Uncovering the genetic basis of the hereditary tooth enamel disorder Amelogenesis Imperfecta

Ummey Hany

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

The University of Leeds
School of Medicine and Health
October 2023

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 2: A targeted smMIPs screen of non-syndromic Amelogenesis Imperfecta.
Hany U, Watson CM, Liu L, Nikolopoulos G, Smith CEL, Poulter JA, Antanaviciute A, Rigby A, Balmer R, Brown CJ, Patel A, De Camargo MGA, Rodd HD, Moffat M, Murillo G, Mudawi A, Jafri H, Inglehearn CF, Mighell AJ.
Status: Final draft stage, to submit to Human Mutation.

This paper describes the development of a smMIPs (single molecule molecular inversion probes) reagent for amelogenesis imperfecta (AI) and reports its success in solving unsolved AI cases. The candidate developed the targeted smMIPs reagent to selectively screen 19 genes involved in non-syndromic (NS) AI and adapted a bioinformatics pipeline to analyse the data for Leeds patients, with the help of her supervisors Professor Chris Inglehearn, Dr Alan Mighell and Dr Christopher Watson. She then used the smMIPs reagent to screen the DNA of 181 unsolved patients. She performed all the laboratory work related to smMIPs sequencing and bioinformatics analysis of the smMIPs data, for all the patients, and wrote the first draft of the manuscript. Lu Liu, a technician, performed Sanger sequencing confirmation of all the variants detected by smMIPs under the supervision of the candidate.

Chapter 3: Novel ameloblastin variants, contrasting Amelogenesis Imperfecta phenotypes.
Hany U, Watson CM, Liu L, Nikolopoulos G, Smith CEL, Poulter JA, Brown CJ, Patel A, Rodd HD, Balmer R, Harfoush A, Al-Jawad M, Inglehearn CF, Mighell AJ. Novel Ameloblastin Variants, Contrasting Amelogenesis Imperfecta Phenotypes. J Dent Res. 2024 Jan;103(1):22-30. doi: 10.1177/00220345231203694. Epub 2023 Dec 6. PMID: 38058155; PMCID: PMC10734210.

This paper summarizes findings from the smMIPs screen, and from whole exome sequencing (WES), specifically relating to AI caused by variants in *AMBN*, including differing phenotypes and modes of inheritance. The candidate performed all the laboratory work related to the sequencing of patients' DNA, data analysis and critical analysis of the results with the help of her supervisory team. Another PhD student, Asmaa Harfoush, performed phenotype analyses of the teeth of patients involved in this study, and Lu Liu performed Sanger sequencing confirmation of all the variants detected by smMIPs, under the supervision of the candidate. The candidate wrote the first draft of the paper and managed the master copy, inputting suggestions from others during the writing process.

Chapter 4: Heterozygous *COL17A1* variants are a frequent cause of Amelogenesis Imperfecta.

Hany U, Watson CM, Liu L, Smith CEL, Harfoush A, Poulter JA, Nikolopoulos G, Balmer R, Brown CJ, Patel A, Simmonds J, Charlton R, Acosta de Camargo MG, Rodd HD, Jafri H, Antanaviciute A, Moffat M, Al-Jawad M, Inglehearn CF, Mighell AJ. Heterozygous COL17A1 variants are a frequent cause of amelogenesis imperfecta. J Med Genet. 2023 Nov 18:jmg-2023-109510. doi: 10.1136/jmg-2023-109510. Epub ahead of print. PMID: 37979963.

This paper describes findings from the smMIPs screen, and whole exome sequencing (WES), that specifically relate to AI caused by variants in *COL17A1*, and potential links to skin and corneal diseases. This paper reports heterozygous variants in the *COL17A1* gene as a common cause of dominant AI. *COL17A1* variants are not listed as a cause of dominant nonsyndromic (NS) AI in genetic disease databases such as OMIM (Online Mendelian Inheritance in Man), so the described results highlight an observation that had been largely overlooked and poorly documented in the current literature. The candidate performed all of the sequencing-related laboratory work, the data analysis for all the patients and the critical analysis of the results with the help of her supervisory team. Once again Asmaa Harfoush did phenotype analyses of the teeth of patients involved in this study, and Lu Liu performed Sanger sequencing confirmation of all the variants detected by smMIPs analysis under the supervision of the candidate.

The candidate wrote the first draft of the paper and managed the master copy, inputting suggestions from others during the writing process.

Acknowledgements

In the name of Allah, the most gracious and the most merciful. All praise belongs to Him, lord of all that exists.

I would like to express my sincere gratitude to all my supervisors, Prof Chris Inglehearn, Dr Alan Mighell and Dr Christopher Watson, for their unwavering guidance and encouragement throughout my doctoral journey. Chris has been a mentor in every aspect of my PhD life, enriched my academic experience and contributed significantly to my personal growth as a researcher. His constant support, constructive criticism, and valuable suggestions have been instrumental in refining my research. I am grateful for the countless hours he dedicated to discussing my ideas, reviewing my drafts and providing constructive feedback, without all these I wouldn't be able to achieve my goals. I am deeply grateful to Chris for believing in me and being an exceptional mentor. My deepest appreciation also goes to Alan for his invaluable input in my research. His knowledge of AI phenotypes is unparalleled. His passion for research, commitment to excellence and willingness to share his clinical expertise have been significant in shaping my academic trajectory. I am grateful for the opportunities to collaborate and learn from his experiences. The mentorship of Christopher Watson has been a source of motivation, his profound knowledge of genetics and technical understanding of various sequencing technologies have been pivotal in helping me overcome challenges and achieve my goals. I am extremely fortunate to be guided by a supervisory team, who were exceptional in the depth of knowledge and proficiency in their respective areas. Their collective experiences and guidance during challenging moments have significantly enriched the quality and impact of this work and will undoubtedly serve me well in my future endeavours.

I want to sincerely thank every member of my group for being a part of my research journey. I am lucky to have had the opportunity to work with such friendly and talented people, who were welcoming to provide a sense of belonging in the group and whose critical feedback and constructive suggestions have left a lasting effect on my academic development.

I am also thankful to the NGS facility staff for their open and responsive communication ensuring seamless and high-quality data were generated which greatly contributed to the success of this project.

I want to extend a special thanks to my family, their encouragement and positivity have been a constant source of strength which made a meaningful impact on my PhD.

Abstract

Enamel is the wear-resistant outer layer of the dental crown and is the hardest tissue in the human body. Enamel is formed from the secretion of an extracellular matrix and subsequent mineralization by secretory ameloblasts, in a process termed amelogenesis. AI refers to a group of rare, Mendelian disorders caused by abnormal amelogenesis. AI can be isolated or part of syndromic conditions; to date, 20 genes are implicated in non-syndromic (NS) AI. To uncover the genetic basis of AI, a DNA sequencing method was developed to target AI-associated genes, using single molecule molecular inversion probes (smMIPs) technology. Using smMIPs, 181 unsolved AI cases were screened, with 36% of these solved using the assay. smMIP-unsolved samples were further analysed by whole exome sequencing, identifying several new candidate AI genes.

Comprehensive genetic analysis of this AI cohort identified variants in *COL17A1*, previously overlooked, as the most frequent cause of AI. Homozygous *COL17A1* variants cause the recessive skin disorder epidermolysis bullosa, highlighting a need for a multidisciplinary approach to the clinical management of these patients. This study also further added to our understanding of the molecular mechanisms underlying contrasting phenotypes in patients with AI due to dominant or recessive mutations in *AMBN*. Another intriguing finding was the detection of a 587 bp homozygous deletion in *PLXNB2* in a syndromic patient, strengthening evidence for *PLXNB2* as a novel gene involved in recessive syndromic AI.

The smMIPs method proved to be an effective first-line screen, offering patients a rapid, low-cost, diagnosis. The technique could now be expanded to investigate whether AI genes contribute to the risk of other dental conditions such as fluorosis or molar incisor hypomineralisation. Cell modelling and functional experiments are planned to demonstrate the biological relevance of the candidate AI-associated genes that were identified. These experiments will advance our understanding of disease mechanisms with the aim of ultimately improving clinical care.

CONTENTS

CHAPTER 1 Introduction....15
1.1 Structure of the Tooth....16
1.1.1 Different layers of a tooth and surrounding tissues....16
1.2 Odontogenesis....20
1.2.1 Initiation (Placode stage)....20
1.2.2 Morphogenesis and Histodifferentiation....21
1.2.2.1 Bud stage....21
1.2.2.2 Cap stage....22
1.2.2.3 Bell stage....24
1.2.2.4 Crown and root formation....26
1.2.2.5 Eruption stage....26
1.2.3 Cementogenesis....26
1.2.4 Dentinogenesis....27
1.2.5 Amelogenesis....29
1.2.5.1 Enamel matrix proteins and proteases....30
1.2.5.2 Stages of Amelogenesis....31
1.2.5.3 Ion transport in the maturation stage of Amelogenesis....37
1.2.5.4. pH balance....38
1.2.5.5 Amelogenesis imperfecta (AI)....40
1.2.5.5.1 Clinical Classification of AI....41
1.2.5.5.2 Genes associated with non-syndromic AI....45
1.2.5.5.3 Syndromes associated with AI....56
1.2.5.5.4 Environmental causes of enamel defects....60
1.2.5.5.5 Impact of AI....62
1.2.5.5.6 Clinical treatments for AI....65
1.3 DNA Sequencing Technologies....68
1.3.1 First generation DNA sequencing....69
1.3.1.1 Chain-termination sequencing....70
1.3.2 Next generation sequencing (NGS)....72
1.3.2.1 Whole exome sequencing....76
1.3.2.2 NGS data analysis....77
1.3.2.3 Single molecule molecular inversion probes (smMIPs)....79
1.3.3 Third generation sequencing....82
1.3.3.1 Nanopore sequencing....83

1.3.3.2 PacBio/SMRT sequencing....85
1.4 Aims....88
1.5 References....90
CHAPTER 2 A targeted smMIPs screen of non-syndromic Amelogenesis Imperfecta....102
2.1 Research Rationale....103
2.2 Research Contribution....105
2.2.1 Designing smMIPs probes using MIPGEN....105
2.2.2 Sample preparation for smMIPs screening....106
2.2.3 Data analysis....106
2.2.4 Additional methodology....107
CHAPTER 3 Novel ameloblastin variants, contrasting Amelogenesis Imperfecta phenotypes....134
3.1 Research Rationale....135
3.2 Research Contribution....136
3.2.1 Additional methodology....136
CHAPTER 4 Heterozygous *COL17A1* variants are a frequent cause of Amelogenesis Imperfecta....162
4.1 Research Rationale....163
4.2 Research Contribution....164
4.2.1 Additional methodology....164
CHAPTER 5 Discussion and Conclusion....205
5.1 Summary of the project, problems encountered and lessons learned....206
5.1.1 Improvement in wet lab methods....207
5.1.2 Selecting appropriate data analysis algorithms....207
5.1.3 Possible explanations for cases screening negative in smMIPs....208
5.2 Main findings....209
5.2.1 COL17A1 variants are a frequent cause of nonsyndromic AI....210
5.2.2 Variants in AMBN cause both dominant and recessive AI with contrasting phenotypes....210
5.2.3 Potential novel candidate genes for AI....211
5.2.3.1 Potential novel CNVs associated with AI....215
5.3 Future of AI research....216
5.3.1 Improving our genetic and genomic understanding of AI....217
5.3.1.1 Potential improvements in the analysis of short read sequencing data 217
5.3.1.2 Long read sequencing technology....220
5.3.2 Disease modelling....221
5.3.3 Clinical research....222

5.3.4 Biobank 223
5.3.5 Shared platform between geneticists and clinicians 223
5.4 Benefit of genetics testing 223
5.4.1 Psychosocial support 224
5.4.2 Personalised medicine 224
5.4.3 AI gene-disease catalogue 225
5.4.4 Support group network 225
5.5 Future of genetics in healthcare 225
5.6 References 230
APPENDICES 234
Appendix 1 Running MIPgen to design MIP probes. 234
Appendix 2 Sequences of MIP backbone, Custom read 1, read 2 and index primers. 235
Appendix 3 MiSeq and NextSeq MIP sequencing protocol using custom primers. 235
Appendix 4: Setting up the MIPVAR pipeline. 237
Appendix 5: Exome library prep using Twist comprehensive exome kit. 242
Appendix 6: Exome data analysis pipeline. 245
Appendix 7: Data processing commands for long read sequencing. 249

List of Figures

Figure 1. 1 Structure of a human tooth. 19
Figure 1. 2 Structure of a tooth germ at cap stage. 23
Figure 1. 3 Histologic slide of tooth in late bell stage. 25
Figure 1. 4 Herringbone pattern in human enamel. 33
Figure 1. 5 Stages of amelogenesis. 36
Figure 1. 6 Model of ion channels and transport mechanisms at the maturation stage of amelogenesis.....39
Figure 1. 7 Clinical images of different types of amelogenesis imperfecta. 44
Figure 1. 8 Schematic of chain termination method.....71
Figure 1. 9 Sequencing by synthesis (SBS) method.....74
Figure 1. 10 Features of a smMIP probe. 81
Figure 1. 11 Schematic representation of nanopore sequencing. 84
Figure 1. 12 Schematic illustration of SMRT bell sequencing. 87

List of Table

Table 1. 1 Genes associated with non-syndromic AI.....53

Abbreviations

ACP4	acid phosphatase 4
ACMG	American college of medical genetics and genomic
AD	autosomal dominant
AE	anion exchanger
AI	amelogenesis imperfecta
AIHHT	AI hypomaturation-hypoplasia type with taurodontism
ALP	alkaline phosphatase
AMBN	ameloblastin
AMEL	amelogenin, refers to both copies present on chromosomes X and Y in humans
AMELX	amelogenin, X linked
AMELY	amelogenin, Y linked
AMTN	amelotin
AR	autosomal recessive
BAM	binary alignment map
BCL	binary base call
BMP	bone morphogenetic protein
BPA	bisphenol A
BWA	Burrows-Wheeler aligner
bp	base pair
Cas9	CRISPR associated protein 9
CF	cystic fibrosis
CFTR	cystic fibrosis transmembrane conductance regulator
CIF	common intermediate format
CRAC	calcium-selective release-activated calcium channel
COL1A1	collagen, type 1, alpha 1
COL7A1	collagen, type 7, alpha 1
COL17A1	collagen, type 17, alpha 1
CNNM4	cyclin and CBS domain divalent metal cation transport mediator 4
CNV	copy number variant
CRAC	Ca^{2+} release activated channel
CRISPR	clustered regularly interspaced short palindromic repeats
DASS	dental anomalies and short stature
DDE	developmental defects of enamel
DEE25	developmental and epileptic encephalopathy-25 with AI
DEJ	dentino-enamel junction
DI	dentinogenesis imperfecta
DLX3	distal-less homeobox 3
DMP1	dentin matrix acidic phosphoprotein
DNA	deoxyribonucleic acid
dNTPs	deoxynucleotide triphosphates
ddNTPs	dideoxynucleotide triphosphates
dRTA	distal renal tubular acidosis
DSPP	dentin sialophosphoprotein
EB	epidermolysis bullosa
ECM	extracellular matrix

EDC	endocrine-disrupting chemical
EMPs	enamel matrix proteins
ENAM	enamelin
ER	endoplasmic reticulum
ERED	epithelial recurrent erosion dystrophy
ERS	enamel renal syndrome
FAM20A	family with sequence similarity 20, member A
FAM20C	family with sequence similarity 20, member C
FAM83H	family with sequence similarity 83, member H
FGF	fibroblast growth factor
FN1	fibronectin 1
GATK	genome analysis toolkit
GJ	gap junction
GINA	genetic information nondiscrimination act
GO	gene ontology
GPR68	G-protein coupled receptor 68
GRCH	genome reference consortium human
GTEx	genotype tissue expression
HAP	hydroxyapatite
hg19	human genome assembly 19
hg20	human genome assembly 20
hiPSC	human-induced pluripotent stem cell
HMLR	heimler syndrome
HOMG5	hypomagnesemia-5, renal with or without ocular involvement
IEE	inner enamel epithelium
IGV	integrative genomics viewer
IMD10	primary immunodeficiency-10
INDELS	insertions/deletions
ITGB6	integrin, beta 6
JEB	junctional epidermolysis bullosa
KLK4	kallikrein related peptidase 4
KTZS	Kohlschutter-Tonz syndrome
LAMA3	laminin, alpha 3
LAMB3	laminin, beta 3
LAMC2	laminin, gamma 2
LINE	long interspersed elements
MAF	minor allele frequency
MSX2	msh homeobox 2
µCT	micro computerised tomography
MELT	mobile element locator tool
MMP20	matrix metalloproteinase 20
MIH	molar incisor hypomineralization
MLPA	multiplex ligation-dependent probe amplification
mRNA	messenger RNA
NaCT	Na+/citrate cotransporter
NaPi2b	sodium-dependent phosphate transport protein 2B

