included in the analysis. Genes that are cooperating in the formation of teeth and enamel could be expected to have co-evolved under the same selective pressures. Codeml, part of the PAML package (Yang, 2007) and SLAC, part of the HyPhy package (Kosakovsky Pond et al., 2020), were selected to conduct the selective pressure analysis, considering amongst site variation and amongst branch variation in selective pressure. The results of both codeml and SLAC for the site-specific models indicate that regions of the majority of the genes are under purifying selection. Sites that were identified as under positive selection do not overlap between codeml and SLAC. Others have reported similar discrepancies between SLAC and codeml output (Kulmuni et al., 2013; Huang et al., 2017). The difference is due to SLAC calculating the ancestral sequence from an input MSA and then using a combination of the ML and counting methods to calculate the dN and dS substitution rates per site of the sequence, while assuming a constant selective pressure across the sequence (Kosakovsky Pond and Frost, 2005). On the contrary codeml doesn't calculate the ancestral sequence, it is not using a counting method and also tests the sequence for both a constant selective pressure and a free ω-ratio model and the analysis is conducted on the best fitting model (Yang, 2007). Each of the methods has its advantages, as discussed previously, and neither can be considered a better option, although codeml is more commonly used in the literature.

Surprisingly, the branch-specific models within codeml showed that genes whose main function is on the formation of teeth or enamel show signatures of species-specific positive selection in toothless and enamel-less species. These genes in toothless and enamel-less species are thought to have been pseudogenised (i.e. internal stop codons rendering the gene non-functional), but the alternative possibility that they are undergoing adaptation to a new function or are still producing a functional peptide despite the internal stop codon should not be dismissed. Other pseudogenes have previously been reported as under positive selection, e.g.: the human olfactory receptor pseudogenes (Gilad et al., 2000) and the mRNAs of the *DSTNP2* and *NAP1L4P1* human pseudogenes (Tan et al., 2021) showing that "pseudogenization" is not synonymous with loss of function. The observed signature of positive selection may be present due to adaptation to a new niche (McGowen et al., 2020) or simply in undergoing rapid evolution (Kulmuni et al., 2013). Traditionally, sites that are indispensable for the function of the protein are under strong purifying selection, which is also confirmed in this study, but sites found to be under positive selection are contributing to the adaptive evolution of the genes (Anisimova et al., 2001; Sanville et al., 2010; Jovanovic et al., 2021), indicating that the positive selection found in pseudogenes might also be a result of adaptive evolution and a possible shift in the function of the pseudogenised gene (Webb et al., 2015; Hyland et al., 2021). The FTR of internal stop codons would also be consistent with this hypothesis, of a gene that is no longer used for amelogenesis or tooth formation altering its function to adapt to different needs for the adaptation of the species, offering a novel, alternative, approach to determining whether internal stop codons should be interpreted as precursors of pseudogenisation in evolution.

5.3 Investigating the potential for stop codon readthrough

In Chapter 4 the internal stop codons in AI associated genes in toothless and / or enamel-less mammals were examined further, to determine if any potential stop codon readthrough occurs. The regions after the stop codons were highly conserved across toothless and enamel-less mammals and all other mammals with teeth and enamel suggesting that the region was not undergoing loss of function and pseudogenization but rather it was visible to selection and was being maintained (Figure 4.4). There were two possibilities for the observed high levels of conservation in the alignments following the stop codons: (1) the regions are retaining their functional peptides by reading through the internal stop codon, or (2) the simpler explanation that it is a sequencing error, produced by the low quality of the genome and a lack of coverage depth.

It is proposed that, in the toothless or enamel-less species, some of the genes that are considered as pseudogenes (Meredith et al., 2011; Kawasaki et al., 2014; Springer et al., 2016; Mu, Huang, et al., 2021) are not following the pseudogenization model (Section 1.4.7), but are still producing a functional peptide via functional translational readthrough (FTR), facilitated by stop codon readthrough. The stop codon readthrough phenomenon has been reported as common in unicellular organisms as it helps with increasing the amount of information contained in a smaller size of genome (Pelham, 1978), but it has also been reported in insects (Jungreis et al., 2011) and mammals (Loughran et al., 2014), including in human (Loughran et al., 2018). The experimental method to confirm the occurrence of stop codon readthrough is by using a dual luciferase assay, in which the internal stop codon and its flanking regions are inserted in an appropriate expression vector (Grentzmann et al., 1998; Loughran et al., 2018). This method has been shown by Loughran *et al* to be effective in detecting readthrough in human genes, with the distinction that Loughran *et al* examined the readthrough capabilities of terminal stop codons in the genes they examined, whereas this study suggests that a similar mechanism is potentially used to read through the internal stop codons found in genes.

Contextual cues have been described to induce FTR in the literature (Williams et al., 2004; Jungreis et al., 2011; Loughran et al., 2014), which lead to the examination of the internal stop codons identified in this study for any nucleotide patterns in the sequences flanking the stop codon. Although there was no specific nucleotide pattern identified in the sequences flanking the internal stop codons (Figure 4.6 and Figure 4.7) this cannot be interpreted as a conclusive result, as the sample size of internal stop codons examined here is too small. Previous studies on FTR have reported that the AUG is the stop codon that facilitates readthrough most (Manuvakhova et al., 2000), however all three stop codons were found in the internal stop codons identified here (Table 4.3), with no clear preference of AUG over the other two. In future studies the internal stop codons with AUG can potentially be prioritised in examining their readthrough potential.

To determine if the internal stop codons were genuine or an artefact of a low-quality genome, samples were obtained for toothless and enamel-less species , i.e. minke whale and zoo samples from X, Y, and Z species (Section 4.2.1), with the purpose of sequencing the regions where internal stop codons were identified). The “zoo samples” were not processed due to time constraints, but the internal stop codons in minke whales were confirmed by Sanger sequencing and were a suitable candidate to examine the possibility of stop codon readthrough. The continuation of this project should ideally be using developing tooth RNA sequencing if possible, or construct-based luciferase assays if not, in future studies. The Dual luciferase assays planned in this thesis were not completed due to constraints imposed by

Covid-19 and supply chains and then time. However, they are currently being completed by collaborators.

5.4 Future prospects

While the number of families analysed for this project is not insignificant, the genetic basis of AI for many families remains unknown. Future studies involving a larger number of recruited families, with more family members sampled and utilising more advanced methodologies could identify the variants causing AI in a shorter time frame and with a higher level of confidence. The UK Biobank (https://www.ukbiobank.ac.uk/) can also be searched for variation in the genes associated with AI to expand the information on the frequency of variants. WES is a powerful technique that can be used to investigate the genetic basis of a disease in a high-throughput manner, but it also has its limitations. WES sequences only the exonic regions, along with a short flanking region for each exon, meaning that it cannot find intronic variants and that its efficiency in identifying splicing or structural variants is limited. Whole genome sequencing (WGS) could resolve most of these limitations, as it is a more powerful method with less of the genome being excluded from the analysis (Belkadi et al., 2015). WGS has been previously used for the identification of pathogenic variants (Nitayavardhana et al., 2020), it is however significantly more expensive, making it more uncommon to be selected over WES, especially when many samples need to be sequenced. An alternative is the smMIPs methodology, which can be used for targeted next-generation sequencing of genes of interest and can be scaled to a high-throughput level of efficiency (Pérez Millán et al., 2018; Bekers et al., 2019; Almomani et al., 2020). As smMIPs can be implemented to provide a faster and more cost-efficient way to screen genes for known and candidate pathogenic variants, it will allow researchers to focus the WES specifically on families that do not have easily identifiable variants in known genes, removing these individuals from the pool of unsolved families. Other than the segregation analysis, the identification of novel variants and new genes that are associated with AI are basing the mutation - phenotype association on the level of conservation of the specific amino acid residue in other, closely related to human, species (Meredith et al., 2011; Kawasaki et al., 2014; Delsuc et al., 2015; Huang et al., 2017; Kim et al., 2019; Mu, Huang, et al., 2021). This methodology needs to be re-examined and re-evaluated, as disease mutations in one species are not necessarily disease mutations in the same ortholog of another species, as the divergence of the sequences of the two orthologs can lead to adaptations that neutralise any possible pathogenic effect of a mutation in one species by introducing other balancing mutations. In essence variants that can be pathogenic in humans could be benign in other species, after being neutralised by adaptive evolution of the sequence. Thus, this comparison of the sequences among different species can provide an indication on the importance of a site for the function of the protein, but should not be taken as evidence that if a site is not conserved in other species then a mutation is tolerated in human, or that if it is conserved it is necessarily pathogenic when altered.

Regarding the evolutionary analysis, the main limitations of this study were the limited number of high-quality genomes available for toothless and enamel-less mammals that limited the number of species that could be examined. This limitation in the number of species with a high-quality genome available will be resolved by the progress made by the Earth Biogenome Project (EBP) which aims to sequence, catalogue and characterise earth's eukaryotic biodiversity (Lewin et al., 2018). Since the start of the project in 2018 almost 9400

species' genomes are planned to have been sequenced by 2023, followed by a long-term goal of sequencing ~1.7 million eukaryotic species by 2030 (Gupta, 2022). Regarding the species of this study, platypus is one of the species that ideally should have been included in the study but had many of the genes of interest missing from the online databases that has since been sequenced by the EBP (Zhou et al., 2021). As many of the genes investigated here are thought to be pseudogenised in Platypus, it is a prime example to research the hypothesis proposed here, that many of the genes thought to be pseudogenised might be under FTR and still produce a functional peptide. As mentioned earlier the dual luciferase assays were not completed, which would provide an answer to whether the internal stop codons are being readthrough. This part of the project, along with the confirmation of the internal stop codons by Sanger sequencing from the zoo samples remains to be concluded at a later time. On a broader scale, with hundreds of genomes soon to be available from the EBP and other projects any follow up studies will have a significantly more robust and diverse dataset of species to examine, identifying numerous genes that would be considered pseudogenes under the current model but would be potential targets for FTR. This should provide the necessary depth of information needed to prove or disprove the hypothesis on the functionality of genes currently considered as pseudogenised, that was proposed here.

Appendix A

A1. List of primers used for segregation analysis

Primer Name	Forward primer	Reverse primer	Expected size
AMBN del789	TGGCACCATCAGATAAGCCA	ATTGTTTATTTTGTGGCATTGGTC	1692
AMBN_del789_2	TCCCATCACAGCCATCCTTG	ACAGGCACATCCCCATAACA	1774
AMBN_ex2	GAGCAGAGACTCAGGCTCAT	TCTATGCGTGGACTCTCAGTC	258
AMBN_ex5	GCGTGCCTGTAGACCCAG	TGTTAATGCTAGGACTTGGCTG	445
AMBN_ex6	TCCTAGCCTCCCTTCCAGAT	GGTTTCAGTCCTGGCTGTTG	193
AMELX ex3	TGAACAATTGCATACTGACTTAATCTC	CATCTGGGATAAAGAATCAACACA	242
AMELX ex5	TGAAGAATGTGTGTGATGGATG	TCCCATTAATGTCTGCATGTG	367
AMELX ex6	GGCAAAGAAAACACTGCTGC	GCACCATTTTCACAGGATTTG	552
AMELX_ex5	GGCAAAGAAAACACTGCTGC	GCACCATTTTCACAGGATTTG	552
COL17A1 ex25	ACAGGAGACACACATGCAGA	CAGCCCTCACCTGGAAGAC	299
COL17A1 ex50	TTTCTTGGACCCCACTTCTG	GCACTTAGACTGCCCTTGCT	498
COL17A1_ex45	CAAAGACAGGAGGGACCCAT	ACTCACAAATGTCACGCAGG	214
COL17A1_ex49+50	TTTCTTGGACCCCACTTCTG	GCACTTAGACTGCCCTTGCT	498
COL17A1_ex54	CTCTCTGCCACCACTTACCT	CTGCTCCAGAGACACCAGTC	293
COL17A1_in34	AGTTCCTGCCCTGGGTTT	GCAGCGATGAGAAGAAG	291
COL17A1_in41	GCAGCTCATAGGTCCTAGCA	AGGAGGTGAGGCAAACTTCT	236
EMILIN1_ex4	CGTCTTGAGGGTGTCTGTGA	TGCAGGGAGCTGTTAAGGG	209
ENAM ex3	TTGACAGACAAGTAGGCTAG	TTTTTACCTTTTTCATAAGAAATA	437
ENAM ex6	TGGCTGAGTTTTAGAGGCTGA	AGTGTGTATATGGGGGTGGC	433
ENAM_ex3_4	GCTAGTACTTAGATAAGTGCAGAGTGC	GACAATTTTCCAATATTCTCCTTTT	388
ENAM_ex9	CCTCGGTGGAACTTCTTTGC	GGGGCTGGTTGGATTCTTTG	297
FAM20A ex5	AGGAAGAAATGCCAGGGAGT	TCTCCTCCCACTTTGCTTGT	391
FAM20A ex5c	TTGAGGGTCTGTCTAGCCAC	CCATCGCCTGTGTGTACATG	299
FAM20A ex6	ACTGATGAATGGAGGGGTGG	GAACAGAGTTGGGATGGTGG	460
FAM20C_ex6	CCTACTACTGCTCCACGGAG	ATCTCACACAGGCCAGCAT	293
FAM83H ex5	GGTGAAGTCATCCGGGTCC	AGACGTTCCTCAGCCACG	176
FAM83H ex5	GGAAGTGGCTGGTCTGGAA	ACTTCCTGTCGGCCTTCC	300
FAM83H ex5K1	ACTTCCTGTCGGCCTTCC	GTAGGAGGCCAAACGCC	713
FAM83H_ex5	GAATGGCCTTCGTGTCATCC	AAGGCTCCTTGCGTCTTAGG	240
FAM83H_ex5_pt1	GGAAGGCCGACAGGAAGT	AGGAGTTCCGCATCCTCTT	249
FAM83H_ex5_pt2	GGTGAAGTCATCCGGGTCC	AGACGTTCCTCAGCCACG	176
ITGB4_ex10	CTCCTGCAGGCTCTGTGATA	AAACGATTCCACTCGGTGTG	296
LAMA3 ex12	GAAACCTAGGGGAAGGGAGA	ACAGCTCTGCCAGAAAGATCT	300
LAMA3 ex35	CAGAACTCCCAGTGGCAATG	ACCTGAACTGCTGGATGCTT	457
LAMB3 ex18	GGTAGGTCTTGGCATAGAGGA	CTGGCTGATGCACTGAACAT	323
LAMB3 ex9	AGAGGGTCAGAGGGCAGTG	AGAGCTTGAATTGTAATGGTGCA	351
LAMB3_ex18	TGGCTGATGCACTGAACATG	GATTCAATCCAGTGCCCAGC	195
LAMC2_ex9	TGTGACCCTGATTTCAGCCC	TCTCTTCCACATGACACCCC	251

MIA3_ex25	TGAACCCTGATCTGTCCACA	CCCACTTCCTGCATGGTTTT	193
MINK1_ex20	TCAGTGACCTTCTTCCACCC	CACAGCCCTTCTCGGTAGAT	229
MMP20 exon 3	TTCAGTACCGGATTATCCCAA	GCGAAGGAGGAGTGTGTGAT	474
MMP20 exon 4	CTGTAATATGATGCGCCCCT	AGTTAAAGGGTGGCTTGGGA	298
MMP20 exon 5	GGGTTAACTGTAATGTGGGCAT	TGCACTTTCTTTAATTCGGAGA	338
MMP20 exon 7	GCAAGAGCAAAGGGCATTTA	ATGACTGGAAAAATGCTGGC	350
MMP20 exon 8	CATTTAAGAGCCTTCATAGAAATCTT	TTCTTTCGTGGAAGGGTTTA	326
PLXNA2_ex22	GCTTGGGGTGGTTCTTGTTC	GCCCAGCTCATCAACAACAA	208
RELT_ex11	AAGGAGAAAGGCATCTGTTGG	CTTCTCGGTCCTCACAGTCC	297
RELT_ex4	CGACCTGGTGAGCATTGC	GCCAGCAGTCTCCACAGA	299
RELT_ex9	AGGCCTGTGTCCAAGTGAG	TGGTCGGGTTAGGCAGAAG	367
SLC26A4_ex10	GCGTCCAAACTCCTGATGTC	TTTCCTCCAGTGCTCTCCTG	287
SNX32_ex5	GCGATGCACGAAGTCTTTCT	GAGCTGTGACACAGGATTGG	215
WDR72_ex15	CTGCCAACTCTACATGCCAA	TCTCACTGATGTTGCCAGGT	265

A.2 Primers used for the microsatellite analysis

Gene	Microsatellites	Seq_F	Seq_R	Position		Product size
MMP20	D11S1339	ATGGCCTTGGAAAAATATC	GGGTGTAACCAGTTCTTCAG	102057334	102057646	122
	D11S4108	TGGCAAGTGGCAGGAT	GCCCATAGATGGATGAGTAGA	102516167	102516411	113
	D11S4951	ATGGGTATACACCCAGCAAA	AACTGTGATTTTAAAAGATAATGCC	104624561	104624826	130
	D11S940	TCATCCCCAAATGCTCAG	GGAATCAAACTTCACATAGGAGG	101473242	101473550	183
	D11S4159	CCGGAGAGCAGTTTGTGT	ATTCGGAGCCACTCCCT	104128850	104129173	180
	D11S1394	CGCCAACAGAGAAACAAGAG	AAATACACTTTTCAGGCCCC	104333311	104333613	246
	D11S898	AGCACCATTTGCTGAGACTG	TGTATTTGTATCGATTAACCAACTT	101056444	101056752	149
RELT	D11S1314	TTGCTACGCACTCCTCTACT	GTGAAGGCAGGAAATGTGAC	72323144	72323418	209-227
	D11S4184	CCCAGCCTTACATATTCC	GCTGATGAGCAGAGGTAG	72670826	72671153	263-277
	D11S916	CAGACTATTCTCATTGCTGC	GGACTTCTAAGCCTCCATAA	73329800	73330060	135-153
	D11S2371	CTGAGGTGGGAGGTTCAGTT	CCCGGCCTTGATTTATTTAA	73505073	73505374	193-213

Appendix B

B.1 Commands for WES analysis, from fastq to file.vep.vcf

Adjust directories appropriately

1: Aligning and calling samples to produce a gvcf.

Trim the adaptors and do QC:

#See https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md

trim_galore -q 20 --fastqc_args "--outdir /data/bs16gn/path/to/outdir" --illumina --gzip -o /data/bs16gn/path/to/outdir --length 20 --paired /data/bs16gn/path/to/file/sample_R1_001.fastq.gz /data/bs16gn/path/to/file/sample_R2_001.fastq.gz

Align the sample to the human genome:

bwa mem -t 12 -M /home/ref/b37/human_g1k_v37.fasta /pathto/val_files_from_previous_step_R1.gz /pathto/val_files_from_previous_step_R2.gz -v 1 -R '@RG\tID:Add_sample_ID\tSM:Add_sample_ID\tPL:Illumina\tPU:HiSeq3000\tLB:$Samplename_exome\' -M | samtools view -Sb - > sample_bwa.bam

Next sort the alignment (alter picard version to the latest one - note which you use):

java -Xmx4g -jar /home/picard/picard-tools-2.5.0/picard.jar SortSam I=sample_bwa.bam O=sample_bwa.sort.bam SO=coordinate CREATE_INDEX=TRUE

remove original bam to save space:

rm -i /path/sample_bwa.bam

Mark PCR duplicates:

java -Xmx4g -jar /home/picard/picard-tools-2.5.0/picard.jar MarkDuplicates I=sample_bwa.sort.bam O=sample_bwa.sort.dedup.bam M=sample_bwa.sort.metrics CREATE_INDEX=TRUE

Delete pre-deduplicated bam to save space:

rm -i sample_bwa.sort.bam

Create indel realigner targets:

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T RealignerTargetCreator -R /home/ref/b37/human_g1k_v37.fasta -known /home/ref/b37/1000G_phase1.indels.b37.vcf -known /home/ref/b37/Mills_and_1000G_gold_standard.indels.b37.sites.vcf -I sample_bwa.sort.dedup.bam -o sample_bwa.sort.dedup.intervals

Perform indel realignment:

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T IndelRealigner -R /home/ref/b37/human_g1k_v37.fasta -known /home/ref/b37/1000G_phase1.indels.b37.vcf -known /home/ref/b37/Mills_and_1000G_gold_standard.indels.b37.sites.vcf -I sample_bwa.sort.dedup.bam -targetIntervals sample_bwa.sort.dedup.intervals -o sample_bwa.sort.dedup.indelrealn.bam

Delete pre-indelrealn bam and gzip interval file to save space:

rm -i sample_bwa.sort.dedup.bam

gzip sample_bwa.sort.dedup.intervals

Perform base quality recalibration

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T BaseRecalibrator -R /home/ref/b37/human_g1k_v37.fasta -knownSites /home/ref/b37/1000G_phase1.indels.b37.vcf -knownSites /home/ref/b37/Mills_and_1000G_gold_standard.indels.b37.sites.vcf -knownSites /home/ref/b37/dbSnp146.b37.vcf.gz -I sample_bwa.sort.dedup.indelrealn.bam -o sample_bwa.sort.dedup.indelrealn.recal.grp -nct 6

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T PrintReads -R /home/ref/b37/human_g1k_v37.fasta -I sample_bwa.sort.dedup.indelrealn.bam -BQSR sample_bwa.sort.dedup.indelrealn.recal.grp -o sample_bwa.sort.dedup.indelrealn.recal.bam

Delete old bam (the non-recal file)

rm -i sample_bwa.sort.dedup.indelrealn.bam

Generate g.vcf file using Haplotype Caller

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T HaplotypeCaller --emitRefConfidence GVCF --variant_index_type LINEAR --variant_index_parameter 128000 -R /home/ref/b37/human_g1k_v37.fasta -D /home/ref/b37/dbSnp146.b37.vcf.gz -stand_call_conf 30 -stand_emit_conf 10 -I sample_bwa.sort.dedup.indelrealn.recal.bam -o sample.g.vcf

Store unused gvcfs, bams and fastqs as .gz files to save space. Delete unneeded files and transfer files from server regularly as server is not backed up.

2: Convert the raw.g.vcf to a raw.vcf: (do for single samples for autozygosity)

java -Xmx8g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T GenotypeGVCFs -R /home/ref/b37/human_g1k_v37.fasta -D /home/ref/b37/dbSnp146.b37.vcf.gz -stand_call_conf 30 -stand_emit_conf 10 -V /data/mededm/SSF_family_results/SSF_family.combined.raw.g.vcf -o /data/mededm/SSF_family_results/SSF_family.combined.raw.vcf -nda --showFullBamList -nt 8

3: Merge g.vcf for filtering:

#The gvcf can be transferred to the server for filtering using (on local terminal):

rsync –trv "/path/to/file" –e ‘ssh –p 4222’ bs16gn@limm-pc4145.leeds.ac.uk:/data/bs16gn/unsolved

#After making the gVCF it would be good to combine them with others, you can use merge gVCF for that:

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T GenotypeGVCFs -R /home/ref/b37/human_g1k_v37.fasta -V /nobackup/bgycels/george/4880.g.vcf -V /nobackup/bgycels/george/5079.g.vcf -V /nobackup/bgycels/george/5167.g.vcf -V /nobackup/bgycels/george/4481.g.vcf -V /nobackup/bgycels/george/4483.g.vcf -o /data/bs16gn/unsolved/combinedAI164_4481_4483.noBED.vcf

Recalibrate indels and SNPs separately from vcf file

Split vcf into indels and snps and recalibrate each separately:

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T SelectVariants -R /home/ref/b37/human_g1k_v37.fasta --variant /data/bs16gn/unsolved/combinedAI63_3630.noBED.vcf -selectType SNP -o /data/bs16gn/unsolved/combinedAI63_3630.noBED.raw-snps.vcf

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T SelectVariants -R /home/ref/b37/human_g1k_v37.fasta --variant /data/bs16gn/unsolved/combinedAI63_3630.noBED.vcf -selectType INDEL -selectType MNP -o /data/bs16gn/unsolved/combinedAI63_3630.noBED.raw-indels.vcf

Perform hard filtering:

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T VariantFiltration -R /home/ref/b37/human_g1k_v37.fasta -V /data/bs16gn/unsolved/combinedAI63_3630.noBED.raw-snps.vcf --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MappingQualityRankSum < -12.5" --filterName "snp_hard_filter" -o /data/bs16gn/unsolved/combinedAI63_3630.noBED.fltd-snps.vcf

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T VariantFiltration -R /home/ref/b37/human_g1k_v37.fasta -V /data/bs16gn/unsolved/combinedAI63_3630.noBED.raw-indels.vcf --filterExpression "QD < 2.0 || FS > 200.0" --filterName "indel_hard_filter" -o /data/bs16gn/unsolved/combinedAI63_3630.noBED.fltd-indels.vcf

Combine files:

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T CombineVariants -R /home/ref/b37/human_g1k_v37.fasta --variant /data/bs16gn/unsolved/combinedAI63_3630.noBED.fltd-snps.vcf --variant /data/bs16gn/unsolved/combinedAI63_3630.noBED.fltd-indels.vcf -o /data/bs16gn/unsolved/AI63_3630.noBED.fltd-combined.vcf --genotypemergeoption UNSORTED

#Then do appropriate gnomAD / ExAC filter, VEP and filter based on mode of inheritance, annotate etc.

Filter with ExAC or gnomAD (1% freq)

perl /home/vcfhacks-v0.2.0/filterVcfOnVcf.pl -i /data/bs16gn/unsolved/AI63_3630.noBED.fltd-combined.vcf -f /home/ref/ExAC/gnomad.exomes.r2.0.1.sites.vcf.gz -o /data/bs16gn/unsolved/AI63_3630.noBED.fltd-combined.gnomAD.vcf -y 0.01 -w

VEP annotation:

/home/variant_effect_predictor/variant_effect_predictor.pl;

perl /home/variant_effect_predictor/variant_effect_predictor.pl --offline --vcf --dir_cache /home/variant_effect_predictor/vep_cache --dir_plugins /home/variant_effect_predictor/vep_cache/Plugins --everything --plugin SpliceConsensus --fasta /home/variant_effect_predictor/fasta/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz -i /data/bs16gn/unsolved/AI63_3630.noBED.fltd-combined.gnomAD.vcf -o /data/bs16gn/unsolved/AI63_3630.noBED.fltd-combined.gnomAD.vep.vcf

B.2 Script for family filtering

```
# Select Biallelic and X linked variants
perl /home/vcfhacks-v0.2.0/findBiallelic.pl \
-i /data/bs16gn/path/to/file.fltd-combined_gnomad.vep.vcf \
--x_linked 2 \
-s sample_ID \
--consensus_splice_site \
-n 1 \
-o /data/bs16gn/path/to/file.fltd-combined_AR_gnomad.vep.hom && \
# -z if looking for common variants in multiple samples
#--check_all_samples if looking at all of them
# Rank on CADD Score -d gives unsorted list
perl /home/vcfhacks-v0.2.0/rankOnCaddScore.pl \
-c /data/shared/cadd/v1.3/*.gz \
-i /data/bs16gn/path/to/file.fltd-combined_AR_gnomad.vep.hom \
-o /data/bs16gn/path/to/file.fltd-combined_AR_gnomad.vep.hom.cadd1.3 \
-n /data/bs16gn/path/to/file.fltd-combined_AR_gnomad.vep.hom.cadd1.3_NOTFOUND.tsv \
--progress -d && \
# GeneAnnotator_vcfhacks
perl /home/vcfhacks-v0.2.0/geneAnnotator.pl \
-d /home/vcfhacks-v0.2.0/data/geneAnnotatorDb \
--i /data/bs16gn/path/to/file.fltd-combined_AR_gnomad.vep.hom.cadd1.3 \
-o /data/bs16gn/path/to/file.fltd-combined_AR_gnomad.vep.hom.cadd1.3.geneanno && \
# AnnovcfToSimple_vcfhacks_xlsx with -f gives only the functional variants
perl /home/vcfhacks-v0.2.0/annovcfToSimple.pl \
-i /data/bs16gn/path/to/file.fltd-combined_AR_gnomad.vep.hom.cadd1.3.geneanno \
--vep --gene_anno --functional \
-o /data/bs16gn/path/to/file.fltd-combined_AR_gnomad.vep.hom.cadd1.3.geneanno.simple.xlsx
&& \
# AnnovcfToSimple_vcfhacks_xlsx with -f gives only the functional variants canonical_only
perl /home/vcfhacks-v0.2.0/annovcfToSimple.pl \
```

```
-i /data/bs16gn/path/to/file.fltd-combined_AR_gnomad.vep.hom.cadd1.3.geneanno \

--vep --gene_anno --canonical_only --functional \

-o /data/bs16gn/path/to/file.fltd-
combined_AR_gnomad.vep.hom.cadd1.3.geneanno.simple.canonicalOnly.xlsx && \

############################

# Select dominant variants#

############################

perl /home/vcfhacks-v0.2.0/getFunctionalVariants.pl \

-i /data/bs16gn/path/to/file.fltd-combined_gnomad.vep.vcf \

--consensus_splice_site \

-s sample_ID \

-o /data/bs16gn/path/to/file.fltd-combined_AD_gnomad.vep.hom && \

# Rank on CADD Score -d gives unsorted list

perl /home/vcfhacks-v0.2.0/rankOnCaddScore.pl \

-c /data/shared/cadd/v1.3/*.gz \

-i /data/bs16gn/path/to/file.fltd-combined_AD_gnomad.vep.hom \

-o /data/bs16gn/path/to/file.fltd-combined_AD_gnomad.vep.hom.cadd1.3 \

-n /data/bs16gn/path/to/file.fltd-combined_AD_gnomad.vep.hom.cadd1.3_NOTFOUND.tsv \

--progress -d && \

# GeneAnnotator_vcfhacks

perl /home/vcfhacks-v0.2.0/geneAnnotator.pl \

-d /home/vcfhacks-v0.2.0/data/geneAnnotatorDb \

--i /data/bs16gn/path/to/file.fltd-combined_AD_gnomad.vep.hom.cadd1.3 \

-o /data/bs16gn/path/to/file.fltd-combined_AD_gnomad.vep.hom.cadd1.3.geneanno && \

# AnnovcfToSimple_vcfhacks_xlsx with only the functional variants

perl /home/vcfhacks-v0.2.0/annovcfToSimple.pl \

-i /data/bs16gn/path/to/file.fltd-combined_AD_gnomad.vep.hom.cadd1.3.geneanno \

--vep --gene_anno --functional \

-o /data/bs16gn/path/to/file.fltd-combined_AD_gnomad.vep.hom.cadd1.3.geneanno.simple.xlsx
&& \
```

Appendix C

C.1 Access IDs from GenBank and Ensembl for the sequence included in this study.

Date of last access December 2019.