NCKX4	potassium dependent sodium/calcium exchanger
NGS	next generation sequencing
NHE	sodium-proton exchanger
NHS	national health service
NMD	nonsense mediated decay
NS	non-syndromic
ODAM	odontogenic, ameloblast associated
ODAPH	odontogenesis associated phosphoprotein
OEE	outer enamel epithelium
OMIM	online Mendelian Inheritance in Man
ONT	Oxford Nanopore Technology
ORAI	calcium release-activated calcium channel protein 1
OSG	online support group
PacBio	Pacific Biosciences
PAX9	paired homeobox 9
PBA	phenylbutyrate
PCDH11	protocadherin 11
PCR	polymerase chain reaction
PEX1	peroxisomal biogenesis factor 1
PEX6	peroxisomal biogenesis factor 6
PEX26	peroxisomal biogenesis factor 26
4-PBA	4-phenylbutyrate
PROMs	patient-reported outcome measures
PRS	polygenic risk score
PTC	premature termination codon
RE-A	ruffle ended ameloblast
RELT	receptor expressed in lymphoid tissues
RELN	reelin
RNA	ribonucleic acid
ROGDI	rogdi atypical leucine zipper
RUNX2	runt-related transcription factor 2
SAM	sequence alignment map
SBS	Sequencing by synthesis
SCPP	secretory calcium-binding phosphoproteins
SE-A	smooth ended ameloblast
SEM	scanning electron microscopy
SHH	sonic hedgehog
SINE	short interspersed element
SSASKS	Short stature, amelogenesis imperfecta, and skeletal dysplasia with scoliosis
SLC24A4	solute carrier family 24 (sodium/potassium/calcium exchanger), member 4
SLC13A5	solute carrier family 13 member 5
smMIP	single molecule molecular inversion probes
SMRT	single molecule real time
SNP	single nucleotide polymorphism
SOCE	store-operated Ca^{2+} entry
SP6	specificity protein 6.

STIM	stromal interaction molecule
TDO	trichodentoosseous syndrome
T2T	telomere to telomere
TIFF	tag image file format
TRPM7	transient receptor potential cation channel, subfamily M, member 7
TUFT1	tuftelin
UV	ultraviolet
UMI	unique molecular identifier
uORFs	upstream open reading frames
UQCRC1	ubiquinol-cytochrome c reductase core protein 1
UTR	untranscribed region
VEP	variant effect predictor
VUS	variant of unknown significance
WDR72	WD repeat-containing protein 72
WES	whole exome sequencing
WGS	whole genome sequencing
ZS	Zellweger syndrome
ZMW	zero mode waveguides

CHAPTER 1 Introduction

1.1 Structure of the Tooth

A tooth is a hard, calcified structure located in the jaws of many vertebrates, including humans. It serves several functions, including cutting and grinding food during the process of chewing, aiding in speech, and contributing to facial aesthetics. Humans usually have two successive sets of teeth during life, referred to as the primary and permanent dentitions. Whether primary or permanent, teeth can be classified morphologically as incisors, canines, pre-molars or molars and each type has unique features distinguishable from each other (Jernvall & Thesleff, 2012). The structure of a typical human tooth can be divided into three main parts: the crown, the neck, and the root. A typical tooth has a crown that is exposed in the oral cavity and below the crown is the root that is buried inside the jawbone. The section connecting the crown and root is the neck. Human teeth have a very limited regeneration capacity, so once formed they are required to last a lifetime.

1.1.1 Different layers of a tooth and surrounding tissues

Teeth are composed of a series of component layers and structures, as displayed in Figure 1. 1. These structures, and the tissues that surround them, are defined as follows (Darling, 1959; Hildebrand et al., 1995; Jernvall & Thesleff, 2012; Taft, 1881):

Enamel: The top hard layer of the crown is the hardest tissue in our body and protects the underlying layers from wear and decay. It is composed primarily of a mineral called hydroxyapatite and is epithelium derived. Hydroxyapatite (HAP) is a type of calcium phosphate mineral and is often found in the form of small, needle-like crystals, which

are the main inorganic component of bone and teeth in vertebrates, providing structural support and rigidity to these tissues.

Dentine: Underneath the enamel is a mesenchyme-derived, less mineralized, softer dentine core, inside which runs the dentinal tubule containing tissue fluid that transmits sensations like temperature and pain to the dental pulp. Dentine provides vital mechanical and biochemical support to the enamel (Pashley, 1989).

The dentino-enamel junction (DEJ): The interface between enamel and dentine is the DEJ. It creates a strong mechanical bond between enamel and dentine. This mechanical bond ensures efficient distribution of forces during biting and chewing. It also enables the tooth to withstand the stresses of everyday use without fracturing or breaking at the junction between enamel and dentine.

Cementum: Cementum and alveolar bone are connected by periodontal ligaments that attach the tooth to the surrounding alveolar bone. Cementum functions to provide a stable and robust attachment of the tooth to the surrounding bone, ensuring the tooth's support and proper function during biting and chewing.

Dental pulp: The dental pulp is in the centre of the tooth and consists of connective tissue, blood vessels and a network of nerves that are responsible for transmitting pain, pressure and temperature sensations. Blood vessels deliver nutrients and oxygen to the odontoblasts during tooth development. Dental pulp occupies the pulp chamber in the crown and extends from the crown to the tip of the root through the root canal.

Periodontal ligament: This is fibrous tissue connecting the tooth root and the alveolar bone. It safeguards the alveolar bone from forces generated while chewing food (Berkovitz, 2004).

Alveolar bone: This is the underlying jawbone within which the teeth are rooted.

Gingiva: Commonly known as gums, is soft tissue that covers the alveolar bone. The top layer of gingiva is composed of epithelial structures covering the underlying fibrous connective tissues (Bartold et al., 2000).

Gingival sulcus: There is a small space between the tooth and the gums which can increase due to inflammation. This is called the periodontal pocket or gingival pocket.

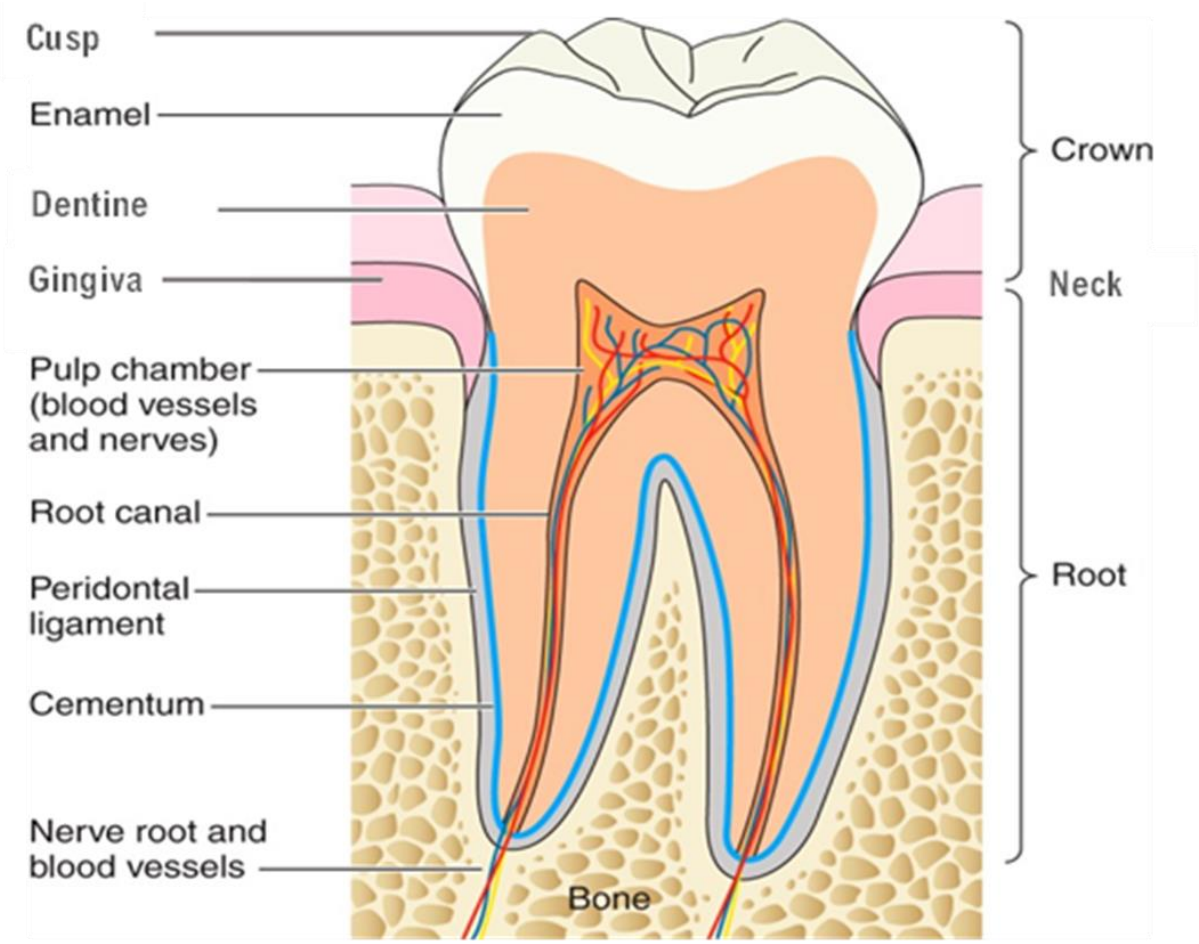


Figure 1. 1 Structure of a human tooth.

Cross section of a human tooth showing crown, neck, and root. The image is also showing the different layers of mineralized and non-mineralized tissues that make up a tooth. Image adapted with permission from Lacruz et al., 2017 under the terms of the Creative Commons CC-BY license.

1.2 Odontogenesis

Odontogenesis is the process of tooth development, which begins with the formation of tooth buds in the embryonic stage and ends with the eruption of teeth into the oral cavity. It occurs in multiple tightly regulated stages, detailed below. These stages are characterized by a complex interaction between dental epithelium and cranial neural crest-derived mesenchyme that controls several critical processes at all stages of tooth development, from initiation to crown morphology. Determination, initiation and differentiation of tooth development are facilitated by evolutionarily conserved transcription factors and signalling pathways to ensure teeth develop in the correct place, at the correct time and with correct morphology. Both primary and permanent teeth undergo the same mechanisms of development (Schour, 1948; Thesleff, 2006; Thesleff et al., 1990).

1.2.1 Initiation (Placode stage)

Between weeks 6 and 7 of human intra-uterine life, the oral epithelium signals to the underlying mesenchyme via bone morphogenic factors (BMPs including BMP2, BMP4), sonic hedgehog (Shh), fibroblast growth factors (Fgfs) and Wnt/β-catenin signalling pathways to initiate the physiologic process of induction. As a result, the oral epithelium thickens, forming a primary epithelial band situated above the connective tissue or mesenchyme. By week seven, the primary epithelial band divides to form,

- A buccal placed vestibular lamina for the development of the vestibule of the mouth (lining of lips, cheeks, and buccal sulcus).

- A lingual placed dental lamina forms the upper and lower dental arches and contributes to the development of ten primary tooth germs in each arch by the end of week 8.

1.2.2 Morphogenesis and Histodifferentiation

After initiation, developing tooth germs undergo several stages of structural and histological changes to develop morphologically different tooth types, incisor, canine, premolar and molar teeth. Each tooth further differentiates to give rise to the functionally different tissues enamel, dentine, dental pulp and periodontal tissues, in varying quantities and distributions. These stages can be classified into bud, cap and bell stages according to the degree of morpho- and histodifferentiation of the enamel organs, as detailed below. The enamel organ is a complex epithelial cell aggregation formed by dental lamina, observed exclusively in the developing tooth, which forms enamel and participates in the formation of the dentine, the enamel crown and the DEJ.

1.2.2.1 Bud stage

At human embryonic week 8, induction of the physiologic process gives rise to a series of epithelial swellings, these being the tooth buds, within the dental lamina, corresponding with the position of the future primary dentition. The enamel organ appears simple, with no distinction evident between different cell types at this stage.

1.2.2.2 Cap stage

The cap stage begins at about the tenth week of human prenatal development. During this stage, unequal proliferation in different parts of the tooth bud give rise to the 3D cap shape of the enamel organ. At approximately week 12, the enamel organ enlarges and early histodifferentiation becomes visible, with distinguishable stellate reticulum, cuboidal outer enamel epithelium (OEE) and columnar inner enamel epithelium (IEE). Mesenchymal cells underneath the IEE aggregate to form dental papillae and these cells further proliferate to form a dental follicle surrounding the enamel organ and dental papilla. By the end of cap stage, three embryonic structures make up the tooth germ: the enamel organ, the dental papilla and the dental follicle as shown in Figure 1. 2.

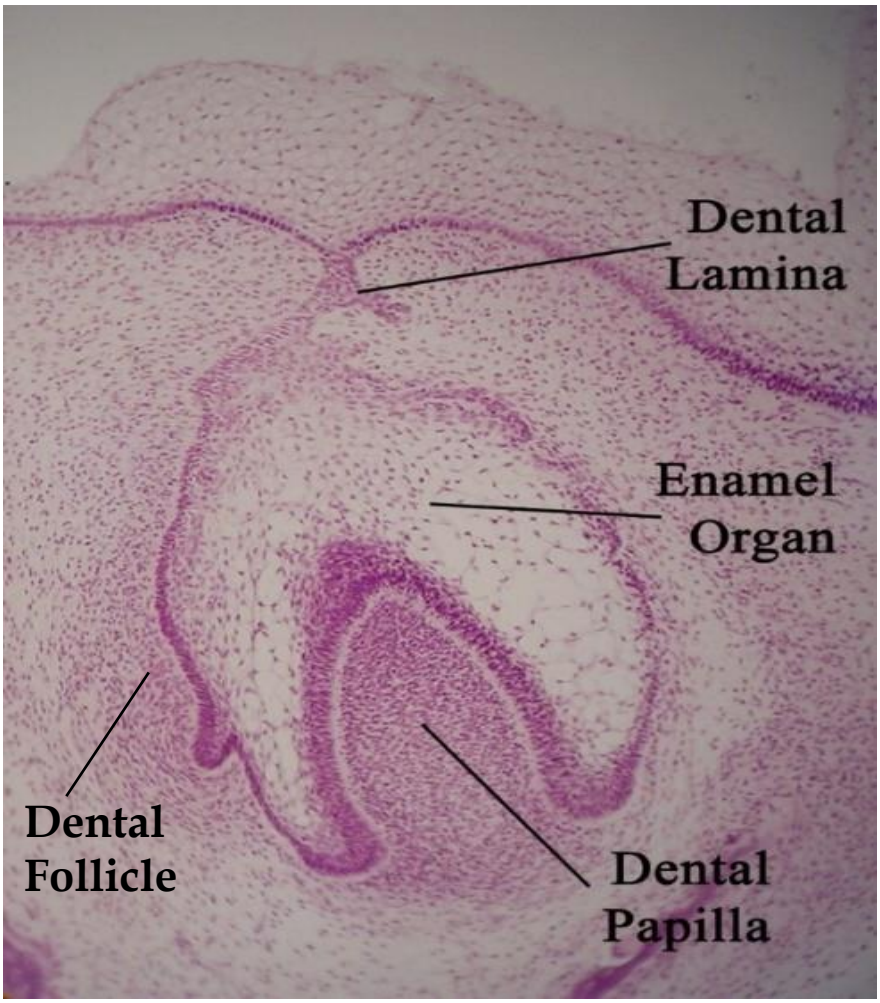


Figure 1. 2 Structure of a tooth germ at cap stage.

The image shows the embryonic structures that make up a tooth germ at the cap stage. The enamel organ looks like a 3D cap, underneath which is the dental papilla, and surrounding the enamel organ and dental papilla is the dental follicle. Image reproduced with permission from Dozenist, available at https://commons.wikimedia.org/w/index.php?curid=427391.

1.2.2.3 Bell stage

By week 14, the enamel organ grows into a recognisable tooth shape. Histodifferentiation at this stage clearly shows four distinctive cell layers of the enamel organ, the IEE, stratum intermedium, stellate reticulum and OEE as detailed in Figure 1. 3. Differential mitosis in the IEE gives rise to a folding characteristic shape of the crown of the tooth by morphodifferentiation. Cusp tips are also formed by cells with no mitotic activity. Each enamel organ develops an extension to the lingual side of each primary tooth germ called the successional lamina, from which the permanent incisor, canine and premolar teeth develop. The dental papilla then undergoes histodifferentiation to form two types of tissues; the outer cells adjacent to the IEE of the dental papilla, which will differentiate into odontoblasts, and the central cells of the dental papilla which will differentiate into pulp tissue.

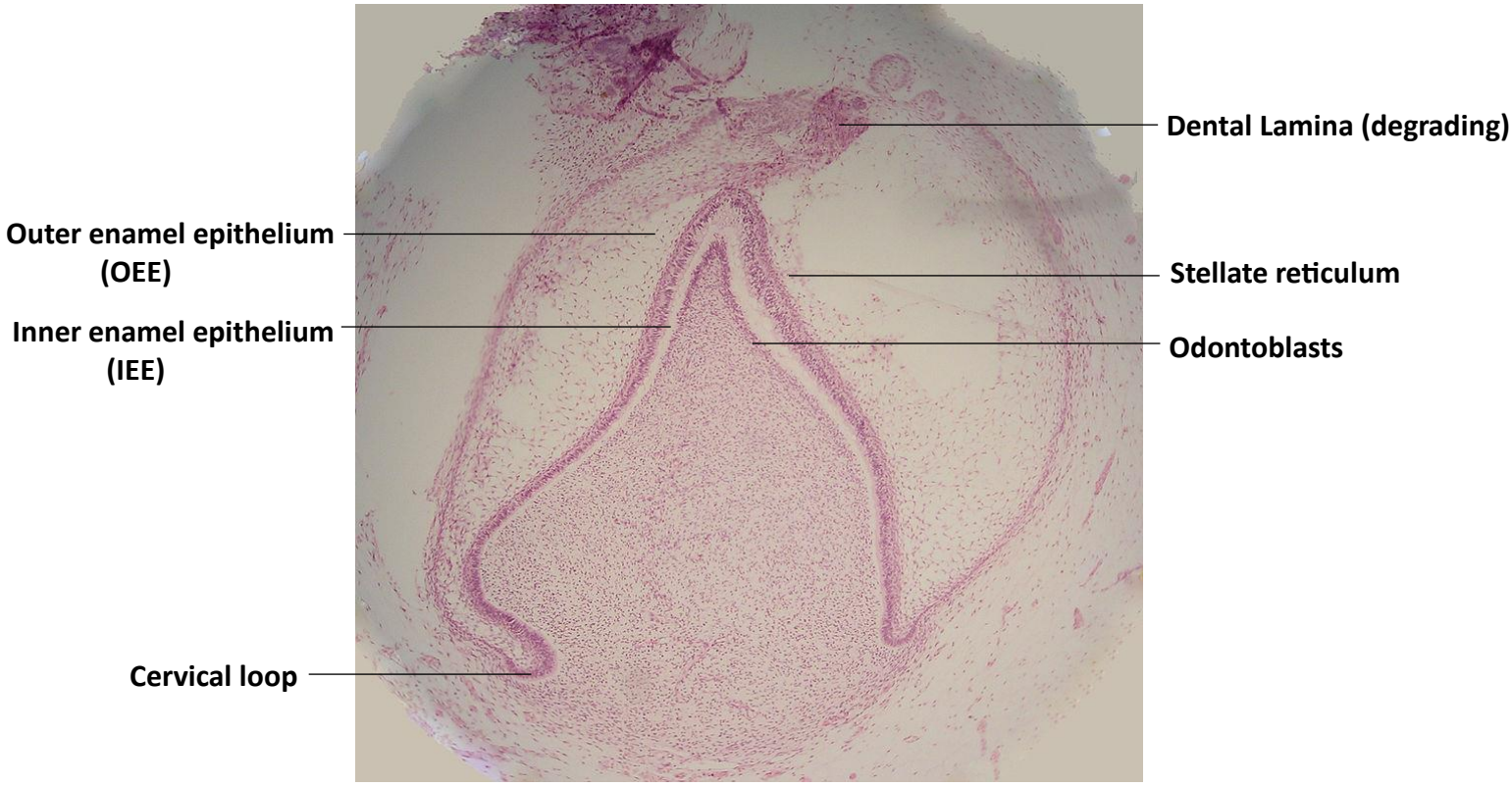


Figure 1. 3 Histologic slide of tooth in late bell stage.

The image shows the structures of a tooth germ at late bell stage. The inner enamel epithelium folds to give rise to the cusp or incisal shape of the crown of the tooth. Histodifferentiation has now occurred, with four layers of the enamel organ clearly visible. The dental lamina disintegrates. The rim of the enamel organ, the cervical loop, now develops. This is where the outer and inner enamel epithelium meet. Image adapted with permission from Dozenist, https://commons.wikimedia.org/w/index.php?curid=427400.

1.2.2.4 Crown and root formation

At late bell stage, which occurs at about the 18th week of human gestation, morphodifferentiation of the crown is complete and the IEE and OEE at the cervical loop of the enamel organ fuse to form a bilayered epithelial sheath termed Hertwig's epithelial root sheath. Cells of the dental follicle are found adjacent to the sheath, and it encloses the dental papilla, outlining the shape of the future root. The sheath grows apically downwards forming the tooth root by interacting with the dental follicle and dental papilla.

1.2.2.5 Eruption stage

Once the crown of the tooth has formed and the root has started to develop, the tooth moves vertically toward the oral cavity so that it can erupt into the correct position. Some of the jawbone above the tooth resorbs and other connective tissues break down to help the tooth move. Depending on the type and position of the tooth, each erupts at different ages.

1.2.3 Cementogenesis

Cementum is less mineralized than dentine and enamel. It is composed approximately 40-45% of an organic matrix primarily made up of type I (90%) and type III collagen, 45-50% of inorganic matrix as a form of calcium HAP mineralized crystals and 10% water by weight. In addition to collagen, the organic matrix also contains the glycosaminoglycans hyaluronic acid, dermatan sulphate, chondroitin sulphate and keratan sulphate. There are many non-collagenous proteins present in cementum, such as bone sialoprotein, osteopontin, osteocalcin, cementum derived growth factor and

cementum attachment protein. These proteins have various functions, including regulating mineralization and facilitating the attachment of cementum to the periodontal ligament (Arzate et al., 2015; Yamamoto et al., 2016).

Cementum formation or Cementogenesis starts when cells of the inner layer of the root sheath induce differentiation of the peripheral cells of the dental papilla into odontoblasts that produce root dentine. Epithelial root sheath breaks up when root dentine is laid down, allowing dental follicle cells to interact with root dentine. These cells differentiate into two distinct cell populations, cementoblasts and fibroblasts. Cementoblasts are responsible for producing cementum, while fibroblasts play a role in forming the periodontal ligament (Nanci & Bosshardt, 2006). Onto the root surface, cementoblasts actively secrete an organic matrix known as cementoid, which eventually forms the unmineralized and soft acellular primary cementum. With the production of primary cementum, further cells of the dental follicle differentiate into cementoblasts, producing cellular cementum. Over time, the primary cementum undergoes mineralization, forming HAP crystals similar to those found in enamel, giving the cementum its hardness.

1.2.4 Dentinogenesis

The content of dentine is 70% mineral, 20% organic matter and 10% water by weight (Yamakoshi, 2008; Goldberg et al., 2011). The major organic materials in dentine are type 1 collagens, synthesized by the *COL1A1* and *COL1A2* genes. The most abundant non-collagenous protein present in dentine is dentine sialophosphoprotein, which is encoded by the *DSPP* gene (Chen et al., 2022). Dentine has a complex structure, the

‘whole dentine’ being composed of several layers of dentine. The peripheral outer layer of the dentine is called mantle dentine, which is atubular and has a thickness of 15-30 mm. Next to the mantle dentine is more mineralized circumpulpal dentine. Circumpulpal dentine may be further subdivided into intertubular and peritubular dentine. While 90% of the protein components of intertubular dentine is type 1 collagen, the peritubular dentin is non-collagenous.

Dentine is formed by specialized cells called odontoblasts, in a process called dentinogenesis. Before the root formation finishes, the IEE cells differentiate to form preameloblasts. This has an inductive effect upon the ectomesenchymal peripheral cells of the dental papilla, which then differentiate into tall columnar preodontoblast cells. Enamel and dentine formation start simultaneously at the tips of the future cusps. Both preodontoblasts and preameloblasts develop into columnar secretory cells with reverse polarity and intracellular secretory organelles. The differentiating ameloblasts degrade the basal lamina, producing further inductive signalling to the odontoblasts to produce and set down dentine matrix, which gets deposited before the enamel matrix. The first dentine layer is mantle dentine, deposited peripherally in the dental papilla under the DEJ. Below mantle dentine is circumpulpal dentine that makes up most of the dentine layer in a regular incremental pattern. Primary dentine is the collective name used for mantle and circumpulpal dentine, which is formed prior to the formation of the root. Predentine is an unmineralized organic matrix that forms next to the pulp tissue during dentinogenesis and continues to exist throughout the lifetime of the tooth (Goldberg et al., 2011). After root formation, the odontoblasts form secondary dentine on the pulpal aspect of primary dentine. Secondary dentine

formation is a continuous process throughout life. Cavities, injuries and wear can give rise to tertiary dentine (Linde & Goldberg, 1993).

1.2.5 Amelogenesis

Fully formed enamel consists of around 96% inorganic content and 4% organic material and water by weight (Bartlett, 2013). The main elements of inorganic content are calcium and phosphate, which form calcium phosphate apatite in the presence of hydroxyl ions ($Ca_{10}(OH)_2(PO4)_6$). Calcium apatite is a crystalline lattice that can bind to several other ions such as strontium, radium, vanadium, and carbonate in exchange for the phosphate in the lattice. There are also several other minor inorganic constituents present in the enamel, including fluoride, zinc, lead, iron, sodium, magnesium and carbon dioxide (Bartlett, 2013; Smith, 1998). The organic component of the enamel consists of extracellular matrix proteins and an array of enzymes.

Enamel formation, or amelogenesis, takes place at the late bell stage before tooth eruption and almost simultaneously with dentinogenesis. It is initiated and regulated by epithelial-mesenchymal interactions along a line that becomes the DEJ. Enamel formation occurs through the secretion of an extracellular matrix (ECM) by secretory ameloblasts and its eventual mineralization. In the early stages of amelogenesis the secreted ECM is highly proteinaceous, consisting of a large amount of enamel matrix proteins (EMPs). At the later stages, ameloblasts stop secreting EMPs and instead switch to secreting proteolytic enzymes that degrade and remove the EMPs, while at the same time secreting the minerals required to form and grow HAP crystals within the ECM, filling the space left by the degrading EMPs. Amelogenesis is tightly regulated

by a hierarchy of transcription factors and cell signalling pathways that work to determine the size and shape of the teeth (Lacruz et al., 2017). Further details on the different stages of amelogenesis are provided in section 1.2.5.2.

1.2.5.1 Enamel matrix proteins and proteases

The EMPs secreted by ameloblasts can be grouped into three categories. In the first group are low molecular weight (5-45 kDa), heterogeneous hydrophobic proteins comprising approximately 90% of the total enamel proteins, called amelogenins (Brookes et al., 1995). Most of the body's amelogenin proteins are generated from the *AMELX* gene located on the X chromosome, while in males only 10% of the total amelogenins is produced by the *AMELY* gene located on the Y chromosome. Amelogenins are rich in proline, histidine, and glutamine. Amelogenins play an important role in the regulation of the size and shape of enamel crystallites and deletion of the amelogenin gene (*AMELX*) in humans results in a thin enamel layer (Lagerström et al., 1990; Salido et al., 1992). The second group, the acidic proteins enamelin (65-kDa) and tuftelin (55-kDa), are crucial for the development of full-length enamel rods (Diekwisch et al., 1997), Crawford et al., 2007, Lacruz et al., 2017). The tuftelin and enamelin proteins are encoded by the *TUFT1* and *ENAM* genes located on chromosomes 1 and 4 respectively (Lacruz et al., 2017). The third group of EMPs are ameloblastins (also known as amelins or sheathlins, 62-70-kDa) which are post-translationally modified and sequentially degraded products of the *AMBN* gene on chromosome 4 and are essential for enamel strength (MacDougall et al., 2000). Proteins in the second and third group comprise approximately 10% of the EMPs (Bartlett 2013; Lacruz et al., 2017). Amelotin (AMTN) and odontogenic ameloblast-

associated protein (ODAM) are also detected in developing enamel (Iwasaki et al., 2005; Moffatt et al., 2006; Park et al., 2007).

Matrix metalloproteinase-20 (MMP20) and kallikrein-related peptidase 4 (KLK4) are two enzymes detected in developing enamel (Bartlett & Simmer, 1999). MMP20 is active in the secretory to maturation stages while KLK4 is active in the transition to maturation stage. MMP20 has broad substrate specificity, and can cleave amelogenin, type XVIII collagen, fibronectin and dentin sialophosphoprotein. It can also cleave KLK4 propeptide to produce catalytically active KLK4. KLK4 can also hydrolyse EMPs and MMP20 for their effective removal as the enamel hardens (Bartlett, 2013).

1.2.5.2 Stages of Amelogenesis

Amelogenesis occurs in several discrete stages, each contributing to the development and maturation of enamel. Different stages of amelogenesis are described below.

Presecretory stage

The presecretory stage starts with the deposition of predentine by odontoblasts at the future DEJ. At this stage, the IEE cells of the enamel organ differentiate into columnar secretory preameloblasts (Bartlett, 2013).

Secretory stage

In the secretory stage, differentiating preameloblasts elongate into tall columnar polarized ameloblasts with a height of 70 µm and a diameter of 5 µm (C.E. Smith, 1998). At this stage the enamel organ has four distinct cell populations. The innermost

layer is the secretory ameloblasts, the next layer is stratum intermedium, the third layer is stellate reticulum and the outermost layer is the OEE (Bartlett 2013, Lacruz et al., 2017). Apart from the ameloblasts, the functional roles of the other cell layers of the enamel organ are poorly understood (Liu et al., 2016).

As ameloblasts enter into the secretory stage, they begin secreting an ECM into the surrounding region. The ECM is highly proteinaceous and consists of a large quantity of EMPs. The EMPs secreted by ameloblasts at this stage are variously modified derivatives of four proteins, amelogenin (AMELX), ameloblastin (AMBN), enamelin (ENAM), and matrix metalloproteinase-20 (MMP20). With the deposition of the ECM, an initial layer of tiny mineral crystals are formed onto the dentine surface by ameloblasts. This first layer forms a scaffold for growth of the eventual mineral crystals. As soon as this initial layer is laid down, highly polarized ameloblasts start moving away from the dentine surface. This phenomenon gives secretory ameloblasts a unique morphology, Tomes' processes, which are triangular-shaped extensions formed by ameloblasts penetrating the enamel matrix, giving the ameloblast monolayer a "picket-fence" appearance under a microscope (Bartlett, 2013; Wakita et al., 1981). Aligning with the movement of ameloblasts, thin HAP-like, hexagonally shaped, enamel crystallites grow and extend at nearly right angles to DEJ. These crystallites grow progressively in parallel with one another to hundreds of micrometers in length, yet are only few nanometers wide. Enamel rods are crystals formed by the matrix secreted from the distal portion of the Tomes process, while crystals oriented at angles relative to the rod crystallites are interrod, formed by the matrix secreted from the proximal portion of the Tomes processes. This alternating arrangement of enamel rods and

interrods gives enamel a characteristic ‘herringbone weave pattern’ for extra strength when mature as shown in Figure 1.4. Protein rich enamel at this stage achieves its full thickness but has a soft cheese like consistency and is comprised of similar amounts of EMPs, mineral and water by weight (Lacruz et al., 2017).