	Human	Chimpanzee	Bushbaby	Squirrel	Mouse	Guinea Pig
	Homo sapiens	*Pan troglodytes*	*Otolemur garnettii*	*Ictidomys tridecemlineatus*	*Mus musculus*	*Cavia porcellus*
ACP4	ENST00000270593	ENSPTRT00000021123	ENSOGAT00000004269	ENSSTOT00000010097	ENSMUST00000118216	ENSCPOT00000027235
AMBN	ENST00000322937	ENSPTRT00000045270	ENSOGAT00000035024	ENSSTOT00000007955	ENSMUST00000198265	ENSCPOT00000001122
AMELX	ENST00000380712	ENSPTRT00000065610	XM_012807134	ENSSTOT00000003773	ENSMUST00000066112	ENSCPOT00000020303
AMTN	ENST00000339336	ENSPTRT00000030049	ENSOGAT00000003035	ENSSTOT00000022909	ENSMUST00000073363	ENSCPOT00000001120
COL17A1	ENST00000353479	ENSPTRT00000005563	ENSOGAT00000017134	ENSSTOT00000013371	ENSMUST00000026045	ENSCPOT00000001005
DLX3	ENST00000434704	ENSPTRT00000017217	ENSOGAT00000005139	ENSSTOT00000010308	ENSMUST00000092768	ENSCPOT00000025370
ENAM	ENST00000396073	ENSPTRT00000091821	ENSOGAT00000003041	XM_005333299	ENSMUST00000031222	XM_003467537
FAM20A	ENST00000592554	ENSPTRT00000017563	ENSOGAT00000015626	ENSSTOT00000023494	ENSMUST00000020938	ENSCPOT00000010541
FAM83H	ENST00000388913	ENSPTRT00000059775	ENSOGAT00000027713	ENSSTOT00000023921	ENSMUST00000170153	ENSCPOT00000013453
GPR68	ENST00000531499	ENSPTRT00000089266	ENSOGAT00000013869	ENSSTOT00000019539	ENSMUST00000110066	ENSCPOT00000031707
ITGB6	ENST00000283249	ENSPTRT00000107705	ENSOGAT00000002899	ENSSTOT00000008227	ENSMUST00000059888	ENSCPOT00000024774
KLK4	ENST00000324041	ENSPTRT00000021136	ENSOGAT00000031952	ENSSTOT00000002320	ENSMUST00000007161	ENSCPOT00000034808
LAMA3	ENST00000313654	ENSPTRT00000018220	ENSOGAT00000014541	ENSSTOT00000006758	ENSMUST00000092070	ENSCPOT00000007693
LAMB3	ENST00000391911	ENSPTRT00000047179	ENSOGAT00000026439	ENSSTOT00000020073	ENSMUST00000194677	ENSCPOT00000004250
MMP20	ENST00000260228	ENSPTRT00000007863	ENSOGAT00000015251	ENSSTOT00000027785	ENSMUST00000034487	ENSCPOT00000033641
ODAPH	ENST00000435974	ENSPTRT00000073664	XM_003790132	XM_005333292	ENSMUST00000178614	XM_013143348
RELT	ENST00000064780	ENSPTRT00000077610	ENSOGAT00000000749	ENSSTOT00000014494	ENSMUST00000008462	ENSCPOT00000027277
RHO	ENST00000296271	ENSPTRT00000028711	ENSOGAT00000010262	ENSSTOT00000025056	ENSMUST00000032471	ENSCPOT00000005038
SLC10A7	ENST00000507030	ENSPTRT00000096714	ENSOGAT00000008694	ENSSTOT00000028234	ENSMUST00000034111	ENSCPOT00000033387
SLC24A4	ENST00000532405	XM_024348873	ENSOGAT00000008612	XM_005339140	ENSMUST00000079020	ENSCPOT00000010402
SP6	ENST00000536300	ENSPTRT00000017152	ENSOGAT00000028661	ENSSTOT00000001603	ENSMUST00000107622	ENSCPOT00000007711
TUBA4A	ENST00000248437	ENSPTRT00000023989	ENSOGAT00000006005	ENSSTOT00000025464	ENSMUST00000186213	ENSCPOT00000004653
WDR72	ENST00000396328	ENSPTRT00000049347	ENSOGAT00000009396	ENSSTOT00000015385	ENSMUST00000055879	ENSCPOT00000011241

	Horse	Minke whale	Dolphin	Cow	Cat	Dog
	Equus caballus	*Balaenoptera acutorostrata*	*Tursiops truncatus*	*Bos taurus*	*Felis catus*	*Canis lupus familiaris*
ACP4	ENSECAT00000021666	XM_007179955	ENSTTRT00000011060	ENSBTAT00000020110	ENSFCAT00000019264	ENSCAFT00000004670
AMBN	ENSECAT00000023127	XM_007193563	ENSTTRT00000008662	ENSBTAT00000053975	ENSFCAT00000009209	ENSCAFT00000004648
AMELX	ENSECAT00000017246	XM_007198396	ENSTTRT00000005725	ENSBTAT00000016034	ENSFCAT00000053505	ENSCAFT00000018403
AMTN	ENSECAT00000000460	XM_007193595	ENSTTRT00000008650	ENSBTAT00000003804	ENSFCAT00000007316	ENSCAFT00000004629
COL17A1	ENSECAT00000008760	XM_007172098	ENSTTRT00000007475	ENSBTAT00000013271	ENSFCAT00000004301	ENSCAFT00000045026
DLX3	ENSECAT00000015627	XM_007176422	-	ENSBTAT00000023142	ENSFCAT00000066707	ENSCAFT00000026875
ENAM	XM_001487894	XM_007193564	XM_019932671	ENSBTAT00000013661	ENSFCAT00000028994	ENSCAFT00000004659
FAM20A	ENSECAT00000025399	XM_007198504	ENSTTRT00000012600	ENSBTAT00000001439	ENSFCAT00000011004	ENSCAFT00000017481
FAM83H	XM_023649125	XM_007166933	ENSTTRT00000015190	ENSBTAT00000019225	ENSFCAT00000029161	ENSCAFT00000049368
GPR68	ENSECAT00000004904	XM_007180867	-	ENSBTAT00000008858	XM_023255960	ENSCAFT00000027822
ITGB6	ENSECAT00000024051	XM_007183159	ENSTTRT00000011042	ENSBTAT00000011972	ENSFCAT00000032085	ENSCAFT00000015738
KLK4	ENSECAT00000019908	XM_007198394	ENSTTRT00000005736	ENSBTAT00000027207	ENSFCAT00000008263	ENSCAFT00000004649
LAMA3	XM_023647569	XM_007197638	ENSTTRT00000008885	ENSBTAT00000060991	ENSFCAT00000009603	ENSCAFT00000028869
LAMB3	XM_023640784	XM_007163957	ENSTTRT00000017004	ENSBTAT00000022005	XM_019821882	ENSCAFT00000018923
MMP20	ENSECAT00000012298	XM_007191222	ENSTTRT00000014736	ENSBTAT00000018936	ENSFCAT00000015792	ENSCAFT00000023926
ODAPH	XM_001490823	-	XM_019932723	XM_002688369	XM_019829356	ENSCAFT00000013233
RELT	ENSECAT00000007605	XM_007173703	ENSTTRT00000004627	ENSBTAT00000021927	ENSFCAT00000039029	ENSCAFT00000009030
RHO	XM_023619934	XM_007192608	ENSTTRT00000011272	ENSBTAT00000001730	ENSFCAT00000000092	ENSCAFT00000007461
SLC10A7	ENSECAT00000029039	XM_007189381	ENSTTRT00000004653	ENSBTAP00000045867	ENSFCAT00000003048	ENSCAFT00000012383
SLC24A4	ENSECAT00000014694	XM_007184262	XM_019918776	ENSBTAT00000008703	ENSFCAT00000007889	ENSCAFT00000017418
SP6	ENSECAT00000013302	XM_007183680	ENSTTRT00000012473	ENSBTAT00000002386	ENSFCAT00000009365	ENSCAFT00000026614
TUBA4A	ENSECAT00000015274	XM_007187970	XM_019938908	ENSBTAT00000003192	ENSFCAT00000040005	ENSCAFT00000024148
WDR72	ENSECAT00000007531	XM_007194904	ENSTTRT00000003537	ENSBTAT00000061231	ENSFCAT00000011845	ENSCAFT00000045029

	Hedgehog	Pangolin	Aardvark	Elephant	Sloth	Armadillo	Platypus
	Erinaceus europaeus	*Manis javanica*	*Orycteropus afer*	*Loxodonta africana*	*Choloepus hoffmanni*	*Dasypus novemcinctus*	*Ornithorhynchus anatinus*
ACP4	XM_007531289	-	XM_007957636	ENSLAFT00000001738	-	XM_023585546	ENSOANT00000002669
AMBN	ENSEEUT00000011199	XM_017655633	XM_007942818	XM_010594187	ENSCHOT00000011631	ENSDNOT00000010867	ENSOANT00000021593
AMELX	ENSEEUT00000011820	XM_017676181	XM_007951637	ENSLAFT00000033942	ENSCHOT00000006301	ENSDNOT00000036350	XM_001515065
AMTN	ENSEEUT00000000206	-	XM_007942817	ENSLAFT00000012608	ENSCHOT00000000414	XM_023584403	-
COL17A1	XM_007530465	XM_017660486	XM_007939867	ENSLAFT00000028366	-	ENSDNOT00000000450	ENSOANT00000002604
DLX3	XM_007525618	XM_017642416	XM_007941926	ENSLAFT00000036461	ENSCHOT00000002771	ENSDNOT00000019409	-
ENAM	XM_007523530	XM_017655632	XM_007942773	ENSLAFT00000026415	-	XM_004477547	XM_007669232
FAM20A	ENSEEUT00000007308	XM_017641924	XM_007959693	ENSLAFT00000000376	ENSCHOT00000013721	ENSDNOT00000017961	ENSOANT00000012832
FAM83H	ENSEEUT00000005203	XM_017646837	XM_007956087	ENSLAFT00000036487	-	XM_004454793	ENSOANT00000000665
GPR68	XM_007518419	XM_017664233	XM_007944490	ENSLAFT00000014724	-	ENSDNOT00000004198	ENSOANT00000006046
ITGB6	ENSEEUT00000013388	XM_017649894	XM_007941832	ENSLAFT00000001853	ENSCHOT00000013138	ENSDNOT00000019549	ENSOANT00000009855
KLK4	ENSEEUT00000013116	-	-	ENSLAFT00000029708	-	-	-
LAMA3	XM_007536377	-	XM_007935580	ENSLAFT00000008286	ENSCHOT00000009404	ENSDNOT00000003516	XM_016227434
LAMB3	XM_016191698	XM_017641500	XM_007953064	ENSLAFT00000005222	ENSCHOT00000006958	ENSDNOT00000004444	XM_007670564
MMP20	XM_007520728	XM_017676338	XM_007946011	ENSLAFT00000010998	ENSCHOT00000004213	ENSDNOT00000006356	-
ODAPH	XM_007519845	XM_017662619	XM_007949023	XM_003414148	-	XM_004472565	-
RELT	XM_016190165	XM_017646709	XM_007941042	ENSLAFT00000037366	ENSCHOT00000013732	ENSDNOT00000042824	-
RHO	XM_007517079	XM_017647349	XM_007956743	ENSLAFT00000001359	-	ENSDNOT00000015131	ENSOANT00000005993
SLC10A7	XM_007528206	XM_017658860	XM_007944113	ENSLAFT00000000501	ENSCHOT00000000114	ENSDNOT00000033700	ENSOANT00000010989
SLC24A4	XM_007529080	XM_017679460	XM_007944500	ENSLAFT00000009800	ENSCHOT00000004634	ENSDNOT00000007808	ENSOANT00000006073
SP6	XM_007530005	XM_017659478	XM_007941991	ENSLAFT00000010700	-	ENSDNOT00000048475	-
TUBA4A	XM_007519711	XM_017657566	XM_007958885	XM_010601880	ENSCHOT00000008202	ENSDNOT00000014135	XM_016227186
WDR72	ENSEEUT00000005322	XM_017669930	XM_007935300	XM_023539510	ENSCHOT00000004979	ENSDNOT00000002162	ENSOANT00000020580

Appendix D

D.1 Example of the parameter file for the codeml analysis, for the M1a model

```
     seqfile = AMELX_MSA.fa            * sequence data file name
    treefile = AMELX_tree.nwk          * tree structure file name

     outfile = AMELX_M1a_mlc           * main result file name
       noisy = 3   * 0,1,2,3,9: how much rubbish on the screen
     verbose = 0   * 1: detailed output, 0: concise output
     runmode = 0   * 0: user tree;  1: semi-automatic;  2: automatic
                   * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

     seqtype = 1   * 1:codons; 2:AAs; 3:codons-->AAs
   CodonFreq = 2   * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
       clock = 0   * 0: no clock, unrooted tree, 1: clock, rooted tree
      aaDist = 0   * 0:equal, +:geometric; -:linear, {1-5:G1974,Miyata,c,p,v}
       model = 0

     NSsites = 1   * 0:one w; 1:NearlyNeutral; 2:PositiveSelection; 3:discrete;
                   * 4:freqs; 5:gamma; 6:2gamma; 7:beta; 8:beta&w; 9:beta&gamma

       icode = 0   * 0:standard genetic code; 1:mammalian mt; 2-10:see below
       Mgene = 0   * 0:rates, 1:separate; 2:pi, 3:kappa, 4:all

   fix_kappa = 0   * 1: kappa fixed, 0: kappa to be estimated
       kappa = .3  * initial or fixed kappa

   fix_omega = 0   * 1: omega or omega_1 fixed, 0: estimate
       omega = 1.3 * initial or fixed omega, for codons or codon-based AAs
       ncatG = 10  * # of categories in the dG or AdG models of rates

        getSE = 0   * 0: don't want them, 1: want S.E.s of estimates
RateAncestor = 0   * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

  Small_Diff = .45e-6
   cleandata = 0  * remove sites with ambiguity data (1:yes, 0:no)
 fix_blength = 0  * 0: ignore, -1: random, 1: initial, 2: fixed
```

D.2 Summary of the results of the codeml analysis and statistical significance

>**ACP4**
Model 0: one-ratio, lnL(ntime: 71 np: 73): -14644.849637 +0.000000
kappa (ts/tv) = 3.56874, omega (dN/dS) = 0.14587
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -14433.386837 +0.000000
kappa (ts/tv) = 3.70251, p: 0.82141 0.17859, w: 0.08619 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -14433.386837 +0.000000
kappa (ts/tv) = 3.70251, p: 0.82141 0.14830 0.03029, w: 0.08619 1.00000 1.00000

LRT_1-0: 2x (-14433.386837 - -14644.849637)= 2x 211.4628= 422.9256
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-14433.386837 - -14433.386837)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -14318.841180 +0.000000
kappa (ts/tv) = 3.67134, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00012 0.00270 0.01144 0.02977 0.06128 0.11022 0.18230 0.28640 0.43967 0.69282
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -14317.016044 +0.000000
kappa (ts/tv) = 3.68182, p: 0.09725 0.09725 0.09725 0.09725 0.09725 0.09725 0.09725 0.09725 0.09725 0.09725 0.02753, w: 0.00021 0.00341 0.01252 0.02976 0.05748 0.09869 0.15788 0.24292 0.37077 0.59990 1.14824
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -14317.162636 +0.000000
kappa (ts/tv) = 3.68014, p: 0.09628 0.09628 0.09628 0.09628 0.09628 0.09628 0.09628 0.09628 0.09628 0.09628 0.03720, w: 0.00023 0.00352 0.01258 0.02936 0.05598 0.09523 0.15136 0.23195 0.35370 0.57578 1.00000

LRT_8-7: 2x (-14317.016044 - -14318.841180)= 2x 1.825136= 3.650272
p-value (df=2): 0.161202
LRT)_8-8a: 2x (-14317.016044 - -14317.162636)= 2x 0.146592= 0.293184
p-value (df=1): 0.588241

>**AMELX**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -4621.434889 +0.000000
kappa (ts/tv) = 2.50333, omega (dN/dS) = 0.60929
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -4464.785210 +0.000000
kappa (ts/tv) = 2.08481, p: 0.63928 0.36072, w: 0.07220 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -4360.423647 +0.000000
kappa (ts/tv) = 2.54522, p: 0.60781 0.28267 0.10952, w: 0.08957 1.00000 9.51903

LRT_1-0: 2x (-4464.785210 - -4621.434889)= 2x 156.649679= 313.299358
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-4360.423647 - -4464.785210)= 2x 104.361563= 208.723126
p-value (df=2): < 0.00001 ***

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -4478.611472 +0.000000

kappa (ts/tv) = 2.12128, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00000 0.00006 0.00215 0.02346 0.13022 0.41624 0.77159 0.95652 0.99714 0.99999
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -4365.876607 +0.000000
kappa (ts/tv) = 2.57015, p: 0.08878 0.08878 0.08878 0.08878 0.08878 0.08878 0.08878 0.08878 0.08878 0.11215, w: 0.00000 0.00043 0.00619 0.03527 0.12407 0.31012 0.57731 0.82621 0.96391 0.99897 9.49647
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -4464.880252 +0.000000
kappa (ts/tv) = 2.08396, p: 0.06407 0.06407 0.06407 0.06407 0.06407 0.06407 0.06407 0.06407 0.06407 0.35932, w: 0.03732 0.04817 0.05552 0.06189 0.06798 0.07423 0.08103 0.08904 0.09969 0.11917 1.00000

LRT_8-7: 2x (-4365.876607 - -4478.611472)= 2x 112.734865= 225.46973
p-value (df=2): < 0.00001 ***
LRT)_8-8a: 2x (-4365.876607 - -4464.880252)= 2x 99.003645= 198.00729
p-value (df=1): < 0.00001 ***

>**AMTN**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -8614.630256 +0.000000
kappa (ts/tv) = 3.04036, omega (dN/dS) = 0.50863
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -8541.331367 +0.000000
kappa (ts/tv) = 3.04395, p: 0.63701 0.36299, w: 0.28140 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -8537.229000 +0.000000
kappa (ts/tv) = 3.08587, p: 0.62957 0.36078 0.00964, w: 0.28490 1.00000 3.09553

LRT_1-0: 2x (-8541.331367 - -8614.630256)= 2x 73.298889= 146.597778
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-8537.229000 - -8541.331367)= 2x 4.102367= 8.204734
p-value (df=2): 0.016534 *

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -8544.423033 +0.000000
kappa (ts/tv) = 3.01456, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.06337 0.16987 0.26945 0.36589 0.46066 0.55466 0.64864 0.74338 0.84012 0.94177
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -8537.826406 +0.000000
kappa (ts/tv) = 3.04380, p: 0.09822 0.09822 0.09822 0.09822 0.09822 0.09822 0.09822 0.09822 0.09822 0.01778, w: 0.07142 0.17648 0.27059 0.36031 0.44814 0.53564 0.62419 0.71548 0.81229 0.92225 2.27861
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -8538.350417 +0.000000
kappa (ts/tv) = 3.02745, p: 0.07224 0.07224 0.07224 0.07224 0.07224 0.07224 0.07224 0.07224 0.07224 0.27764, w: 0.10475 0.16826 0.21489 0.25649 0.29675 0.33791 0.38223 0.43305 0.49764 0.60443 1.00000

LRT_8-7: 2x (-8537.826406 - -8544.423033)= 2x 6.596627= 13.193254
p-value (df=2): 0.001365 **
LRT)_8-8a: 2x (-8537.826406 - -8538.350417)= 2x 0.524011= 1.048022
p-value (df=1): 0.305968

>**COL17A1**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -54217.402989 +0.000000
kappa (ts/tv) = 2.71510, omega (dN/dS) = 0.26385
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -53278.239205 +0.000000
kappa (ts/tv) = 2.79742, p: 0.74336 0.25664, w: 0.13147 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -53278.239205 +0.000000
kappa (ts/tv) = 2.79742, p: 0.74336 0.22844 0.02820, w: 0.13147 1.00000 1.00000

LRT_1-0: 2x (-53278.239205 - -54217.402989)= 2x 939.163784= 1878.327568
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-53278.239205 - -53278.239205)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -53122.048729 +0.000000
kappa (ts/tv) = 2.74703, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000
0.10000 0.10000, w: 0.00118 0.01368 0.04277 0.09070 0.15924 0.25000 0.36462 0.50509
0.67418 0.87748
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -53085.206555 +0.000000
kappa (ts/tv) = 2.76411, p: 0.09387 0.09387 0.09387 0.09387 0.09387 0.09387 0.09387 0.09387
0.09387 0.09387 0.06130, w: 0.00279 0.01909 0.04716 0.08649 0.13765 0.20212 0.28276
0.38497 0.52061 0.72917 1.40127
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -53096.321921 +0.000000
kappa (ts/tv) = 2.75396, p: 0.08778 0.08778 0.08778 0.08778 0.08778 0.08778 0.08778 0.08778
0.08778 0.08778 0.12221, w: 0.00382 0.02036 0.04509 0.07733 0.11760 0.16733 0.22930
0.30898 0.41912 0.60734 1.00000

LRT_8-7: 2x (-53085.206555 - -53122.048729)= 2x 36.842174= 73.684348
p-value (df=2): < 0.00001 ***
LRT)_8-8a: 2x (-53085.206555 - -53096.321921)= 2x 11.115366= 22.230732
p-value (df=1): < 0.00001 ***

>**DLX3**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -5145.929810 +0.000000
kappa (ts/tv) = 3.64887, omega (dN/dS) = 0.02127
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -5138.083112 +0.000000
kappa (ts/tv) = 3.65334, p: 0.99405 0.00595, w: 0.01739 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -5138.083112 +0.000000
kappa (ts/tv) = 3.65332, p: 0.99405 0.00595 0.00000, w: 0.01739 1.00000 8.35864

LRT_1-0: 2x (-5138.083112 - -5145.929810)= 2x 7.846698= 15.693396
p-value (df=1): 0.000074 ***
LRT_2-1: 2x (-5138.083112 - -5138.083112)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -5113.826652 +0.000000

kappa (ts/tv) = 3.74311, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00000 0.00004 0.00026 0.00099 0.00270 0.00612 0.01241 0.02383 0.04599 0.10441
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -5113.137708 +0.000000
kappa (ts/tv) = 3.74720, p: 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.00393, w: 0.00000 0.00004 0.00027 0.00097 0.00252 0.00553 0.01092 0.02052 0.03889 0.08703 1.00000
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -5113.137708 +0.000000
kappa (ts/tv) = 3.74720, p: 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.09961 0.00393, w: 0.00000 0.00004 0.00027 0.00097 0.00252 0.00553 0.01092 0.02052 0.03889 0.08703 1.00000

LRT_8-7: 2x (-5113.137708 - -5113.826652)= 2x 0.688944= 1.377888
p-value (df=2): 0.502128
LRT)_8-8a: 2x (-5113.137708 - -5113.137708)= 2x 0= 0
p-value (df=1): 1

>**ENAM**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -44846.157924 +0.000000
kappa (ts/tv) = 3.26044, omega (dN/dS) = 0.42713
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -44423.877812 +0.000000
kappa (ts/tv) = 3.31457, p: 0.65997 0.34003, w: 0.25150 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -44406.943263 +0.000000
kappa (ts/tv) = 3.34452, p: 0.65672 0.33166 0.01163, w: 0.25513 1.00000 3.92932

LRT_1-0: 2x (-44423.877812 - -44846.157924)= 2x 422.280112= 844.560224
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-44406.943263 - -44423.877812)= 2x 16.934549= 33.869098
p-value (df=2): < 0.00001 ***

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -44371.479276 +0.000000
kappa (ts/tv) = 3.23931, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.03143 0.10914 0.19515 0.28669 0.38274 0.48289 0.58707 0.69557 0.80925 0.93083
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -44341.851283 +0.000000
kappa (ts/tv) = 3.27051, p: 0.09799 0.09799 0.09799 0.09799 0.09799 0.09799 0.09799 0.09799 0.09799 0.09799 0.02007, w: 0.03607 0.11418 0.19634 0.28194 0.37108 0.46414 0.56193 0.66585 0.77878 0.90913 2.83199
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -44359.263536 +0.000000
kappa (ts/tv) = 3.24288, p: 0.08230 0.08230 0.08230 0.08230 0.08230 0.08230 0.08230 0.08230 0.08230 0.08230 0.17703, w: 0.04502 0.10797 0.16558 0.22276 0.28173 0.34443 0.41330 0.49239 0.59045 0.73888 1.00000

LRT_8-7: 2x (-44341.851283 - -44371.479276)= 2x 29.627993= 59.255986
p-value (df=2): < 0.00001 ***
LRT)_8-8a: 2x (-44341.851283 - -44359.263536)= 2x 17.412253= 34.824506
p-value (df=1): < 0.00001 ***

>**FAM20A**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -14646.957679 +0.000000
kappa (ts/tv) = 2.95305, omega (dN/dS) = 0.13049
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -14339.550468 +0.000000
kappa (ts/tv) = 3.20159, p: 0.82059 0.17941, w: 0.06693 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -14339.550468 +0.000000
kappa (ts/tv) = 3.20159, p: 0.82059 0.14543 0.03398, w: 0.06693 1.00000 1.00000

LRT_1-0: 2x (-14339.550468 - -14646.957679)= 2x 307.407211= 614.814422
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-14339.550468 - -14339.550468)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -14199.167682 +0.000000
kappa (ts/tv) = 3.02258, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000
0.10000 0.10000, w: 0.00001 0.00044 0.00318 0.01170 0.03104 0.06806 0.13214 0.23694
0.40568 0.69532
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -14199.167750 +0.000000
kappa (ts/tv) = 3.02258, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000
0.10000 0.10000 0.00001, w: 0.00001 0.00044 0.00318 0.01170 0.03104 0.06806 0.13213
0.23693 0.40566 0.69528 1.00000
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -14199.167750 +0.000000
kappa (ts/tv) = 3.02258, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000
0.10000 0.10000 0.00001, w: 0.00001 0.00044 0.00318 0.01170 0.03104 0.06806 0.13213
0.23693 0.40566 0.69529 1.00000

LRT_8-7: 2x (-14199.167750 - -14199.167682)= 2x -0.000068= 0
p-value (df=2): 1
LRT)_8-8a: 2x (-14199.167750 - -14199.167750)= 2x 0= 0
p-value (df=1): 1

>**FAM83H**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -37611.289059 +0.000000
kappa (ts/tv) = 3.75128, omega (dN/dS) = 0.10193
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -37027.392167 +0.000000
kappa (ts/tv) = 3.86791, p: 0.89311 0.10689, w: 0.06628 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -37027.392167 +0.000000
kappa (ts/tv) = 3.86793, p: 0.89311 0.07048 0.03642, w: 0.06628 1.00000 1.00000

LRT_1-0: 2x (-37027.392167 - -37611.289059)= 2x 583.896892= 1167.793784
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-37027.392167 - -37027.392167)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -36701.408356 +0.000000

kappa (ts/tv) = 3.81034, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00003 0.00089 0.00466 0.01395 0.03191 0.06260 0.11182 0.18919 0.31485 0.55661
Model 8: beta&w>1 (11 categories)
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -36692.900663 +0.000000
kappa (ts/tv) = 3.81986, p: 0.09741 0.09741 0.09741 0.09741 0.09741 0.09741 0.09741 0.09741 0.09741 0.02594, w: 0.00006 0.00138 0.00578 0.01501 0.03098 0.05626 0.09476 0.15356 0.24900 0.44377 1.00000

LRT_8-7: 2x (-26477.948788 - -36701.408356)= 2x 24.741381= 49.482762
p-value (df=2): < 0.00001 ***
LRT)_8-8a: 2x (-26477.948788 - -36692.900663)= 2x 17.668226= 35.336452
p-value (df=1): < 0.00001 ***

>**GPR68**
Model 0: one-ratio
lnL(ntime: 73 np: 75): -10283.964401 +0.000000
kappa (ts/tv) = 4.30854, omega (dN/dS) = 0.05980
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 73 np: 76): -10017.200564 +0.000000
kappa (ts/tv) = 4.42100, p: 0.85498 0.14502, w: 0.03607 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 73 np: 78): -10010.409648 +0.000000
kappa (ts/tv) = 4.46802, p: 0.85164 0.13774 0.01062, w: 0.03625 1.00000 4.44281

LRT_1-0: 2x (-10017.200564 - -10283.964401)= 2x 266.763837= 533.527674
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-10010.409648 - -10017.200564)= 2x 6.790916= 13.581832
p-value (df=2): 0.001124 **

Model 7: beta (10 categories)
lnL(ntime: 73 np: 76): -9808.155161 +0.000000
kappa (ts/tv) = 4.60716, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00000 0.00003 0.00042 0.00233 0.00842 0.02364 0.05652 0.12196 0.24984 0.52990
Model 8: beta&w>1 (11 categories)
lnL(ntime: 73 np: 78): -9784.990112 +0.000000
kappa (ts/tv) = 4.67084, p: 0.09853 0.09853 0.09853 0.09853 0.09853 0.09853 0.09853 0.09853 0.09853 0.01466, w: 0.00000 0.00006 0.00058 0.00261 0.00808 0.02017 0.04401 0.08880 0.17501 0.37879 2.93399
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 73 np: 77): -9794.661498 +0.000000
kappa (ts/tv) = 4.63595, p: 0.09773 0.09773 0.09773 0.09773 0.09773 0.09773 0.09773 0.09773 0.09773 0.02268, w: 0.00000 0.00007 0.00060 0.00260 0.00781 0.01900 0.04068 0.08091 0.15806 0.34299 1.00000

LRT_8-7: 2x (-9784.990112 - -9808.155161)= 2x 23.165049= 46.330098
p-value (df=2): < 0.00001 ***
LRT)_8-8a: 2x (-9784.990112 - -9794.661498)= 2x 9.671386= 19.342772
p-value (df=1): 0.000011 ***

>**ITGB6**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -20628.455757 +0.000000

kappa (ts/tv) = 2.89885, omega (dN/dS) = 0.13914
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -20310.960920 +0.000000
kappa (ts/tv) = 3.05155, p: 0.89958 0.10042, w: 0.08448 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -20308.592743 +0.000000
kappa (ts/tv) = 3.06840, p: 0.89935 0.09830 0.00235, w: 0.08502 1.00000 2.68165

LRT_1-0: 2x (-20310.960920 - -20628.455757)= 2x 317.494837= 634.989674
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-20308.592743 - -20310.960920)= 2x 2.368177= 4.736354
p-value (df=2): 0.093654

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -20227.306235 +0.000000
kappa (ts/tv) = 2.92188, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00004 0.00127 0.00642 0.01880 0.04222 0.08144 0.14293 0.23689 0.38320 0.64223
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -20208.541424 +0.000000
kappa (ts/tv) = 2.93013, p: 0.09758 0.09758 0.09758 0.09758 0.09758 0.09758 0.09758 0.09758 0.09758 0.02419, w: 0.00020 0.00287 0.01006 0.02326 0.04412 0.07493 0.11935 0.18415 0.28502 0.48182 1.27467
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -20210.270931 +0.000000
kappa (ts/tv) = 2.91659, p: 0.09627 0.09627 0.09627 0.09627 0.09627 0.09627 0.09627 0.09627 0.09627 0.03729, w: 0.00028 0.00339 0.01089 0.02380 0.04337 0.07146 0.11119 0.16853 0.25770 0.43508 1.00000

LRT_8-7: 2x (-20208.541424 - -20227.306235)= 2x 18.764811= 37.529622
p-value (df=2): < 0.00001 ***
LRT)_8-8a: 2x (-20208.541424 - -20210.270931)= 2x 1.729507= 3.459014
p-value (df=1): 0.062908

>**KLK4**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -10388.324635 +0.000000
kappa (ts/tv) = 2.41082, omega (dN/dS) = 0.22039
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -10142.739489 +0.000000
kappa (ts/tv) = 2.70143, p: 0.67233 0.32767, w: 0.10950 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -10142.739489 +0.000000
kappa (ts/tv) = 2.70143, p: 0.67233 0.23890 0.08877, w: 0.10950 1.00000 1.00000

LRT_1-0: 2x (-10142.739489 - -10388.324635)= 2x 245.585146= 491.170292
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-10142.739489 - -10142.739489)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -10047.780520 +0.000000
kappa (ts/tv) = 2.57389, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00147 0.01402 0.04021 0.08100 0.13763 0.21196 0.30682 0.42707 0.58228 0.80020
Model 8: beta&w>1 (11 categories)

lnL(ntime: 71 np: 76): -10046.012109 +0.000000
kappa (ts/tv) = 2.58457, p: 0.09643 0.09643 0.09643 0.09643 0.09643 0.09643 0.09643 0.09643
0.09643 0.09643 0.03567, w: 0.00201 0.01541 0.04008 0.07601 0.12404 0.18587 0.26458
0.36591 0.50242 0.71589 1.19287
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -10046.368777 +0.000000
kappa (ts/tv) = 2.57874, p: 0.09471 0.09471 0.09471 0.09471 0.09471 0.09471 0.09471 0.09471
0.09471 0.09471 0.05286, w: 0.00210 0.01527 0.03888 0.07281 0.11783 0.17563 0.24931
0.34473 0.47497 0.68504 1.00000

LRT_8-7: 2x (-10046.012109 - -10047.780520)= 2x 1.768411= 3.536822
p-value (df=2): 0.170606
LRT)_8-8a: 2x (-10046.012109 - -10046.368777)= 2x 0.356668= 0.713336
p-value (df=1): 0.39835

>**LAMA3**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -108237.391365 +0.000000
kappa (ts/tv) = 3.04535, omega (dN/dS) = 0.27741
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -106641.066941 +0.000000
kappa (ts/tv) = 3.24132, p: 0.75730 0.24270, w: 0.15456 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -106641.066941 +0.000000
kappa (ts/tv) = 3.24132, p: 0.75730 0.20558 0.03712, w: 0.15456 1.00000 1.00000

LRT_1-0: 2x (-106641.066941 - -108237.391365)= 2x 1596.324424= 3192.648848
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-106641.066941 - -106641.066941)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -106283.082663 +0.000000
kappa (ts/tv) = 3.09632, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000
0.10000 0.10000, w: 0.00285 0.02221 0.05795 0.10952 0.17717 0.26179 0.36510 0.49025
0.64364 0.84333
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -106233.626516 +0.000000
kappa (ts/tv) = 3.12178, p: 0.09707 0.09707 0.09707 0.09707 0.09707 0.09707 0.09707 0.09707
0.09707 0.09707 0.02932, w: 0.00445 0.02650 0.06139 0.10784 0.16604 0.23713 0.32356
0.42999 0.56673 0.76726 1.49303
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -106248.586323 +0.000000
kappa (ts/tv) = 3.10560, p: 0.09116 0.09116 0.09116 0.09116 0.09116 0.09116 0.09116 0.09116
0.09116 0.09116 0.08842, w: 0.00595 0.02819 0.05908 0.09767 0.14435 0.20049 0.26875
0.35435 0.46931 0.65744 1.00000

LRT_8-7: 2x (-106233.626516 - -106283.082663)= 2x 49.456147= 98.912294
p-value (df=2): < 0.00001 ***
LRT)_8-8a: 2x (-106233.626516 - -106248.586323)= 2x 14.959807= 29.919614
p-value (df=1): < 0.00001 ***

>**LAMB3**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -39239.469028 +0.000000

kappa (ts/tv) = 3.69245, omega (dN/dS) = 0.17955
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -38754.215595 +0.000000
kappa (ts/tv) = 3.87366, p: 0.83758 0.16242, w: 0.12355 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -38754.215595 +0.000000
kappa (ts/tv) = 3.87366, p: 0.83758 0.14146 0.02096, w: 0.12355 1.00000 1.00000

LRT_1-0: 2x (-38754.215595 - -39239.469028)= 2x 485.253433= 970.506866
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-38754.215595 - -38754.215595)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -38548.026158 +0.000000
kappa (ts/tv) = 3.74498, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000
0.10000 0.10000, w: 0.00276 0.01711 0.04051 0.07251 0.11370 0.16563 0.23123 0.31623
0.43380 0.63184, Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -38535.133498 +0.000000
kappa (ts/tv) = 3.75635, p: 0.09607 0.09607 0.09607 0.09607 0.09607 0.09607 0.09607 0.09607
0.09607 0.09607 0.03928, w: 0.00455 0.02072 0.04282 0.07037 0.10390 0.14475 0.19543
0.26092 0.35318 0.51941 1.00000
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -38535.133498 +0.000000
kappa (ts/tv) = 3.75636, p: 0.09607 0.09607 0.09607 0.09607 0.09607 0.09607 0.09607 0.09607
0.09607 0.09607 0.03928, w: 0.00455 0.02072 0.04282 0.07037 0.10390 0.14475 0.19543
0.26092 0.35319 0.51942 1.00000

LRT_8-7: 2x (-38535.133498 - -38548.026158)= 2x 12.89266= 25.78532
p-value (df=2): < 0.00001 ***
LRT)_8-8a: 2x (-38535.133498 - -38535.133498)= 2x 0= 0
p-value (df=1): 1

>**MMP20**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -12589.910027 +0.000000
kappa (ts/tv) = 3.18536, omega (dN/dS) = 0.14708
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -12319.979326 +0.000000
kappa (ts/tv) = 3.39077, p: 0.83529 0.16471, w: 0.06139 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -12319.979326 +0.000000
kappa (ts/tv) = 3.39077, p: 0.83529 0.14909 0.01561, w: 0.06139 1.00000 1.00000