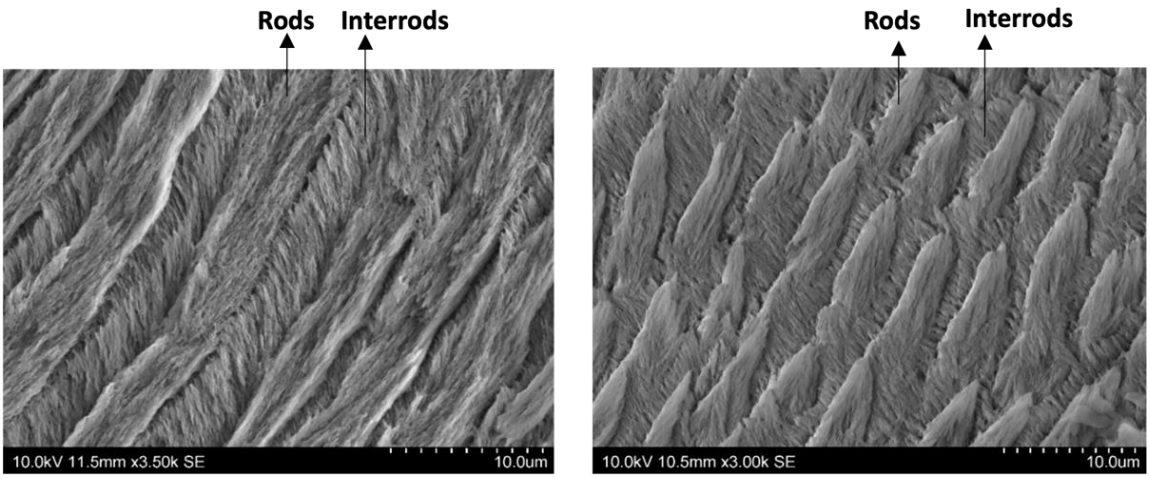


Figure 1. 4 Herringbone pattern in human enamel.

Scanning electron microscope (SEM) images showing structural arrangement of rods and interrods in human enamel, appearing as a series of interlocking and branching lines, resembling the zigzag pattern of a herringbone weave. This pattern contributes to the strength and durability of teeth, helping them withstand the stresses of everyday use.

Transition stage

During the Transition stage, ameloblasts undergo significant morphological changes. They become shorter, lose their secretory Tomes' processes and reduce in number, possibly by apoptosis. Also, the genes encoding the EMPs are downregulated, while genes involved in ion transport, proteolysis and pH homeostasis are upregulated (Lacruz et al., 2017).

Maturation stage

During the Maturation stage, ameloblasts undergo two main processes; degradation and removal of organic matrix, and mineralization to support crystal growth. Accumulation of minerals in the enamel matrix increases while proteins and water bulk are reduced, until the enamel becomes almost completely mineralised. This is accomplished by a process called modulation, where morphologically, ameloblasts appear to undergo cyclic alteration between ruffled-ended and smooth-ended types. Ruffle-ended ameloblasts (RE-A) have distal tight junctions and proximal leaky junctions and dominate in this stage (80% of total cells), whereas smooth-ended ameloblasts (SE-A) are completely the opposite, having distal leaky junctions and proximal tight junctions (20% of total cells). The RE-A degrade EMPs by secreting kallikrein-related peptidase-4 (KLK4). The degraded proteins are removed by diffusing into the leaky proximal junctions of the SE-A and also engulfed by RE-A by endocytosis. A considerable amount of bicarbonate ions (HCO_3^-) and calcium ions (Ca^{2+}) are also secreted by RE-A to maintain favourable pH for mineral growth (C. E. Smith, 1998). In this way, enamel crystals mature, reducing the protein content of the enamel

progressively down to 4% while achieving 96% mineral content. At the end of this stage, 50% of the original ameloblast cell population that participated in enamel crystal formation die by apoptosis (C. E. Smith, 1998).

Protective Stage

Following the completion of enamel maturation, ameloblasts undergo a final stage known as the protective stage. During this stage, ameloblasts create a protective layer called the reduced enamel epithelium. The reduced enamel epithelium covers the enamel surface until the tooth erupts into the oral cavity. It helps protect the enamel from potential damage during eruption and prevents contact with the surrounding connective tissues and oral bacteria (C. E. Smith, 1998; Ten Cate, 1996). Different stages of amelogenesis are depicted in Figure 1. 5.

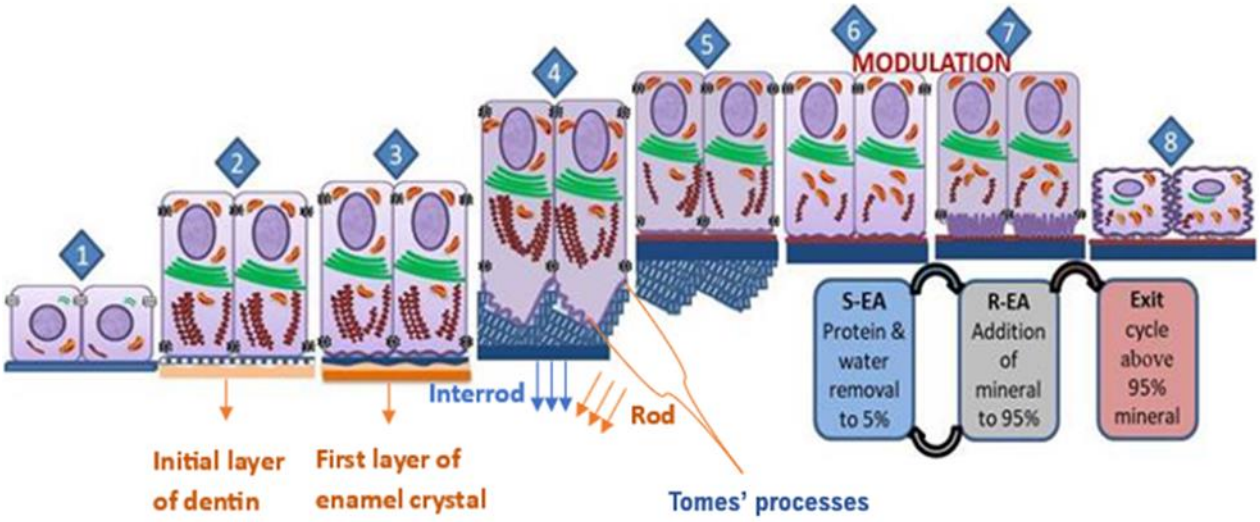


Figure 1. 5 Stages of amelogenesis.

A schematic illustration of the different stages of amelogenesis. (1) Morphogenetic stage, (2) Differentiation stage, (3) Secretory stage (aprismatic enamel), (4) Secretory stage (prismatic enamel), (5) Transitional stage, (6-7) Maturation Stage, (8) Protective stage. S-EA: smooth ended ameloblast, R-EA: ruffle ended ameloblasts. Image adapted with permission from Mandana Donoghue available at, https://commons.wikimedia.org/w/index.php?curid=76184792.

1.2.5.3 Ion transport in the maturation stage of Amelogenesis

Ionic composition of enamel includes mainly Ca^{2+} and PO_4^{3-}, and trace amount of other ions such as HCO_3^-, Cl^-, Na^+, F^-, K^+ and Mg^{2+} (Patel & Brown, 1975). To maintain a proper environment for enamel crystals to grow, ameloblasts control the movement of these elements from the blood to the enamel crystals during amelogenesis (Lacruz et al., 2013). The major entry pathway for Ca^{2+} is store-operated Ca^{2+} entry (SOCE) from the endoplasmic reticulum (ER), mediated by CRAC (Ca^{2+} release activated channel) channels (Meerim K Nurbaeva et al., 2015). The components of CRAC channels are ER transmembrane proteins STIM1 and STIM2 (stromal interaction molecule) and a plasma membrane protein ORAI1 (calcium release-activated calcium channel protein 1) (Prakriya & Lewis, 2015). ER Ca^{2+} depletion triggers activation of STIM1 and STIM2 to form a multimer with ORAI1. This interaction opens up ORAI1, allowing Ca^{2+} influx into the cytoplasm. One study reported that an increase in cytoplasmic Ca^{2+} ion concentration upregulated the expression of *Amelx, Ambn, Enam* and *Mmp20* in ~4 week-old Sprague-Dawley rats (M. K. Nurbaeva et al., 2015). Cytosolic Ca^{2+} ions are removed by the potassium dependent sodium/calcium exchanger NCKX4 (encoded by the *SLC24A4* gene), which co-transports one intracellular Ca^{2+} and one K^+ ion in exchange for four extracellular Na^+ ions (Li et al., 2002). Studies show that the solute carrier gene family *Slc34a2* that encodes a pH dependent Na^+/ PO_4^{3-} transporter NaPi2b (sodium-dependent phosphate transport protein 2B) localizes in secretory and maturation ameloblasts in mice (Lacruz et al., 2012). NaPi2b transports three Na^+ with one PO_4^{3-}across plasma membrane, and may play a role in phosphate transportation in amelogenesis (Lacruz, 2017). Another study reported the expression of *TRPM7* (transient receptor potential cation channel, subfamily M, member 7) was significantly

upregulated in maturation-stage ameloblasts (Nakano et al., 2016). TRPM7 is a protein kinase predicted to play role in the regulation of Mg^{2+} homeostasis. Deletion of exons 32-36 (the kinase domain) of the *Trpm7* gene in mice caused a severe hypomineralised enamel phenotype (Nakano et al., 2016). This observation may indicate that TRPM7 participates in transporting Mg^{2+} to the cytosol in developing teeth. *CNNM4* (cyclin and CBS domain divalent metal cation transport mediator 4) encodes a Mg^{2+} transporter, and may be involved in the removal of Mg^{2+} from cells, as a study reported humans with *CNNM4* loss of function mutation presented with severe hypomineralization of enamel, as well as inherited blindness (Parry et al., 2009).

1.2.5.4. pH balance

At the maturation stage, with the progressive growth of HAP crystals, free protons (H^+) are released into the extracellular area that can lower the pH, drastically disrupting the crystal growth (Lacruz, 2017; Smith et al., 2005). Ameloblasts neutralize this acidic environment by either secreting AMELX that can bind H^+ or by pumping out the HCO_3^- into the extracellular compartment (Bori et al., 2016). It has been observed that there is an increased expression of sodium bicarbonate co-transporter NBCe1 (encoded by the *SLC4A4* gene) at the maturation stage (Rauth et al., 2009). Another further study reported localization of AE2 (anion exchange protein 2, encoded by the *SLC4A2* gene) and NHE1 (sodium-proton exchanger, encoded by *SLC9A1*) proteins at the basolateral pole of maturation ameloblasts, which may suggest they have a role in maintaining intracellular pH homeostasis (Josephsen et al., 2010). It is proposed that AE2 transports HCO^{3-} across the basolateral plasma membrane in exchange for extracellular Cl^-, while on the other hand H^+ is removed from the cell by NHE1 (Bronckers, 2017; Josephsen et

al., 2010). Potential ion transport routes and their likely locations on a maturation stage ameloblast are depicted in Figure 1.6.

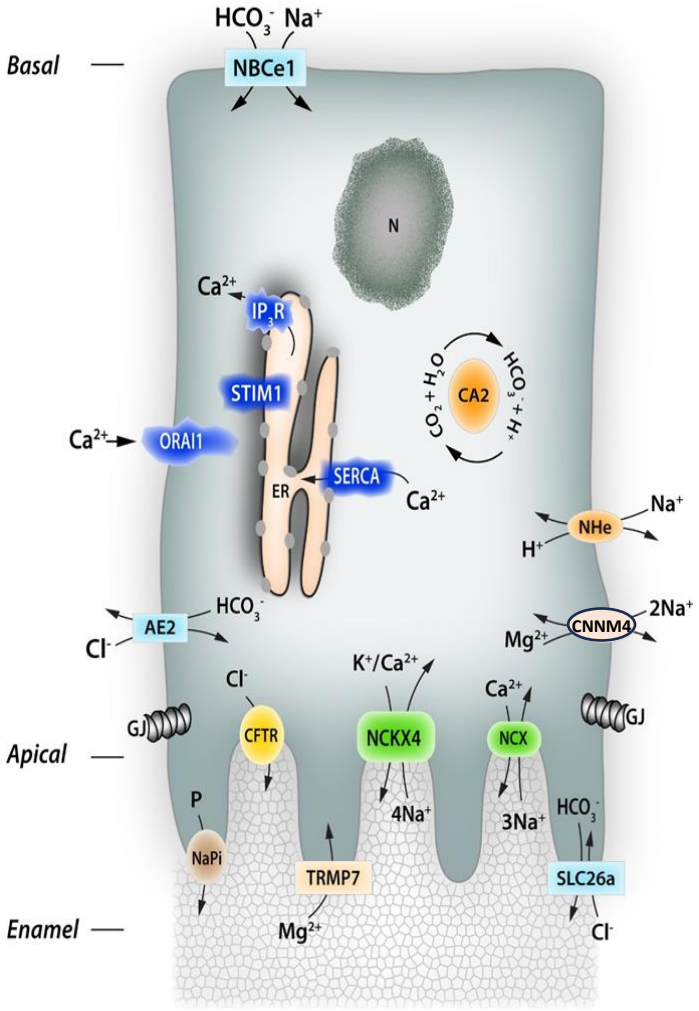


Figure 1. 6 Model of ion channels and transport mechanisms at the maturation stage of amelogenesis.

A schematic illustration of putative ion transport mechanisms in a maturation-stage ruffled ended ameloblast. This model is based on reports of mRNA and protein expression as well as cellular localization, so not all the functional roles of the proteins have been established. N: nucleus. ER: endoplasmic reticulum. GJ: gap junctions. Image adapted with permission from Lacruz et al., 2017 under the terms of the Creative Commons CC-BY license.

1.2.5.5 Amelogenesis imperfecta (AI)

Amelogenesis is governed by a complex network of genes, regulatory proteins, and signalling pathways to develop this highly sophisticated tissue. There are over 10,000 genes and hundreds of regulatory elements thought to contribute to amelogenesis (reviewed by Hu et al., 2015; Yin et al., 2017). Defects in key genes, epigenetic regulators or negative influence from environmental stressors at any stage of amelogenesis may affect normal enamel formation.

AI is a genetic disorder resulting from the failure of amelogenesis. It is a common endpoint for a clinically and genetically heterogeneous group of inherited diseases. AI affects the structure and clinical appearance of the dental enamel of all teeth, in both the primary and permanent dentitions. It can occur in isolation or in conjunction with defects in other dental, oral and extraoral tissues. AI was first described as 'brown teeth' with a familial history a century ago (Spokes, 1890). Later, in 1912, another researcher reported two cases of hereditary hypoplasia of the teeth across five generations of the same family (Turner, 1912). However, there was little interest in AI thereafter until in 1945, when researchers introduced the term 'amelogenesis imperfecta' to describe hypoplastic and hypocalcified types of hereditary abnormalities of the enamel (Weinmann et al., 1945). The reported prevalence of AI ranges from 1 in 233 in Turkey (Altug-Atac & Erdem, 2007), 1 in 700 in Sweden (Bäckman & Holm, 1986), 1 in 1,000 in Argentina (Sedano, 1975), 1 in 8,000 in Israel (Chosack et al., 1979),) to 1 in 14,000 in USA (Witkop, 1988). According to these values, the average global prevalence of AI is 1 in 2000 (Gadhia et al., 2012).

AI starts in infancy upon eruption of the first teeth and worsens with further tooth eruptions through childhood, adolescence, and early adulthood. AI leads to discoloured teeth and weak enamel that breaks down easily. It is often accompanied by pain, infection, early tooth loss and malocclusion. AI patients also suffer adverse psychosocial impacts with high levels of distress, social avoidance, discomfort, isolation, and emotional problems. AI may be inherited as an X-linked, autosomal dominant, or autosomal recessive genetic trait, and can be syndromic or non-syndromic. In the disease database OMIM, currently 20 genes are associated with AI that do not have any associated conditions (non-syndromic, NS), and 95 syndromes are reported to have enamel phenotypes as part of their symptoms (Wright, 2023). Both syndromic and NS AI follow a Mendelian inheritance pattern (Bäckman & Holmgren, 1988; Gadhia et al., 2012).