LRT_1-0: 2x (-12319.979326 - -12589.910027)= 2x 269.930701= 539.861402
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-12319.979326 - -12319.979326)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -12239.858821 +0.000000
kappa (ts/tv) = 3.18820, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000
0.10000 0.10000, w: 0.00000 0.00008 0.00100 0.00523 0.01809 0.04874 0.11143 0.22699
0.42544 0.75341
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -12238.390360 +0.000000

kappa (ts/tv) = 3.20117, p: 0.09767 0.09767 0.09767 0.09767 0.09767 0.09767 0.09767 0.09767 0.09767 0.09767 0.02334, w: 0.00000 0.00015 0.00139 0.00613 0.01862 0.04547 0.09661 0.18751 0.34581 0.64288 1.12673
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -12238.507100 +0.000000
kappa (ts/tv) = 3.19827, p: 0.09620 0.09620 0.09620 0.09620 0.09620 0.09620 0.09620 0.09620 0.09620 0.09620 0.03798, w: 0.00000 0.00018 0.00153 0.00631 0.01827 0.04304 0.08903 0.16976 0.31113 0.58906 1.00000

LRT_8-7: 2x (-12238.390360 - -12239.858821)= 2x 1.468461= 2.936922
p-value (df=2): 0.230282
LRT)_8-8a: 2x (-12238.390360 - -12238.507100)= 2x 0.11674= 0.23348
p-value (df=1): 0.629014

>**RHO**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -8124.556486 +0.000000
kappa (ts/tv) = 4.73429, omega (dN/dS) = 0.03586
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -8041.345568 +0.000000
kappa (ts/tv) = 5.02819, p: 0.95551 0.04449, w: 0.02639 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -8041.345568 +0.000000
kappa (ts/tv) = 5.02819, p: 0.95551 0.04449 0.00000, w: 0.02639 1.00000 28.74400

LRT_1-0: 2x (-8041.345568 - -8124.556486)= 2x 83.210918= 166.421836
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-8041.345568 - -8041.345568)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -7897.502941 +0.000000
kappa (ts/tv) = 4.86632, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00000 0.00000 0.00005 0.00041 0.00184 0.00617 0.01713 0.04236 0.10017 0.26345
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -7896.709833 +0.000000
kappa (ts/tv) = 4.87896, p: 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.00268, w: 0.00000 0.00000 0.00006 0.00044 0.00189 0.00615 0.01666 0.04041 0.09416 0.24579 1.00000
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -7896.709833 +0.000000
kappa (ts/tv) = 4.87897, p: 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.09973 0.00268, w: 0.00000 0.00000 0.00006 0.00044 0.00189 0.00615 0.01666 0.04041 0.09416 0.24579 1.00000

LRT_8-7: 2x (-7896.709833 - -7897.502941)= 2x 0.793108= 1.586216
p-value (df=2): 0.45244
LRT)_8-8a: 2x (-7896.709833 - -7896.709833)= 2x 0= 0
p-value (df=1): 1

>**SLC24A4**

Model 0: one-ratio
lnL(ntime: 71 np: 73): -15469.627083 +0.000000

kappa (ts/tv) = 3.56689, omega (dN/dS) = 0.08397
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -15346.328270 +0.000000
kappa (ts/tv) = 3.76218, p: 0.92119 0.07881, w: 0.06071 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -15346.343225 +0.000000
kappa (ts/tv) = 3.75933, p: 0.92120 0.07878 0.00003, w: 0.06072 1.00000 15.27355

LRT_1-0: 2x (-15346.328270 - -15469.627083)= 2x 123.298813= 246.597626
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-15346.343225 - -15346.328270)= 2x -0.014955= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -15217.760989 +0.000000
kappa (ts/tv) = 3.65883, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00002 0.00070 0.00359 0.01059 0.02402 0.04693 0.08386 0.14282 0.24177 0.44786
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -15212.179308 +0.000000
kappa (ts/tv) = 3.65950, p: 0.09933 0.09933 0.09933 0.09933 0.09933 0.09933 0.09933 0.09933 0.09933 0.09933 0.00675, w: 0.00004 0.00094 0.00422 0.01147 0.02448 0.04567 0.07873 0.13040 0.21638 0.39838 1.40811
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -15212.818427 +0.000000
kappa (ts/tv) = 3.65852, p: 0.09889 0.09889 0.09889 0.09889 0.09889 0.09889 0.09889 0.09889 0.09889 0.09889 0.01112, w: 0.00004 0.00100 0.00435 0.01157 0.02435 0.04493 0.07679 0.12636 0.20871 0.38386 1.00000

LRT_8-7: 2x (-15212.179308 - -15217.760989)= 2x 5.581681= 11.163362
p-value (df=2): 0.003767 **
LRT)_8-8a: 2x (-15212.179308 - -15212.818427)= 2x 0.639119= 1.278238
p-value (df=1): 0.258234

>**SP6**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -6896.139727 +0.000000
kappa (ts/tv) = 3.51131, omega (dN/dS) = 0.04117
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -6828.092220 +0.000000
kappa (ts/tv) = 3.65622, p: 0.94179 0.05821, w: 0.02160 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -6828.092220 +0.000000
kappa (ts/tv) = 3.65606, p: 0.94179 0.05821 0.00000, w: 0.02160 1.00000 146.01167

LRT_1-0: 2x (-6828.092220 - -6896.139727)= 2x 68.047507= 136.095014
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-6828.092220 - -6828.092220)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -6777.224891 +0.000000
kappa (ts/tv) = 3.56488, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00000 0.00000 0.00000 0.00002 0.00025 0.00163 0.00771 0.02949 0.09946 0.34097
Model 8: beta&w>1 (11 categories)

lnL(ntime: 71 np: 76): -6777.226193 +0.000000
kappa (ts/tv) = 3.56489, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.00001, w: 0.00000 0.00000 0.00000 0.00002 0.00025 0.00163 0.00771 0.02949 0.09945 0.34094 1.00000
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -6777.226193 +0.000000
kappa (ts/tv) = 3.56489, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.00001, w: 0.00000 0.00000 0.00000 0.00002 0.00025 0.00163 0.00771 0.02949 0.09945 0.34094 1.00000

LRT_8-7: 2x (-6777.226193 - -6777.224891)= 2x -0.001302= 0
p-value (df=2): 1
LRT)_8-8a: 2x (-6777.226193 - -6777.226193)= 2x 0= 0
p-value (df=1): 1

>**TUBA4A**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -7070.652107 +0.000000
kappa (ts/tv) = 3.40597, omega (dN/dS) = 0.00675
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -6994.999033 +0.000000
kappa (ts/tv) = 3.42930, p: 0.98643 0.01357, w: 0.00185 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -6988.935823 +0.000000
kappa (ts/tv) = 3.46259, p: 0.98741 0.00552 0.00707, w: 0.00201 1.00000 13.83270

LRT_1-0: 2x (-6994.999033 - -7070.652107)= 2x 75.653074= 151.306148
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-6988.935823 - -6994.999033)= 2x 6.06321= 12.12642
p-value (df=2): 0.002327 **

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -7017.293106 +0.000000
kappa (ts/tv) = 3.40436, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000, w: 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.23871
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -6988.215368 +0.000000
kappa (ts/tv) = 3.42304, p: 0.09895 0.09895 0.09895 0.09895 0.09895 0.09895 0.09895 0.09895 0.09895 0.01053, w: 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00003 0.02594 1.74183
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -6990.091777 +0.000000
kappa (ts/tv) = 3.40370, p: 0.09885 0.09885 0.09885 0.09885 0.09885 0.09885 0.09885 0.09885 0.09885 0.01151, w: 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000 0.02388 1.00000

LRT_8-7: 2x (-6988.215368 - -7017.293106)= 2x 29.077738= 58.155476
p-value (df=2): < 0.00001 ***
LRT)_8-8a: 2x (-6988.215368 - -6990.091777)= 2x 1.876409= 3.752818
p-value (df=1): 0.052719

>**WDR72**
Model 0: one-ratio
lnL(ntime: 71 np: 73): -36112.594119 +0.000000

kappa (ts/tv) = 3.07762, omega (dN/dS) = 0.25978
Model 1: NearlyNeutral (2 categories)
lnL(ntime: 71 np: 74): -35491.221790 +0.000000
kappa (ts/tv) = 3.33396, p: 0.72528 0.27472, w: 0.12822 1.00000
Model 2: PositiveSelection (3 categories)
lnL(ntime: 71 np: 76): -35491.221790 +0.000000
kappa (ts/tv) = 3.33394, p: 0.72528 0.22874 0.04599, w: 0.12822 1.00000 1.00000

LRT_1-0: 2x (-35491.221790 - -36112.594119)= 2x 621.372329= 1242.744658
p-value (df=1): < 0.00001 ***
LRT_2-1: 2x (-35491.221790 - -35491.221790)= 2x 0= 0
p-value (df=2): 1

Model 7: beta (10 categories)
lnL(ntime: 71 np: 74): -35337.365363 +0.000000
kappa (ts/tv) = 3.11203, p: 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000 0.10000
0.10000 0.10000, w: 0.00112 0.01283 0.03986 0.08434 0.14808 0.23299 0.34136 0.47639
0.64344 0.85523
Model 8: beta&w>1 (11 categories)
lnL(ntime: 71 np: 76): -35332.409174 +0.000000
kappa (ts/tv) = 3.13280, p: 0.09401 0.09401 0.09401 0.09401 0.09401 0.09401 0.09401 0.09401
0.09401 0.09401 0.05995, w: 0.00192 0.01535 0.04068 0.07807 0.12838 0.19334 0.27600
0.38195 0.52320 0.73858 1.08364
Model 8a: beta&w>1 (11 categories), omega = 1.000 fixed
lnL(ntime: 71 np: 75): -35332.677517 +0.000000
kappa (ts/tv) = 3.12935, p: 0.09216 0.09216 0.09216 0.09216 0.09216 0.09216 0.09216 0.09216
0.09216 0.09216 0.07840, w: 0.00206 0.01546 0.03988 0.07527 0.12243 0.18307 0.26026
0.35978 0.49436 0.70680 1.00000

LRT_8-7: 2x (-35332.409174 - -35337.365363)= 2x 4.956189= 9.912378
p-value (df=2): 0.00704 **
LRT)_8-8a: 2x (-35332.409174 - -35332.677517)= 2x 0.268343= 0.536686
p-value (df=1): 0.463845

Appendix E

E.1 Length of MSAs for each gene

Gene	Length of MSA
ACP4	1656
AMBN	1650
AMELX	621
AMTN	735
COL17A1	5439
DLX3	882
ENAM	5034
FAM20A	1719
FAM83H	4053
GPR68	1236
ITGB6	2439
KLK4	960
LAMA3	10707
LAMB3	3600
MMP20	1491
ODAPH	561
RELT	1392
RHO	1092
SLC24A4	1941
SP6	1161
TUBA4A	1407
WDR72	3651

E.2 Length of each gene per species

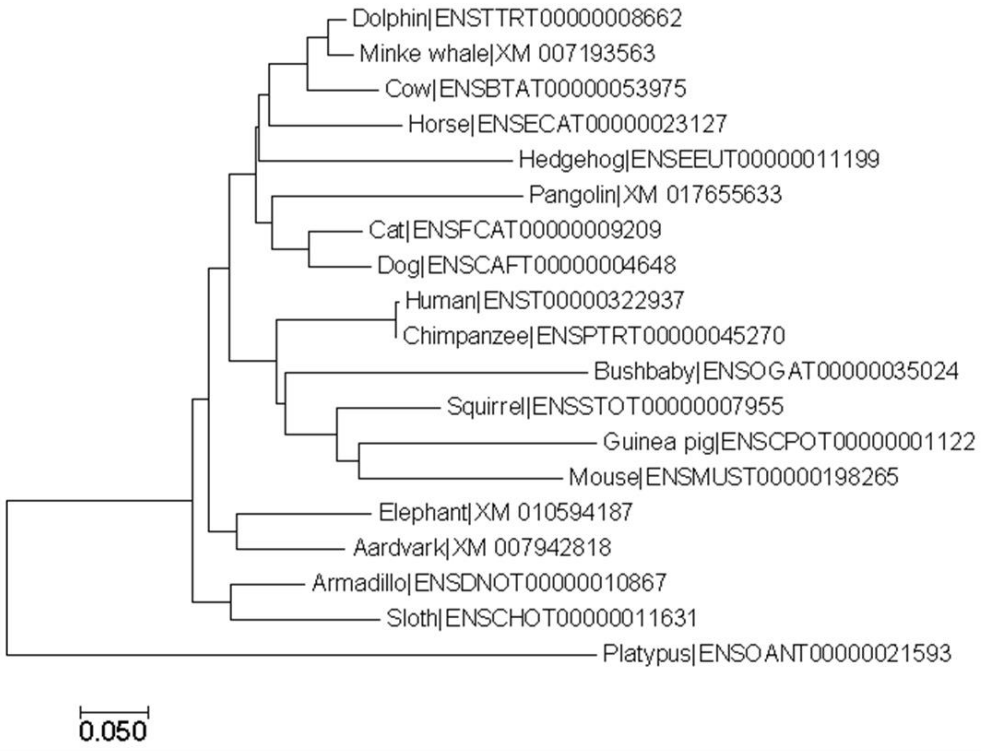
	Human	Chimpanzee	Bushbaby	Squirrel	Mouse	Guinea Pig	Horse	Minke whale	Dolphin	Cow	Cat	Dog
ACP4	1278	1278	1113	1266	1275	1272	972	1230	1266	1293	1263	1278
AMBN	1341	1341	1332	1269	1266	1269	1341	1263	1257	1191	1245	1308
AMELX	615	615	663	627	657	345	576	513	549	639	615	576
AMTN	627	627	603	636	639	636	621	741	645	636	627	639
COL17A1	4491	4491	4473	4407	4410	4362	4431	4431	4452	4419	4407	4791
DLX3	861	861	861	861	861	846	861	861		861	861	861
ENAM	3426	3426	3435	3426	3822	3417	3432	3435	3234	3435	3453	3444
FAM20A	1623	1623	1209	1614	1623	1626	1542	1518	1563	1575	1611	1221
FAM83H	3537	3537	3528	3501	3627	3525	3540	3312	3474	3543	3480	3453
GPR68	1095	1095	1095	1095	1095	1089	1083	1083		1083	1083	1080
ITGB6	2364	2103	2361	2358	2361	2364	2364	2364	2364	2364	2364	2364
KLK4	762	762	765	765	765	759	681	552	762	768	762	762
LAMA3	9999	9999	9870	9888	9990	9786	10002	2226	10008	9726	9942	10011
LAMB3	3516	3513	3489	3519	3504	3519	3516	3516	3516	3516	3516	3513
MMP20	1449	1449	1449	1449	1446	1449	1449	1452	1449	1443	1449	1449
ODAPH	528	528	258	420	378	441	462		429	285	459	399
RELT	1290	1290	1269	1290	1287	1275	1287	1287	1287	1290	1290	1290
RHO	1044	1044	1044	1044	1044	1044	1044	1044	1044	1044	1044	1074
SLC24A4	1866	1539	1698	1866	1815	1815	1620	1848	1611	1815	1866	1617
SP6	1128	1128	1128	1128	1128	1128	1128	1137	810	1125	1128	1128
TUBA4A	1344	1344	1344	1344	1344	1344	1344	1344	1401	1344	1344	1344
WDR72	3306	3306	3162	3339	3342	3213	3156	3312	3312	3159	3300	3300

	Hedgehog	Pangolin	Aardvark	Elephant	Sloth	Armadillo	Platypus
ACP4	1272		903	1152		936	348
AMBN	1338	1017	1014	855	918	1263	1371
AMELX	618	522	597	615	582	480	597
AMTN	627		357	639	495	630	
COL17A1	4431	2010	4404	4446		4443	4392
DLX3	861	861	861	876	513	864	
ENAM	3441	3282	3441	3429		3486	4086
FAM20A	1218	1209	1200	1221	978	1617	1194
FAM83H	3519	3549	3591	2604		546	753
GPR68	1062	1083	1098	1098		1092	1209
ITGB6	2358	2160	2364	2364	2223	2400	2331
KLK4	747			831			
LAMA3	5178		5007	9594	9978	10011	6267
LAMB3	2667	1629	3516	3510	3381	3507	2751
MMP20	1449	792	1467	1449	1260	1431	
ODAPH	399	441	399	399		399	
RELT	1269	1290	1323	1047	720	1278	
RHO	1041	1044	1044	1044		1035	1059
SLC24A4	1707	669	1878	1815	1728	1485	1689
SP6	1122	1128	1128	1128		1128	
TUBA4A	1344	873	1344	1344	1164	1344	1248
WDR72	3099	3294	3246	2076	3039	3024	3042

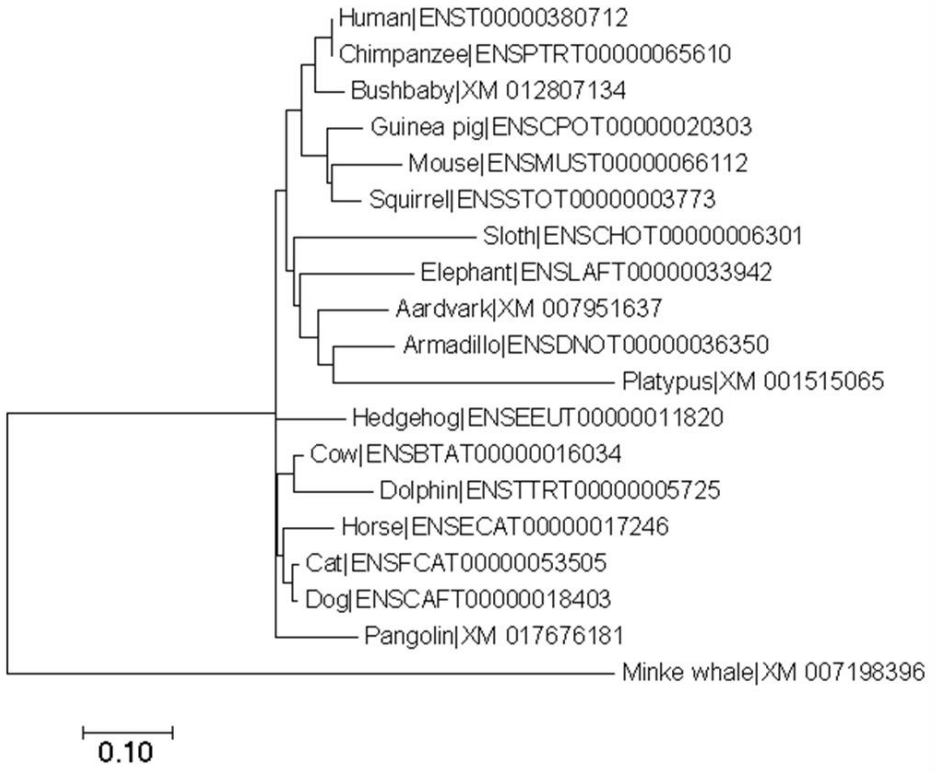
Appendix F

F.1 Phylogenetic trees constructed with the maximum likelihood (ML) method
The branch lengths correspond to the nucleotide distance between the species.

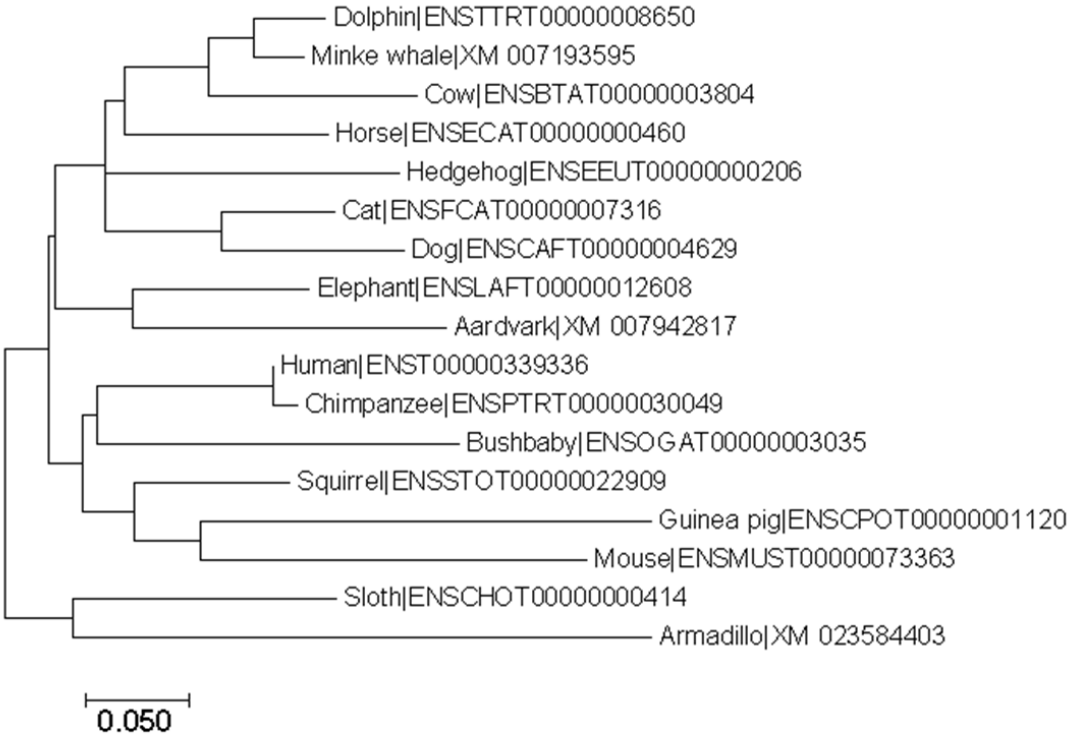
AMBN



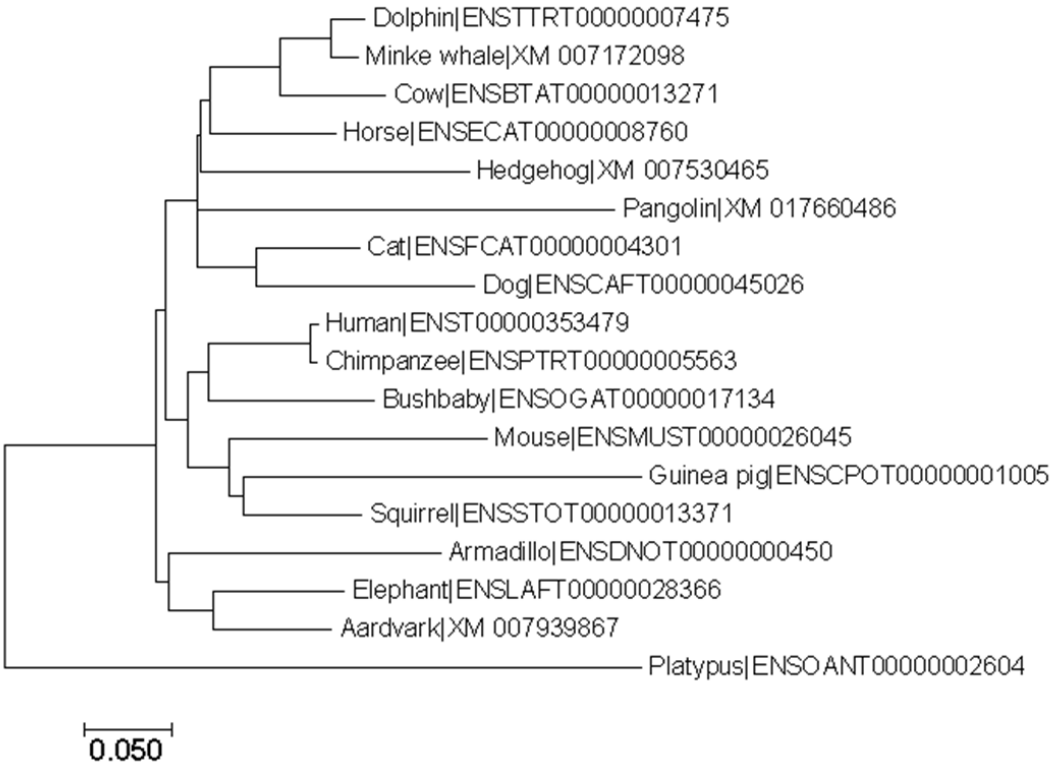
AMELX



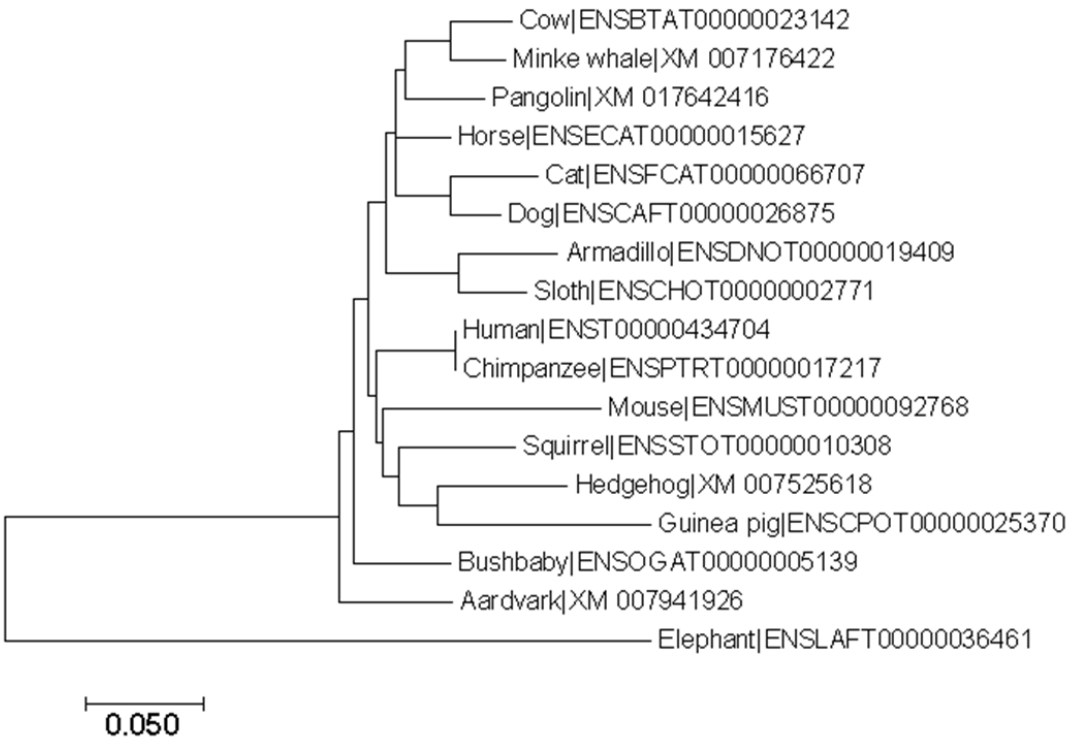
AMTN



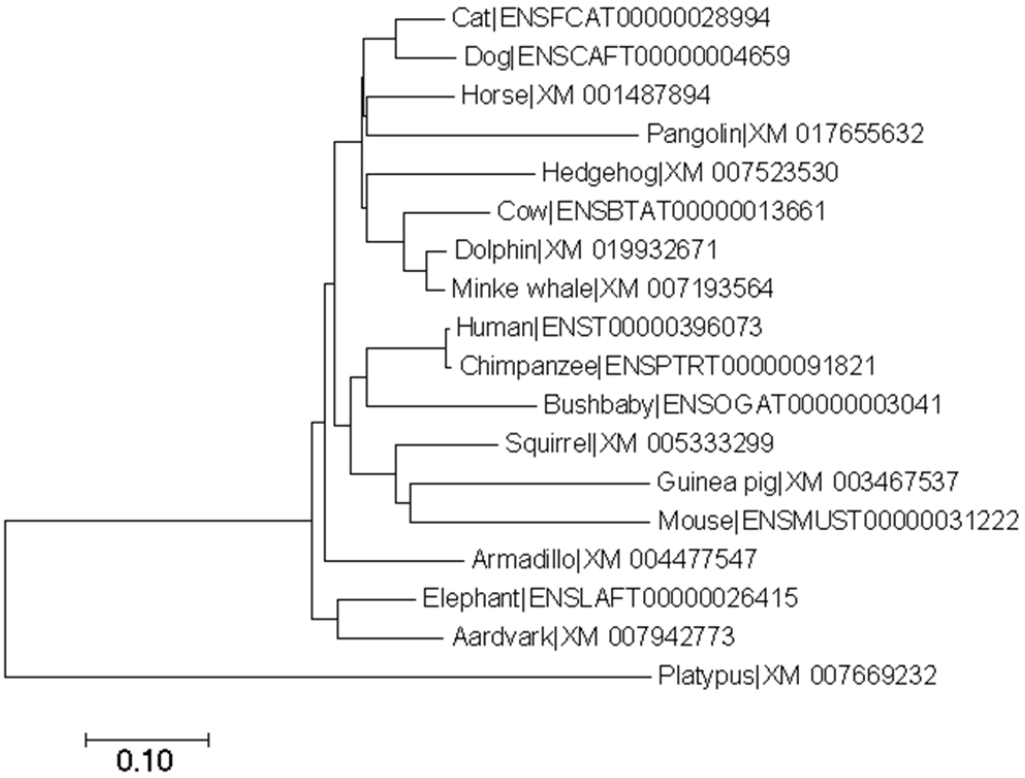
COL17A1



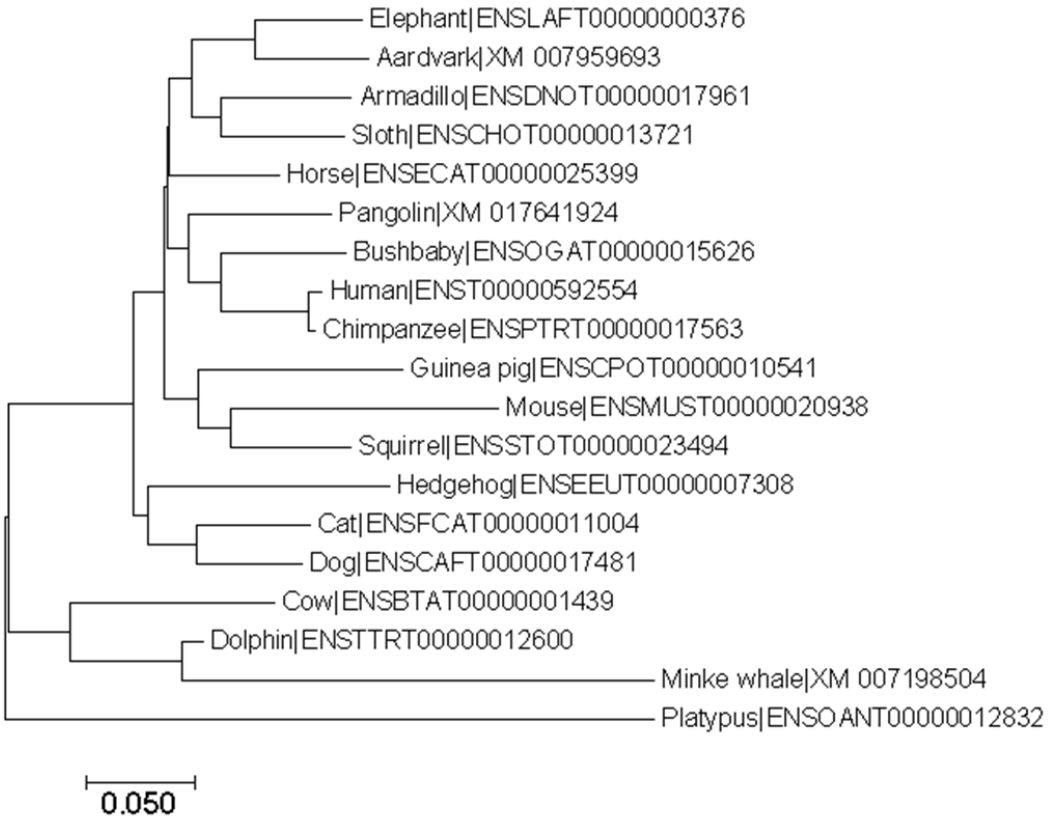
DLX3


Cow|ENSBTAT00000023142
Minke whale|XM 007176422
Pangolin|XM 017642416
Horse|ENSECAT00000015627
Cat|ENSFCAT00000066707
Dog|ENSCAFT00000026875
Armadillo|ENSDNOT00000019409
Sloth|ENSCHOT00000002771
Human|ENST00000434704
Chimpanzee|ENSPTRT00000017217
Mouse|ENSMUST00000092768
Squirrel|ENSSTOT00000010308
Hedgehog|XM 007525618
Guinea pig|ENSCPOT00000025370
Bushbaby|ENSOGAT00000005139
Aardvark|XM 007941926
Elephant|ENSLAFT00000036461
0.050

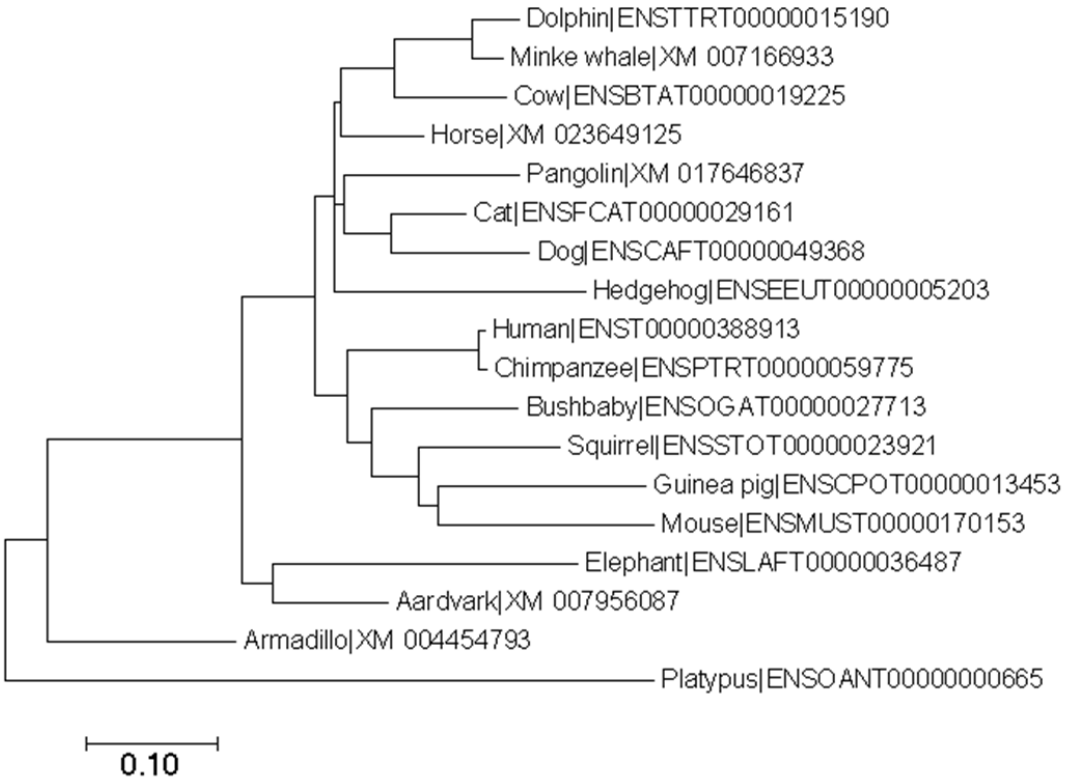
ENAM


Cat|ENSFCAT00000028994
Dog|ENSCAFT00000004659
Horse|XM 001487894
Pangolin|XM 017655632
Hedgehog|XM 007523530
Cow|ENSBTAT00000013661
Dolphin|XM 019932671
Minke whale|XM 007193564
Human|ENST00000396073
Chimpanzee|ENSPTRT00000091821
Bushbaby|ENSOGAT00000003041
Squirrel|XM 005333299
Guinea pig|XM 003467537
Mouse|ENSMUST00000031222
Armadillo|XM 004477547
Elephant|ENSLAFT00000026415
Aardvark|XM 007942773
Platypus|XM 007669232
0.10