1.2.5.5.1 Clinical Classification of AI

AI presents with a spectrum of clinical appearance that is dependent on the inheritance pattern, the gene, and specific mutation(s) involved. Severity of the disease phenotype varies with the particular stage affected in amelogenesis. Specific clinical sub-types can also be discerned with defects in different cellular or biochemical pathways. Environmental factors, including nutrition and oral hygiene practices, can influence the severity and progression of AI. There is currently no universally accepted classification system for AI. Clinicians still use the classification introduced by Dr. Charles J. Witkop in 1988 (Witkop, 1988). It is important to note that the Witkop classification is primarily based on clinical and visual characteristics of AI and does not consider the genetic or molecular basis of the condition, which was unknown at the

time. Witkop classification system categorizes AI into four major groups described below. However, the boundaries between the categories are becoming increasingly unclear, and given recent advances in the genetic understanding of AI, it would be more accurate, relevant and clinically useful now to classify cases by their molecular diagnoses.

Type I Hypoplastic AI is characterized by inadequate enamel which may have resulted from abnormal deposition of enamel matrix at the secretory stage of amelogenesis, leading to thin or, in extreme cases, completely absent enamel exposing the underlying dentine. Though enamel appears thin in this type, it may have near-normal hardness. The reduced enamel thickness makes the teeth more sensitive to temperature changes, touch, and certain foods or beverages. The enamel surface may appear rough, irregular, or pitted due to incomplete enamel formation. The roughness can make the teeth more prone to staining, plaque accumulation, and tooth decay. The appearance of the affected teeth may be discoloured or appear dull due to incomplete enamel formation, as shown in Figure 1. 7(i-ii).

Type II Hypomaturation/Hypomineralised AI presents as an open bite and creamy white to yellow brown roughly surfaced teeth, along with fragile enamel, as shown in Figure 1. 7(iii-iv). In this type of AI, a defect in maturation stage amelogenesis is thought to result in enamel that is of full thickness but fails to fully mineralize or mature, resulting in softer and less resistant enamel that can be chipped away or scraped (Gadhia et al., 2012). This type of AI can be characterized by surface pit and damage to or underdevelopment of the enamel prism structure, as shown in Figure 1.

7(v-vi) (Bäckman and Holmgren, 1988, Gadhia et al., 2012). A more severe form of hypomaturation enamel may be termed as hypomineralized AI, which presents as brown, discoloured enamel, as shown in Figure 1. 7(vii) (Crawford et al., 2007; Sundell & Koch, 1985).

Type III Hypocalcified AI is characterized by very soft and easily eroded enamel. The enamel affected often appears opaque and has a chalky white or creamy yellowish colour, as seen in Figure 1. 7(viii). This discoloration is thought to be due to the incomplete mineralization of the enamel after a failure of maturation, and the resultant higher porosity of the enamel. Hypocalcification (soft enamel) is caused by insufficient transport of calcium ions into the developing enamel, because of which the enamel consists of around 30% less mineral content than usual.

Type IV Hypomaturation/Hypoplasia/Taurodontism AI teeth exhibit thin enamel, with pitting and discolouration, in addition to enlarged pulp chambers and shortened roots due to taurodontism (Aldred et al. 2003).

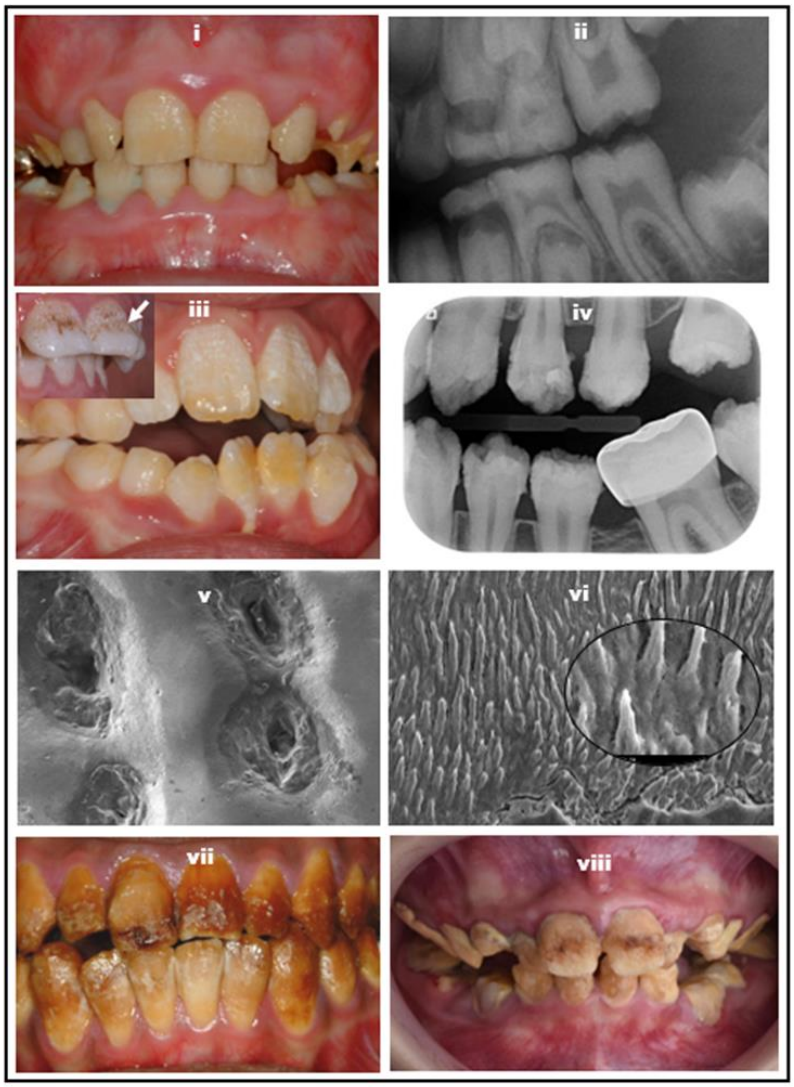


Figure 1. 7 Clinical images of different types of amelogenesis imperfecta.

(i-ii) Yellow hypoplastic AI reflects the absence of any meaningful enamel on dental radiography. (iii) A predominantly hypomaturation AI phenotype with some surface irregularities, (inset) hypomaturation enamel is combined with more exaggerated surface pits merging into grooves with mid-third crown regional hypoplasia (arrow). (iv) Intraoral radiograph illustrating near normal enamel thickness in Hypomaturation AI, but with enamel irregularities and a lesser difference in radiodensity between enamel and dentine than would be expected. (v) SEM image of surface pits in hypomaturation AI. (vi) Generally disrupted and poorly formed prismatic microstructure, with difficulty distinguishing between rods and interrods, observed in a hypomaturation AI (inset). (vii) Brown discolouration and early post-eruptive enamel loss is typical of hypomineralized AI. (viii) Hypocalcified AI phenotype displaying brittle enamel.

1.2.5.5.2 Genes associated with non-syndromic AI

Though AI was first described nearly 80 years ago, the genes and mutations involved have only been identified in the last 20 years, the majority as a result of the advent of next generation sequencing technology (NGS). Different clinically defined sub-phenotypes of AI (hypoplastic, hypocalcified, and hypomaturation) typically correlate with mutations in different genes, each of which encode proteins that contribute to distinct stages of and processes essential in the enamel synthesis. However, the AI phenotypes observed do not always fit neatly into these phenotypic categories, and variation has been observed between individuals in families that carry the same mutation. To-date mutations causing NS AI have been identified in 20 genes. These genes code for proteins performing diverse and critical functions in amelogenesis, including EMPs, enamel matrix proteases, transcription factors, cell-cell and cell-matrix adhesion, pH sensing and ion transport. It was observed that while some gene mutations may cause an isolated or NS form, other mutations in the same gene may cause more complex syndromic forms by affecting multiple biological processes or pathways. Molecular mechanisms and disease phenotypes caused by mutations in these genes are detailed below and in Table 1.1.

Genes encoding EMPs and matrix proteases

A relatively common cause of AI is mutations in the genes that encode EMPs and proteases. The roles of EMPs and proteases in amelogenesis are discussed in more detail in section 1.2.5.1. A 5kb deletion in the *AMELX* gene was the first mutation detected causing X-linked dominant hypomineralized AI (OMIM: 301200) in males (Lagerström et al., 1991). A mixed phenotype of hypoplastic and hypomineralized AI

was also observed in patients, suggesting that mutation in the *AMELX* gene may impact both the mineralization process and the deposition of enamel matrix (Aldred et al., 1992). The molecular mechanism of most of the *AMELX* mutations causing AI was predicted to be loss-of-function. A mutation in the signal peptide coding sequence was also reported to cause AI by defective translocation of the protein (Lagerström-Fermér et al., 1995).

Mutations in *ENAM* are reported to cause both AD and AR AI. The first mutation reported in *ENAM* was a splice site mutation causing dominant AI (OMIM: 104500), predicted to be a result of read through followed by NMD (nonsense mediated decay) or exon skipping caused by aberrant splicing. These patients presented with a severe, smooth hypoplastic AI phenotype (Rajpar et al., 2001). Another study reported a homozygous 2-bp insertion in the *ENAM* coding sequence leading to a frameshift that introduced premature termination codon (PTC), causing recessive AI (OMIM: 204650). These patients presented with a severe generalized hypoplastic phenotype with open-bite, malocclusion, while the heterozygous carriers had only a mild localised enamel pitting phenotype (Hart et al., 2003).

Mutations in the *AMBN* gene were reported to cause AR hypoplastic AI. The first mutation reported in *AMBN* was a large, in-frame deletion causing recessive hypoplastic AI (OMIM: 616270) (Poulter, Murillo, et al., 2014). The molecular mechanism underlying recessive AI caused by *AMBN* variants is predicted to be complete loss of function. Several further recessive mutations were reported afterwards (Liang et al., 2019). In addition, a heterozygous missense mutation in *AMBN*

reported to cause AD AI in a putative dominant family presented with generalized hypominerailzed phenotype (Lu et al., 2018). This result was queried by other researchers who sought further evidence of the disease mechanism caused by heterozygous *AMBN* mutation (Liang et al., 2019). This thesis presents further evidence of both AD and AR forms of AI due to *AMBN* variants (see chapter 3).

Another EMP, AMTN, localizes to the ameloblast basal lamina and mutations in the *AMTN* gene are reported to cause AD hypomineralized AI (OMIM: 617607) (C. E. L. Smith et al., 2016).

Mutation in the *MMP20* and *KLK4,* encoding the enamel matrix proteases, are known to cause AR AI. Enamel phenotypes of the patients carrying mutation in these two genes share many similarities, with a phenotype that is usually described as pigmented, hypomaturation AI (OMIM: 612529, 204700) (Kim et al., 2005).

Cell-cell and cell-matrix adhesion

In amelogenesis, several adhesion molecules function in a coordinated fashion to maintain cell-cell and cell-matrix contact. Mutations in the genes encoding different types of adhesion molecules are reported to cause AI, and these are largely the same genes implicated in the recessively inherited skin disorder epidermolysis bullosa. For example, ITGB6, is a member of a large family of cell surface heteromeric glycoproteins and is highly expressed during the maturation stage of amelogenesis. Loss-of-function mutations in the *ITGB6* gene are reported to cause AR AI (OMIM: 616221). The enamel phenotype associated with *ITGB6* mutation is hypoplastic and hypomineralized that

may be rough, pitted, and/or discoloured (Poulter, Brookes, et al., 2014; S. K. Wang et al., 2014).

LM332 (laminin 332), a heterotrimeric protein, is a major constituent of the basement membrane and has a central role in the assembly and stability of hemidesmosomes (Colognato & Yurchenco, 2000). The three subunits of LM332 are encoded by the *LAMB3, LAMA3* and *LAMC2* genes. Collagen XVII is a hemidesmosome protein, encoded by *COL17A1*, is a ligand for LM332. Mutations in the genes *LAMB3, LAMA3, LAMC2* and *COL17A1* are all reported to cause AD hypoplastic AI (OMIM: 104530) (Poulter, El-Sayed, et al., 2014; Wang et al., 2022).

FAM83H is an intracellular protein expressed in presecretory and secretory stages of amelogenesis (Lee et al., 2008). It plays an important role in the structural development and calcification of tooth enamel, though the exact mechanism of its function is not fully understood. A study reported that a novel *Fam83h* c.1186C>T (p.Q396*) knock-in mutant mouse presented with severe inflammation, disturbed iron deposition, and enamel abnormalities. Histological cross-sections of the lower incisors of the mutant mouse displayed gaps between adjacent ameloblasts. RNA-sequencing and GO (gene ontology) analyses showed cell adhesion pathway was the most affected in this mouse. Desmoglein 3, a transmembrane glycoprotein protein, is a component of desmosomes was downregulated in the ameloblasts of the mutated mouse. All these findings indicated that FAM83H may play a role in cell-cell adhesion which may have disrobed due to the mutation producing truncated-FAM83H in the mutant mouse

(Zheng et al., 2023). Most mutations identified in *FAM83H* genes are frameshift leading to PTC causing AD hypocalcified AI in human (OMIM 130900) (Kim et al., 2008).

ODAPH localizes to the ameloblast basal lamina and mutations in *ODAPH* are reported to cause AR hypomineralized AI (OMIM 614832) (Parry et al., 2012). In *Odaph*$^{-/-}$ mice, the integrity of the atypical basal lamina was impaired. It was observed that ameloblasts lost cell polarity, became short and flattened in early maturation stage in the mutant mice. Additionally, the enamel matrix was not degraded and most EMPs were retained in late maturation stage. Further, the expression of LAMC2 and AMTN were downregulated in mutant mice. According to the above observations, this study concluded ODAPH may play a vital role in maintaining the integrity of basal lamina (Ji et al., 2021).

Ion transport and pH balance

Ameloblasts facilitate and regulate the flow of numerous proteins and ions to the enamel crystals during amelogenesis. More details of the ion transport mechanism were provided in section 1.2.5.3. Mutations affecting the functions of ion transport channel proteins and vesicle trafficking proteins reported to cause various forms of AI. One example would be WDR72, which is predicted to be an intracellular vesicle coat protein and is expressed in the maturation stage of amelogenesis. WDR72 may function in regulating microtubule structure and assembly in ameloblast which may critical for ameloblast modulation between SE and RE forms, endocytosis and vesicle trafficking for pH homeostasis (Katsura et al., 2022). One study reported a homozygous nonsense mutation in *WDR72* caused AR hypomaturation AI (OMIM: 613211) (El-Sayed

et al., 2009). Mutations in *WDR72* have also been associated with other syndromic conditions. A study reported observing hypomaturation enamel and relatively short stature caused by a homozygous mutation in *WDR72* in two patients (Kuechler et al., 2012). Another study reported compound heterozygous mutations in *WDR72* caused hereditary distal renal tubular acidosis (dRTA) along with hypoplastic AI (Rungroj et al., 2018).

SLC24A4 is upregulated in maturation stage ameloblasts and is predicted to be responsible for the active transport of Ca^{2+} ions from ameloblasts into the enamel matrix (S. Wang et al., 2014). Homozygous mutations in the *SLC24A4* gene were reported to cause AR hypomineralised AI (OMIM: 615887) (Parry et al., 2013).

GPR68 is expressed in ameloblasts throughout all stages of amelogenesis. GPR68 is thought to function as a pH sensor, predicted to direct ameloblasts to switch between the ruffle ended and smooth ended conformations during the maturation stage for maintaining pH balance. Loss of function mutations in *GPR68* were reported to cause AR hypomineralized AI (OMIM: 617217) (Parry, Smith, et al., 2016; Sato et al., 2020).

Master controllers of amelogenesis

Genes in this group control expression of other genes or affect function of other proteins involved in amelogenesis. Although the exact regulatory mechanisms of these proteins in amelogenesis are still not clear, mutations in these genes have been linked to a number of AI-related disorders.

FAM20A is a pseudokinase localized in secretory and maturation stage ameloblasts. Mutations in *FAM20A* are associated with both non-syndromic and syndromic AI. One study reported homozygous mutations in *FAM20A* caused AR hypoplastic AI (OMIM 204690) in patients with no other health problems (Cho et al., 2012). The disease mechanism was predicted to be loss of function of the FAM20A protein in these patients. Mutations in *FAM20A* has also been associated with a syndromic condition known as enamel renal syndrome (ERS). ERS is characterized by hypoplastic enamel on the primary and secondary dentition, pulp stones, delayed or failed eruption of the secondary dentition, gingival overgrowth, and nephrocalcinosis (MacGibbon, 1972; Martelli-Júnior et al., 2008). More details of ERS are discussed in section 1.2.5.5.3.

DLX3 is member of a family of six DLX transcription factors and is highly expressed in ameloblasts during the late secretory stage (Zhang et al., 2015). Heterozygous mutations in *DLX3* are known to cause Trichodentoosseous syndrome (TDO, OMIM 190320). The TDO patients present with kinked hair, alteration in tooth size and increased density/thickening of the bones. Heterozygous mutations in *DLX3* can also cause AD hypomaturation-hypoplasia type AI with taurodontism (AIHHT, OMIM 104510) (Crawford et al., 1988). The dental phenotypes in the TDO and AIHHT patients appear similar, but TDO presents a more highly variable clinical phenotype involving hair and bone.

Heterozygous mutations in *SP6* reported to cause AD hypoplastic AI (OMIM: 620104). The enamel phenotype was characterized by generalized hypoplastic AI with an irregular surface involving all teeth (Smith et al., 2020). SP6 is expressed in secretory

stage ameloblasts and is predicted to regulate expression of *Amtn*, *Fst* and *Rock1* genes in mice (Ruspita et al., 2008; Utami et al., 2011).

Genes encoding proteins of unknown function

Although the function of RELT is not known in amelogenesis, homozygous mutations in the *RELT* gene were reported as a cause hypocalcified AR AI (OMIM: 611211). The enamel of these patients was rough and yellow-brown in colour (Kim et al., 2019).

ACP4 is expressed in the secretory stage of amelogenesis, the molecular functions of ACP4 during amelogenesis is not fully elucidated yet. Homozygous mutations in *ACP4* are reported to cause rough hypoplastic AR AI (OMIM: 61729) (Seymen et al., 2016).

Table 1. 1 Genes associated with non-syndromic AI.

Types	Gene (OMIM)	Protein	Function	Inheritance
Enamel matrix proteins	*AMELX* (300391)	Amelogenin. Secretory calcium-binding phosphoprotein (SCPP).	AMEL and its MMP20 cleavage products bind, separate, and support the mineral ribbons, and guide their transition into apatite (Brookes et al., 1995).	XLD
	AMBN (601259)	Ameloblastin. Secretory calcium-binding phosphoprotein (SCPP).	AMBN influences differentiation and proliferation of ameloblasts. It also plays role in cell adhesion and enamel mineralisation (Fukumoto et al., 2005).	AD, AR
	ENAM (606585)	Enamelin. Secretory calcium-binding phosphoprotein (SCPP).	ENAM maintains ameloblast integrity and plays a role in crystal formation, to achieve prism structural organization, and optimal enamel thickness (Hu et al., 2008).	AD, AR
	AMTN (610912)	Amelotin. Secretory calcium-binding phosphoprotein (SCPP).	AMTN promotes enamel mineralization and plays critical role in the formation of the compact aprismatic enamel surface layer during the maturation stage of amelogenesis (Abbarin et al., 2015).	AD
Proteases	*MMP20* (612529)	Matrix metalloproteinase 20. Zinc-dependent endopeptidases.	MMP20 activates EMPs by proteolysis at the secretory stage of amelogenesis (Guan & Bartlett, 2013).	AR
	KLK4 (204700)	kallikrein-related peptidase 4. Serine protease.	KLK4 degrades EMPs in the maturation stage, processed by MMP20 previously, and can function over a wide pH range (Bartlett, 2013).	AR
Cell-cell and cell-matrix adhesion	*LAMB3* (150310)	Laminin beta 3. LAMB3, LAMA3 and LAMC2 constitute three subunits of the heterotrimeric protein laminin 332 (LM332), which plays a central role in the assembly and stability of hemidesmosomes.	Laminin 332, a heterotrimeric protein, is a component of the basement membrane that separates the differentiating presecretory ameloblasts from the forming mantle dentine and associated odontoblasts (Nanci, 2008).	AD, AR
	LAMA3 (600805)	Laminin alpha 3. See LAMB3.	Cell adhesion ligand for integrins.	AD
	COL17A1 (113811)	Collagen XVII alpha chain 1. Binding ligand of laminin 332.	COL17A1 is a hemidesmosome protein expressed all through amelogenesis.	AD

	ITGB6 (147558)	Integrin beta 6. Heterodimer, cell surface glycoproteins.	ITGB6 is a cell-surface adhesion receptors that bind to extracellular matrices (ECM) and mediates cell-ECM interactions (Alberts et al., 2008).	AR
	FAM83H (611927)	Family with sequence similarity 83.	FAM83H functions in cell-cell adhesion, iron deposition, calcium transportation, and may also have a role in transportation and secretion of AMELX (Zheng et al., 2023).	AD
	ODAPH (614829)	Odontogenesis-Associated Phosphoprotein.	Phosphorylated ODAPH has the capacity to promote nucleation of hydroxyapatite. It maintains the integrity of the atypical basal lamina in maturation stage (Parry et al., 2012).	AR
Ion transport, pH balance	*WDR72* (613214)	Tryptophan-aspartate repeat domain 72. 4–8 repeating units of approximately 44–60 amino acids ending in tryptophan (W) and aspartic acid (D).	WDR72 functions in endocytic vesicle trafficking of matrix proteins and subsequent enamel mineralization. It may play a role in the removal of amelogenin during enamel maturation (K. Katsura et al., 2022).	AR
	SLC24A4 (609840)	Solute carrier family 24, (Na/K/Ca exchanger) member 4.	SLC24A4 transports calcium ions out of the cell and into the enamel matrix (Parry et al., 2013).	AR
	GPR68 (601404)	G-protein-Coupled Receptor 68	GPR68 plays important role as a proton sensor in the enamel organ at all stages of amelogenesis (Parry, Smith, et al., 2016).	AR
Master controllers of amelogenesis	*FAM20A* (611062)	Family with sequence similarity 20 member A. Golgi associated secretory pathway pseudokinase.	FAM20A binds to FAM20C to form a functional complex that phosphorylates EMPs (Li et al., 2019).	AR
	DLX3 (600525)	Distal-less homeobox 3. Transcription factor.	DLX3 functions as a pH regulator during enamel maturation, is involved in proper enamel rod decussation, and regulates myosin II and associated protein production in the enamel organ (Duverger & Morasso, 2018). The highest expression of DLX3 is detected in late secretory phase ameloblasts.	AD
	SP6 (608613)	Specificity protein 6. Transcription factor consisting of zinc finger DNA-binding domains.	Msx2 and Sp6 work in a coordinated fashion to control follistatin's production and promote enamel deposition. SP6 is expressed in the secretory phase of amelogenesis (Smith et al., 2020).	AD

Unknown function	*RELT* (611211)	Receptor expressed in lymphoid tissues. A transmembrane protein, is a member of the TNFR (Tumour necrosis factor receptor) family	RELT is expressed in the secretory phase, but its function is unclear in amelogenesis. It takes part in apoptosis, cell differentiation in other tissues, (Yao et al., 2021).	AR
	ACP4 (606362)	Acid Phosphatase 4. Membrane-bound acid phosphatase.	EMPs may be dephosphorylated by ACP4 during endocytosis, releasing phosphate into the matrix for mineralization (Simmer et al., 2021).	AR

1.2.5.5.3 Syndromes associated with AI

AI can occur in conjunction with other systemic or developmental conditions. There are several known syndromes reported in the literature that present AI as one of their features. Some examples of these conditions are provided below.

Jalili Syndrome (OMIM: 217080): In Jalili syndrome patients suffer from cone-rod dystrophy and AI. Patients present with hypomineralized AI along with early, severe visual impairment, loss of colour vision, night blindness and loss of peripheral visual fields. Jalili syndrome is caused by homozygous or compound heterozygous mutations in the *CNNM4* gene (Parry et al., 2009). CNNM4 is a Mg^{2+} transporter thought to function in maintaining Mg^{2+} homeostasis in the developing enamel and elsewhere in the body. Mutations reported were missense, PTC, deletions and insertions detected in the conserved regions of the protein and were predicted to result in loss of function.

Trichodentoosseous syndrome (TDO, OMIM: 190320): TDO is an autosomal dominant disorder with complete penetrance. It is caused by heterozygous mutations in the *DLX3* gene, and is characterized by abnormalities involving hair, teeth, and bones (Nguyen et al., 2013; Wright et al., 1997).

Heimler Syndrome (HMLR, OMIM: 234580, 616617): HMLR is a rare autosomal recessive disorder characterized by sensorineural hearing loss, enamel hypoplasia of the secondary dentition and nail abnormalities (Heimler et al., 1991). Case report of a HMLR patient also noted the development of adult-onset macular dystrophy (Lima et al., 2011). Several studies reported that HMLR is caused by homozygous or compound heterozygous mutations in the

PEX1, PEX6 or *PEX26* (peroxisomal biogenesis factor) genes (Ratbi et al., 2015; C. E. Smith et al., 2016). PEX proteins play important roles in controlling peroxisomal size, and functions. *PEX1, PEX6*, and *PEX26* encodes peroxisomal AAA–ATPase complex that prevents pexophagy (Law et al., 2017). Variants in these three *PEX* genes, along with 12 other *PEX* genes, also contribute to the much severe Zellweger syndrome (ZS, OMIM: 214100). Most ZS patient die within the first year of life. The characteristic features of ZS include severe neurologic dysfunction, craniofacial abnormalities, and liver dysfunction (Yik et al., 2009). Phenotypic comparison of HMLR to ZS shows that HMLR comprises the mildest end of the ZS phenotype spectrum, caused by variants leading to partial loss of peroxisomal function due to hypomorphic alleles (Ratbi et al., 2015; C. E. Smith et al., 2016).

Kohlschutter-Tonz syndrome (KTZS OMIM: 226750): KTZS is a condition characterized by hypocalcified AI along with severe global developmental delay, early-onset intractable seizures and spasticity. In severe cases patients can suffer from profound mental retardation, never acquire speech, and become bedridden early in life. KTZS is caused by homozygous or compound heterozygous mutations in the *ROGDI* gene. All the mutations reported so far predicted to cause complete loss of protein function (Mory et al., 2012). The exact function of the ROGDI is unknown, though study suggested it may function in regulating exocytosis in ameloblasts (Riemann et al., 2017).

Enamel-renal syndrome (ERS OMIM: 204690): ERS is caused by homozygous or compound heterozygous mutations in the *FAM20A* gene. This syndrome is characterised by hypoplastic AI along with pulp stones, delayed or failed eruption of the secondary dentition, gingival overgrowth (Gingival fibromatosis syndrome), and nephrocalcinosis. ERS has been always

associated with intrapulpal calcification on both erupted and failed-to-erupt teeth, with variable severity (de la Dure-Molla et al., 2014). A study reported, loss of function *FAM20A* pathogenic variants caused ERS with extremely thin enamel in a patient. Transcriptome analysis of dental pulp tissues of this patient revealed significant upregulation of *DSPP, MMP20* and canonical BMP signalling pathway associated genes. This study concluded that pathogenic variants in *FAM20A* causes increased process of dentinogenesis that may cause intrapulpal calcification in the ERS patient (S. K. Wang et al., 2023). Mutations in *FAM20A* can also cause non-syndromic autosomal recessive AI (Jaureguiberry et al., 2012). It is unknown whether certain mutations only affect teeth or whether all patients are at risk of kidney problems in later life.

Dental anomalies and short stature (DASS OMIM: 601216): DASS is caused by homozygous or compound heterozygous loss of function mutations in the *LTBP3* gene. LTBP3 is an extracellular matrix protein may function in regulating TGF-beta secretion, trapping and activation. Alongside hypoplastic AI, DASS patients also display significantly shorter stature and brachyolmia (Bertola et al., 2009). Heterozygous mutations in *LTBP3* gene are also known to cause acromicric dysplasia and geleophysic dysplasia 3 (OMIM: 617809), characterized by disproportionate short stature with a short trunk, distinctive facial features, and heart diseases (McInerney-Leo et al., 2016). The dental phenotype of acromicric dysplasia patients shows a significant reduction in the microhardness in both enamel and dentine, with deeper grooves surrounding the enamel prisms than are seen in equivalent control teeth (Tantibhaedhyangkul et al., 2023).

Short stature, amelogenesis imperfecta, and skeletal dysplasia with scoliosis (SSASKS, OMIM: 618363): SSASKS is caused by homozygous or compound heterozygous mutations in the *SLC10A7* gene. The function of SLC10A7 in human is not known. The molecular mechanism of the disease is predicted to be loss of function of the protein. The characteristic features of SSASKS are disproportionate short stature, dental decay with discoloured hypoplastic enamel, dental crowding and skeletal dysplasia (Ashikov et al., 2018).

Developmental and epileptic encephalopathy-25 with amelogenesis imperfecta (DEE25 OMIM: 615905): DEE25 is caused by homozygous or compound heterozygous mutations in the *SLC13A5* gene. *SLC13A5* encodes a transmembrane protein NaCT (Na+/citrate cotransporter) that transports citrate from the circulation to the tissues like bone, enamel (Markovich & Murer, 2004). Citrate may play a role in stabilizing the thin mineral ribbons as several biophysical studies demonstrated that citrate can initiate and stabilise apatite crystallization in aqueous solutions (Jiang et al., 2009). DEEF25 is characterized by seizures, global developmental delay with intellectual disability and poor speech and communication, and enamel hypoplasia or hypodontia (Hardies et al., 2015). The disorder shows phenotypic similarities to Kohlschutter-Tonz syndrome (KTZS; 226750) which is caused by mutations in the *ROGDI* gene. However, a study reported detecting biallelic mutations in *SLC13A5* in clinically diagnosed KTZS patients. No mutation was detected in the *ROGDI* gene in these patients, suggesting *SLC13A5* is a second major gene causing KTSZ (Schossig et al., 2017).

Hypomagnesemia-5, renal with or without ocular involvement (HOMG5 OMIM 248190): HOMG5 is caused by homozygous or compound heterozygous mutations in the *CLDN19*

gene. Claudins are transmembrane tight junction proteins. CLDN19 forms a cation-selective tight junction complex with CLDN16 controlling selective paracellular ion movement between ECM and enamel organ at amelogenesis (Hou et al., 2008). HOMG5 is characterized by severe renal magnesium wasting, progressive renal failure, nephrocalcinosis and severe visual impairment (Konrad et al., 2006). The characteristic features of enamel in HOMG5 patients include yellow-brownish hypomaturation, hypoplastic enamel with pits and grooves as well as areas of total enamel absence (Yamaguti et al., 2017).

Primary immunodeficiency-10 (IMD10 OMIM: 612783): IMD10 is a primary immunodeficiency syndrome caused by homozygous mutations in the *STIM1* gene. STIM1, a plasma membrane protein, is a component of SOCE. The function of STIM1 in amelogenesis has been detailed in section 1.2.5.3 Ion transport in the maturation stage of Amelogenesis. IMD10 is characterized by recurrent infections, impaired T- and NK-cell function and decreased T-cell production of cytokines. Affected individuals may also have hypotonia, hypohidrosis, or AI (Parry, Holmes, et al., 2016).