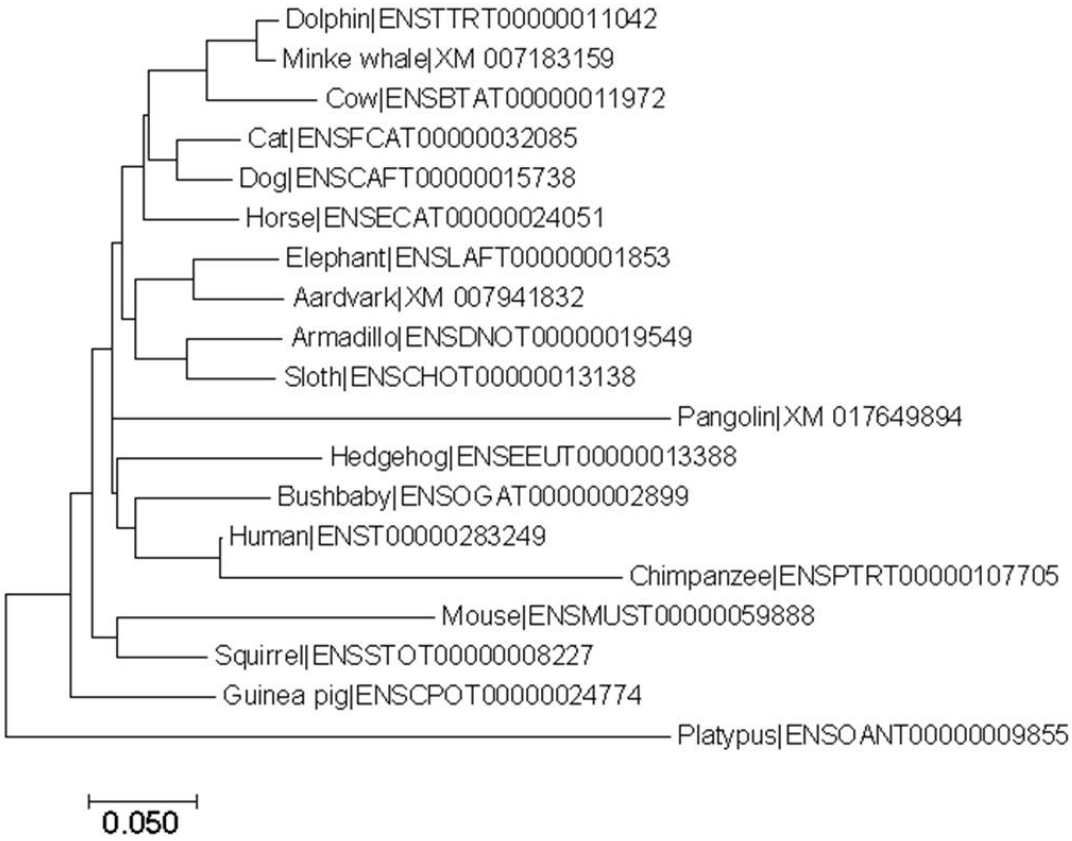
FAM20A



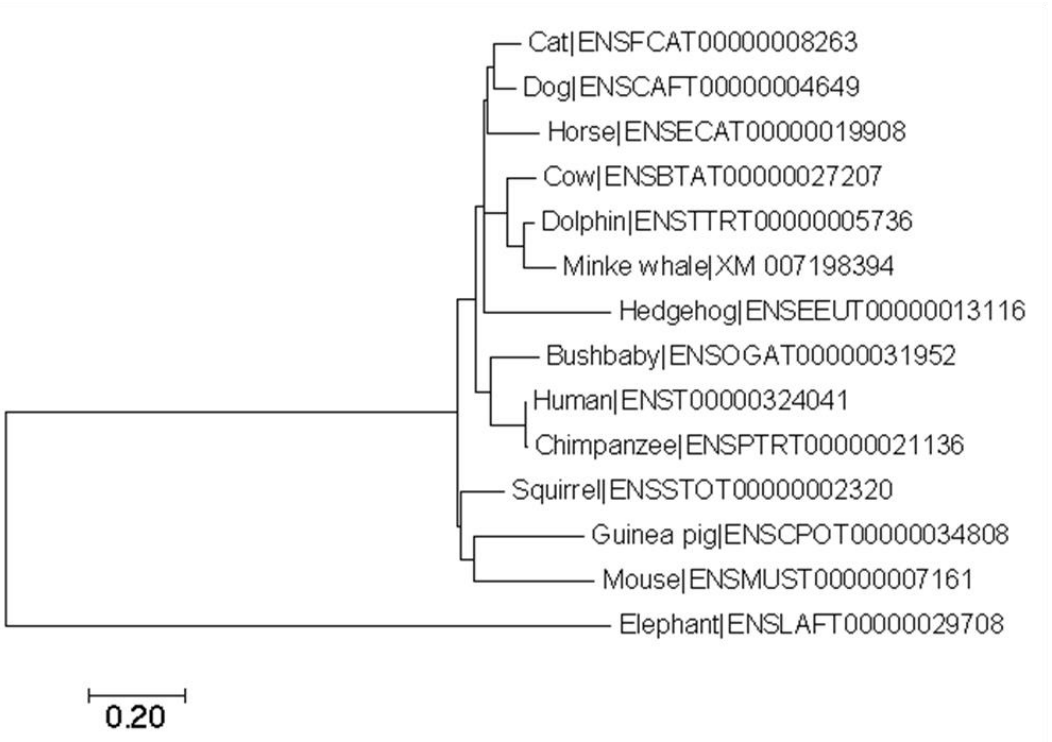
FAM83H



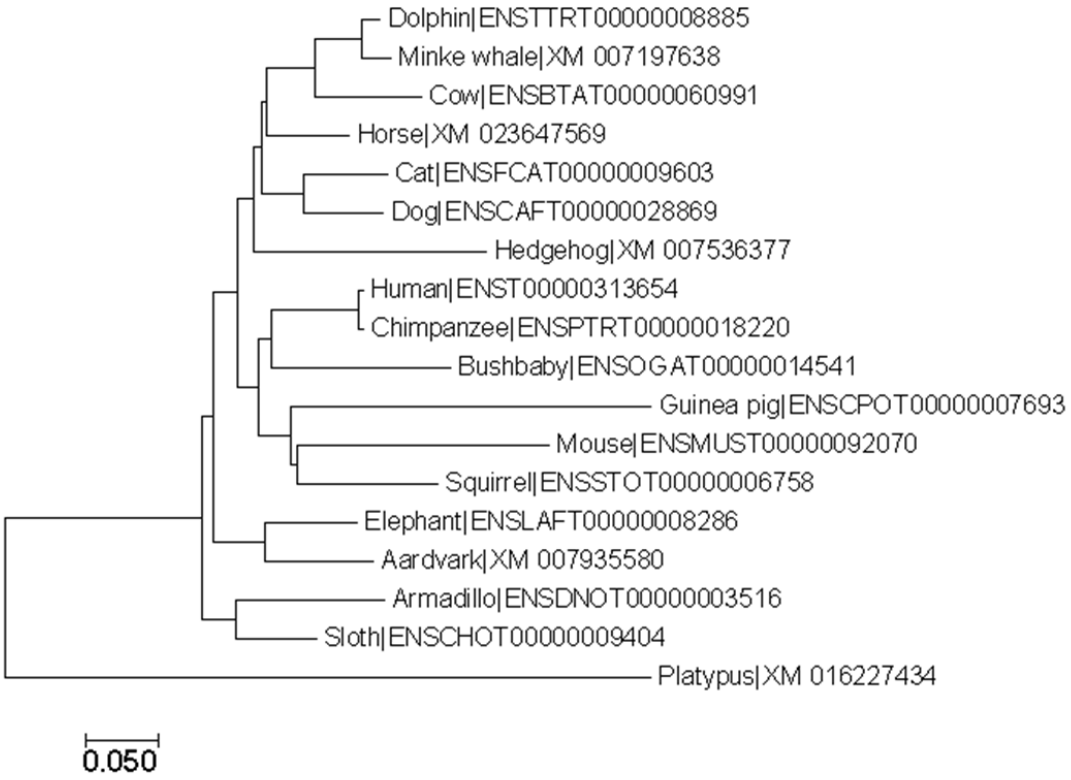
ITGB6


Dolphin|ENSTTRT00000011042
Minke whale|XM 007183159
Cow|ENSBTAT00000011972
Cat|ENSFCAT00000032085
Dog|ENSCAFT00000015738
Horse|ENSECAT00000024051
Elephant|ENSLAFT00000001853
Aardvark|XM 007941832
Armadillo|ENSDNOT00000019549
Sloth|ENSCHOT00000013138
Pangolin|XM 017649894
Hedgehog|ENSEEUT00000013388
Bushbaby|ENSOGAT00000002899
Human|ENST00000283249
Chimpanzee|ENSPTRT00000107705
Mouse|ENSMUST00000059888
Squirrel|ENSSTOT00000008227
Guinea pig|ENSCPOT00000024774
Platypus|ENSOANT00000009855
0.050

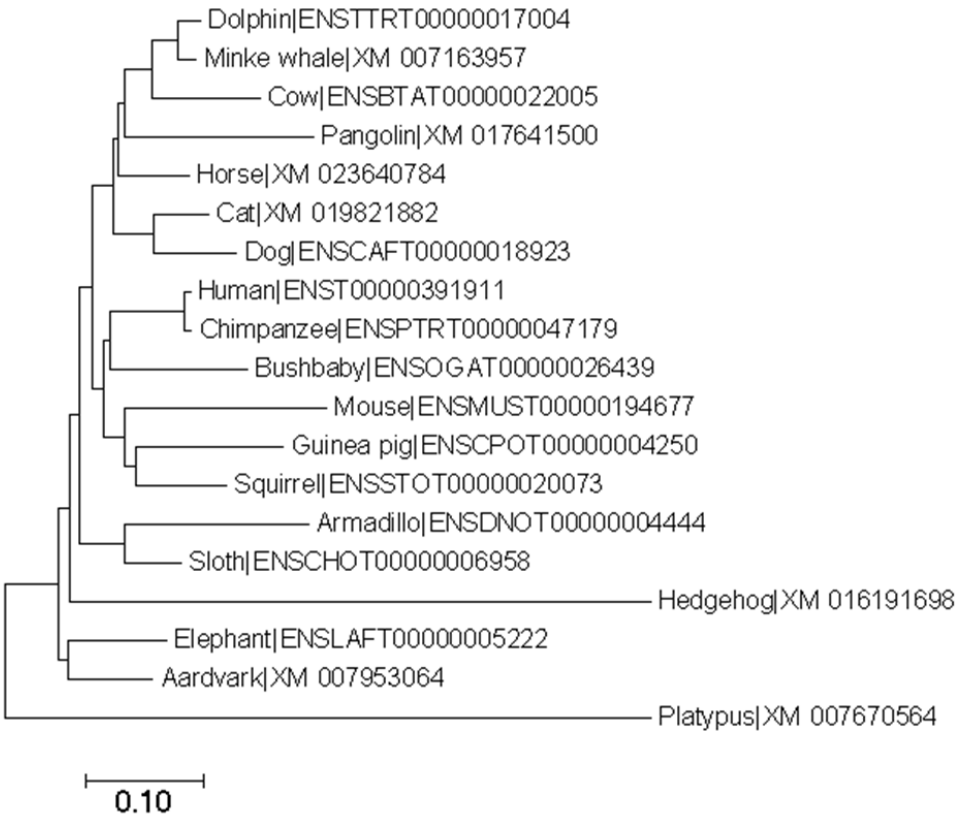
KLK4


Cat|ENSFCAT00000008263
Dog|ENSCAFT00000004649
Horse|ENSECAT00000019908
Cow|ENSBTAT00000027207
Dolphin|ENSTTRT00000005736
Minke whale|XM 007198394
Hedgehog|ENSEEUT00000013116
Bushbaby|ENSOGAT00000031952
Human|ENST00000324041
Chimpanzee|ENSPTRT00000021136
Squirrel|ENSSTOT00000002320
Guinea pig|ENSCPOT00000034808
Mouse|ENSMUST00000007161
Elephant|ENSLAFT00000029708
0.20

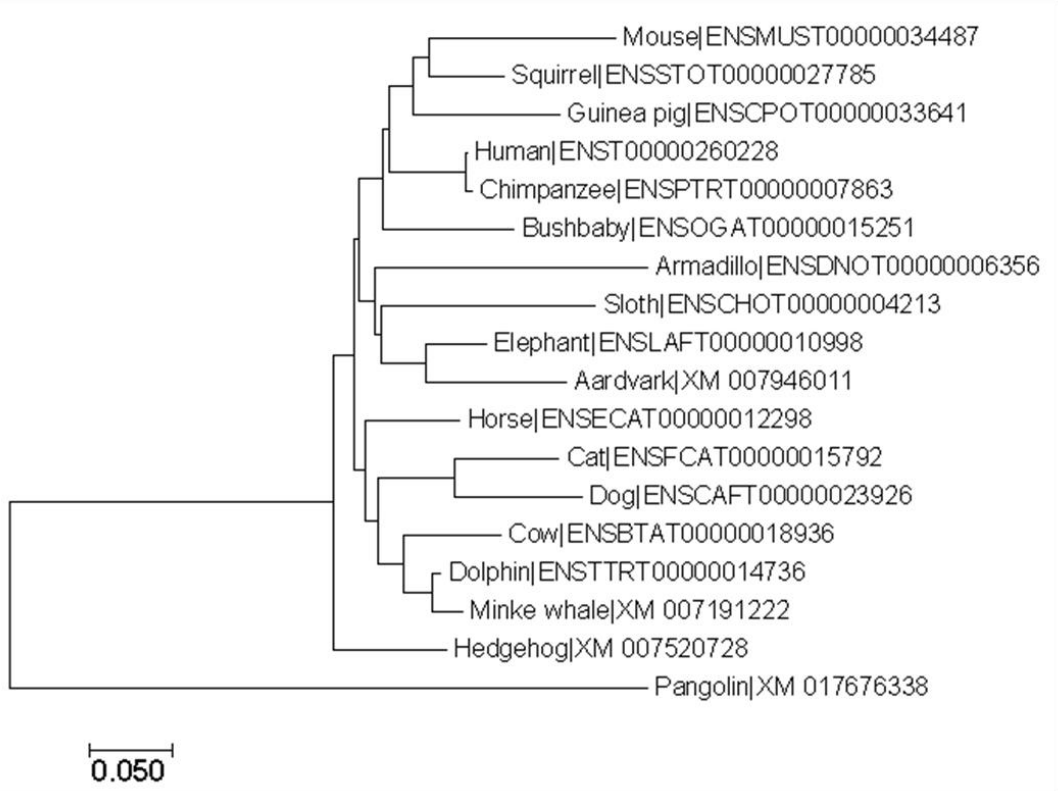
LAMA3


Dolphin|ENSTTRT00000008885
Minke whale|XM 007197638
Cow|ENSBTAT00000060991
Horse|XM 023647569
Cat|ENSFCAT00000009603
Dog|ENSCAFT00000028869
Hedgehog|XM 007536377
Human|ENST00000313654
Chimpanzee|ENSPTRT00000018220
Bushbaby|ENSOGAT00000014541
Guinea pig|ENSCPOT00000007693
Mouse|ENSMUST00000092070
Squirrel|ENSSTOT00000006758
Elephant|ENSLAFT00000008286
Aardvark|XM 007935580
Armadillo|ENSDNOT00000003516
Sloth|ENSCHOT00000009404
Platypus|XM 016227434
0.050

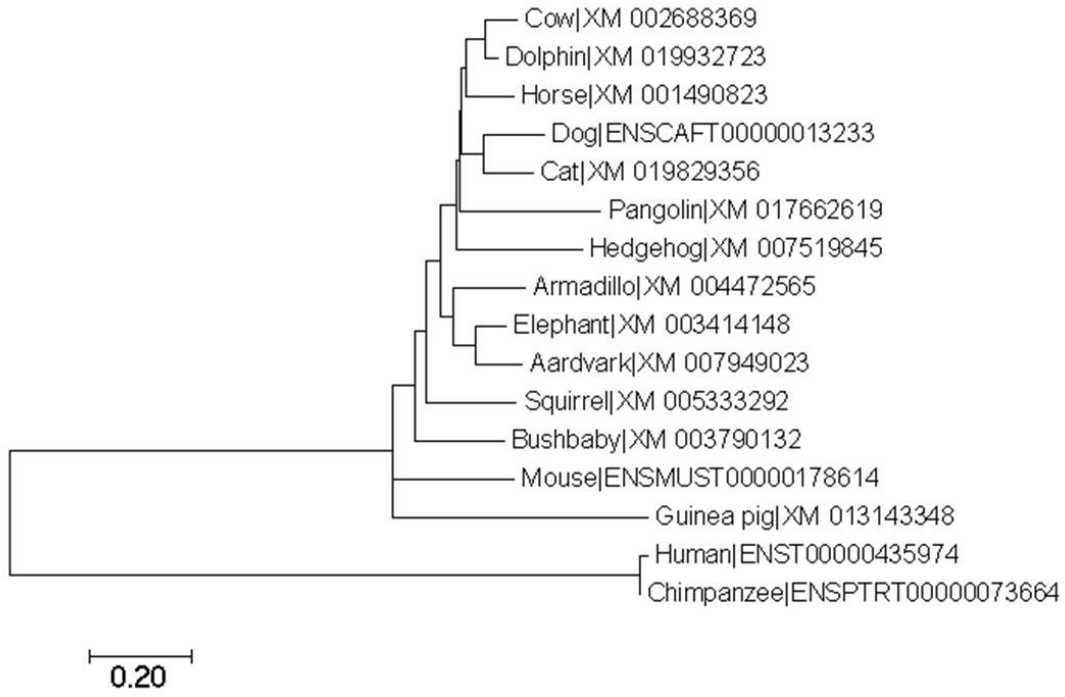
LAMB3


Dolphin|ENSTTRT00000017004
Minke whale|XM 007163957
Cow|ENSBTAT00000022005
Pangolin|XM 017641500
Horse|XM 023640784
Cat|XM 019821882
Dog|ENSCAFT00000018923
Human|ENST00000391911
Chimpanzee|ENSPTRT00000047179
Bushbaby|ENSOGAT00000026439
Mouse|ENSMUST00000194677
Guinea pig|ENSCPOT00000004250
Squirrel|ENSSTOT00000020073
Armadillo|ENSDNOT00000004444
Sloth|ENSCHOT00000006958
Hedgehog|XM 016191698
Elephant|ENSLAFT00000005222
Aardvark|XM 007953064
Platypus|XM 007670564
0.10

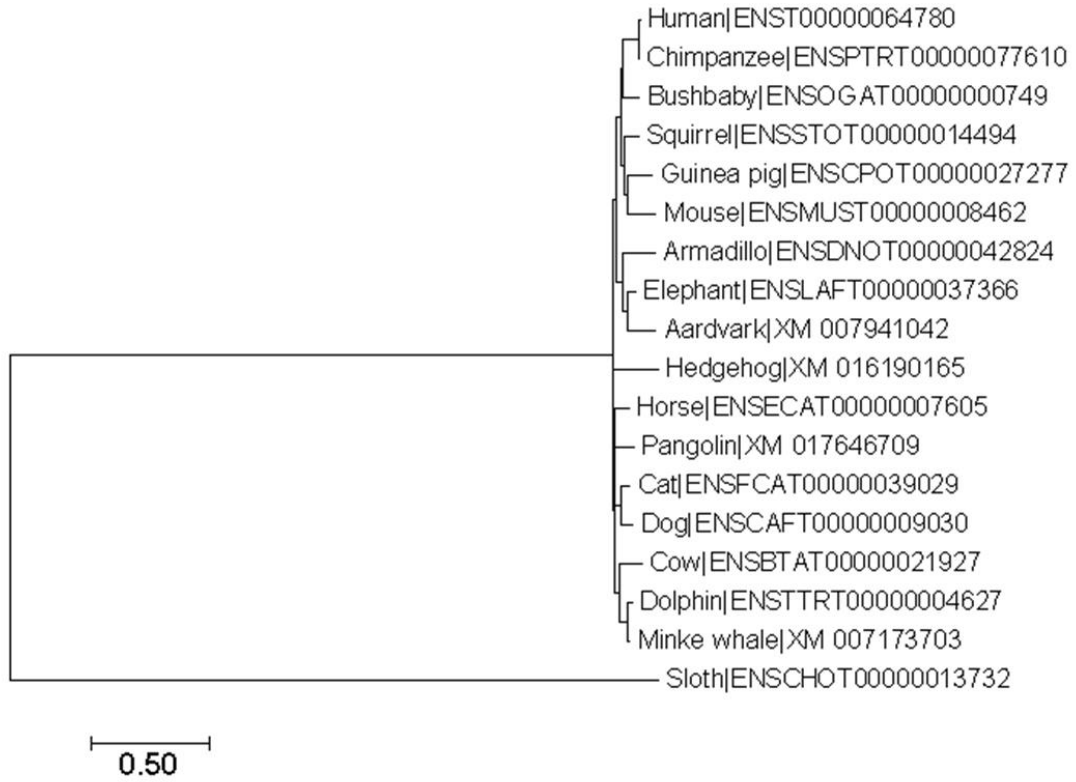
MMP20


Mouse|ENSMUST00000034487
Squirrel|ENSSTOT00000027785
Guinea pig|ENSCPOT00000033641
Human|ENST00000260228
Chimpanzee|ENSPTRT00000007863
Bushbaby|ENSOGAT00000015251
Armadillo|ENSDNOT00000006356
Sloth|ENSCHOT00000004213
Elephant|ENSLAFT00000010998
Aardvark|XM 007946011
Horse|ENSECAT00000012298
Cat|ENSFCAT00000015792
Dog|ENSCAFT00000023926
Cow|ENSBTAT00000018936
Dolphin|ENSTTRT00000014736
Minke whale|XM 007191222
Hedgehog|XM 007520728
Pangolin|XM 017676338
0.050

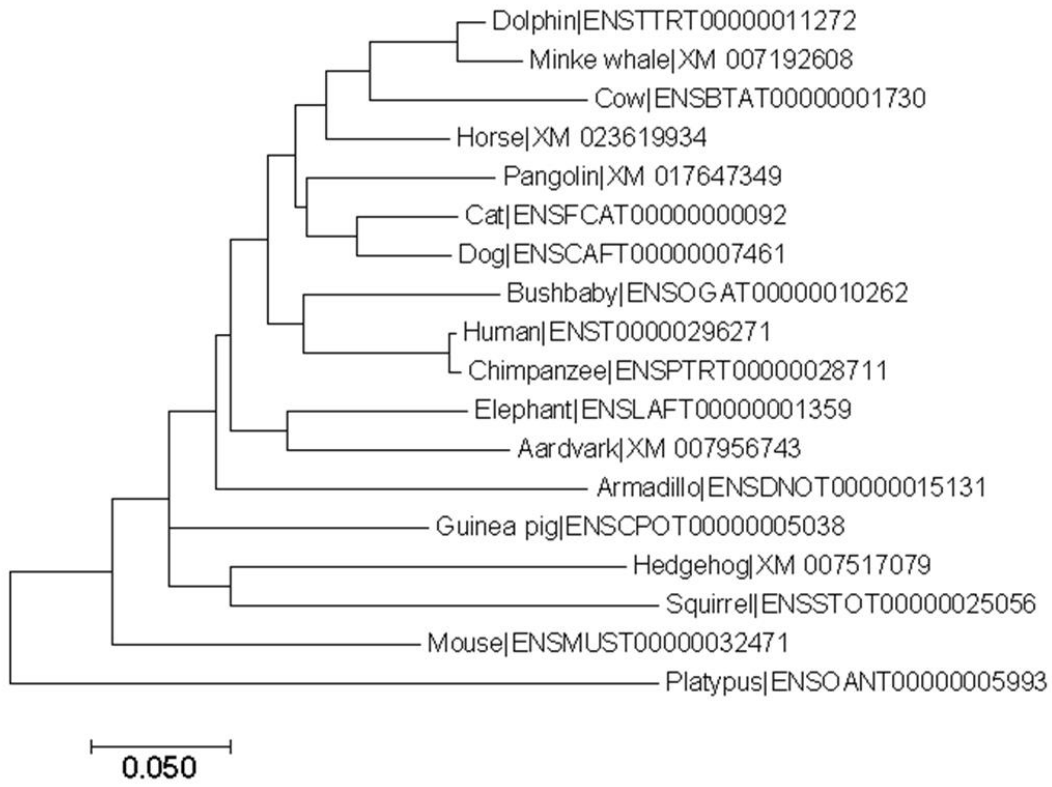
ODAPH


Cow|XM 002688369
Dolphin|XM 019932723
Horse|XM 001490823
Dog|ENSCAFT00000013233
Cat|XM 019829356
Pangolin|XM 017662619
Hedgehog|XM 007519845
Armadillo|XM 004472565
Elephant|XM 003414148
Aardvark|XM 007949023
Squirrel|XM 005333292
Bushbaby|XM 003790132
Mouse|ENSMUST00000178614
Guinea pig|XM 013143348
Human|ENST00000435974
Chimpanzee|ENSPTRT00000073664
0.20

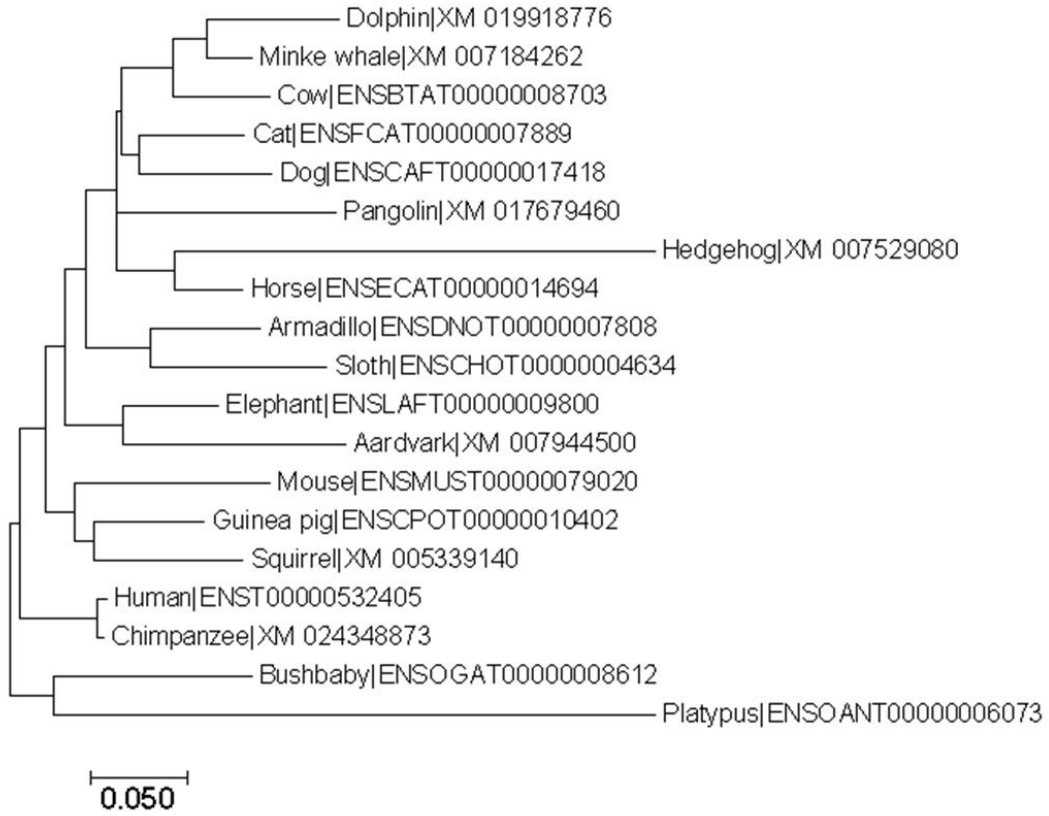
RELT


Human|ENST00000064780
Chimpanzee|ENSPTRT00000077610
Bushbaby|ENSOGAT00000000749
Squirrel|ENSSTOT00000014494
Guinea pig|ENSCPOT00000027277
Mouse|ENSMUST00000008462
Armadillo|ENSDNOT00000042824
Elephant|ENSLAFT00000037366
Aardvark|XM 007941042
Hedgehog|XM 016190165
Horse|ENSECAT00000007605
Pangolin|XM 017646709
Cat|ENSFCAT00000039029
Dog|ENSCAFT00000009030
Cow|ENSBTAT00000021927
Dolphin|ENSTTRT00000004627
Minke whale|XM 007173703
Sloth|ENSCHOT00000013732
0.50

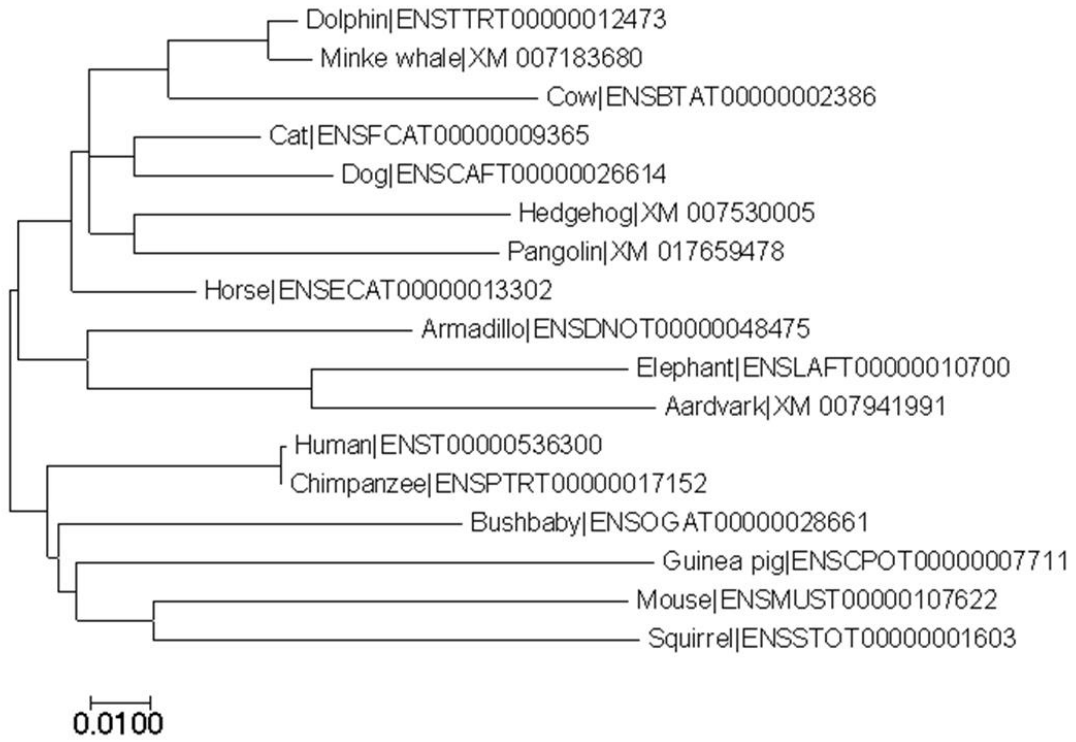
RHO


Dolphin|ENSTTRT00000011272
Minke whale|XM 007192608
Cow|ENSBTAT00000001730
Horse|XM 023619934
Pangolin|XM 017647349
Cat|ENSFCAT00000000092
Dog|ENSCAFT00000007461
Bushbaby|ENSOGAT00000010262
Human|ENST00000296271
Chimpanzee|ENSPTRT00000028711
Elephant|ENSLAFT00000001359
Aardvark|XM 007956743
Armadillo|ENSDNOT00000015131
Guinea pig|ENSCPOT00000005038
Hedgehog|XM 007517079
Squirrel|ENSSTOT00000025056
Mouse|ENSMUST00000032471
Platypus|ENSOANT00000005993
0.050