1.2.5.5.4 Environmental causes of enamel defects

A recent review listed 114 environmental risk factors associated with developmental defects of enamel (DDE) of various aetiologies (Collignon et al., 2022). Environmental stressors affecting enamel development can be grouped by the timing of the exposure to the risk factors and the specific cellular processes affected by the risk factors. A higher frequency of DDE was associated with children who suffered from intrauterine malnutrition, maternal infections like toxaemia, Zika virus infection, rubella embryopathy, urinary tract infection, vitamin D deficiency, low calcaemia, gestational diabetes or maternal consumption of

alcohol, cigarettes or antibiotics (Purvis et al., 1973; Thomaz É et al., 2015; Via & Churchill, 1959). DDE was also found to be correlated with a number of neonatal health factors, such as delivery difficulties, intrapartum haemorrhage, caesarean delivery, and low birth weight (Pitiphat et al., 2014; Via & Churchill, 1959). Severe diseases like chicken pox, hypocalcaemia, renal disorders, liver failure, or cancer and associated medications and treatment procedures, between the age of 0-3 years were also found to increase the risk of DDE in children (Rugg-Gunn et al., 1998; Silva et al., 2020). Chemical exposures of radioactive elements, chronic lead poisoning and excess fluorides were also reported to cause DDE. Putative biological mechanisms affected by environmental stressors leading to DDE are predicted to be phosphocalcic metabolism, inadequate blood supply or oxygen deficiency, infections affecting functions of immune system, conditions like asthma, severe allergies causing respiratory acidosis and abnormal oxygen level (Fraser & Nikiforuk, 1982; Silva et al., 2020).

One of the most common and well-researched risk factors is excessive exposure to high fluoride levels during amelogenesis, leading to pitted, rough, or discoloured enamel which may also be more porous or chalky in texture (Wright, 2023). The prevalence of fluorosis is variable around the world, and is reported to affect around 60% of the population in the United States (Wiener et al., 2018). Depending on the amount of fluoride exposure, mineral composition and structure of the enamel can be affected to varying degrees, causing a mild to severe hypomineralised phenotype. A study in rats suggests fluorosis affects Wnt-βcatenin, Hedgehog and Notch signalling pathways (Qiao et al., 2021). Studies conducted on ameloblast-like cell line provided evidence that fluoride exposure affected ER calcium signalling pathways and also caused ER stress (Aulestia et al., 2020; Kubota et al., 2005).

MIH (Molar incisor hypomineralisation) is another enamel abnormality thought to be caused by a combination of genetic and environmental factors affecting permanent molars and incisor teeth in children (Wright, 2023). The enamel phenotype in MIH is characterized by mild to severe hypomineralization with enamel loss, but only specific teeth are affected, distinguishing it from AI which affects all teeth (Lygidakis et al., 2010). It is reported that childhood illness from birth to three years of age, prenatal and postnatal stressors including maternal illness, prematurity, caesarean section birth, kidney diseases, urinary tract infections, gastric disorders all can be associated with MIH (Garot et al., 2022; Ghanim et al., 2013). Bisphenol A (BPA), an endocrine-disrupting chemical (EDC), has been reported to increase the risk of developing MIH-like symptoms in rats (Jedeon et al., 2013). Genetic association studies reported that variants in a number of the genes involved in amelogenesis were found to be significantly correlated to susceptibility to develop MIH (Elzein et al., 2022; Jeremias et al., 2016). Another genetic association study also reported observing significant association between polymorphisms in immune response related genes and amelogenesis genes on the risk of developing MIH (Bussaneli et al., 2019).

1.2.5.5.5 Impact of AI

AI is a rare genetic disorder that severely affects the quality of oral health. The impact of AI can be significant on patients, their families, clinicians, and society as a whole. These impacts are reviewed below.

The most direct impact of AI is on the patient's dental health. AI makes enamel thin and weak, leaving the teeth prone to pain and infection. Patients experience difficulties in eating, drinking, and maintaining oral hygiene in everyday life (Lakhani, 2021). There are also well-

documented adverse psychosocial impacts of AI, with high levels of distress, social avoidance, embarrassment, isolation, and emotional problems (Pousette Lundgren et al., 2016). The family of an AI child, especially the parents, also suffer from the increased financial burden of ongoing, often expensive treatment, and their everyday normal routine is significantly affected (Poulsen et al., 2008). They may experience guilt and shame for passing on a hereditary disorder to their children. Both parents and the affected child need support to face this challenge. It was suggested that there is a need for an anonymous OSG (online support group) for adolescents with AI and for their parents, which could be helpful in providing them with emotional and informational support to cope with the disease. However, more research is also needed to understand the adverse psychosocial effects of AI on children, adolescents and their parents, in order to formularize the functions and elements of an OSG (Sneller et al., 2014).

There are currently no UK guidelines for the management of AI in children and adolescents. Studies have been conducted on the use of patient reported outcome measures (PROMs) in a group of children and young people with AI to assess the range of issues they experience in their daily life (Appelstrand et al., 2022; Lyne et al., 2021). PROMs are assessments completed by patients to provide information about their health, symptoms, quality of life, and treatment experiences. PROMs prioritize the patient's perspective and experiences, shifting the focus from disease-centred care to patient-centred care. The main outcome of these studies was that each patient experiences unique issues, many of which are not directly linked to the clinical presentation of AI, and early treatment strategies may improve patients’ satisfaction. These studies therefore recommended that clinicians incorporate

PROMs into routine clinical practice in order to explore all the individual's issues and concerns.

The impact of treating AI on the national health service (NHS) and clinicians can be significant due to the complexities and costs associated with managing this rare condition. AI patients often require a multi-disciplinary approach to treatment, including restorative dentistry, orthodontics, and periodontics. Coordinating these treatments can be challenging for dentists (Ayers et al., 2004). Treating AI patients often requires innovative dental materials and techniques to restore and protect teeth with compromised enamel. Dentists may also need to provide emotional support to patients who may be dealing with self-esteem issues, psychological challenges and financial burdens (Lafferty et al., 2021).

Research on enamel is also challenging primarily due to the unique properties and characteristics of enamel tissue. Enamel is acellular; it lacks the ability to regenerate which limits the scope to study amelogenesis through traditional cell biology techniques. Amelogenesis in humans occurs during early gestational stages, which makes it inaccessible to study. Rodents are used as experimental models to study various aspects of amelogenesis. Rodents make excellent model for studying amelogenesis because they share evolutionary conserved EMPs, proteases and other proteins with humans, and rodents have constantly growing incisors and their genetics are well characterized (Pugach & Gibson, 2014). However, there are several ethical issues associated with the use of animals in research. Scientists and public hold strong opposition to using animal in research for causing them pain and sufferings (Kiani et al., 2022). It is recommended that whenever possible, scientists should explore non-animal alternatives, such as cell cultures, computer modelling,

or in vitro studies, to replace or reduce the use of animals. Advances in cell culture techniques, stem-cell research, 3D tissue models, and organ-on-a-chip systems are improving the capabilities of in vitro experiments to mimic physiological processes more accurately, making them even more useful as alternatives to animal research (Swaters et al., 2022).

1.2.5.5.6 Clinical treatments for AI

The management of patients with AI is often complex and demanding for the patient and their family, as well as for the clinicians. This is intensified by care being extended over many years, starting as young children and extending into adulthood. The evidence base for clinical decision-making in AI is extremely limited (Strauch & Hahnel, 2018). Accordingly, treatment choices can be dependent on the experience and opinion of the clinician providing care.

While there is no cure for AI, various treatments are available to manage its symptoms and improve the appearance and function of affected teeth. The specific treatment approach depends on the AI type, the underlying enamel defect, the age of the patient and how severely the enamel has been damaged whilst in the mouth. Treatments that rely on bonding to enamel, such as sealants and composites are most likely to be effective where the enamel is well mineralised, such as in some forms of hypoplastic AI. By contrast, restorations that depend on bonding are unlikely to have good longevity where the enamel is poorly mineralised or includes protein, such as hypocalcified and hypomaturation AI. In these instances, full dental coverage or crowns may be a better option (Brunton et al., 2013).

Recently, the design of bioinspired peptides for tissue engineering and repair of enamel has attracted much interest among researchers in clinical dentistry, particularly in treating caries. A study was conducted on the application of amelogenin-inspired synthetic peptides with retained functional domains to mediate aprismatic enamel mineralization in vitro. This experiment showed that these peptides were able to promote oriented nucleation of layers of HAP crystals on human molar tooth enamel. However, further research is needed to reach the level of endurance, structural hierarchy, and scalability (from microns to millimetres) required for clinical applications of these peptides (Mukherjee et al., 2018).

Once the enamel is fully developed, ameloblasts undergo apoptosis leaving enamel no capacity for cellular repair. This has prompted researchers to speculate as to whether it may be possible to intervene during enamel formation to prevent defects from developing at this early stage. It was observed that the p.Tyr64His mutation in *AMELX* resulted in a hypomaturation AI phenotype in mice similar to the phenotype observed in human X-linked AI. The disease phenotype in the mice was predicted to be caused by malfunctioning of the ameloblast secretory pathway as evidenced from the accumulation of amelogenins intracellularly. These intracellular amelogenins induced ER stress leading to ameloblast apoptosis in the mutant mice. This study further showed that oral application of 4-phenylbutyrate (4-PBA) rescued the enamel phenotype in female mice that were heterozygous for the mutation but failed to do so in the male mice. As, females have two X chromosomes, the normal copy of the gene had partially compensated for the mutant copy, therefore, application of 4-PBA had improved the disease phenotype in females significantly. On the other hand, male mice being hemizygous for amelogenin produced exclusively mutant protein couldn’t be rescued by 4-PBA (Brookes et al., 2014). 4-PBA is an FDA

approved drug that acts as a chaperone protein in the treatment of human urea cycle disorders. 4-PBA can assist in protein folding and help maintain proper protein structure. These properties of 4-PBA have been shown to reduce ER stress response associated with the accumulation of unfolded or misfolded proteins. Although this finding raises the possibility of using chaperone therapy as a treatment for AI, the potential side effects associated with the dose of 4-PBA required to produce a therapeutic effect were such that other more effective chaperones need to be identified before this approach can be used clinically.

Recently, researchers successfully differentiated hiPSC (human-induced pluripotent stem cell) derived cells with ameloblast-like characteristics. The differentiated cells demonstrated high ALP (alkaline phosphatase) activity and elevated expression of several odontogenesis specific genes *COL1A1, RUNX2*, *Osterix, DSPP* and *DMP1*. Success in differentiating ameloblast-like cells holds potential for modelling tooth-related diseases in order to better comprehend the disease mechanisms (Kim et al., 2023).

1.3 DNA Sequencing Technologies

The primary aim of this project was to use new and improving DNA technologies to analyse a cohort of previously unsolved AI cases. DNA sequencing is the process of determining the type and order of nucleotides in a DNA molecule. Before the deciphering of the genetic code, scientists relied on various methods and approaches to study genetics and the heredity of traits. Classical genetics, pioneered by Gregor Mendel, relied on the observation of phenotypic traits and the principles of inheritance to understand genetic mechanisms. Selective breeding was used for centuries to improve the traits of plants and animals. In the early 20th century, Thomas Hunt Morgan and his colleagues created genetic maps by studying the inheritance patterns of genes on different chromosomes in fruit fly Drosophila melanogaster (Morgan, 1910). This work led to the identification of linked genes and the concept of recombination. In the mid-20th century, George Beadle and Edward Tatum's "one gene-one enzyme" hypothesis highlighted the connection between genes and enzymes, which was a significant step towards our understanding of the role of DNA in encoding genetic information (Beadle & Tatum, 1941). The Hershey-Chase experiment in the 1950s provided strong evidence for the role of DNA in carrying genetic information (Hershey & Chase, 1952). The discovery of the DNA double helix structure by James Watson, Francis Crick, Rosalind Franklin and Maurice Wilkins laid the foundation for understanding the structure of DNA (Watson & Crick, 1953). Then in the early 1960s Marshall Nirenberg, Har Govind Khorana and Robert William Holley worked together unravelling the genetic code by which DNA encodes information for the synthesis of proteins ("Nobel Prizes for Medicine, 1968," 1968). The elucidation of the genetic code was a critical step in our understanding of how the information stored in DNA is translated into the proteins that carry out the functions of life. The breakthrough in DNA sequencing came in 1970s with the development

of methods that could directly decode the genetic information contained in DNA molecules. After that, DNA sequencing technologies have evolved significantly, especially over the last decade, offering various methods with differing principles, read lengths, costs, and applications. A summary of the key milestones in the history of DNA sequencing are outlined below.

1.3.1 First generation DNA sequencing

Frederick Sanger is considered to be the pioneer of DNA sequencing. He developed a method to determine the type and order of nucleotides in a DNA molecule, based on the selective incorporation of dideoxynucleotides (ddNTP), into the DNA molecule as base-specific chain terminators (Sanger et al., 1977). At the same time, Allan Maxam and Walter Gilbert also introduced a DNA sequencing method based on chemical modification of DNA followed by cleavage at specific bases (Maxam & Gilbert, 1977). However, the Maxam-Gilbert method lost popularity soon afterwards because it was lengthy, required the use of hazardous materials and only allowed sequencing of up to 500 base pairs. On the other hand, the chain termination method developed by Frederick Sanger gained popularity and is still widely used in research areas that require sequencing of single small stretches of DNA quickly (Daniels et al., 2021; Sanger et al., 1977b). Improvements were made to the technique over time, and several more advanced sequencing technologies were built based on the chain termination method. The mechanism of the chain termination method is described below.

1.3.1.1 Chain-termination sequencing

In the chain-termination method of DNA sequencing, the DNA template is denatured and a DNA primer is annealed to a specific region of the DNA template to provide a starting point for DNA synthesis. DNA polymerase synthesizes a new DNA strand from the 3' end of the primer and proceeds in the 5' to 3' direction along the template strand in the presence of regular deoxynucleotidetriphosphates (dNTPs: A, C, T, G), and dideoxynucleotidetriphosphates (ddNTPs). The ddNTPs lack the 3'-OH group that forms the phosphodiester bond between adjacent nucleotides and are fluorescently labelled. Therefore, when a ddNTP gets incorporated in the growing DNA strand, the DNA extension ceases (Sanger et al., 1977a). The labelled DNA fragments are size separated by gel electrophoresis and visualized using a laser or ultraviolet (UV) light source. The fluorescence emitted by the labelled nucleotides is detected and recorded, revealing the sequence of bases in each fragment. The output file, generated from specialist basecalling software, produces a chromatogram which shows corresponding nucleotide (A, T, C, G) at each position along the sequence (Slatko et al., 2011). The trace in the chromatogram is often referred to as an electropherogram, representing the fluorescence intensity over time, as the DNA fragments pass through the detector. A schematic illustration of chain termination method is provided in Figure 1. 8.

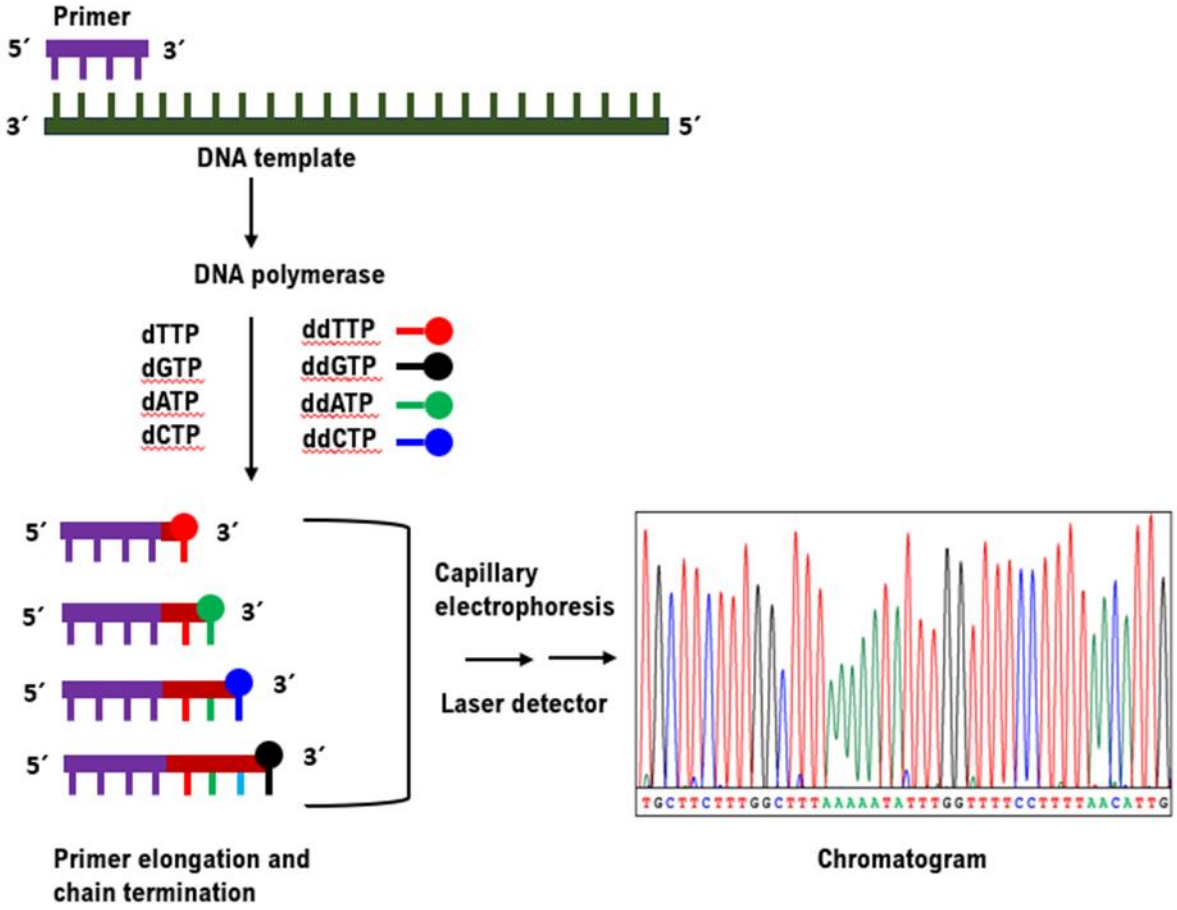


Figure 1. 8 Schematic of chain termination method.