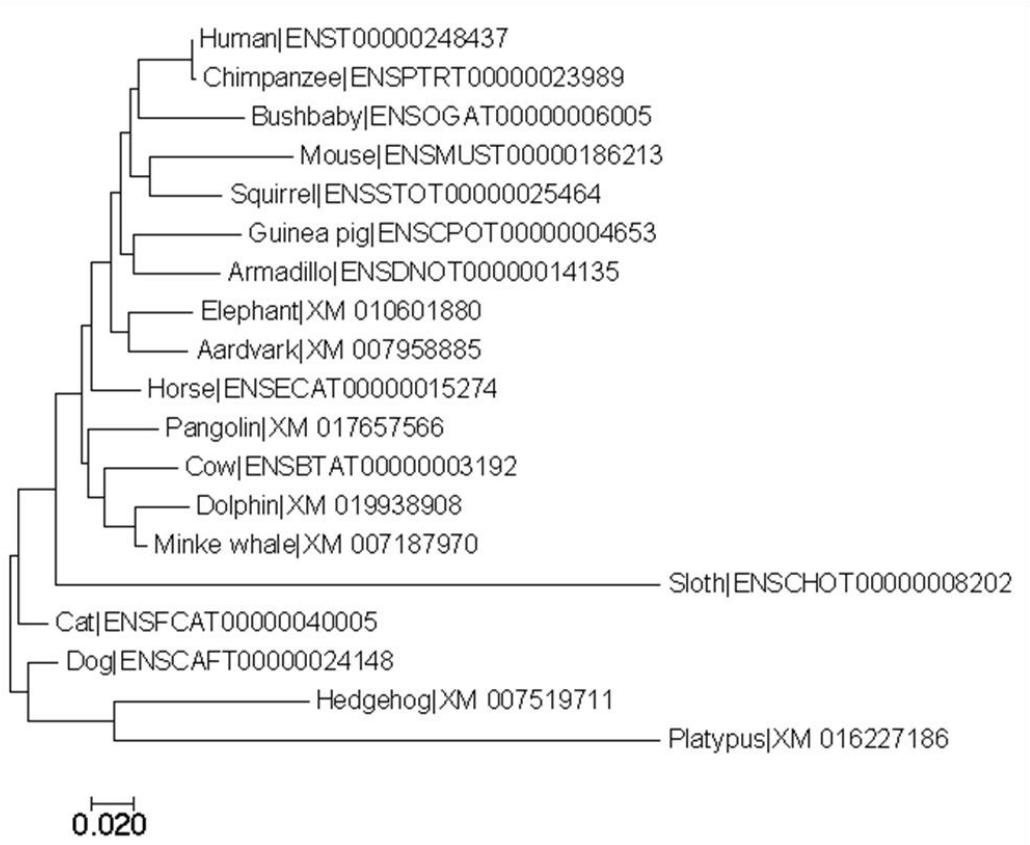
SLC24A4


Dolphin|XM 019918776
Minke whale|XM 007184262
Cow|ENSBTAT00000008703
Cat|ENSFCAT00000007889
Dog|ENSCAFT00000017418
Pangolin|XM 017679460
Hedgehog|XM 007529080
Horse|ENSECAT00000014694
Armadillo|ENSDNOT00000007808
Sloth|ENSCHOT00000004634
Elephant|ENSLAFT00000009800
Aardvark|XM 007944500
Mouse|ENSMUST00000079020
Guinea pig|ENSCPOT00000010402
Squirrel|XM 005339140
Human|ENST00000532405
Chimpanzee|XM 024348873
Bushbaby|ENSOGAT00000008612
Platypus|ENSOANT00000006073
0.050

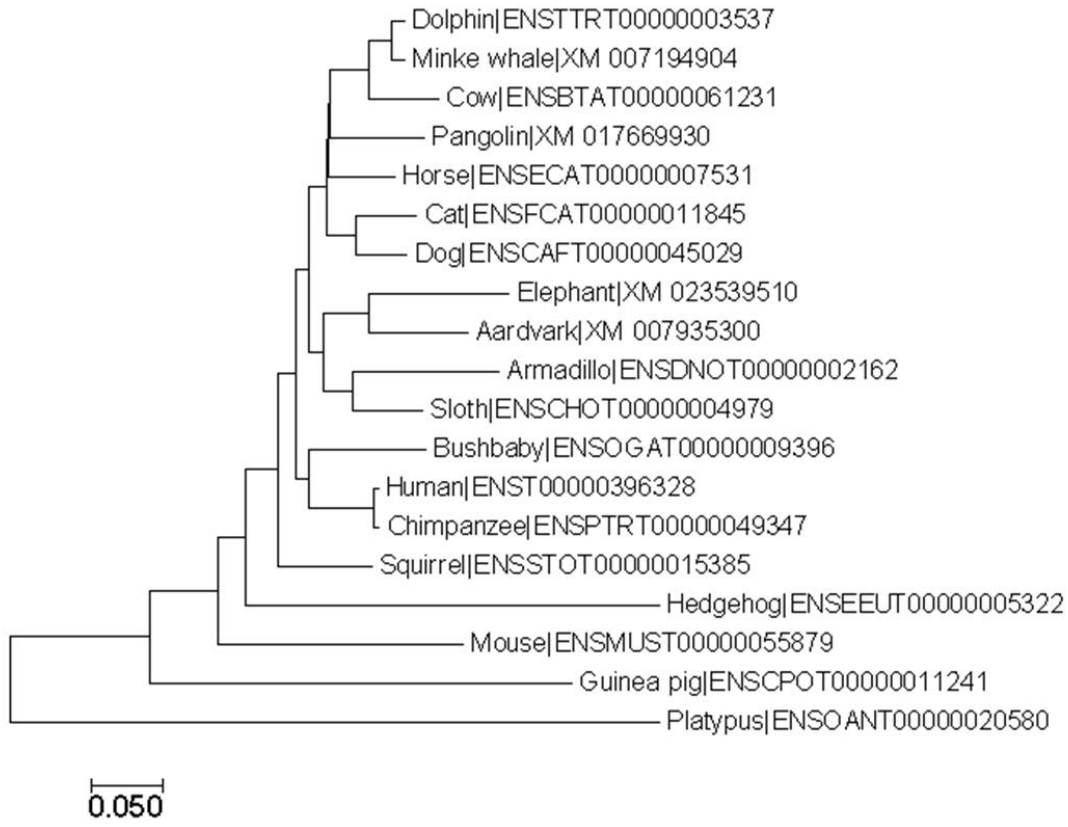
SP6


Dolphin|ENSTTRT00000012473
Minke whale|XM 007183680
Cow|ENSBTAT00000002386
Cat|ENSFCAT00000009365
Dog|ENSCAFT00000026614
Hedgehog|XM 007530005
Pangolin|XM 017659478
Horse|ENSECAT00000013302
Armadillo|ENSDNOT00000048475
Elephant|ENSLAFT00000010700
Aardvark|XM 007941991
Human|ENST00000536300
Chimpanzee|ENSPTRT00000017152
Bushbaby|ENSOGAT00000028661
Guinea pig|ENSCPOT00000007711
Mouse|ENSMUST00000107622
Squirrel|ENSSTOT00000001603
0.0100

TUBA4A



WDR72



F.2 Newick format of the phylogenetic trees presented in this study

Mammal Phylogeny (Morgan et al., 2013)
(((((Human:0.00306573,Chimpanzee:0.00466271):0.05951125,Bushbaby:0.11272390):0.00912071,(Squirrel:0.14859060,(Mouse:0.23603460,Guinea_pig:0.21377600):0.02024196):0.00802100):0.01283980,(((Horse:0.06712910,((((Minke_whale:0.00000000,Humpback_whale:0.00000000):0.00000000,Bowhead_whale:0.00000000):0.00000000,Dolphin:0.00000000):0.00000000,Cow:0.10169049):0.02807200):0.00397700,(Cat:0.07481390,Dog:0.06427730):0.03160440):0.00368924,(Hedgehog:0.21702700,Pangolin:0.00000000):0.00000000):0.01390640):0.01217960,((Aardvark:0.11693500,Elephant:0.06466650):0.05560330,(Sloth:0.07060950,Armadillo:0.07136050):0.03432370):0.00267120,Platypus:0.54085430);
ACP4
(((((((Cat|ENSFCAT00000019264:0.04899194,Dog|ENSCAFT00000004670:0.04615329):0.03902644,Horse|ENSECAT00000021666:0.31809084):0.01024049,(Cow|ENSBTAT00000020110:0.06236809,(Dolphin|ENSTTRT00000011060:0.02186318,Minke_whale|XM_007179955:0.10124710):0.03294680):0.04365256):0.03234982,(Elephant|ENSLAFT00000001738:0.03782331,Aardvark|XM_007957636:0.26307233):0.06796463):0.01569373,(Human|ENST00000270593:0.00369612,Chimpanzee|ENSPTRT00000021123:0.00218791):0.07775180):0.00000000,Hedgehog|XM_007531289:0.22845796):0.00569742,(Squirrel|ENSSTOT00000010097:0.10712281,(Guinea_pig|ENSCPOT00000027235:0.14906722,Mouse|ENSMUST00000118216:0.13862669):0.01786748):0.03079353):0.00305002,(Bushbaby|ENSOGAT00000004269:0.05668202,Platypus|ENSOANT00000002669:0.73840441):0.03721235,Armadillo|XM_023585546:0.80705651);
AMBN
(((((((Dolphin|ENSTTRT00000008662:0.01274338,Minke_whale|XM_007193563:0.01872823):0.01507939,Cow|ENSBTAT00000053975:0.05198400):0.02850077,Horse|ENSECAT00000023127:0.09821184):0.00695853,Hedgehog|ENSEEUT00000011199:0.18692174):0.00259616,(Pangolin|XM_017655633:0.18514833,(Cat|ENSFCAT00000009209:0.03947137,Dog|ENSCAFT00000004648:0.04566828):0.02747591):0.01163903):0.01900743,((Human|ENST00000322937:0.00152788,Chimpanzee|ENSPTRT00000045270:0.00000000):0.08842827,(Bushbaby|ENSOGAT00000035024:0.22264005,(Squirrel|ENSSTOT00000007955:0.07571927,(Guinea_pig|ENSCPOT00000001122:0.17542850,Mouse|ENSMUST00000198265:0.15029657):0.01607517):0.03793748):0.00695927):0.03431696):0.01586038,(Elephant|XM_010594187:0.09922675,Aardvark|XM_007942818:0.07985706):0.02104448):0.01181625,(Armadillo|ENSDNOT00000010867:0.05471858,Sloth|ENSCHOT00000011631:0.11019256):0.02770952,Platypus|ENSOANT00000021593:0.57255079);
AMELX
((((((Human|ENST00000380712:0.00000000,Chimpanzee|ENSPTRT00000065610:0.00000000):0.01875783,Bushbaby|XM_012807134:0.03250904):0.01550408,(Guinea_pig|ENSCPOT00000020303:0.03961246,(Mouse|ENSMUST00000066112:0.07795757,Squirrel|ENSSTOT00000003773:0.03309848):0.00462321):0.02995436):0.01637713,(Sloth|ENSCHOT00000006301:0.20504887,(Elephant|ENSLAFT00000033942:0.12708678,(Aardvark|XM_007951637:0.07873427,(Armadillo|ENSDNOT00000036350:0.06853020,Platypus|XM_001515065:0.31574904):0.01735235):0.01974395):0.00761517):0.00831859):0.01172174,(Hedgehog|ENSEEUT00000011820:0.07726124,((Cow|ENSBTAT00000016034:0.01000301,Dolphin|ENSTTRT00000005725:0.08848140):0.01894271,(Horse|ENSECAT00000017246:0.05622958,(Cat|ENSFCAT00000053505:0.00670634,Dog|ENSCAFT00000018403:0.00554400):0.00974705):0.00667125):0.00150092):0.00000000):0.00000000,Pangolin|XM_017676181:0.09108068,Minke_whale|XM_007198396:0.98545038);
AMTN
(((((((Dolphin|ENSTTRT00000008650:0.03477087,Minke_whale|XM_007193595:0.02418201):0.02253991,Cow|ENSBTAT00000003804:0.10156064):0.04074075,Horse|ENSECAT00000000460:0.09933500):0.00954995,Hedgehog|ENSEEUT00000000206:0.14336706):0.00000000,(Cat|ENSFCAT00000007316:0.05449356,Dog|ENSCAFT00000004629:0.08845306):0.05685459):0.02416425,(Elephant|ENSLAFT00000012608:0.08616243,Aardvark|XM_007942817:0.15256479):0.03756198):0.00298184,(((Human|ENST00000339336:0.00000000,Chimpanzee|ENSPTRT00000030049:0.01159218):0.08585899,Bushbaby|ENSOGAT00000003035:0.17622891):0.00736754,(Squirrel|ENSSTOT00000022909:0.07593300,(Guinea_pig|ENSCPOT0000

0001120:0.21974536,Mouse|ENSMUST00000073363:0.18795720):0.03265880):0.02518712):0.01594973,(Sloth|ENSCHOT00000000414:0.12838917,Armadillo|XM_023584403:0.28203361):0.05485444);
COL17A1
(((((((Dolphin|ENSTTRT00000007475:0.01911031,Minke_whale|XM_007172098:0.01523198):0.02931487,Cow|ENSBTAT00000013271:0.06017265):0.03966043,Horse|ENSECAT00000008760:0.07153069):0.00530523,Hedgehog|XM_007530465:0.15208801):0.00118690,(Pangolin|XM_017660486:0.23585700,(Cat|ENSFCAT00000004301:0.05875406,Dog|ENSCAFT00000045026:0.12363505):0.03295127):0.00000000):0.01792810,(((Human|ENST00000353479:0.00407399,Chimpanzee|ENSPTRT00000005563:0.00349819):0.05799525,Bushbaby|ENSOGAT00000017134:0.09392667):0.01155443,(Mouse|ENSMUST00000026045:0.14566473,(Guinea_pig|ENSCPOT00000001005:0.22495424,Squirrel|ENSSTOT00000013371:0.06694944):0.00791902):0.02342829):0.01236165):0.00592987,(Armadillo|ENSDNOT00000000450:0.15318827,(Elephant|ENSLAFT00000028366:0.07392790,Aardvark|XM_007939867:0.06685780):0.02498233):0.00753643,Platypus|ENSOANT00000002604:0.44524964);
DLX3
(((((((Cow|ENSBTAT00000023142:0.02610937,Minke_whale|XM_007176422:0.02321994):0.01935758,Pangolin|XM_017642416:0.03382682):0.00372091,Horse|ENSECAT00000015627:0.02301752):0.00000000,(Cat|ENSFCAT00000066707:0.03706163,Dog|ENSCAFT00000026875:0.02116357):0.02341132):0.00388720,(Armadillo|ENSDNOT00000019409:0.04207278,Sloth|ENSCHOT00000002771:0.02865405):0.03083375):0.00797613,((Human|ENST00000434704:0.00000000,Chimpanzee|ENSPTRT00000017217:0.00000000):0.03381924,(Mouse|ENSMUST00000092768:0.09249450,(Squirrel|ENSSTOT00000010308:0.04917784,(Hedgehog|XM_007525618:0.05516871,Guinea_pig|ENSCPOT00000025370:0.09065732):0.01640058):0.00698428):0.00293442):0.00351842):0.00588907,Bushbaby|ENSOGAT00000005139:0.04111748):0.00627251,Aardvark|XM_007941926:0.04792516,Elephant|ENSLAFT00000036461:0.41720711);
ENAM
((((((Cat|ENSFCAT00000028994:0.03826433,Dog|ENSCAFT00000004659:0.04847747):0.02622024,(Horse|XM_001487894:0.07031872,Pangolin|XM_017655632:0.22031634):0.00237722):0.00114791,(Hedgehog|XM_007523530:0.13671592,(Cow|ENSBTAT00000013661:0.07020973,(Dolphin|XM_019932671:0.01583018,Minke_whale|XM_007193564:0.01395722):0.01921352):0.03027851):0.00305843):0.02236729,(((Human|ENST00000396073:0.00277806,Chimpanzee|ENSPTRT00000091821:0.00494637):0.06412507,Bushbaby|ENSOGAT00000003041:0.13822442):0.01314406,(Squirrel|XM_005333299:0.08255515,(Guinea_pig|XM_003467537:0.19313886,Mouse|ENSMUST00000031222:0.19432938):0.01215801):0.03736667):0.01214003):0.00838927,Armadillo|XM_004477547:0.11223642):0.00997562,(Elephant|ENSLAFT00000026415:0.06285672,Aardvark|XM_007942773:0.08545938):0.02067262,Platypus|XM_007669232:0.77342654);
FAM20A
(((((((Elephant|ENSLAFT00000000376:0.04893423,Aardvark|XM_007959693:0.05191114):0.02877345,(Armadillo|ENSDNOT00000017961:0.06021667,Sloth|ENSCHOT00000013721:0.05672382):0.01286295):0.01067990,Horse|ENSECAT00000025399:0.05065940):0.00084865,(Pangolin|XM_017641924:0.06588853,(Bushbaby|ENSOGAT00000015626:0.05771418,(Human|ENST00000592554:0.00596070,Chimpanzee|ENSPTRT00000017563:0.00276921):0.04080865):0.01496948):0.00956904):0.00154326,(Guinea_pig|ENSCPOT00000010541:0.09391931,(Mouse|ENSMUST00000020938:0.12290549,Squirrel|ENSSTOT00000023494:0.05519808):0.01514769):0.01539940):0.01422008,(Hedgehog|ENSEEUT00000007308:0.11098002,(Cat|ENSFCAT00000011004:0.05174941,Dog|ENSCAFT00000017481:0.04822906):0.02257368):0.00675735):0.05701145,(Cow|ENSBTAT00000001439:0.09388455,(Dolphin|ENSTTRT00000012600:0.00925028,Minke_whale|XM_007198504:0.21664012):0.05172227):0.02797252,Platypus|ENSOANT00000012832:0.29992312);
FAM83H
(((((((Dolphin|ENSTTRT00000015190:0.03630194,Minke_whale|XM_007166933:0.02360264):0.05854550,Cow|ENSBTAT00000019225:0.08463762):0.04119063,Horse|XM_023649125:0.06290811):0.00551990,(Pangolin|XM_017646837:0.13362255,(Cat|ENSFCAT00000029161:0.05659758,Dog|ENSCAFT00000049368:0.10516615):0.03620170):0.00807437):0.00000010,Hedgehog|ENSEEUT00000005203:0.19169430):0.01467213,((Human|ENST00000388913:0.00490670,Chimpanzee|ENSPTRT00000059775:0.00638380):0.0

9969596,(Bushbaby|ENSOGAT00000027713:0.11169564,(Squirrel|ENSSTOT00000023921:0.10755321,(Guinea_pig|ENSCPOT00000013453:0.15829154,Mouse|ENSMUST00000170153:0.16471139):0.01451311):0.03613114):0.01840620):0.02463092):0.05579046,(Elephant|ENSLAFT00000036487:0.23241819,Aardvark|XM_007956087:0.08711122):0.02423571):0.14773343,Armadillo|XM_004454793:0.14324401,Platypus|ENSOANT00000000665:0.52676689);
GPR68
(((((Elephant|ENSLAFT00000014724:0.05782125,Aardvark|XM_007944490:0.03374587):0.03246045,Bushbaby|ENSOGAT00000013869:0.07378818):0.01059216,(Squirrel|ENSSTOT00000019539:0.03931625,(Guinea_pig|ENSCPOT00000031707:0.09258814,Mouse|ENSMUST00000110066:0.11880836):0.00771062):0.00816759):0.00847655,((Human|ENST00000531499:0.00045362,Chimpanzee|ENSPTRT00000089266:0.00578679):0.05021878,(Horse|ENSECAT00000004904:0.05214355,Pangolin|XM_017664233:0.03883021):0.00804364):0.00376742):0.00000154,(Armadillo|ENSDNOT00000004198:0.13134130,((Cow|ENSBTAT00000008858:0.04868368,Minke_whale|XM_007180867:0.05478720):0.01322983,(Hedgehog|XM_007518419:0.37268047,(Cat|XM_023255960:0.03894804,Dog|ENSCAFT00000027822:0.04461625):0.02901552):0.00656971):0.01632731):0.00810120,Platypus|ENSOANT00000006046:0.56264917);
ITGB6
(((((((Dolphin|ENSTTRT00000011042:0.00939003,Minke_whale|XM_007183159:0.00803454):0.02305669,Cow|ENSBTAT00000011972:0.05062519):0.02653958,(Cat|ENSFCAT00000032085:0.02916615,Dog|ENSCAFT00000015738:0.02501302):0.01268801):0.00269467,Horse|ENSECAT00000024051:0.04373223):0.01018045,((Elephant|ENSLAFT00000001853:0.03873142,Aardvark|XM_007941832:0.04147328):0.02685337,(Armadillo|ENSDNOT00000019549:0.04364302,Sloth|ENSCHOT00000013138:0.04065566):0.02382888):0.00634953):0.00456095,(Pangolin|XM_017649894:0.25669011,(Hedgehog|ENSEEUT00000013388:0.09447985,(Bushbaby|ENSOGAT00000002899:0.06178589,(Human|ENST00000283249:0.00082312,Chimpanzee|ENSPTRT00000107705:0.18459441):0.03891327):0.00865649):0.00204238):0.00024775):0.00911149,(Mouse|ENSMUST00000059888:0.14615748,Squirrel|ENSSTOT00000008227:0.04119850):0.01170749):0.00935417,Guinea_pig|ENSCPOT00000024774:0.06620688,Platypus|ENSOANT00000009855:0.33568531);
KLK4
((((((Cat|ENSFCAT00000008263:0.05521994,Dog|ENSCAFT00000004649:0.04396480):0.01302142,Horse|ENSECAT00000019908:0.10794068):0.00613219,(Cow|ENSBTAT00000027207:0.05863085,(Dolphin|ENSTTRT00000005736:0.01859299,Minke_whale|XM_007198394:0.06285084):0.03639919):0.04627538):0.00021084,Hedgehog|ENSEEUT00000013116:0.26147868):0.01818407,(Bushbaby|ENSOGAT00000031952:0.09901395,(Human|ENST00000324041:0.00000000,Chimpanzee|ENSPTRT00000021136:0.00261343):0.07260022):0.03009339):0.03754026,(Squirrel|ENSSTOT00000002320:0.08875401,(Guinea_pig|ENSCPOT00000034808:0.22856092,Mouse|ENSMUST00000007161:0.25069013):0.02537496):0.00662614,Elephant|ENSLAFT00000029708:2.19825775);
LAMA3
(((((((Dolphin|ENSTTRT00000008885:0.01236952,Minke_whale|XM_007197638:0.01917892):0.03327738,Cow|ENSBTAT00000060991:0.07436684):0.03298649,Horse|XM_023647569:0.05724488):0.00314560,(Cat|ENSFCAT00000009603:0.05783230,Dog|ENSCAFT00000028869:0.05530440):0.02836632):0.00628196,Hedgehog|XM_007536377:0.16145097):0.01147150,(((Human|ENST00000313654:0.00335749,Chimpanzee|ENSPTRT00000018220:0.00325115):0.06039748,Bushbaby|ENSOGAT00000014541:0.12394186):0.00961370,(Guinea_pig|ENSCPOT00000007693:0.24997782,(Mouse|ENSMUST00000092070:0.17396422,Squirrel|ENSSTOT00000006758:0.09704282):0.00493455):0.02225974):0.01439772):0.01669813,(Elephant|ENSLAFT00000008286:0.06335782,Aardvark|XM_007935580:0.07458064):0.03631492):0.00758868,(Armadillo|ENSDNOT00000003516:0.10231467,Sloth|ENSCHOT00000009404:0.05535861):0.02377788,Platypus|XM_016227434:0.58267073);
LAMB3
(((((((((Dolphin|ENSTTRT00000017004:0.01910323,Minke_whale|XM_007163957:0.01532862):0.02258331,Cow|ENSBTAT00000022005:0.09146302):0.02321508,Pangolin|XM_017641500:0.16026768):0.00444703,Horse|XM_023640784:0.05957873):0.00468491,(Cat|XM_019821882:0.04598667,Dog|ENSCAFT00000018923:0.06972650):0.03499352):0.01797841,(((Human|ENST00000391911:0.00456025,Chimpanzee|

ENSPTRT00000047179:0.00547627):0.06395257,Bushbaby|ENSOGAT00000026439:0.11688157):0.00443134,(Mouse|ENSMUST00000194677:0.17097484,(Guinea_pig|ENSCPOT00000004250:0.12529513,Squirrel|ENSSTOT00000020073:0.07724255):0.00951008):0.01687255):0.01043815):0.01079464,(Armadillo|ENSDNOT00000004444:0.15474023,Sloth|ENSCHOT00000006958:0.04721117):0.03926377):0.00822587,Hedgehog|XM_016191698:0.49221720):0.00876544,(Elephant|ENSLAFT00000005222:0.08222103,Aardvark|XM_007953064:0.07064144):0.00865703,Platypus|XM_007670564:0.59207277);
MMP20
(((((((Mouse|ENSMUST00000034487:0.11336756,Squirrel|ENSSTOT00000027785:0.04588275):0.00976064,Guinea_pig|ENSCPOT00000033641:0.08912727):0.01391126,(Human|ENST00000260228:0.00148133,Chimpanzee|ENSPTRT00000007863:0.00434743):0.04592706):0.00437381,Bushbaby|ENSOGAT00000015251:0.08000858):0.01403398,(Armadillo|ENSDNOT00000006356:0.16513373,(Sloth|ENSCHOT00000004213:0.13024502,(Elephant|ENSLAFT00000010998:0.03604020,Aardvark|XM_007946011:0.08471575):0.02750512):0.00337963):0.00995682):0.00264671,(Horse|ENSECAT00000012298:0.05785269,((Cat|ENSFCAT00000015792:0.06342398,Dog|ENSCAFT00000023926:0.07750240):0.04602014,(Cow|ENSBTAT00000018936:0.05924551,(Dolphin|ENSTTRT00000014736:0.00535521,Minke_whale|XM_007191222:0.01858093):0.01753092):0.01498344):0.00848406):0.00608167):0.01251140,Hedgehog|XM_007520728:0.06851681,Pangolin|XM_017676338:0.58382448);
ODAPH
((((((((((Cow|XM_002688369:0.06220934,Dolphin|XM_019932723:0.02394218):0.03674040,Horse|XM_001490823:0.09086039):0.00997596,(Dog|ENSCAFT00000013233:0.11883463,Cat|XM_019829356:0.09595705):0.04282285):0.00387966,Pangolin|XM_017662619:0.27490101):0.00735810,Hedgehog|XM_007519845:0.24537225):0.02991149,(Armadillo|XM_004472565:0.14166332,(Elephant|XM_003414148:0.06056426,Aardvark|XM_007949023:0.09133986):0.04361970):0.02316331):0.02821277,Squirrel|XM_005333292:0.17468413):0.02171409,Bushbaby|XM_003790132:0.17259710):0.04544041,Mouse|ENSMUST00000178614:0.23743482):0.00000011,Guinea_pig|XM_013143348:0.49776138,(Human|ENST00000435974:0.01546012,Chimpanzee|ENSPTRT00000073664:0.00000000):1.97324313);
RELT
((((((Human|ENST00000064780:0.00447356,Chimpanzee|ENSPTRT00000077610:0.00109489):0.06493899,Bushbaby|ENSOGAT00000000749:0.06894154):0.00828618,(Squirrel|ENSSTOT00000014494:0.06057878,(Guinea_pig|ENSCPOT00000027277:0.10884368,Mouse|ENSMUST00000008462:0.12167262):0.01240428):0.01163185):0.02086787,(Armadillo|ENSDNOT00000042824:0.14206931,(Elephant|ENSLAFT00000037366:0.03497040,Aardvark|XM_007941042:0.12111924):0.02239567):0.02452588):0.01754356,Hedgehog|XM_016190165:0.19585973):0.00000000,((Horse|ENSECAT00000007605:0.06198381,Pangolin|XM_017646709:0.07989480):0.00261273,((Cat|ENSFCAT00000039029:0.03606974,Dog|ENSCAFT00000009030:0.05030624):0.03027884,(Cow|ENSBTAT00000021927:0.09117178,(Dolphin|ENSTTRT00000004627:0.02222652,Minke_whale|XM_007173703:0.01248149):0.02932618):0.02523621):0.00080029):0.00724076,Sloth|ENSCHOT00000013732:5.31594609);
RHO
((((((((((Dolphin|ENSTTRT00000011272:0.00984727,Minke_whale|XM_007192608:0.02297314):0.03135114,Cow|ENSBTAT00000001730:0.07772625):0.01564123,Horse|XM_023619934:0.04418946):0.01088197,(Pangolin|XM_017647349:0.06785045,(Cat|ENSFCAT00000000092:0.03597936,Dog|ENSCAFT00000007461:0.03370194):0.01827288):0.00364462):0.01019347,(Bushbaby|ENSOGAT00000010262:0.07025297,(Human|ENST00000296271:0.00245118,Chimpanzee|ENSPTRT00000028711:0.00422635):0.05212756):0.01293289):0.01356587,(Elephant|ENSLAFT00000001359:0.06472374,Aardvark|XM_007956743:0.07970750):0.02038960):0.00520843,Armadillo|ENSDNOT00000015131:0.13336045):0.01639822,Guinea_pig|ENSCPOT00000005038:0.09270408):0.00000000,(Hedgehog|XM_007517079:0.14181048,Squirrel|ENSSTOT00000025056:0.15359381):0.02160650):0.02040793,Mouse|ENSMUST00000032471:0.11009863,Platypus|ENSOANT00000005993:0.26895303);
SLC24A4
(((((((((Dolphin|XM_019918776:0.06833785,Minke_whale|XM_007184262:0.02363043):0.01678340,Cow|ENSBTAT00000008703:0.04945045):0.02645885,(Cat|ENSFCAT00000007889:0.05409723,Dog|ENSCAFT00000017418:0.06816225):0.00889386):0.00297946,Pangolin|XM_017679460:0.11249986):0.00000000,

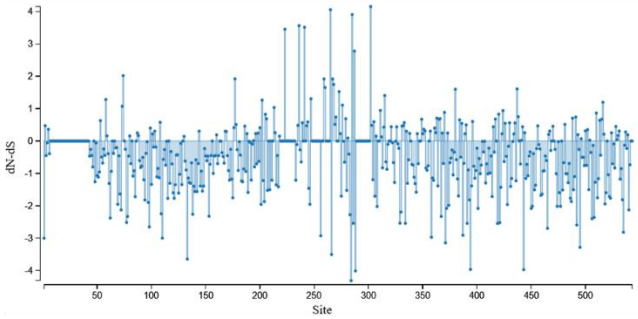
(Hedgehog|XM_007529080:0.24629399,Horse|ENSECAT00000014694:0.03495387):0.03009714):0.0160 1389,(Armadillo|ENSDNOT00000007808:0.05604677,Sloth|ENSCHOT00000004634:0.09055760):0.03334 166):0.01036027,(Elephant|ENSLAFT00000009800:0.04882843,Aardvark|XM_007944500:0.11401057):0. 02956979):0.01042313,(Mouse|ENSMUST00000079020:0.10002410,(Guinea_pig|ENSCPOT00000010402 :0.05690198,Squirrel|XM_005339140:0.07688784):0.00958546):0.01528260):0.01278411,(Human|ENST 00000532405:0.00461845,Chimpanzee|XM_024348873:0.00286952):0.03996460,(Bushbaby|ENSOGAT0 0000008612:0.10135354,Platypus|ENSOANT00000006073:0.30855236):0.02777212);
SP6
((((((Dolphin|ENSTTRT00000012473:0.00485616,Minke_whale|XM_007183680:0.00732720):0.01651903 ,Cow|ENSBTAT00000002386:0.06112135):0.01299986,(Cat|ENSFCAT00000009365:0.02098026,Dog|ENS CAFT00000026614:0.03292623):0.00735022):0.00000000,(Hedgehog|XM_007530005:0.06220382,Pango lin|XM_017659478:0.06038766):0.00731002):0.00295440,Horse|ENSECAT00000013302:0.02048318):0.0 0894148,(Armadillo|ENSDNOT00000048475:0.05392853,(Elephant|ENSLAFT00000010700:0.05237068,A ardvark|XM_007941991:0.05684823):0.03711980):0.01138865,((Human|ENST00000536300:0.00098229 ,Chimpanzee|ENSPTRT00000017152:0.00000000):0.03858607,(Bushbaby|ENSOGAT00000028661:0.0669 9231,(Guinea_pig|ENSCPOT00000007711:0.09567833,(Mouse|ENSMUST00000107622:0.07860841,Squir rel|ENSSTOT00000001603:0.08047967):0.01266449):0.00306877):0.00164348):0.00760044);
TUBA4A
(((((((((Human|ENST00000248437:0.00000000,Chimpanzee|ENSPTRT00000023989:0.00157474):0.02570 566,Bushbaby|ENSOGAT00000006005:0.05023835):0.00450312,(Mouse|ENSMUST00000186213:0.0689 1760,Squirrel|ENSSTOT00000025464:0.03506651):0.00938103):0.00401897,(Guinea_pig|ENSCPOT00000 004653:0.05145604,Armadillo|ENSDNOT00000014135:0.04171111):0.00583605):0.00469136,(Elephant| XM_010601880:0.03114583,Aardvark|XM_007958885:0.02871705):0.00795093):0.00998692,Horse|ENS ECAT00000015274:0.02391431):0.00418395,(Pangolin|XM_017657566:0.03323794,(Cow|ENSBTAT0000 0003192:0.03453412,(Dolphin|XM_019938908:0.02593169,Minke_whale|XM_007187970:0.00563291): 0.01431231):0.00785157):0.00309057):0.01267608,Sloth|ENSCHOT00000008202:0.29102113):0.018089 30,Cat|ENSFCAT00000040005:0.01433126,(Dog|ENSCAFT00000024148:0.01427109,(Hedgehog|XM_007 519711:0.09328584,Platypus|XM_016227186:0.26251583):0.04161674):0.01241956);
WDR72
((((((((((Dolphin|ENSTTRT00000003537:0.00839303,Minke_whale|XM_007194904:0.00909923):0.01575 978,Cow|ENSBTAT00000061231:0.04809528):0.02727611,Pangolin|XM_017669930:0.06523822):0.0007 0935,Horse|ENSECAT00000007531:0.04582337):0.00090983,(Cat|ENSFCAT00000011845:0.04186371,Do g|ENSCAFT00000045029:0.03509913):0.01978513):0.01296833,((Elephant|XM_023539510:0.09639606, Aardvark|XM_007935300:0.06749319):0.03127598,(Armadillo|ENSDNOT00000002162:0.10098689,Sloth |ENSCHOT00000004979:0.04815656):0.01980803):0.01060185):0.00866648,(Bushbaby|ENSOGAT00000 009396:0.08119075,(Human|ENST00000396328:0.00315757,Chimpanzee|ENSPTRT00000049347:0.0040 5107):0.04508584):0.00831686):0.01203308,Squirrel|ENSSTOT00000015385:0.06408512):0.02260928,H edgehog|ENSEEUT00000005322:0.28573753):0.01879029,Mouse|ENSMUST00000055879:0.16839693):0 .04769515,Guinea_pig|ENSCPOT00000011241:0.29187526,Platypus|ENSOANT00000020580:0.54492693);

Appendix G

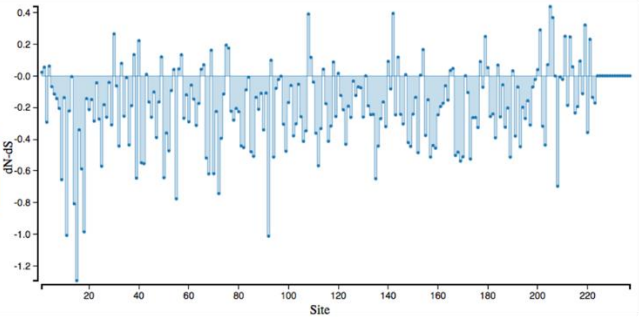
G.1 SLAC analysis results

The x axis represents the position on the sequence, the y axis shows the dN-dS value for each site. Positive difference indicates positive selection, and a negative value indicates purifying selection. Detailed values are in the electronic Appendix.

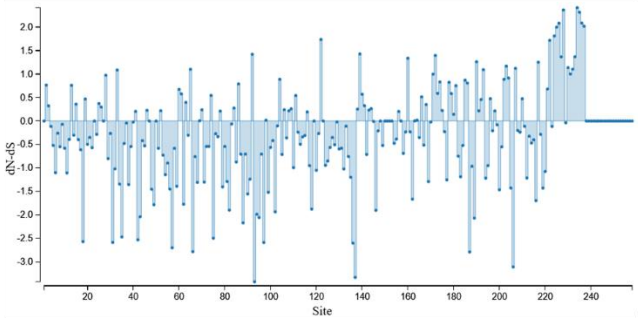
AMBN



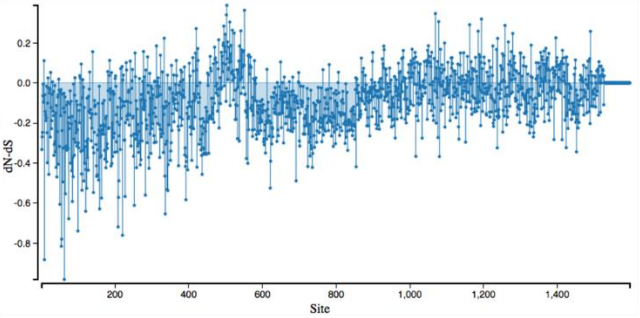
AMELX



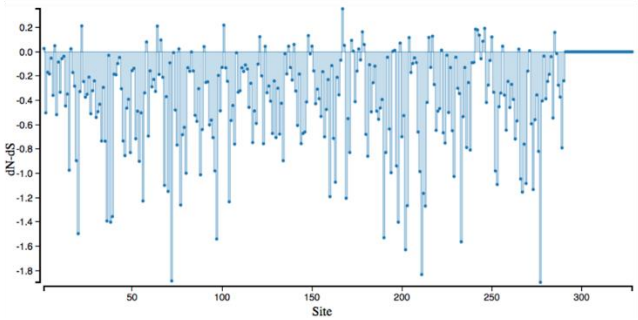
AMTN



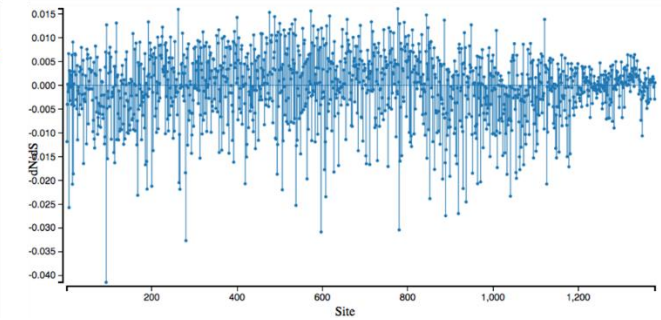
COL17A1



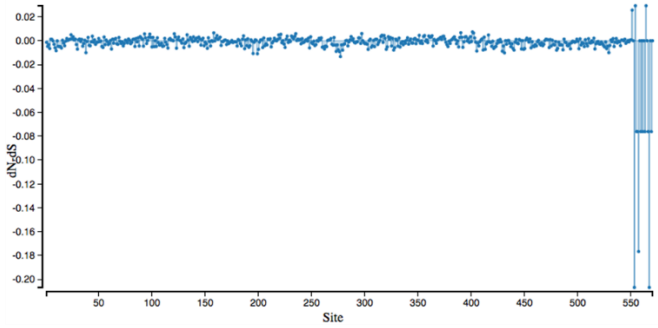
DLX3



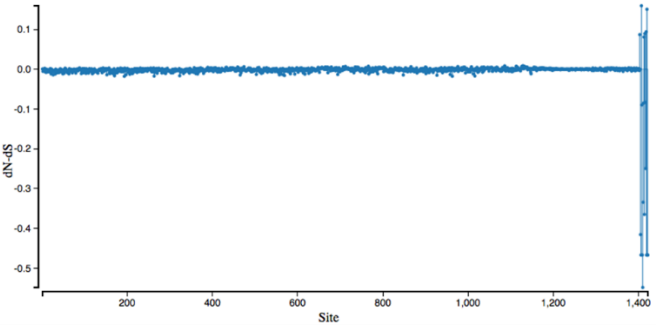
ENAM



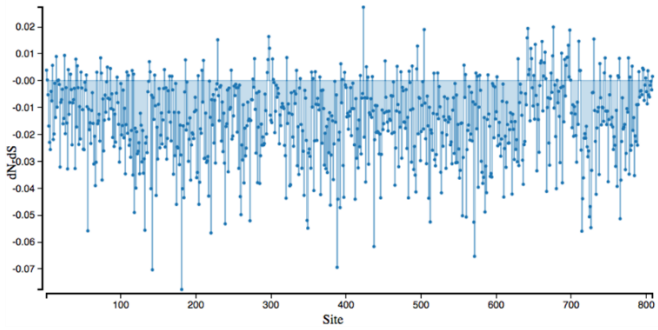
FAM20A



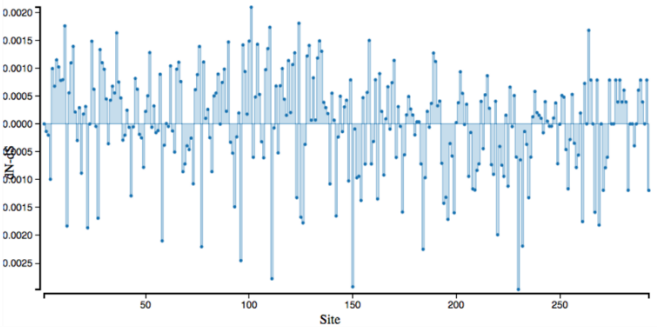
FAM83H



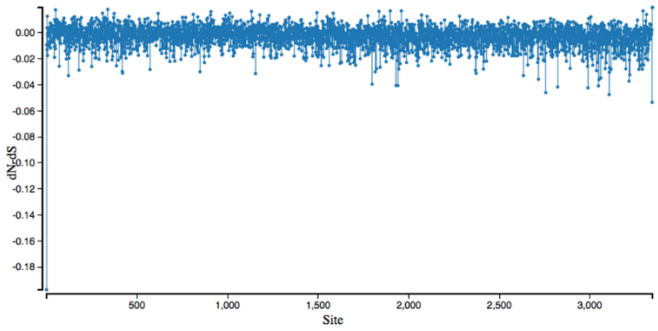
ITGB6



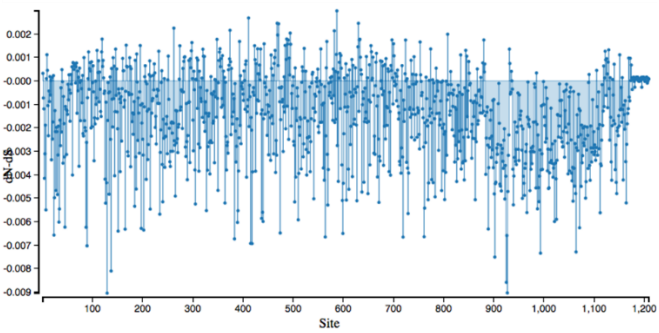
KLK4



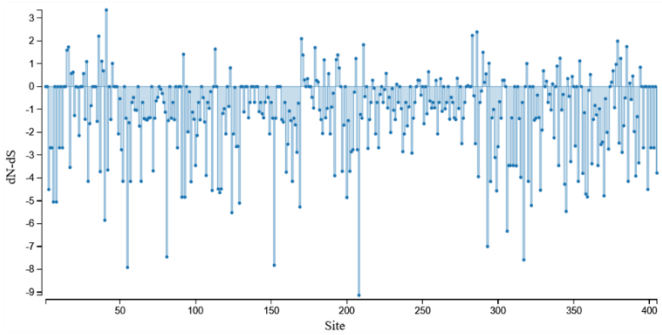
LAMA3



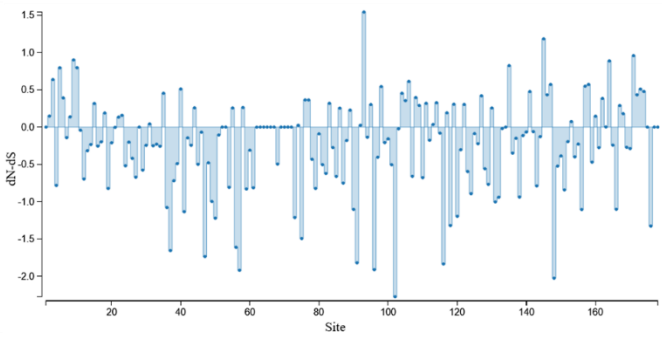
LAMB3



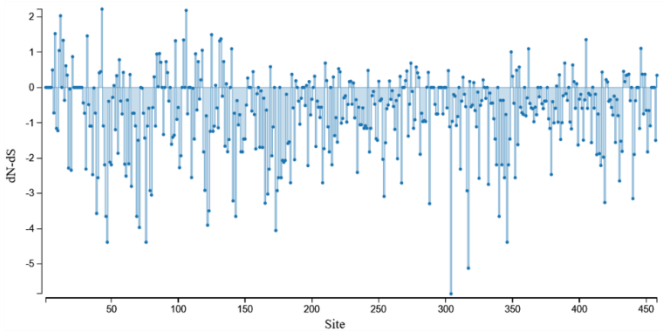
MMP20



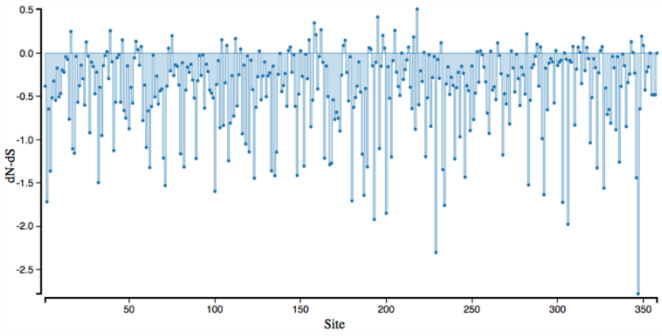
ODAPH



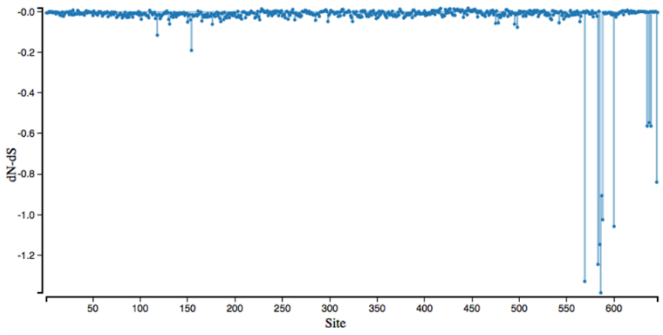
RELT



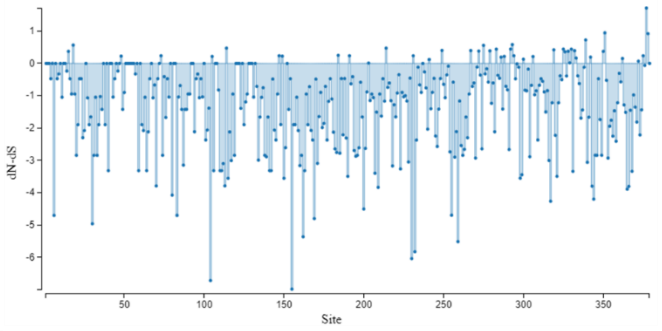
RHO



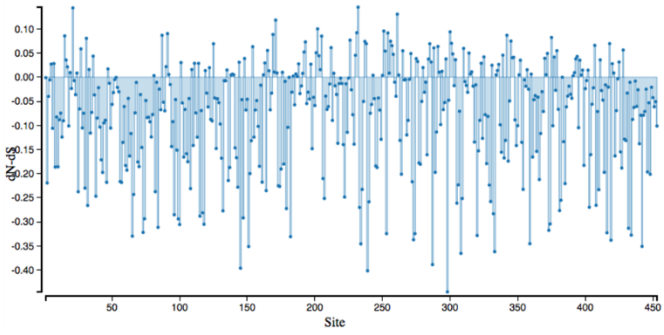
SLC24A4



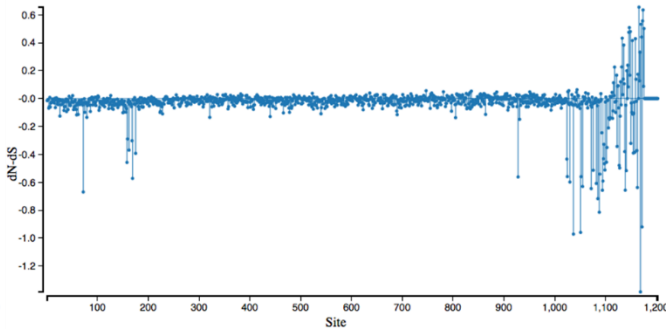
SP6



TUBA4A



WDR72



Appendix H

H.1 Certificate of donation of samples from Chester Zoo

CHESTER ZOO

12th February 2021

Georgios Nikolopoulos

CERTIFICATE DONATION

This is to certify that the following faecal sample ownership is being transferred to:

Georgios Nikolopoulos
Level 8 Welcome trust Brenner Building
St James's University Hospital
Beckett Street
Leeds, LS9 7TF

I can confirm that these samples have come from animals (aardvark, two-toed sloth, giant anteater) which currently reside at The North of England Zoological Society.

This is a **DONATION** from the North of England Zoological Society to the University of Leeds who now own these samples. The North of England Zoological Society request that the samples only be used in the study to understanding the convergent evolution of tooth/enamel loss in mammals at the molecular level.

I can confirm that no money is changing hands in this non-commercial transaction.

John O'Hanlon
Laboratory Technician
Sciences Division

Appendix I

I.1 Inserts constructed for the dual luciferase analysis, attempt 1

ACP4	S	agcttAGAGAAGGCCTAGCTGTCTGGGGGTa
	AS	gatctACCCCCAGACAGCTAGGCCTTCTCTa
	UGG-S	agcttAGAGAAGGCCTGGCTGTCTGGGGGTa
	UGG-AS	gatctACCCCCAGACAGCCAGGCCTTCTCTa
	Stop	agcttAGAGAAGGCCTAGTAATCTGGGGGTa
	Stop-AS	gatctACCCCCAGATTACTAGGCCTTCTCTa
AQP4	S	agcttATCAGTATGACTAGAAGATCGCa
	AS	gatctGCGATCTTCTAGTCATACTGATa
	UGG-S	agcttATCAGTATGGCTAGAAGATCGCa
	UGG-AS	gatctGCGATCTTCTAGCCATACTGATa
	Stop	agcttATCAGTATGATAAGAAGATCGCa
	Stop-AS	gatctGCGATCTTCTTATCATACTGATa
ENAM-1	S	agcttTGGGCCTTTTTACTGAAATCAACAAATTa
	AS	gatctAATTTGTTGATTTCAGTAAAAAGGCCCAa
	UGG-S	agcttTGGGCCTTTTTACTGGAATCAACAAATTa
	UGG-AS	gatctAATTTGTTGATTCCAGTAAAAAGGCCCAa
ENAM-2	S	agcttTGGACTTAATAAATGAAACTGTAAACTGa
	AS	gatctCAGTTTACAGTTTCATTTATTAAGTCCAa
	UGG-S	agcttTGGACTTAATAAATGGAACTGTAAACTGa
	UGG-AS	gatctCAGTTTACAGTTCCATTTATTAAGTCCAa
MMP20-1	S	agcttCCTCTTCCCAGGTTAACCCAAATGGAAAa
	AS	gatctTTTCCATTTGGGTTAACCTGGGAAGAGGa
	UGG-S	agcttCCTCTTCCCAGGTTAACCCAAATGGAAAa
	UGG-AS	gatctTTTCCATTTGGGTTAACCTGGGAAGAGGa
MMP20-2	S	agcttCTGTGACTCCAGCTAAGCCTTTGATGCTa
	AS	gatctAGCATCAAAGGCTTAGCTGGAGTCACAGa
	UGG-S	agcttCTGTGACTCCAGCTAAGCCTTTGATGCTa
	UGG-AS	gatctAGCATCAAAGGCTTAGCTGGAGTCACAGa

References

Abhiman, S. and Sonnhammer, E.L.L. 2005. Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins: Structure, Function and Genetics*. **60**(4), pp.758–768.

Ackermann, R.R. and Cheverud, J.M. 2004. Detecting genetic drift versus selection in human evolution. *Proceedings of the National Academy of Sciences of the United States of America*. **101**(52), pp.17946–17951.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. 2010. A method and server for predicting damaging missense mutations. *Nature Methods*. **7**(4), pp.248–249.

Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Juettemann, T., Keenan, S., Laird, M.R., Lavidas, I., Maurel, T., McLaren, W., Moore, B., Murphy, D.N., Nag, R., Newman, V., Nuhn, M., Ong, C.K., Parker, A., Patricio, M., Riat, H.S., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Wilder, S.P., Zadissa, A., Kostadima, M., Martin, F.J., Muffato, M., Perry, E., Ruffier, M., Staines, D.M., Trevanion, S.J., Cunningham, F., Yates, A., Zerbino, D.R. and Flicek, P. 2017. Ensembl 2017. *Nucleic Acids Research*. **45**(D1), pp.D635–D642.

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., Weinstock, G.M. and Gibbs, R.A. 2007. Direct selection of human genomic loci by microarray hybridization. *Nature Methods*. **4**(11), pp.903–905.

Almomani, R., Marchi, M., Sopacua, M., Lindsey, P., Salvi, E., de Koning, B., Santoro, S., Magri, S., Smeets, H.J.M., Boneschi, F.M., Malik, R.R., Ziegler, D., Hoeijmakers, J.G.J., Bönhof, G., Dib-Hajj, S., Waxman, S.G., Merkies, I.S.J., Lauria, G., Faber, C.G. and Gerrits, M.M. 2020. Evaluation of molecular inversion probe versus TruSeq® custom methods for targeted next-generation sequencing K. Y. K. Chan, ed. *PLoS ONE*. **15**(9 September), p.e0238467.

Almuallem, Z. and Busuttil-Naudi, A. 2018. Molar incisor hypomineralisation (Mih) – an overview. *British Dental Journal*. **225**(7), pp.601–609.

Altug-Atac, A.T. and Erdem, D. 2007. Prevalence and distribution of dental anomalies in orthodontic patients. *American Journal of Orthodontics and Dentofacial Orthopedics*. **131**(4), pp.510–514.

Alvarez, J.A., Rezende, K.M.P.C., Marocho, S.M.S., Alves, F.B.T., Celiberti, P. and Ciamponi, A.L. 2009. Dental fluorosis: Exposure, prevention and management. *Medicina Oral, Patologia Oral y Cirugia Bucal*. **14**(2), pp.E103-7.

An, Z., Sabalic, M., Bloomquist, R.F., Fowler, T.E., Streelman, T. and Sharpe, P.T. 2018. A quiescent cell population replenishes mesenchymal stem cells to drive accelerated growth in mouse incisors. *Nature Communications*. **9**(1), p.378.

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*. **18**(8), pp.1585–1592.

Anisimova, M. and Gascuel, O. 2006. Approximate likelihood-ratio test for branches: A fast,

accurate, and powerful alternative J. Sullivan, ed. *Systematic Biology*. **55**(4), pp.539–552.

Aoba, T. and Fejerskov, O. 2002. Dental fluorosis: Chemistry and biology. *Critical Reviews in Oral Biology and Medicine*. **13**(2), pp.155–170.

Areal, H., Abrantes, J. and Esteves, P.J. 2011. Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. *BMC evolutionary biology*. **11**(1), p.368.

Arendt, Y., Banci, L., Bertini, I., Cantini, F., Cozzi, R., Del Conte, R. and Gonnelli, L. 2007. Catalytic domain of MMP20 (Enamelysin) - The NMR structure of a new matrix metalloproteinase. *FEBS Letters*. **581**(24), pp.4723–4726.

Arnittali, M., Rissanou, A.N. and Harmandaris, V. 2019. Structure of Biomolecules Through Molecular Dynamics Simulations. *Procedia Computer Science*. **156**, pp.69–78.

Avery, J.K., Steele, P.F. and Avery, N. 2002. Oral development and histology: Thieme.

Bäckman, B. and Holm, A. -K 1986. Amelogenesis imperfecta: prevalence and incidence in a northern Swedish county. *Community Dentistry and Oral Epidemiology*. **14**(1), pp.43–47.

Bai, R.Q., He, W. Bin, Peng, Q., Shen, S.H., Yu, Q.Q., Du, J., Tan, Y.Q., Wang, Y.H. and Liu, B.J. 2022. A novel FAM83H variant causes familial amelogenesis imperfecta with incomplete penetrance. *Molecular Genetics and Genomic Medicine*. **10**(4).

Bartlett, J.D. 2013. Dental Enamel Development: Proteinases and Their Enamel Matrix Substrates. *ISRN Dentistry*. **2013**, pp.1–24.

Beaumont, M.A. and Balding, D.J. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*. **13**(4), pp.969–980.

Bekers, E.M., Eijkelenboom, A., Rombout, P., Van Zwam, P., Mol, S., Ruijter, E., Scheijen, B. and Flucke, U. 2019. Identification of novel GNAS mutations in intramuscular myxoma using next-generation sequencing with single-molecule tagged molecular inversion probes. *Diagnostic Pathology*. **14**(1), p.15.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.L. and Abel, L. 2015. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences of the United States of America*. **112**(17), pp.5473–5478.

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.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. 2013. GenBank. *Nucleic Acids Research*. **41**(D1), pp.36–42.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Research*. **28**(1), pp.235–242.

Biagini, T., Chillemi, G., Mazzoccoli, G., Grottesi, A., Fusilli, C., Capocefalo, D., Castellana, S., Vescovi, A.L. and Mazza, T. 2018. Molecular dynamics recipes for genome research. *Briefings in bioinformatics*. **19**(5), pp.853–862.

Biswas, S. and Akey, J.M. 2006. Genomic insights into positive selection. *Trends in Genetics*. **22**(8), pp.437–446.

Blais, S.A., MacKenzie, L.A. and Wilson, M.V.H. 2011. Tooth-like scales in Early Devonian eugnathostomes and the 'outside-in' hypothesis for the origins of teeth in vertebrates. *Journal of Vertebrate Paleontology*. **31**(6), pp.1189–1199.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S. and SanCristobal, M. 2010. Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*. **186**(1), pp.241–262.

Boskey, A.L. and Roy, R. 2008. Cell culture systems for studies of bone and tooth mineralization. *Chemical Reviews*. **108**(11), pp.4716–4733.

Bosshardt, D.D. and Lang, N.P. 2005. The junctional epithelium: From health to disease. *Journal of Dental Research*. **84**(1), pp.9–20.

Botella, H., Blom, H., Dorka, M., Ahlberg, P.E. and Janvier, P. 2007. Jaws and teeth of the earliest bony fishes. *Nature*. **448**(7153), pp.583–586.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J. and Rolf, B. 1998. Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. *American Journal of Human Genetics*. **62**(6), pp.1408–1415.

Bromham, L. 2011. The genome as a life-history character: Why rate of molecular evolution varies between mammal species. *Philosophical Transactions of the Royal Society B: Biological Sciences*. **366**(1577), pp.2503–2513.

Bronckers, A.L.J.J. 2017. Ion Transport by Ameloblasts during Amelogenesis. *Journal of Dental Research*. **96**(3), pp.243–253.

Bronckers, A.L.J.J., Lyaruu, D.M. and Denbesten, P.K. 2009. Critical review in oral biology and medicine: The impact of fluoride on ameloblasts and the mechanisms of enamel fluorosis. *Journal of Dental Research*. **88**(10), pp.877–893.

Brookes, S.J., Barron, M.J., Boot-Handford, R., Kirkham, J. and Dixon, M.J. 2014. Endoplasmic reticulum stress in amelogenesis imperfecta and phenotypic rescue using 4-phenylbutyrate. *Human Molecular Genetics*. **23**(9), pp.2468–2480.

Brookes, S.J., Barron, M.J., Smith, C.E.L., Poulter, J.A., Mighell, A.J., Inglehearn, C.F., Brown, C.J., Rodd, H., Kirkham, J. and Dixon, M.J. 2017. Amelogenesis imperfecta caused by N-terminal enamelin point mutations in mice and men is driven by endoplasmic reticulum stress. *Human Molecular Genetics*. **26**(10), pp.1863–1876.

Brooks, D.A., Muller, V.J. and Hopwood, J.J. 2006. Stop-codon read-through for patients affected by a lysosomal storage disorder. *Trends in Molecular Medicine*. **12**(8), pp.367–373.

Der Brugghen, W. Van and Janvier, P. 1993. Denticles in thelodonts [6]. *Nature*. **364**(6433), p.107.

Case, D.A., Ben-Shalom, I.Y., Brozell, S.R., Cerutti, D.S., Cheatham III, T.E., Cruzeiro, V.W.D., Darden, T.A., Duke, R.E., Ghoreishi, D. and Gilson, M.K. 2018. AMBER 2018: San Francisco.

Catón, J. and Tucker, A.S. 2009. Current knowledge of tooth development: Patterning and mineralization of the murine dentition. *Journal of Anatomy*. **214**(4), pp.502–515.

Charlesworth, J. and Eyre-Walker, A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. *Molecular Biology and Evolution*. **25**(6), pp.1007–1015.

Chen, R., Im, H. and Snyder, M. 2015. Whole-exome enrichment with the agilent sureselect human all exon platform. *Cold Spring Harbor Protocols*. **2015**(7), pp.626–633.

Chenais, B. 2015. Transposable Elements in Cancer and Other Human Diseases. *Current Cancer Drug Targets*. **15**(3), pp.227–242.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. and Chan, A.P. 2012. Predicting the Functional Effect of Amino Acid Substitutions and Indels A. G. de Brevern, ed. *PLoS ONE*. **7**(10), p.e46688.

Chosack, A., Eidelman, E., Wisotski, I. and Cohen, T. 1979. Amelogenesis imperfecta among Israeli Jews and the description of a new type of local hypoplastic autosomal recessive amelogenesis imperfecta. *Oral Surgery, Oral Medicine, Oral Pathology*. **47**(2), pp.148–156.

Close, R.A., Friedman, M., Lloyd, G.T. and Benson, R.B.J. 2015. Evidence for a mid-Jurassic adaptive radiation in mammals. *Current Biology*. **25**(16), pp.2137–2142.

Coffield, K.D., Phillips, C., Brady, M., Roberts, M.W., Strauss, R.P. and Wright, J.T. 2005. The psychosocial impact of developmental dental defects in people with hereditary amelogenesis imperfecta. *Journal of the American Dental Association*. **136**(5), pp.620–630.

Collins, M.A., Mauriello, S.M., Tyndall, D.A. and Wright, J.T. 1999. Dental anomalies associated with amelogenesis imperfecta: A radiographic assessment. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontics*. **88**(3), pp.358–364.

Conrad, B. and Antonarakis, S.E. 2007. Gene duplication: A drive for phenotypic diversity and cause of human disease. *Annual Review of Genomics and Human Genetics*. **8**(1), pp.17–35.

Creevey, C.J. and McInerney, J.O. 2009. Trees from trees: Construction of phylogenetic supertrees using clann *In*: *Methods in Molecular Biology*., pp.139–161.

Crombie, F., Manton, D. and Kilpatrick, N. 2009. Aetiology of molar-incisor hypomineralization: A critical review. *International Journal of Paediatric Dentistry*. **19**(2), pp.73–83.

Csibra, E., Brierley, I. and Irigoyen, N. 2014. Modulation of Stop Codon Read-Through Efficiency and Its Effect on the Replication of Murine Leukemia Virus. *Journal of Virology*. **88**(18), pp.10364–10376.

Cui, J., Zhu, Q., Zhang, H., Cianfrocco, M.A., Leschziner, A.E., Dixon, J.E. and Xiao, J. 2017. Structure of Fam20A reveals a pseudokinase featuring a unique disulfide pattern and inverted ATP-binding. *eLife*. **6**, pp.1–16.

Cvijović, I., Good, B.H. and Desai, M.M. 2018. The effect of strong purifying selection on genetic diversity. *Genetics*. **209**(4), pp.1235–1278.

Dabrowski, M., Bukowy-Bieryllo, Z. and Zietkiewicz, E. 2018. Advances in therapeutic use of a drug-stimulated translational readthrough of premature termination codons. *Molecular Medicine*. **24**(1), pp.1–15.

Dabrowski, M., Bukowy-Bieryllo, Z. and Zietkiewicz, E. 2015. Translational readthrough potential of natural termination codons in eucaryotes – The impact of RNA sequence. *RNA Biology*. **12**(9), pp.950–958.

Davit-Béal, T., Tucker, A.S. and Sire, J.Y. 2009. Loss of teeth and enamel in tetrapods: Fossil

record, genetic data and morphological adaptations. *Journal of Anatomy*. **214**(4), pp.477–501.

Delsuc, F., Gasse, B. and Sire, J.Y. 2015. Evolutionary analysis of selective constraints identifies ameloblastin (AMBN) as a potential candidate for amelogenesis imperfecta. *BMC Evolutionary Biology*. **15**(1), pp.1–18.

Deméré, T.A., McGowen, M.R., Berta, A. and Gatesy, J. 2008. Morphological and molecular evidence for a stepwise evolutionary transition from teeth to baleen in mysticete whales. *Systematic Biology*. **57**(1), pp.15–37.

Desmet, F.O., Hamroun, D., Lalande, M., Collod-Bëroud, G., Claustres, M. and Béroud, C. 2009. Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*. **37**(9), pp.e67–e67.

Dever, T.E., Dinman, J.D. and Green, R. 2018. Translation elongation and recoding in eukaryotes. *Cold Spring Harbor Perspectives in Biology*. **10**(8), pp.1–20.

Dong, X., Zhao, B., Iacob, R.E., Zhu, J., Koksal, A.C., Lu, C., Engen, J.R. and Springer, T.A. 2017. Force interacts with macromolecular structure in activation of TGF-β. *Nature*. **542**(7639), pp.55–59.

Döring, K., Ahmed, N., Riemer, T., Suresh, H.G., Vainshtein, Y., Habich, M., Riemer, J., Mayer, M.P., O'Brien, E.P., Kramer, G. and Bukau, B. 2017. Profiling Ssb-Nascent Chain Interactions Reveals Principles of Hsp70-Assisted Folding. *Cell*. **170**(2), pp.298-311.e20.

Doronina, V.A. and Brown, J.D. 2006. When nonsense makes sense and vice versa: Noncanonical decoding events at stop codons in eukaryotes. *Molecular Biology*. **40**(4), pp.654–663.

Duarte, M.A., Fernandes, C.R., Heckel, G., Mathias, M. da L. and Bastos-Silveira, C. 2021. Variation and selection in the putative sperm-binding region of zp3 in muroid rodents: A comparison between cricetids and murines. *Genes*. **12**(9).

Dubail, J., Huber, C., Chantepie, S., Sonntag, S., Tüysüz, B., Mihci, E., Gordon, C.T., Steichen-Gersdorf, E., Amiel, J., Nur, B., Stolte-Dijkstra, I., van Eerde, A.M., van Gassen, K.L., Breugem, C.C., Stegmann, A., Lekszas, C., Maroofian, R., Karimiani, E.G., Bruneel, A., Seta, N., Munnich, A., Papy-Garcia, D., De La Dure-Molla, M. and Cormier-Daire, V. 2018. SLC10A7 mutations cause a skeletal dysplasia with amelogenesis imperfecta mediated by GAG biosynthesis defects. *Nature Communications*. **9**(1), p.3087.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. **32**(5), pp.1792–1797.

Eichler, E.E. 2019. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *New England Journal of Medicine*. **381**(1), pp.64–74.

El-Sayed, W., Parry, D.A., Shore, R.C., Ahmed, M., Jafri, H., Rashid, Y., Al-Bahlani, S., Al Harasi, S., Kirkham, J., Inglehearn, C.F. and Mighell, A.J. 2009. Mutations in the Beta Propeller WDR72 Cause Autosomal-Recessive Hypomaturation Amelogenesis Imperfecta. *American Journal of Human Genetics*. **85**(5), pp.699–705.

El-Sayed, W., Shore, R.C., Parry, D.A., Inglehearn, C.F. and Mighell, A.J. 2011. Hypomaturation amelogenesis imperfecta due to WDR72 mutations: A novel mutation and ultrastructural analyses of deciduous teeth. *Cells Tissues Organs*. **194**(1), pp.60–66.

El-Sayed, W., Shore, R.C., Parry, D.A., Inglehearn, C.F. and Mighell, A.J. 2010. Ultrastructural

analyses of deciduous teeth affected by hypocalcified amelogenesis imperfecta from a family with a novel Y458X FAM83H nonsense mutation. *Cells Tissues Organs*. **191**(3), pp.235–239.

Eng, B., Ainsworth, P. and Waye, J.S. 1994. Anomalous Migration of PCR Products Using Nondenaturing Polyacrylamide Gel Electrophoresis: The Amelogenin Sex-Typing System. *Journal of Forensic Sciences*. **39**(6), p.13724J.

Eyre-Walker, A.C. 1991. An analysis of codon usage in mammals: Selection or mutation bias? *Journal of Molecular Evolution*. **33**(5), pp.442–449.

Fejerskov, O., Manji, F. and Baelum, V. 1990. The nature and mechanisms of dental fluorosis in man. *Journal of Dental Research*. **69**(SPEC. ISS. FEB.), pp.692–700.

Feng, P., Zheng, J., Rossiter, S.J., Wang, D. and Zhao, H. 2014. Massive losses of taste receptor genes in toothed and baleen whales. *Genome Biology and Evolution*. **6**(6), pp.1254–1265.

Fernald, R.D. 1997. The Evolution of Eyes. *Brain, Behavior and Evolution*. **50**(4), pp.253–259.

Firth, A.E., Wills, N.M., Gesteland, R.F. and Atkins, J.F. 2011. Stimulation of stop codon readthrough: Frequent presence of an extended 3' RNA structural element. *Nucleic Acids Research*. **39**(15), pp.6679–6691.

Fisher, R.A. 1930. The genetical theory of natural selection. Clarendon.

Fitch, W.M. 2000. Homology: a personal view on some of the problems. *Trends in Genetics*. **16**(5), pp.227–231.

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*. **47**(3), pp.272–275.

Fu, Y.X. and Li, W.H. 1993. Statistical tests of neutrality of mutations. *Genetics*. **133**(3), pp.693–709.

Gadhia, K., McDonald, S., Arkutu, N. and Malik, K. 2012. Amelogenesis imperfecta: An introduction. *British Dental Journal*. **212**(8), pp.377–379.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glemin, S., Bierne, N. and Duret, L. 2018. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion N. Singh, ed. *Molecular Biology and Evolution*. **35**(5), pp.1092–1103.

Gasse, B., Karayigit, E., Mathieu, E., Jung, S., Garret, A., Huckert, M., Morkmued, S., Schneider, C., Vidal, L., Hemmerlé, J., Sire, J.Y. and Bloch-Zupan, A. 2013. Homozygous and compound heterozygous MMP20 mutations in amelogenesis imperfecta. *Journal of Dental Research*. **92**(7), pp.598–603.

Gasse, B., Prasad, M., Delgado, S., Huckert, M., Kawczynski, M., Garret-Bernardin, A., Lopez-Cazaux, S., Bailleul-Forestier, I., Manière, M.C., Stoetzel, C., Bloch-Zupan, A. and Sire, J.Y. 2017. Evolutionary analysis predicts sensitive positions of MMP20 and validates newly- and previously-identified MMP20 mutations causing amelogenesis imperfecta. *Frontiers in Physiology*. **8**(JUN).

Gasse, B., Silvent, J. and Sire, J.Y. 2012. Evolutionary analysis suggests that AMTN is enamel-

specific and a candidate for AI. *Journal of Dental Research*. **91**(11), pp.1085–1089.

Gazzo, A., Raimondi, D., Daneels, D., Moreau, Y., Smits, G., Van Dooren, S. and Lenaerts, T. 2017. Understanding mutational effects in digenic diseases. *Nucleic Acids Research*. **45**(15), pp.e140–e140.

Gemayel, R., Cho, J., Boeynaems, S. and Verstrepen, K.J. 2012. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes*. **3**(3), pp.461–480.

Genereux, D.P., Serres, A., Armstrong, J., Johnson, J., Marinescu, V.D., Murén, E., Juan, D., Bejerano, G., Casewell, N.R., Chemnick, L.G., Damas, J., Di Palma, F., Diekhans, M., Fiddes, I.T., Garber, M., Gladyshev, V.N., Goodman, L., Haerty, W., Houck, M.L., Hubley, R., Kivioja, T., Koepfli, K.P., Kuderna, L.F.K., Lander, E.S., Meadows, J.R.S., Murphy, W.J., Nash, W., Noh, H.J., Nweeia, M., Pfenning, A.R., Pollard, K.S., Ray, D.A., Shapiro, B., Smit, A.F.A., Springer, M.S., Steiner, C.C., Swofford, R., Taipale, J., Teeling, E.C., Turner-Maier, J., Alfoldi, J., Birren, B., Ryder, O.A., Lewin, H.A., Paten, B., Marques-Bonet, T., Lindblad-Toh, K. and Karlsson, E.K. 2020. A comparative genomics multitool for scientific discovery and conservation. *Nature*. **587**(7833), pp.240–245.

Gesteland, R.F., Weiss, R.B. and Atkins, J.F. 1992. Recoding: Reprogrammed genetic decoding. *Science*. **257**(5077), pp.1640–1641.

Gharib, W.H. and Robinson-Rechavi, M. 2013. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Molecular Biology and Evolution*. **30**(7), pp.1675–1686.

Gibson, C.W., Yuan, Z.A., Hall, B., Longenecker, G., Chen, E., Thyagarajan, T., Sreenath, T., Wright, J.T., Decker, S., Piddington, R., Harrison, G. and Kulkarni, A.B. 2001. Amelogenin-deficient Mice Display an Amelogenesis Imperfecta Phenotype. *Journal of Biological Chemistry*. **276**(34), pp.31871–31875.

Gilad, Y., Segré, D., Skorecki, K., Nachman, M.W., Lancet, D. and Sharon, D. 2000. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nature Genetics*. **26**(2), pp.221–224.

Grentzmann, G., Ingram, J.A., Kelly, P.J., Gesteland, R.F. and Atkins, J.F. 1998. A dual-luciferase reporter system for studying recoding signals. *Rna*. **4**(4), pp.479–486.

Gruebele, M. 2002. Protein folding: The free energy surface. *Current Opinion in Structural Biology*. **12**(2), pp.161–168.

Guazzi, G., Palmeri, S., Malandrini, A., Ciacci, G., Di Perri, R., Mancini, G., Messina, C. and Salvadori, C. 1994. Ataxia, mental deterioration, epilepsy in a family with dominant enamel hypoplasia: A variant of Kohlschutter-Tonz syndrome? *American Journal of Medical Genetics*. **50**(1), pp.79–83.

Guijarro-Clarke, C., Holland, P.W.H. and Paps, J. 2020. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nature Ecology and Evolution*. **4**(4), pp.519–523.

Gupta, P.K. 2022. Earth Biogenome Project: present status and future plans. *Trends in Genetics*. **38**(8), pp.811–820.

Hallström, B.M. and Janke, A. 2008. Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. *BMC Evolutionary Biology*. **8**(1), p.162.

Hallström, B.M., Kullberg, M., Nilsson, M.A. and Janke, A. 2007. Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups. *Molecular Biology and Evolution*. **24**(9), pp.2059–2068.

Hart, P.S., Aldred, M.J., Crawford, P.J.M., Wright, N.J., Hart, T.C. and Wright, J.T. 2002. Amelogenesis imperfecta phenotype-genotype correlations with two amelogenin gene mutations. *Archives of Oral Biology*. **47**(4), pp.261–265.

Hart, P.S., Hart, T.C., Michalec, M.D., Ryu, O.H., Simmons, D., Hong, S. and Wright, J.T. 2004. Mutation in kallikrein 4 causes autosomal recessive hypomaturation amelogenesis imperfecta. *Journal of Medical Genetics*. **41**(7), pp.545–549.

Hart, S., Hart, T., Gibson, C. and Wright, J.T. 2000. Mutational analysis of X-linked amelogenesis imperfecta in multiple families. *Archives of Oral Biology*. **45**(1), pp.79–86.

Hart, T.C., Hart, P.S., Gorry, M.C., Michalec, M.D., Ryu, O.H., Uygur, C., Ozdemir, D., Firatli, S., Aren, G. and Firatli, E. 2003. Novel ENAM mutation responsible for autosomal recessive amelogenesis imperfecta and localised enamel defects. *Journal of Medical Genetics*. **40**(12), pp.900–906.

Hautier, L., Gomes Rodrigues, H., Billet, G. and Asher, R.J. 2016. The hidden teeth of sloths: Evolutionary vestiges and the development of a simplified dentition. *Scientific Reports*. **6**(February), pp.1–9.

Hawkins, J.A., Kaczmarek, M.E., Müller, M.A., Drosten, C., Press, W.H. and Sawyer, S.L. 2019. A metaanalysis of bat phylogenetics and positive selection based on genomes and transcriptomes from 18 species. *Proceedings of the National Academy of Sciences of the United States of America*. **166**(23), pp.11351–11360.

Hedrick, P.W. 2007. Balancing selection. *Current Biology*. **17**(7), pp.R230–R231.

Hentschel, J., Tatun, D., Parkhomchuk, D., Kurth, I., Schimmel, B., Heinrich-Weltzien, R., Bertzbach, S., Peters, H. and Beetz, C. 2016. Identification of the first multi-exonic WDR72 deletion in isolated amelogenesis imperfecta, and generation of a WDR72-specific copy number screening tool. *Gene*. **590**(1), pp.1–4.

Ho, S.S., Urban, A.E. and Mills, R.E. 2020. Structural variation in the sequencing era. *Nature Reviews Genetics*. **21**(3), pp.171–189.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J. and McCombie, W.R. 2007. Genome-wide in situ exon capture for selective resequencing. *Nature Genetics*. **39**(12), pp.1522–1527.

Hoffman, K.S., Crnković, A. and Söll, D. 2018. Versatility of synthetic tRNAs in genetic code expansion. *Genes*. **9**(11).

Huang, W., Yang, M., Wang, C. and Song, Y. 2017. Evolutionary analysis of FAM83H in vertebrates. *PLoS ONE*. **12**(7), pp.1–12.

Humphrey, W., Dalke, A. and Schulten, K. 1996. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*. **14**(1), pp.33–38.

Hurst, L.D. and Pál, C. 2001. Evidence for purifying selection acting on silent sites in BRCA1. *Trends in Genetics*. **17**(2), pp.62–65.

Hyland, E.M., Webb, A.E., Kennedy, K.F., Gerek Ince, Z.N., Loscher, C.E. and O'Connell, M.J. 2021. Adaptive Evolution in TRIF Leads to Discordance between Human and Mouse Innate Immune Signaling A. Betancourt, ed. *Genome biology and evolution*. **13**(12),

pp.1–11.

Hyun, H.K., Lee, S.K., Lee, K.E., Kang, H.Y., Kim, E.J., Choung, P.H. and Kim, J.W. 2009. Identification of a novel FAM83H mutation and microhardness of an affected molar in autosomal dominant hypocalcified amelogenesis imperfecta. *International Endodontic Journal*. **42**(11), pp.1039–1043.

Innan, H. and Kim, Y. 2008. Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics*. **179**(3), pp.1713–1720.

Ishikawa, H.O., Xu, A., Ogura, E., Manning, G. and Irvine, K.D. 2012. The raine syndrome protein FAM20C is a golgi kinase that phosphorylates bio-mineralization proteins E. Giniger, ed. *PLoS ONE*. **7**(8), p.e42988.

Iwasaki, K., Bajenova, E., Somogyi-Ganss, E., Miller, M., Nguyen, V., Nourkeyhani, H., Gao, Y., Wendel, M. and Ganss, B. 2005. Amelotin - A novel secreted, ameloblast-specific protein. *Journal of Dental Research*. **84**(12), pp.1127–1132.

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., Chow, E.D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S.J. and Farh, K.K.H. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*. **176**(3), pp.535-548.e24.

Jaureguiberry, G., De La Dure-Molla, M., Parry, D., Quentric, M., Himmerkus, N., Koike, T., Poulter, J., Klootwijk, E., Robinette, S.L., Howie, A.J., Patel, V., Figueres, M.L., Stanescu, H.C., Issler, N., Nicholson, J.K., Bockenhauer, D., Laing, C., Walsh, S.B., McCredie, D.A., Povey, S., Asselin, A., Picard, A., Coulomb, A., Medlar, A.J., Bailleul-Forestier, I., Verloes, A., Le Caignec, C., Roussey, G., Guiol, J., Isidor, B., Logan, C., Shore, R., Johnson, C., Inglehearn, C., Al-Bahlani, S., Schmittbuhl, M., Clauss, F., Huckert, M., Laugel, V., Ginglinger, E., Pajarola, S., Spartà, G., Bartholdi, D., Rauch, A., Addor, M.C., Yamaguti, P.M., Safatle, H.P., Acevedo, A.C., Martelli-Júnior, H., Dos Santos Netos, P.E., Coletta, R.D., Gruessel, S., Sandmann, C., Ruehmann, D., Langman, C.B., Scheinman, S.J., Ozdemir-Ozenen, D., Hart, T.C., Hart, P.S., Neugebauer, U., Schlatter, E., Houillier, P., Gahl, W.A., Vikkula, M., Bloch-Zupan, A., Bleich, M., Kitagawa, H., Unwin, R.J., Mighell, A., Berdal, A. and Kleta, R. 2013. Nephrocalcinosis (enamel renal syndrome) caused by autosomal recessive FAM20A mutations. *Nephron - Physiology*. **122**(1–2), pp.1–6.

Jheon, A.H., Seidel, K., Biehs, B. and Klein, O.D. 2013. From molecules to mastication: The development and evolution of teeth. *Wiley Interdisciplinary Reviews: Developmental Biology*. **2**(2), pp.165–182.

Johri, P., Stephan, W. and Jensen, J.D. 2022. Soft selective sweeps: Addressing new definitions, evaluating competing models, and interpreting empirical outliers J. C. Fay, ed. *PLoS Genetics*. **18**(2), p.e1010022.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E. and Taipale, J. 2015. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*. **527**(7578), pp.384–388.

Jovanovic, V.M., Sarfert, M., Reyna-Blanco, C.S., Indrischek, H., Valdivia, D.I., Shelest, E. and Nowick, K. 2021. Positive Selection in Gene Regulatory Factors Suggests Adaptive Pleiotropic Changes During Human Evolution. *Frontiers in Genetics*. **12**.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J.,

Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**(7873), pp.583–589.

Jungreis, I., Chan, C.S., Waterhouse, R.M., Fields, G., Lin, M.F. and Kellis, M. 2016. Evolutionary dynamics of abundant stop codon readthrough. *Molecular Biology and Evolution*. **33**(12), pp.3108–3132.

Jungreis, I., Lin, M.F., Spokony, R., Chan, C.S., Negre, N., Victorsen, A., White, K.P. and Kellis, M. 2011. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Research*. **21**(12), pp.2096–2113.

Kallenbach, E. 1977. Fine structure of secretory ameloblasts in the kitten. *American Journal of Anatomy*. **148**(4), pp.479–511.

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*. **14**(6), pp.587–589.

Kantaputra, P.N., Bongkochwilawan, C., Lubinsky, M., Pata, S., Kaewgahya, M., Tong, H.J., Ketudat Cairns, J.R., Guven, Y. and Chaisrisookumporn, N. 2017. Periodontal disease and FAM20A mutations. *Journal of Human Genetics*. **62**(7), pp.679–686.

Karplus, M. and Kuriyan, J. 2005. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America*. **102**(19), pp.6679–6685.

Katoh, K. and Toh, H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*. **9**(4), pp.286–298.

Kawasaki, K., Hu, J.C.C. and Simmer, J.P. 2014. Evolution of Klk4 and enamel maturation in eutherians. *Biological Chemistry*. **395**(9), pp.1003–1013.

Kim, J.W., Lee, S.K., Lee, Z.H., Park, J.C., Lee, K.E., Lee, M.H., Park, J.T., Seo, B.M., Hu, J.C.C. and Simmer, J.P. 2008. FAM83H Mutations in Families with Autosomal-Dominant Hypocalcified Amelogenesis Imperfecta. *American Journal of Human Genetics*. **82**(2), pp.489–494.

Kim, J.W., Seymen, F., Lee, K.E., Ko, J., Yildirim, M., Tuna, E.B., Gencay, K., Shin, T.J., Kyun, H.K., Simmer, J.P. and Hu, J.C.C. 2013. LAMB3 mutations causing autosomal-dominant amelogenesis imperfecta. *Journal of Dental Research*. **92**(10), pp.899–904.

Kim, J.W., Simmer, J.P., Hart, T.C., Hart, P.S., Ramaswami, M.D., Bartlett, J.D. and Hu, J.C.C. 2005. MMP-20 mutation in autosomal recessive pigmented hypomaturation amelogenesis imperfecta. *Journal of Medical Genetics*. **42**(3), pp.271–275.

Kim, J.W., Zhang, H., Seymen, F., Koruyucu, M., Hu, Y., Kang, J., Kim, Y.J., Ikeda, A., Kasimoglu, Y., Bayram, M., Zhang, C., Kawasaki, K., Bartlett, J.D., Saunders, T.L., Simmer, J.P. and Hu, J.C.C. 2019. Mutations in RELT cause autosomal recessive amelogenesis imperfecta. *Clinical Genetics*. **95**(3), pp.375–383.

Kim, Y. and Stephan, W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*. **160**(2), pp.765–777.

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press.

Kimura, M. and Ohta, T. 1974. On Some Principles Governing Molecular Evolution.

Proceedings of the National Academy of Sciences. **71**(7), pp.2848–2852.

Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M. and Shendure, J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*. **46**(3), pp.310–315.

Knudsen, B. and Miyamoto, M.M. 2009. Accurate and fast methods to estimate the population mutation rate from error prone sequences. *BMC Bioinformatics*. **10**(1), p.247.

Koonin, E. V., Wolf, Y.I. and Karev, G.P. 2002. The structure of the protein universe and genome evolution. *Nature*. **420**(6912), pp.218–223.

Korneliussen, T.S., Moltke, I., Albrechtsen, A. and Nielsen, R. 2013. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*. **14**(1), p.289.

Kosakovsky Pond, S.L. and Frost, S.D.W. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*. **22**(5), pp.1208–1222.

Kosakovsky Pond, S.L., Poon, A.F.Y., Velazquez, R., Weaver, S., Hepler, N.L., Murrell, B., Shank, S.D., Magalis, B.R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S.J., Frost, S.D.W. and Muse, S. V. 2020. HyPhy 2.5 - A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies K. Crandall, ed. *Molecular Biology and Evolution*. **37**(1), pp.295–299.

Kosiol, C., Vinař, T., Da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R. and Siepel, A. 2008. Patterns of positive selection in six mammalian genomes M. H. Schierup, ed. *PLoS Genetics*. **4**(8), p.e1000144.

Kuga, T., Sasaki, M., Mikami, T., Miake, Y., Adachi, J., Shimizu, M., Saito, Y., Koura, M., Takeda, Y., Matsuda, J., Tomonaga, T. and Nakayama, Y. 2016. FAM83H and casein kinase i regulate the organization of the keratin cytoskeleton and formation of desmosomes. *Scientific Reports*. **6**(May), pp.1–15.

Kulmuni, J., Wurm, Y. and Pamilo, P. 2013. Comparative genomics of chemosensory protein genes reveals rapid evolution and positive selection in ant-specific duplicates. *Heredity*. **110**(6), pp.538–547.

Kumar, S., Stecher, G. and Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*. **33**(7), pp.1870–1874.

Kwiatkowski, D.P. 2005. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *The American Journal of Human Genetics*. **77**(2), pp.171–192.

Lacruz, R.S. and Feske, S. 2015. Diseases caused by mutations in ORAI1 and STIM1. *Annals of the New York Academy of Sciences*. **1356**(1), pp.45–79.

Lacruz, R.S., Nakayama, Y., Holcroft, J., Nguyen, V., Somogyi-Ganss, E., Snead, M.L., White, S.N., Paine, M.L. and Ganss, B. 2012. Targeted overexpression of amelotin disrupts the microstructure of dental enamel. *PLoS ONE*. **7**(4), pp.1–11.

Lee, C., Li, X., Hechmer, A., Eisen, M., Biggin, M.D., Venters, B.J., Jiang, C., Li, J., Pugh, B.F. and Gilmour, D.S. 2008. NELF and GAGA Factor Are Linked to Promoter-Proximal Pausing at Many Genes in Drosophila . *Molecular and Cellular Biology*. **28**(10), pp.3290–3300.

Lee, S.K., Hu, J.C.C., Bartlett, J.D., Lee, K.E., Lin, B.P.J., Simmer, J.P. and Kim, J.W. 2008. Mutational spectrum of FAM83H: the C-terminal portion is required for tooth enamel calcification. *Human mutation*. **29**(8), pp.E95–E99.

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M.T.P., Goldstein, M.M., Grigoriev, I. V., Hackett, K.J., Haussler, D., Jarvis, E.D., Johnson, W.E., Patrinos, A., Richards, S., Castilla-Rubio, J.C., Van Sluys, M.A., Soltis, P.S., Xu, X., Yang, H. and Zhang, G. 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*. **115**(17), pp.4325–4333.

Lewontin, R.C. and Krakauer, J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*. **74**(1), pp.175–195.

Li, Z., Yu, M. and Tian, W. 2013. An inductive signalling network regulates mammalian tooth morphogenesis with implications for tooth regeneration. *Cell Proliferation*. **46**(5), pp.501–508.

Lillegraven, J.A., Thompson, S.D., McNab, B.K. and Patton, J.L. 1987. The origin of eutherian mammals. *Biological Journal of the Linnean Society*. **32**(3), pp.281–336.

Lincoln, V., Cogan, J., Hou, Y., Hirsch, M., Hao, M., Alexeev, V., De Luca, M., De Rosa, L., Bauer, J.W., Woodley, D.T. and Chen, M. 2018. Gentamicin induces LAMB3 nonsense mutation readthrough and restores functional laminin 332 in junctional epidermolysis bullosa. *Proceedings of the National Academy of Sciences of the United States of America*. **115**(28), pp.E6536–E6545.

Linde, A. and Goldberg, M. 1994. Dentinogenesis. *Critical Reviews in Oral Biology and Medicine*. **4**(5), pp.679–728.

Liu, Y., Cotton, J.A., Shen, B., Han, X., Rossiter, S.J. and Zhang, S. 2010. Convergent sequence evolution between echolocating bats and dolphins. *Current Biology*. **20**(2), pp.R53–R54.

Lopes Dias, J., Borges, A. and Lima Rego, R. 2016. Primary intraosseous squamous cell carcinoma of the mandible: a case with atypical imaging features. *BJR|case reports*. **2**(4), p.20150276.

Loughran, G., Chou, M.Y., Ivanov, I.P., Jungreis, I., Kellis, M., Kiran, A.M., Baranov, P. V. and Atkins, J.F. 2014. Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Research*. **42**(14), pp.8928–8938.

Loughran, G., Jungreis, I., Tzani, I., Power, M., Dmitriev, R.I., Ivanov, I.P., Kellis, M. and Atkins, J.F. 2018. Stop codon readthrough generates a C-terminally extended variant of the human Vitamin D receptor with reduced calcitriol response. *Journal of Biological Chemistry*. **293**(12), pp.4434–4444.

Lowe, C.B., Kellis, M., Siepel, A., Raney, B.J., Clamp, M., Salama, S.R., Kingsley, D.M., Lindblad-Toh, K. and Haussler, D. 2011. Three periods of regulatory innovation during vertebrate evolution. *Science*. **333**(6045), pp.1019–1024.

Lu, T., Li, M., Xu, X., Xiong, J., Huang, C., Zhang, X., Hu, A., Peng, L., Cai, D., Zhang, L., Wu, B. and Xiong, F. 2018. Whole exome sequencing identifies an AMBN missense mutation causing severe autosomal-dominant amelogenesis imperfecta and dentin disorders. *International Journal of Oral Science*. **10**(3), p.26.

Ludwig, M.G., Vanek, M., Guerini, D., Gasser, J.A., Jones, C.E., Junker, U., Hofstetter, H., Wolf, R.M. and Seuwen, K. 2003. Proton-sensing G-protein-coupled receptors. *Nature*.

425(6953), pp.93–98.

Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., Garcia, C.A., Chakravarti, A. and Patel, P.I. 1991. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*. **66**(2), pp.219–232.

Lynch, M. 2007. The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics*. **8**(10), pp.803–813.

Lynch, M., Ackerman, M.S., Gout, J.F., Long, H., Sung, W., Thomas, W.K. and Foster, P.L. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*. **17**(11), pp.704–714.

Machado, J.P., Philip, S., Maldonado, E., ÖBrien, S.J., Johnson, W.E. and Antunes, A. 2016. Positive selection linked with generation of novel mammalian dentition patterns. *Genome Biology and Evolution*. **8**(9), pp.2748–2759.

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

Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E. and Simmerling, C. 2015. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*. **11**(8), pp.3696–3713.

Manuvakhova, M., Keeling, K. and Bedwell, D.M. 2000. Aminoglycoside antibiotics mediate context-dependent suppression of termination codons in a mammalian translation system. *Rna*. **6**(7), pp.1044–1055.

Mårdh, C.K., Bäckman, B., Holmgren, G., Hu, J.C.C., Simmer, J.P. and Forsman-Semb, K. 2002. A nonsense mutation in the enamelin gene causes local hypoplastic autosomal dominant amelogenesis imperfecta (AIH2). *Human Molecular Genetics*. **11**(9), pp.1069–1074.

Mariotti, M., Ridge, P.G., Zhang, Y., Lobanov, A. V., Pringle, T.H., Guigo, R., Hatfield, D.L. and Gladyshev, V.N. 2012. Composition and evolution of the vertebrate and mammalian selenoproteomes. *PLoS ONE*. **7**(3).

Matsui, C., Wang, C.K., Nelson, C.F., Bauer, E.A. and Hoeffler, W.K. 1995. The assembly of laminin-5 subunits. *Journal of Biological Chemistry*. **270**(40), pp.23496–23503.

Matsuo, S., Ichikawa, H., Wakisaka, S. and Akai, M. 1992. Changes of cytochemical properties in the Golgi apparatus during in vivo differentation of the ameloblast in developing rat molar tooth germs. *The Anatomical Record*. **234**(4), pp.469–478.

McCarl, C.A., Picard, C., Khalil, S., Kawasaki, T., Röther, J., Papolos, A., Kutok, J., Hivroz, C., LeDeist, F., Plogmann, K., Ehl, S., Notheis, G., Albert, M.H., Belohradsky, B.H., Kirschner, J., Rao, A., Fischer, A. and Feske, S. 2009. ORAI1 deficiency and lack of store-operated Ca2+ entry cause immunodeficiency, myopathy, and ectodermal dysplasia. *Journal of Allergy and Clinical Immunology*. **124**(6), pp.1311-1318.e7.

McDonald, J.H. and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*. **351**(6328), pp.652–654.

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 B. Chang, ed. *Molecular Biology and Evolution*. **37**(7), pp.2069–2083.

McGrath, J.A., Gatalica, B., Li, K., Dunnill, M.G.S., McMillan, J.R., Christiano, A.M., Eady, R.A.J.

and Uitto, J. 1996. Compound heterozygosity for a dominant glycine substitution and a recessive internal duplication mutation in the type XVII collagen gene results in junctional epidermolysis bullosa and abnormal dentition. *American Journal of Pathology*. **148**(6), pp.1787–1796.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. **20**(9), pp.1297–1303.

McKusick, V.A. 1998. *Mendelian inheritance in man: a catalog of human genes and genetic disorders*. JHU Press.

McNabb, D.S., Reed, R. and Marciniak, R.A. 2005. Dual luciferase assay system for rapid assessment of gene expression in Saccharomyces cerevisiae. *Eukaryotic Cell*. **4**(9), pp.1539–1549.

Meredith, R.W., Gatesy, J., Cheng, J. and Springer, M.S. 2011. Pseudogenization of the tooth gene enamelysin (MMP20) in the common ancestor of extant baleen whales. *Proceedings of the Royal Society B: Biological Sciences*. **278**(1708), pp.993–1002.

Messer, P.W. and Petrov, D.A. 2013. Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America*. **110**(21), pp.8615–8620.

Miller, R.F., Cloutier, R. and Turner, S. 2003. The oldest articulated chondrichthyan from the Early Devonian period. *Nature*. **425**(6957), pp.501–504.

Moffatt, P., Wazen, R.M., Dos Santos Neves, J. and Nanci, A. 2014. Characterisation of secretory calcium-binding phosphoprotein-proline-glutamine-rich 1: a novel basal lamina component expressed at cell-tooth interfaces. *Cell and Tissue Research*. **358**(3), pp.843–855.

Møller, P., Clark, N. and Mæhle, L. 2011. A simplified method for segregation analysis (SISA) to determine penetrance and expression of a genetic variant in a family. *Human Mutation*. **32**(5), pp.568–571.

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*. **30**(9), pp.2145–2156.

Mu, Y., Huang, X., Liu, R., Gai, Y., Liang, N., Yin, D., Shan, L., Xu, S. and Yang, G. 2021. ACPT gene is inactivated in mammalian lineages that lack enamel or teeth. *PeerJ*. **9**, p.e10219.

Mu, Y., Tian, R., Xiao, L., Sun, D., Zhang, Z., Xu, S. and Yang, G. 2021. Molecular Evolution of Tooth-Related Genes Provides New Insights into Dietary Adaptations of Mammals. *Journal of Molecular Evolution*. **89**(7), pp.458–471.

Muller, J., Creevey, C.J., Thompson, J.D., Arendt, D. and Bork, P. 2010. AQUA: Automated quality improvement for multiple sequence alignments. *Bioinformatics*. **26**(2), pp.263–265.

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A. and O'Brien, S.J. 2001.

Molecular phylogenetics and the origins of placental mammals. *Nature*. **409**(6820), pp.614–618.

Murray, A.W. 2020. Can gene-inactivating mutations lead to evolutionary novelty? *Current Biology*. **30**(10), pp.R465–R471.

Nair, A.K. and Baier, L.J. 2018. Using luciferase reporter assays to identify functional variants at disease-associated loci *In*: *Methods in Molecular Biology*., pp.303–319.

Nalbant, D., Youn, H., Nalbant, S.I., Sharma, S., Cobos, E., Beale, E.G., Du, Y. and Williams, S.C. 2005. FAM20: An evolutionarily conserved family of secreted proteins expressed in hematopoietic cells. *BMC Genomics*. **6**, pp.1–21.

Nanci, A. 2017. *Ten Cate's Oral Histology-e-book: development, structure, and function*. Elsevier Health Sciences.

Nasoori, A. 2020. Tusks, the extra-oral teeth. *Archives of oral biology*. **117**(March), p.104835.

Nei, M. and Tajima, F. 1981. Genetic drift and estimation of effective population size. *Genetics*. **98**(3), pp.625–640.

Neuhaus, C., Eisenberger, T., Decker, C., Nagl, S., Blank, C., Pfister, M., Kennerknecht, I., Müller-Hofstede, C., Charbel Issa, P., Heller, R., Beck, B., Rüther, K., Mitter, D., Rohrschneider, K., Steinhauer, U., Korbmacher, H.M., Huhle, D., Elsayed, S.M., Taha, H.M., Baig, S.M., Stöhr, H., Preising, M., Markus, S., Moeller, F., Lorenz, B., Nagel-Wolfrum, K., Khan, A.O. and Bolz, H.J. 2017. Next-generation sequencing reveals the mutational landscape of clinically diagnosed Usher syndrome: copy number variations, phenocopies, a predominant target for translational read-through, and PEX26 mutated in Heimler syndrome. *Molecular Genetics and Genomic Medicine*. **5**(5), pp.531–552.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., Bamshad, M., Nickerson, D.A. and Shendure, J. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. **461**(7261), pp.272–276.

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*. **32**(1), pp.268–274.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G. and Bustamante, C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Research*. **15**(11), pp.1566–1575.

Nikolopoulos, G., Smith, C.E.L., Brookes, S.J., El-Asrag, M.E., Brown, C.J., Patel, A., Murillo, G., O'Connell, M.J., Inglehearn, C.F. and Mighell, A.J. 2020. New missense variants in RELT causing hypomineralised amelogenesis imperfecta. *Clinical Genetics*. **97**(5), pp.688–695.

Nikolopoulos, G., Smith, C.E.L., Poulter, J.A., Murillo, G., Silva, S., Lamb, T., Berry, I.R., Brown, C.J., Day, P.F., Soldani, F., Al-Bahlani, S., Harris, S.A., O'Connell, M.J., Inglehearn, C.F. and Mighell, A.J. 2021. Spectrum of pathogenic variants and founder effects in amelogenesis imperfecta associated with MMP20. *Human Mutation*. **42**(5), pp.567–576.

Nishihara, H., Okada, N. and Hasegawa, M. 2007. Rooting the eutherian tree: The power and pitfalls of phylogenomics. *Genome Biology*. **8**(9), p.R199.

Nitayavardhana, I., Theerapanon, T., Srichomthong, C., Piwluang, S., Wichadakul, D., Porntaveetus, T. and Shotelersuk, V. 2020. Four novel mutations of FAM20A in

amelogenesis imperfecta type IG and review of literature for its genotype and phenotype spectra. *Molecular Genetics and Genomics*. **295**(4), pp.923–931.

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*. **5**(1), p.140.

Nweeia, M.T., Eichmiller, F.C., Hauschka, P. V., Tyler, E., Mead, J.G., Potter, C.W., Angnatsiak, D.P., Richard, P.R., Orr, J.R. and Black, S.R. 2012. Vestigial Tooth Anatomy and Tusk Nomenclature for Monodon Monoceros. *Anatomical Record*. **295**(6), pp.1006–1016.

O'Sullivan, J., Bitu, C.C., Daly, S.B., Urquhart, J.E., Barron, M.J., Bhaskar, S.S., Martelli-Júnior, H., Dos Santos Neto, P.E., Mansilla, M.A., Murray, J.C., Coletta, R.D., Black, G.C.M. and Dixon, M.J. 2011. Whole-exome sequencing identifies FAM20A mutations as a cause of amelogenesis imperfecta and gingival hyperplasia syndrome. *American Journal of Human Genetics*. **88**(5), pp.616–620.

Ohta, T. and Gillespie, J.H. 1996. Development of neutral nearly neutral theories. *Theoretical Population Biology*. **49**(2), pp.128–142.

Ong, K.R., Visram, S., McKaig, S. and Brueton, L.A. 2006. Sensorineural deafness, enamel abnormalities and nail abnormalities: A case report of Heimler syndrome in identical twin girls. *European Journal of Medical Genetics*. **49**(2), pp.187–193.

Orr, H.A. 1998. Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. *Genetics*. **149**(4), pp.2099–2104.

Ørvig, T. 1967. Phylogeny of tooth tissue: evolution of some calcified tissues in early vertebrates. *Structural and chemical organisation of teeth*., pp.45–110.

Owen, R. 1848. *On the archetype and homologies of the vertebrate skeleton*. author.

Pandurangan, A.P., Ochoa-Montaño, B., Ascher, D.B. and Blundell, T.L. 2017. SDM: A server for predicting effects of mutations on protein stability. *Nucleic Acids Research*. **45**(W1), pp.W229–W235.

Parker, J., Tsagkogeorga, G., Cotton, J.A., Liu, Y., Provero, P., Stupka, E. and Rossiter, S.J. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*. **502**(7470), pp.228–231.

Parry, D.A., Brookes, S.J., Logan, C. V., Poulter, J.A., El-Sayed, W., Al-Bahlani, S., Al harasi, S., Sayed, J., Raïf, E.M., Shore, R.C., Dashash, M., Barron, M., Morgan, J.E., Carr, I.M., Taylor, G.R., Johnson, C.A., Aldred, M.J., Dixon, M.J., Wright, J.T., Kirkham, J., Inglehearn, C.F. and Mighell, A.J. 2012. Mutations in C4orf26, encoding a peptide with in vitro hydroxyapatite crystal nucleation and growth activity, cause amelogenesis imperfecta. *American Journal of Human Genetics*. **91**(3), pp.565–571.

Parry, D.A., Holmes, T.D., Gamper, N., El-Sayed, W., Hettiarachchi, N.T., Ahmed, M., Cook, G.P., Logan, C. V., Johnson, C.A., Joss, S., Peers, C., Prescott, K., Savic, S., Inglehearn, C.F. and Mighell, A.J. 2016. A homozygous STIM1 mutation impairs store-operated calcium entry and natural killer cell effector function without clinical immunodeficiency. *Journal of Allergy and Clinical Immunology*. **137**(3), pp.955-957.e8.

Parry, D.A., Poulter, J.A., Logan, C. V., Brookes, S.J., Jafri, H., Ferguson, C.H., Anwari, B.M., Rashid, Y., Zhao, H., Johnson, C.A., Inglehearn, C.F. and Mighell, A.J. 2013. Identification

of mutations in SLC24A4, encoding a potassium-dependent sodium/calcium exchanger, as a cause of amelogenesis imperfecta. *American Journal of Human Genetics*. **92**(2), pp.307–312.

Parry, D.A., Smith, C.E.L., El-Sayed, W., Poulter, J.A., Shore, R.C., Logan, C. V., Mogi, C., Sato, K., Okajima, F., Harada, A., Zhang, H., Koruyucu, M., Seymen, F., Hu, J.C.C., Simmer, J.P., Ahmed, M., Jafri, H., Johnson, C.A., Inglehearn, C.F. and Mighell, A.J. 2016. Mutations in the pH-Sensing G-protein-Coupled Receptor GPR68 Cause Amelogenesis Imperfecta. *American Journal of Human Genetics*. **99**(4), pp.984–990.

Parveen, A., Mirza, M.U., Vanmeert, M., Akhtar, J., Bashir, H., Khan, S., Shehzad, S., Froeyen, M., Ahmed, W., Ansar, M. and Wasif, N. 2019. A novel pathogenic missense variant in CNNM4 underlying Jalili syndrome: Insights from molecular dynamics simulations. *Molecular Genetics and Genomic Medicine*. **7**(9).

Pattabiraman, N., Ward, K.B. and Fleming, P.J. 1995. Occluded molecular surface: Analysis of protein packing. *Journal of Molecular Recognition*. **8**(6), pp.334–344.

Payseur, B.A. and Cutter, A.D. 2006. Integrating patterns of polymorphism at SNPs and STRs. *Trends in Genetics*. **22**(8), pp.424–429.

Peery, Z.M., Kirby, R., Reid, B.N., Stoelting, R., Doucet-Bëer, E., Robinson, S., Vásquez-Carrillo, C., Pauli, J.N. and Palsboll, P.J. 2012. Reliability of genetic bottleneck tests for detecting recent population declines. *Molecular Ecology*. **21**(14), pp.3403–3418.

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K.A., Lin, G.N., Nam, H.J., Mort, M., Cooper, D.N., Sebat, J., Iakoucheva, L.M., Mooney, S.D. and Radivojac, P. 2020. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications*. **11**(1).

Pelham, H.R.B. 1978. Leaky UAG termination codon in tobacco mosaic virus RNA. *Nature*. **272**(5652), pp.469–471.

Pérez Millán, M.I., Vishnopolska, S.A., Daly, A.Z., Bustamante, J.P., Seilicovich, A., Bergadá, I., Braslavsky, D., Keselman, A.C., Lemons, R.M., Mortensen, A.H., Marti, M.A., Camper, S.A. and Kitzman, J.O. 2018. Next generation sequencing panel based on single molecule molecular inversion probes for detecting genetic variants in children with hypopituitarism. *Molecular Genetics and Genomic Medicine*. **6**(4), pp.514–525.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. 2004. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*. **25**(13), pp.1605–1612.

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.

Plagnol, V., Curtis, J., Epstein, M., Mok, K.Y., Stebbings, E., Grigoriadou, S., Wood, N.W., Hambleton, S., Burns, S.O., Thrasher, A.J., Kumararatne, D., Doffinger, R. and Nejentsev, S. 2012. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*. **28**(21), pp.2747–2754.

Poli, G., Barravecchia, I., Demontis, G.C., Sodi, A., Saba, A., Rizzo, S., Macchia, M. and Tuccinardi, T. 2022. Predicting potentially pathogenic effects of h RPE65 missense mutations: a computational strategy based on molecular dynamics simulations . *Journal of Enzyme Inhibition and Medicinal Chemistry*. **37**(1), pp.1765–1772.

Ponte, I., Romero, D., Yero, D., Suau, P. and Roque, A. 2017. Complex evolutionary history of

the mammalian histone H1.1-H1.5 gene family. *Molecular Biology and Evolution*. **34**(3), pp.545–558.

Poulter, J.A., Brookes, S.J., Shore, R.C., Smith, C.E.L., Abi farraj, L., Kirkham, J., Inglehearn, C.F. and Mighell, A.J. 2014. A missense mutation in ITGB6 causes pitted hypomineralized amelogenesis imperfecta. *Human Molecular Genetics*. **23**(8), pp.2189–2197.

Poulter, J.A., El-Sayed, W., Shore, R.C., Kirkham, J., Inglehearn, C.F. and Mighell, A.J. 2014. Whole-exome sequencing, without prior linkage, identifies a mutation in LAMB3 as a cause of dominant hypoplastic amelogenesis imperfecta. *European Journal of Human Genetics*. **22**(1), pp.132–135.

Poulter, J.A., Murillo, G., Brookes, S.J., Smith, C.E.L., Parry, D.A., Silva, S., Kirkham, J., Inglehearn, C.F. and Mighell, A.J. 2014. Deletion of ameloblastin exon 6 is associated with amelogenesis imperfecta. *Human molecular genetics*. **23**(20), pp.5317–5324.

Prasad, M.K., Geoffroy, V., Vicaire, S., Jost, B., Dumas, M., Le Gras, S., Switala, M., Gasse, B., Laugel-Haushalter, V., Paschaki, M., Leheup, B., Droz, D., Dalstein, A., Loing, A., Grollemund, B., Muller-Bolla, M., Lopez-Cazaux, S., Minoux, M., Jung, S., Obry, F., Vogt, V., Davideau, J.L., Davit-Beal, T., Kaiser, A.S., Moog, U., Richard, B., Morrier, J.J., Duprez, J.P., Odent, S., Bailleul-Forestier, I., Rousset, M.M., Merametdijan, L., Toutain, A., Joseph, C., Giuliano, F., Dahlet, J.C., Courval, A., El Alloussi, M., Laouina, S., Soskin, S., Guffon, N., Dieux, A., Doray, B., Feierabend, S., Ginglinger, E., Fournier, B., de la Dure Molla, M., Alembik, Y., Tardieu, C., Clauss, F., Berdal, A., Stoetzel, C., Manière, M.C., Dollfus, H. and Bloch-Zupan, A. 2015. A targeted next-generation sequencing assay for the molecular diagnosis of genetic disorders with orodental involvement. *Journal of Medical Genetics*. **53**(2), pp.98–110.

Price, J.A., Bowden, D.W., Wright, J.T., Pettenati, M.J. and Hart, T.C. 1998. Identification of a mutation in DLX3 associated with tricho-dento-osseous (TDO) syndrome. *Human Molecular Genetics*. **7**(3), pp.563–569.

Provine, W.B. 2004. Ernst Mayr: Genetics and speciation. *Genetics*. **167**(3), pp.1041–6.

Quintana-Murci, L. and Clark, A.G. 2013. Population genetic tools for dissecting innate immunity in humans. *Nature Reviews Immunology*. **13**(4), pp.280–293.

Raine, J., Winter, R.M., Davey, A. and Tucker, S.M. 1989. Unknown syndrome: Microcephaly, hypoplastic nose, exophthalmos, gum hyperplasia, cleft palate, low set ears, and osteosclerosis. *Journal of Medical Genetics*. **26**(12), pp.786–788.

Rajput, B., Pruitt, K.D. and Murphy, T.D. 2019. RefSeq curation and annotation of stop codon recoding in vertebrates. *Nucleic Acids Research*. **47**(2), pp.594–606.

Rajter, Ľ. and Vďačný, P. 2018. Selection and paucity of phylogenetic signal challenge the utility of alpha-tubulin in reconstruction of evolutionary history of free-living litostomateans (Protista, Ciliophora). *Molecular Phylogenetics and Evolution*. **127**, pp.534–544.

Ratbi, I., Falkenberg, K.D., Sommen, M., Al-Sheqaih, N., Guaoua, S., Vandeweyer, G., Urquhart, J.E., Chandler, K.E., Williams, S.G., Roberts, N.A., El Alloussi, M., Black, G.C., Ferdinandusse, S., Ramdi, H., Heimler, A., Fryer, A., Lynch, S.A., Cooper, N., Ong, K.R., Smith, C.E.L., Inglehearn, C.F., Mighell, A.J., Elcock, C., Poulter, J.A., Tischkowitz, M., Davies, S.J., Sefiani, A., Mironov, A.A., Newman, W.G., Waterham, H.R. and Van Camp, G. 2015. Heimler Syndrome Is Caused by Hypomorphic Mutations in the Peroxisome-Biogenesis Genes PEX1 and PEX6. *American Journal of Human Genetics*. **97**(4), pp.535–

545.

Reith, E.J. 1970. The stages of amelogenesis as observed in molar teeth of young rats. *Journal of Ultrasructure Research*. **30**(1–2), pp.111–151.

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. and Kircher, M. 2019. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*. **47**(D1), pp.D886–D894.

Reuter, J.A., Spacek, D. V. and Snyder, M.P. 2015. High-Throughput Sequencing Technologies. *Molecular Cell*. **58**(4), pp.586–597.

Riley, B.T., Ilyichova, O., Costa, M.G.S., Porebski, B.T., De Veer, S.J., Swedberg, J.E., Kass, I., Harris, J.M., Hoke, D.E. and Buckle, A.M. 2016. Direct and indirect mechanisms of KLK4 inhibition revealed by structure and dynamics. *Scientific Reports*. **6**(September), pp.1–14.

Robinson, C., Kirkham, J., Brookes, S.J., Bonass, W.A. and Shore, R.C. 1995. Int. J. Dev. Biol. *The chemistry of enamel development.* **39**, pp.145–152.

Robinson, D.F. and Foulds, L.R. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*. **53**(1–2), pp.131–147.

Rodnina, M. V., Korniy, N., Klimova, M., Karki, P., Peng, B.Z., Senyushkina, T., Belardinelli, R., Maracci, C., Wohlgemuth, I., Samatova, E. and Peske, F. 2020. Survey and summary: Translational recoding: Canonical translation mechanisms reinterpreted. *Nucleic Acids Research*. **48**(3), pp.1056–1067.

Rogers, Y.H. and Venter, J.C. 2005. Genomics: Massively parallel sequencing. *Nature*. **437**(7057), pp.326–327.

Romiguier, J., Ranwez, V., Douzery, E.J.P. and Galtier, N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Research*. **20**(8), pp.1001–1009.

Roscito, J.G., Sameith, K., Kirilenko, B.M., Hecker, N., Winkler, S., Dahl, A., Rodrigues, M.T. and Hiller, M. 2022. Convergent and lineage-specific genomic differences in limb regulatory elements in limbless reptile lineages. *Cell Reports*. **38**(3), p.110280.

Roy, B., Leszyk, J.D., Mangus, D.A. and Jacobson, A. 2015. Nonsense suppression by near-cognate tRNAs employs alternative base pairing at codon positions 1 and 3. *Proceedings of the National Academy of Sciences of the United States of America*. **112**(10), pp.3038–3043.

Sabbavarapu, N.M., Pieńko, T., Zalman, B.H., Trylska, J. and Baasov, T. 2018. Exploring eukaryotic: Versus prokaryotic ribosomal RNA recognition with aminoglycoside derivatives. *MedChemComm*. **9**(3), pp.503–508.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J. V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. and Lander, E.S. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. **419**(6909), pp.832–837.

Saleh, A., Macia, A. and Muotri, A.R. 2019. Transposable elements, inflammation, and neurological disease. *Frontiers in Neurology*. **10**(AUG).

Sandve, S.R., Rohlfs, R. V. and Hvidsten, T.R. 2018. Subfunctionalization versus

neofunctionalization after whole-genome duplication. *Nature Genetics*. **50**(7), pp.908–909.

Sanger, F., Nicklen, S. and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*. **74**(12), pp.5463–5467.

Sanville, B., Dolan, M.A., Wollenberg, K., Yan, Y., Martin, C., Yeung, M.L., Strebel, K., Buckler-White, A. and Kozak, C.A. 2010. Adaptive evolution of Mus Apobec3 includes retroviral insertion and positive selection at two clusters of residues flanking the substrate groove M. Farzan, ed. *PLoS Pathogens*. **6**(7), pp.1–14.

Schilling, T.F., Piotrowski, T., Grandel, H., Brand, M., Heisenberg, C.P., Jiang, Y.J., Beuchle, D., Hammerschmidt, M., Kane, D.A., Mullins, M.C., Van Eeden, F.J.M., Kelsh, R.N., Furutani-Seiki, M., Granato, M., Haffter, P., Odenthal, J., Warga, R.M., Trowe, T. and Nüsslein-Volhard, C. 1996. Jaw and branchial arch mutants in zebrafish I: Branchial arches. *Development*. **123**(1), pp.329–344.

Schossig, A., Wolf, N.I., Fischer, C., Fischer, M., Stocker, G., Pabinger, S., Dander, A., Steiner, B., Tönz, O., Kotzot, D., Haberlandt, E., Amberger, A., Burwinkel, B., Wimmer, K., Fauth, C., Grond-Ginsbach, C., Koch, M.J., Deichmann, A., Von Kalle, C., Bartram, C.R., Kohlschütter, A., Trajanoski, Z. and Zschocke, J. 2012. Mutations in ROGDI cause Kohlschütter-Tönz syndrome. *American Journal of Human Genetics*. **90**(4), pp.701–707.

Schueren, F. and Thoms, S. 2016. Functional Translational Readthrough: A Systems Biology Perspective J. Brosius, ed. *PLoS Genetics*. **12**(8), p.e1006196.

Schwarz, J.M., Cooper, D.N., Schuelke, M. and Seelow, D. 2014. Mutationtaster2: Mutation prediction for the deep-sequencing age. *Nature Methods*. **11**(4), pp.361–362.

Sedano, H.O. 1975. Congenital oral anomalies in Argentinian children. *Community Dentistry and Oral Epidemiology*. **3**(2), pp.61–63.

Seymen, F., Kim, Youn Jung, Lee, Y.J., Kang, J., Kim, T.H., Choi, H., Koruyucu, M., Kasimoglu, Y., Tuna, E.B., Gencay, K., Shin, T.J., Hyun, H.K., Kim, Young Jae, Lee, S.H., Lee, Z.H., Zhang, H., Hu, J.C.C., Simmer, J.P., Cho, E.S. and Kim, J.W. 2016. Recessive Mutations in ACPT, Encoding Testicular Acid Phosphatase, Cause Hypoplastic Amelogenesis Imperfecta. *American Journal of Human Genetics*. **99**(5), pp.1199–1205.

Seymen, F., Lee, K.E., Koruyucu, M., Gencay, K., Bayram, M., Tuna, E.B., Lee, Z.H. and Kim, J.W. 2014. ENAM mutations with incomplete penetrance. *Journal of Dental Research*. **93**(10), pp.988–992.

Seymen, F., Lee, K.E., Tran Le, C.G., Yildirim, M., Gencay, K., Lee, Z.H. and Kim, J.W. 2014. Exonal deletion of SLC24A4 causes hypomaturation amelogenesis imperfecta. *Journal of Dental Research*. **93**(4), pp.366–370.

Shakhnovich, B.E. and Koonin, E. V. 2006. Origins and impact of constraints in evolution of gene families. *Genome Research*. **16**(12), pp.1529–1536.

Sharma, V., Hecker, N., Roscito, J.G., Foerster, L., Langer, B.E. and Hiller, M. 2018. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nature Communications*. **9**(1), pp.1–9.

Sharma, V., Srinivasan, A., Nikolajeff, F. and Kumar, S. 2021. Biomineralization process in hard tissues: The interaction complexity within protein and inorganic counterparts. *Acta Biomaterialia*. **120**, pp.20–37.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. 2001. DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*. **29**(1), pp.308–311.

Shi, R., Wang, X., Lu, X., Zhu, Z., Xu, Q., Wang, H., Song, L. and Zhu, C. 2020. A systematic review of the clinical and genetic characteristics of Chinese patients with cystic fibrosis. *Pediatric Pulmonology*. **55**(11), pp.3005–3011.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*. **51**(3), pp.492–508.

Sica, G.L., Zhu, G., Tamada, K., Liu, D., Ni, J. and Chen, L. 2001. RELT, a new member of the tumor necrosis factor receptor superfamily, is selectively expressed in hematopoietic tissues and activates transcription factor NF-κB. *Blood*. **97**(9), pp.2702–2707.

Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G. and Ng, P.C. 2012. SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*. **40**(W1), pp.W452–W457.

Simmer, J. and Hu, J. 2002. Expression, Structure, and Function of Enamel Proteinases. *Connective Tissue Research*. **43**(2), pp.441–449.

Simmer, J.P. and Fincham, A.G. 1995. Molecular mechanisms of dental enamel formation. *Critical Reviews in Oral Biology and Medicine*. **6**(2), pp.84–108.

Simmer, J.P., Richardson, A.S., Smith, C.E., Hu, Y. and Hu, J.C.C. 2011. Expression of kallikrein-related peptidase 4 in dental and non-dental tissues. *European Journal of Oral Sciences*. **119**(SUPPL.1), pp.226–233.

Simpson, M.A., Hsu, R., Keir, L.S., Hao, J., Sivapalan, G., Ernst, L.M., Zackai, E.H., Al-Gazali, L.I., Hulskamp, G., Kingston, H.M., Prescott, T.E., Ion, A., Patton, M.A., Murday, V., George, A. and Crosby, A.H. 2007. Mutations in FAM20C are associated with lethal osteosclerotic bone dysplasia (Raine syndrome), highlighting a crucial molecule in bone development. *American Journal of Human Genetics*. **81**(5), pp.906–912.

Sire, J.Y., Davit-Béal, T., Delgado, S. and Gu, X. 2007. The origin and evolution of enamel mineralization genes. *Cells Tissues Organs*. **186**(1), pp.25–48.

Sire, J.Y. and Huysseune, A. 2003. Formation of dermal skeletal and dental tissues in fish: A comparative and evolutionary approach. *Biological Reviews of the Cambridge Philosophical Society*. **78**(2), pp.219–249.

Skuzeski, J.M., Nichols, L.M., Gesteland, R.F. and Atkins, J.F. 1991. The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons. *Journal of Molecular Biology*. **218**(2), pp.365–373.

Sloan, A.J. and Smith, A.J. 2007. Stem cells and the dental pulp: Potential roles in dentine regeneration and repair. *Oral Diseases*. **13**(2), pp.151–157.

Smale, S.T. 2010. Luciferase assay. *Cold Spring Harbor Protocols*. **5**(5), pdb.prot5421.

Smith, C.E. 1998. Cellular and chemical events during enamel maturation. *Critical Reviews in Oral Biology and Medicine*. **9**(2), pp.128–161.

Smith, C.E. and Warshawsky, H. 1977. Quantitative analysis of cell turnover in the enamel organ of the rat incisor. Evidence for ameloblast death immediately after enamel matrix secretion. *The Anatomical Record*. **187**(1), pp.63–97.

Smith, C.E.L., Kirkham, J., Day, P.F., Soldani, F., McDerra, E.J., Poulter, J.A., Inglehearn, C.F., Mighell, A.J. and Brookes, S.J. 2017. A fourth KLK4 mutation is associated with enamel hypomineralisation and structural abnormalities. *Frontiers in Physiology*. **8**(MAY).

Smith, C.E.L., Murillo, G., Brookes, S.J., Poulter, J.A., Silva, S., Kirkham, J., Inglehearn, C.F. and Mighell, A.J. 2016. Deletion of amelotin exons 3-6 is associated with amelogenesis imperfecta. *Human Molecular Genetics*. **25**(16), pp.3578–3587.

Smith, C.E.L., Poulter, J.A., Antanaviciute, A., Kirkham, J., Brookes, S.J., Inglehearn, C.F. and Mighell, A.J. 2017. Amelogenesis imperfecta; genes, proteins, and pathways. *Frontiers in Physiology*. **8**(JUN), p.435.

Smith, C.E.L., Poulter, J.A., Brookes, S.J., Murillo, G., Silva, S., Brown, C.J., Patel, A., Hussain, H., Kirkham, J., Inglehearn, C.F. and Mighell, A.J. 2019. Phenotype and Variant Spectrum in the LAMB3 Form of Amelogenesis Imperfecta. *Journal of Dental Research*. **98**(6), pp.698–704.

Smith, C.E.L., Whitehouse, L.L.E., Poulter, J.A., Wilkinson Hewitt, L., Nadat, F., Jackson, B.R., Manfield, I.W., Edwards, T.A., Rodd, H.D., Inglehearn, C.F. and Mighell, A.J. 2020. A missense variant in specificity protein 6 (SP6) is associated with amelogenesis imperfecta. *Human Molecular Genetics*. **29**(9), pp.1417–1425.

Smith, H.F. 2011. The Role of Genetic Drift in Shaping Modern Human Cranial Evolution: A Test Using Microevolutionary Modeling. *International Journal of Evolutionary Biology*. **2011**, pp.1–11.

Smith, M.M. and Coates, M.I. 1998. Evolutionary origins of the vertebrate dentition: Phylogenetic patterns and developmental evolution. *European Journal of Oral Sciences*. **106**(1 SUPPL.), pp.482–500.

Smith, N.G.C. and Eyre-Walker, A. 2002. Adaptive protein evolution in Drosophila. *Nature*. **415**(6875), pp.1022–1024.

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

Springer, M.S., Starrett, J., Morin, P.A., Lanzetti, A., Hayashi, C. and Gatesy, J. 2016. Inactivation of C4orf26 in toothless placental mammals. *Molecular Phylogenetics and Evolution*. **95**, pp.34–45.

Stakkestad, Ø., Heyward, C., Lyngstadaas, S.P., Medin, T., Vondrasek, J., Lian, A.M., Pezeshki, G. and Reseland, J.E. 2018. An ameloblastin C-terminus variant is present in human adipose tissue. *Heliyon*. **4**(12), p.e01075.

Staple, D.W. and Butcher, S.E. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biology*. **3**(6), pp.0956–0959.

Steenkamp, G. 2021. Elephant Dentistry *In*: *Zoo and Wild Animal Dentistry*. Wiley, pp.65–78.

De Summa, S., Malerba, G., Pinto, R., Mori, A., Mijatovic, V. and Tommasi, S. 2017. GATK hard filtering: Tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC Bioinformatics*. **18**(Suppl 5).

Suzuki, Y. and Gojobori, T. 1999. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*. **16**(10), pp.1315–1328.

Swenson, K.M. and El-Mabrouk, N. 2012. Gene trees and species trees: irreconcilable differences. *BMC bioinformatics*. **13 Suppl 1**(Suppl 19).

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J. and Von Mering, C. 2017. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*. **45**(D1), pp.D362–D368.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. **123**(3), pp.585–595.

Tam, B., Sinha, S. and Wang, S.M. 2020. Combining Ramachandran plot and molecular dynamics simulation for structural-based variant classification: Using TP53 variants as model. *Computational and Structural Biotechnology Journal*. **18**, pp.4033–4039.

Tan, L., Cheng, W., Liu, F., Wang, D.O., Wu, L., Cao, N. and Wang, J. 2021. Positive natural selection of N6-methyladenosine on the RNAs of processed pseudogenes. *Genome Biology*. **22**(1), p.180.

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*. **8**(2), pp.330–344.

Tasanen, K., Floeth, M., Schumann, H. and Bruckner-Tuderman, L. 2000. Hemizygosity for a glycine substitution in collagen XVII: Unfolding and degradation of the ectodomain. *Journal of Investigative Dermatology*. **115**(2), pp.207–212.

Tataru, P., Simonsen, M., Bataillon, T. and Hobolth, A. 2017. Statistical inference in the Wright-Fisher model using allele frequency data. *Systematic Biology*. **66**(1), pp.e30–e46.

Taylor, G.D. 2017. Molar incisor hypomineralisation. *Evidence-Based Dentistry*. **18**(1), pp.15–16.

Teare, M.D. and Barrett, J.H. 2005. Genetic Epidemiology 2: Genetic linkage studies. *Lancet*. **366**(9490), pp.1036–1044.

Teeling, E.C. and Hedges, S.B. 2013. Making the impossible possible: Rooting the tree of placental mammals. *Molecular Biology and Evolution*. **30**(9), pp.1999–2000.

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.

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.

Tjäderhane, L., Carrilho, M.R., Breschi, L., Tay, F.R. and Pashley, D.H. 2009. Dentin basic structure and composition-an overview. *Endodontic Topics*. **20**(1), pp.3–29.

Tompkins, K. 2006. Molecular mechanisms of cytodifferentiation in mammalian tooth development. *Connective Tissue Research*. **47**(3), pp.111–118.

Touat-Hamici, Z., Legrain, Y., Bulteau, A.L. and Chavatte, L. 2014. Selective up-regulation of human selenoproteins in response to oxidative stress. *Journal of Biological Chemistry*.

289(21), pp.14750–14761.

Tucker, T., Marra, M. and Friedman, J.M. 2009. Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. *American Journal of Human Genetics*. **85**(2), pp.142–154.

Tziafas, D. 1994. Mechanisms controlling secondary initiation of dentinogenesis: a review. *International Endodontic Journal*. **27**(2), pp.61–74.

Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T. and Posthuma, D. 2021. Genome-wide association studies. *Nature Reviews Methods Primers*. **1**(1), p.59.

Ullah, F., Hamilton, M., Reddy, A.S.N. and Ben-Hur, A. 2018. Exploring the relationship between intron retention and chromatin accessibility in plants. *BMC genomics*. **19**(1), p.21.

Ungar, P.S. 2010. *Mammal teeth: origin, evolution, and diversity*. JHU Press.

Urzúa, B., Martínez, C., Ortega-Pinto, A., Adorno, D., Morales-Bozo, I., Riadi, G., Jara, L., Plaza, A., Lefimil, C., Lozano, C. and Reyes, M. 2015. Novel missense mutation of the FAM83H gene causes retention of amelogenin and a mild clinical phenotype of hypocalcified enamel. *Archives of Oral Biology*. **60**(9), pp.1356–1367.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. and Velankar, S. 2022. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*. **50**(D1), pp.D439–D444.

Vieira, M.L.C., Santini, L., Diniz, A.L. and Munhoz, C. de F. 2016. Microsatellite markers: What they mean and why they are so useful. *Genetics and Molecular Biology*. **39**(3), pp.312–328.

Vizcaíno, S.F. 2009. The teeth of the “toothless”: novelties and key innovations in the evolution of xenarthrans (Mammalia, Xenarthra). *Paleobiology*. **35**(3), pp.343–366.

Vogel, P., Hansen, G.M., Read, R.W., Vance, R.B., Thiel, M., Liu, J., Wronski, T.J., Smith, D.D., Jeter-Jones, S. and Brommage, R. 2012. Amelogenesis Imperfecta and Other Biomineralization Defects in Fam20a and Fam20c Null Mice. *Veterinary Pathology*. **49**(6), pp.998–1017.

Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. 2006. A map of recent positive selection in the human genome L. Hurst, ed. *PLoS Biology*. **4**(3), pp.0446–0458.

Wang, S.K., Choi, M., Richardson, A.S., Reid, B.M., Lin, B.P., Wang, S.J., Kim, J.W., Simmer, J.P. and Hu, J.C.C. 2014. ITGB6 loss-of-function mutations cause autosomal recessive amelogenesis imperfecta. *Human Molecular Genetics*. **23**(8), pp.2157–2163.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. and Watson, M. 2015. Exome sequencing: Current and future perspectives. *G3: Genes, Genomes, Genetics*. **5**(8), pp.1543–1550.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., De Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R. and Schwede, T. 2018. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*. **46**(W1), pp.W296–W303.

Watson, J.D. and Crick, F.H.C. 1953. THE STRUCTURE OF DNA. *Cold Spring Harbor Symposia on Quantitative Biology*. **18**, pp.123–131.

Webb, A.E., Gerek, Z.N., Morgan, C.C., Walsh, T.A., Loscher, C.E., Edwards, S. V., O'Connell, M.J. and O'Connell, M.J. 2015. Adaptive Evolution as a Predictor of Species-Specific Innate Immune Response. *Molecular Biology and Evolution*. **32**(7), pp.1717–1729.

Webb, A.E., Walsh, T.A. and O'Connell, M.J. 2017. VESPA: Very large-scale evolutionary and selective pressure analyses. *PeerJ Computer Science*. **2017**(6).

Weerheijm, K.L., Jälevik, B. and Alaluusua, S. 2001. Molar-Incisor Hypomineralisation. *Caries Research*. **35**(5), pp.390–391.

Wei, L., Liu, Y., Dubchak, I., Shon, J. and Park, J. 2002. Comparative genomics approaches to study organism similarities and differences. *Journal of Biomedical Informatics*. **35**(2), pp.142–150.

Welch, J.J., Bininda-Emonds, O.R.P. and Bromham, L. 2008. Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evolutionary Biology*. **8**(1), p.53.

Wheeler, H.E., Metter, E.J., Tanaka, T., Absher, D., Higgins, J., Zahn, J.M., Wilhelmy, J., Davis, R.W., Singleton, A., Myers, R.M., Ferrucci, L. and Kim, S.K. 2009. Sequential use of transcriptional profiling, expression quantitative trait mapping, and gene association implicates MMP20 in human kidney aging G. Gibson, ed. *PLoS Genetics*. **5**(10), p.e1000685.

Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A.H., Karczewski, K., Ing, A.Y., Barton, P.J.R., Funke, B., Cook, S.A., Macarthur, D. and Ware, J.S. 2017. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine*. **19**(10), pp.1151–1158.

Whitney, M.R. and Sidor, C.A. 2019. Histological and developmental insights into the herbivorous dentition of tapinocephalid therapsids. *PLoS ONE*. **14**(10), pp.1–21.

Williams, E.J.B. and Bowles, D.J. 2004. Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Research*. **14**(6), pp.1060–1067.

Williams, I., Richardson, J., Starkey, A. and Stansfield, I. 2004. Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae. *Nucleic Acids Research*. **32**(22), pp.6605–6616.

Witkop, C.J. and Sauk, J.J. 1976. Heritable Defects of Enamel *In*: *Oral Facial Genetics*. Mosby, pp.151-226.

Wright, J.T., Johnson, L.B. and Fine, J.D. 1993. Developmental defects of enamel in humans with hereditary epidermolysis bullosa. *Archives of Oral Biology*. **38**(11), pp.945–955.

Wright, J.T., Torain, M., Long, K., Seow, K., Crawford, P., Aldred, M.J., Hart, P.S. and Hart, T.C. 2011. Amelogenesis imperfecta: Genotype-phenotype studies in 71 families. *Cells Tissues Organs*. **194**(2–4), pp.279–283.

Wright, S. 1931. Evolution in Mendelian Populations. *Genetics*. **16**(2), pp.97–159.

Wroe, S. and Milne, N. 2007. Convergence and remarkably consistent constraint in the evolution of carnivore skull shape. *Evolution*. **61**(5), pp.1251–1260.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. and Zhang, Y. 2014. The I-TASSER suite: Protein structure and function prediction. *Nature Methods*. **12**(1), pp.7–8.

Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., Hardison, M., Person, R., Bekheirnia, M.R., Leduc, M.S., Kirby, A., Pham, P., Scull, J., Wang, M., Ding, Y., Plon, S.E., Lupski, J.R., Beaudet, A.L., Gibbs, R.A. and Eng, C.M. 2013. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New England Journal of Medicine*. **369**(16), pp.1502–1511.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*. **24**(8), pp.1586–1591.

Yuen, W.Y., Pasmooij, A.M.G., Stellingsma, C. and Jonkman, M.F. 2012. Enamel defects in carriers of a novel LAMA3 mutation underlying epidermolysis bullosa. *Acta Dermato-Venereologica*. **92**(6), pp.695–696.

Zhang, Z., Tian, H., Lv, P., Wang, W., Jia, Z., Wang, S., Zhou, C. and Gao, X. 2015. Transcriptional factor DLX3 promotes the gene expression of enamel matrix proteins during amelogenesis. *PLoS ONE*. **10**(3), pp.1–15.

Zhao, H., Feng, J., Seidel, K., Shi, S., Klein, O., Sharpe, P. and Chai, Y. 2014. Secretion of shh by a neurovascular bundle niche supports mesenchymal stem cell homeostasis in the adult mouse incisor. *Cell Stem Cell*. **14**(2), pp.160–173.

Zheng, W., Zhang, C., Li, Y., Pearce, R., Bell, E.W. and Zhang, Y. 2021. Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods*. **1**(3), p.100014.

Zhou, Y., Shearwin-Whyatt, L., Li, J., Song, Z., Hayakawa, T., Stevens, D., Fenelon, J.C., Peel, E., Cheng, Y., Pajpach, F., Bradley, N., Suzuki, H., Nikaido, M., Damas, J., Daish, T., Perry, T., Zhu, Z., Geng, Y., Rhie, A., Sims, Y., Wood, J., Haase, B., Mountcastle, J., Fedrigo, O., Li, Q., Yang, H., Wang, J., Johnston, S.D., Phillippy, A.M., Howe, K., Jarvis, E.D., Ryder, O.A., Kaessmann, H., Donnelly, P., Korlach, J., Lewin, H.A., Graves, J., Belov, K., Renfree, M.B., Grutzner, F., Zhou, Q. and Zhang, G. 2021. Platypus and echidna genomes reveal mammalian biology and evolution. *Nature*. **592**(7856), pp.756–762.