The schematic details the steps of the chain termination method developed by Frederick Sanger. In this method, DNA synthesis is performed using a mixture of dNTPs and ddNTPs. DNA synthesis continues, using dNTPs, until there is incorporation of a ddNTP, at which point extension ceases. The terminal ddNTP labelled with a fluorophore that fluoresces at a wavelength that is specific to the corresponding base (A, T, C, G). A detector records the emitted fluorescent signal as peaks on a chromatogram. The order and timing of these peaks represents the sequence of bases in the original DNA strand. The sequence is determined by reading the DNA fragments from the smallest to the largest. Image reproduced with permission from Estevezj, https://commons.wikimedia.org/w/index.php?curid=23264166).

1.3.2 Next generation sequencing (NGS)

Three decades after the Sanger sequencing was first described, a significant technological breakthrough in DNA sequencing occurred with the introduction and subsequent ubiquitous adoption of next generation sequencing (NGS) or massively parallel sequencing. This approach allows the concurrent sequencing of millions of short (typically 150-bp) DNA fragments from both ends (paired-end), at relatively low cost and high accuracy, with the potential for large-scale multiplexing (Metzker, 2010).

Library preparation

The predominant approach of most NGS platforms is sequencing-by-synthesis (SBS), which allows the detection of a cluster of bases as they are incorporated into DNA strands that are immobilised on a fixed surface (Bentley et al., 2008). DNA libraries are prepared by random fragmentation of the double stranded DNA template followed by the addition of sequencing adaptors at each end of their double stranded terminal ends. The DNA libraries are typically PCR amplified to increase their molarity prior to sequencing. A unique DNA barcode is also incorporated into each fragment to allow multiple samples to be analysed in a single run (Linnarsson, 2010). The sequencing involves generating two sequences for each DNA fragment one from each end of the fragment, called paired-end sequencing. In the early days of NGS, several competing platforms emerged offering different technical versions of massively parallel sequencing. Ultimately, SBS, using platforms manufactured by Illumina, now dominate the field.

Sequencing

Illumina short-read paired-end sequencing starts with denaturing double stranded DNA libraries into single strands which are loaded onto a patterned flowcell. Patterned flow cells contain billions of nano-wells at fixed locations, embedded with oligonucleotides complementary to the sequencing adapters. Once a single stranded DNA molecule binds to a seeding primer within a single nanowell, it starts cluster generation (making many copies of itself) immediately and rapidly before a second molecule is seeded in the same location. This process, called exclusion amplification, enables clonal amplification from a single DNA template. After cluster generation finishes, sequencing primers are hybridized to the DNA strands in each cluster. A sequencing reaction mix containing all four fluorescently labelled reversible terminator-bound dNTPs and DNA polymerase are allowed to flow over the flowcell. When a dNTP is incorporated in a cycle, the attached flurorophore is excited and images are captured by a camera or detector from each cluster or spot. All fluorescently labelled dNTPs include a block that prevents addition of more than one nucleotide in one cycle. After the image is captured in a cycle, the incorporated dNTP is deblocked, the fluorescent component is removed and another cycle begins. This repeated process of incorporation, excision and imaging continues for a fixed number of cycles (typically 150) to generate the first single read, called ‘read 1’. This process is also repeated to generate a second read, ‘read 2’ by sequencing the other end of the same DNA fragment. The wavelength of the fluorescent signal indicates which base has been incorporated into the growing DNA strand cluster in each nano well. This process is massively parallel across millions of fragments (Yoshinaga et al., 2018). A schematic representation of sequencing by synthesis (SBS) technology is illustrated in Figure 1. 9.

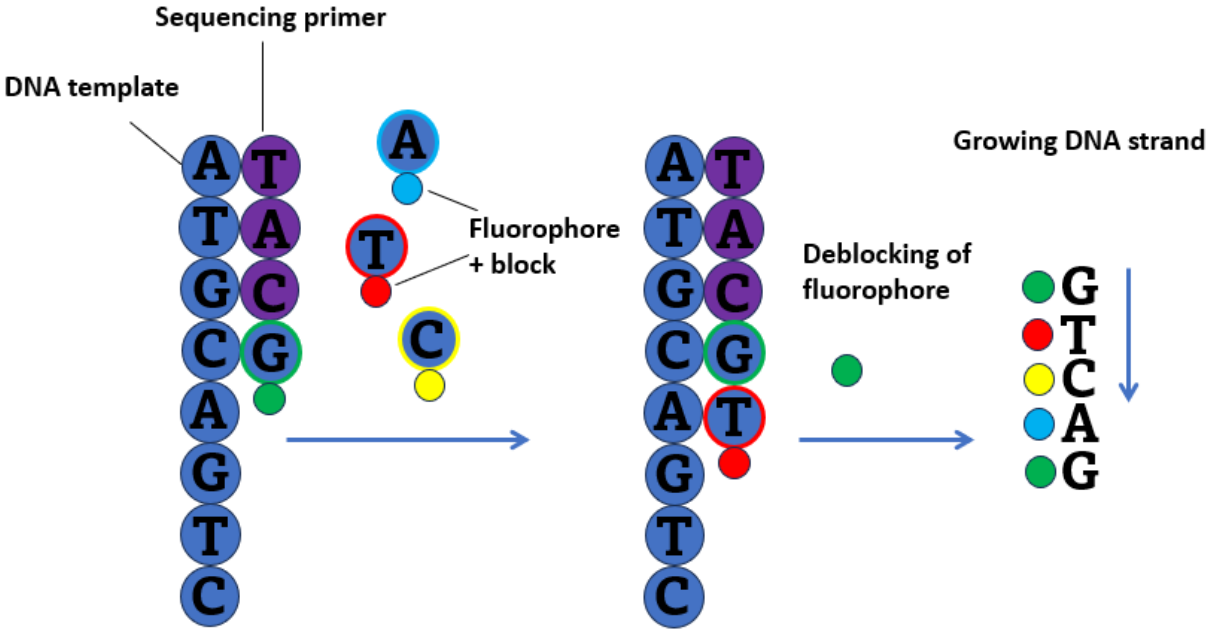


Figure 1. 9 Sequencing by synthesis (SBS) method.

The schematic details the steps of SBS technology used by Illumina platforms. In the first step, fluorescently tagged nucleotides are incorporated into the growing DNA strand. Each of the four dNTPs have a unique fluorescent label that, when excited, emits a characteristic wavelength when added to the template DNA. A computer records all of the emissions, and from this data, base calls are made.

Base calling

After sequencing has finished, the instrument control software processes the raw fluorescence images saved as a TIFF (tag image file format) files into CIF (common intermediate format) files. The control software then extracts the signal intensities of each unique wavelength from the CIF file, converting them into a series of base calls for each cluster into a BCL (binary base call) file. Once the signals are processed, a basecalling algorithm assigns a base (A, T, C, G) to each cluster or spot, for each cycle, based on the intensity and characteristics of the fluorescent signals. In addition to assigning a base, the basecalling software generates per-base quality scores. The quality score is often a Phred-scaled quality value that reflects the probability of an incorrect base call. Higher Phred scores indicate higher confidence in the base call. For example, a Phred score of 20 corresponds to an accuracy rate of 99%, while a Phred score of 30 corresponds to an accuracy rate of 99.9%. The emitted file is generated in FASTQ format, a four line repeating file structure for each reads (Wang et al., 2015). FASTQ files are used for further analysis, using bioinformatics tools and software, to align and interpret the sequence, identify variations, and extract meaningful biological information.

Since it was initially introduced, the sequence yield from NGS instruments has been iteratively increased, library preparation workflows have been simplified and instrument run times have decreased (owing to improvements in sequencing chemistry and improved image capture hardware). As a result, NGS has been universally adopted to fit a wide range of applications for both small and large scale research projects.

1.3.2.1 Whole exome sequencing

NGS has a wide range of applications in various medical fields, enabling more precise diagnoses, personalized treatments, and advancements in research. Some key NGS applications are, whole genome sequencing (WGS), RNA sequencing (RNAseq), whole exome sequencing (WES) and methylation sequencing.

WES is a targeted approach focuses exclusively on sequencing the exome, which represents approximately 1% of the genome. The exome describes the transcribed regions of genes and includes the protein-coding regions where the majority of disease-causing mutations are located. WES employs a hybridization-based target enrichment method that uses a pool of biotinylated oligonucleotide (RNA) probes complimentary to a predefined set of exons. DNA libraries are made by adding platform specific sequencing adapters to the DNA by PCR, which also generates enough materials for subsequent capture hybridization. Biotinylated probes are hybridized to target genomic sequences, allowing subsequent capture of the exons by magnetic streptavidin beads. A post-capture PCR is performed to increase the number of captured DNA libraries for high-throughput DNA sequencing. The basic workflow for WGS library preparation is similar to that of WES, however it lacks a hybridization capture target enrichment step, allowing sequencing the entire genome including both coding regions (exons) and non-coding regions (introns, intergenic regions). Generally, WES is more cost-effective than WGS because it sequences only a fraction of the genome. WGS generates a larger volume of data due to sequencing the entire genome and requires more storage and computational resources for data analysis. On the other hand, WES generates a smaller volume of data, making it more manageable for analysis and storage.

A custom target enrichment probe technique is also available provided by some commercial vendors to selectively capture any coding or non-coding regions of the genome. This allows researchers to design custom panels or probes to selectively sequence particular genes, genomic regions, or regions associated with a specific research question or disease (Li et al., 2010).

1.3.2.2 NGS data analysis

Analysis of next-generation sequencing (NGS) data is a complex, multi-step, process that requires a number of consecutively arranged components. Data analysis is computationally intensive, requiring a high-performance computing system, adequate data storage capacity and a reliable data management system (S. Pabinger et al., 2014). General guidelines on the different steps involved in NGS data analysis are described below. The workflow may vary depending on the specific objectives of the study or clinical applications. Custom analysis scripts and workflows are typically deployed to address specific research questions. WES specific data analysis pipeline can be found in Appendix 6.

Data analysis begins by assessing the quality of raw sequencing data using tools such as FastQC to determine the proportion of low-quality reads, adapter contamination, and overrepresented sequences. Adaptor trimming and low-quality data filtering are essential preprocessing steps in NGS data analysis to ensure the accuracy and reliability of downstream analyses. A bioinformatic tool called Cutadapt (short form of cut adaptor) can identify and remove adaptors, primers and low quality bases from the NGS data. The next step is read alignment and mapping to a suitable reference genome. BWA (Burrows-Wheeler Aligner) is a bioinformatics tool used for aligning DNA sequencing reads to a reference

genome. BWA starts by finding short, exact matches between the sequencing read and the reference genome. Then it extends the matching process to create longer alignment candidates. BWA assigns a score to each alignment candidate, taking into account factors like mismatches, gap openings, and quality scores of the sequencing read. The alignment with the highest score is reported, which includes information about the location in the reference genome where the read aligns, as well as any mismatches or gaps. BWA typically generates output files in the Sequence Alignment/Map (SAM) or Binary Alignment/Map (BAM file) format. Example of other aligners are Bowtie2 and minimap2.

The next step in a data processing pipeline is the removal of PCR duplicates from the mapped reads. PCR duplicates are copies of the original DNA fragment and are defined as having identical mapping coordinates between mapped read pairs. They occur due to the PCR amplification step during library preparation. Identifying and removing these duplicates is crucial to avoid biases and inaccuracies in variant calling. Samtools or Picard are programmes that are commonly used for the manipulation of high throughput sequencing data and include functions such as file format conversion, and the sorting of alignment files by mapping coordinate. Picard can identify and mark PCR duplicates in the sorted BAM file allowing their removal from downstream analysis steps.

The identification of non-reference base (variant calling) is often performed using the HaplotypeCaller a tool in the Genome Analysis Toolkit (GATK). The HaplotypeCaller is a robust tool for identifying single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) from high-throughput sequencing data. Variants can be annotated using tools such as VEP (Variant Effect Predictor) to predict functional

consequences and determine allele frequencies in population databases (Stephan Pabinger et al., 2014).

The Integrative Genomics Viewer (IGV) is a powerful and widely used software tool for visualizing and exploring genomic data in a graphical and interactive way (Robinson et al., 2022). IGV allows the viewing of aligned sequence reads (BAM files) and is platform agnostic. This allows it to be used as the end-point interface for a variety of data processing pipelines.

1.3.2.3 Single molecule molecular inversion probes (smMIPs)

Despite improvements to WES, both customized and non-customized targeted hybridisation capture workflows are still complex, time-intensive, have high per-sample reagent costs, and have no or limited flexibility to reformulate the protocol if the regions of interest change over time. Furthermore, targeted hybridisation sequencing remains poor for the reliable detection of subclonal variations that have allele frequency less than 1% (a parameter that is of particular interest to the oncology field). Hiatt and colleagues (Hiatt et al., 2013) optimized a simple and scalable, massively parallel sequencing-based method for the detection of low frequency somatic variation in cancer causing genes. The method is called the molecular inversion probes (MIP) technique and was originally developed for detecting SNPs for patients with immunoglobin nephropathy or Berger's disease (Hardenbol et al., 2003). The technique was further improved by incorporating a series of random nucleotides, the unique molecular identifiers (UMIs), to the MIP backbone to capture single molecular events and distinguish them from PCR duplicates (Turner et al., 2009). PCR duplicates result in an overrepresentation of sequences in the sequence data. This skews the representation of genomic regions, potentially masking low-abundance variants and affecting the accuracy

of variant calling, leading to the identification of false-positive variants or an inaccurate estimation of variant allele frequencies. The combination of MIPs and single molecule tagging created an ultra-sensitive assay called single molecule molecular inversion probes (smMIPs). The smMIPs approach uses a PCR based targeted sequencing method to selectively capture many genomic positions of interest in hundreds of samples simultaneously. The library preparation method allows target amplification without the need for the target-capture steps required by WES. The method is also cost-effective and can be used to analyse small quantities of degraded DNA (Cantsilieris et al., 2017; Hiatt et al., 2013).

The smMIPs method involves designing oligonucleotide probes consisting of a common DNA backbone flanked by target-specific sequences located at the 3’ and 5’-end called extension and ligation arms. These arms hybridize to the complementary sequence at the target DNA in the same way as PCR primers, but they are physically linked. A DNA polymerase fills the gap between the arms, this being the target region of interest. Finally, the probe is circularized with the addition of DNA ligase. The circular target is then amplified by PCR using universal primers complementary to the smMIP backbone. The primers contain sequencer specific sequences which are added to the target during the PCR, making the final PCR product ready for sequencing. The integration of UMIs uniquely tags each individual clone in the starting library to track single molecular events and removes PCR duplicates from the bioinformatics analysis, providing sensitive detection of variants at low frequencies (Hardenbol et al., 2003; Hiatt et al., 2013). Features of a smMIP probe is presented in Figure 1. 10.

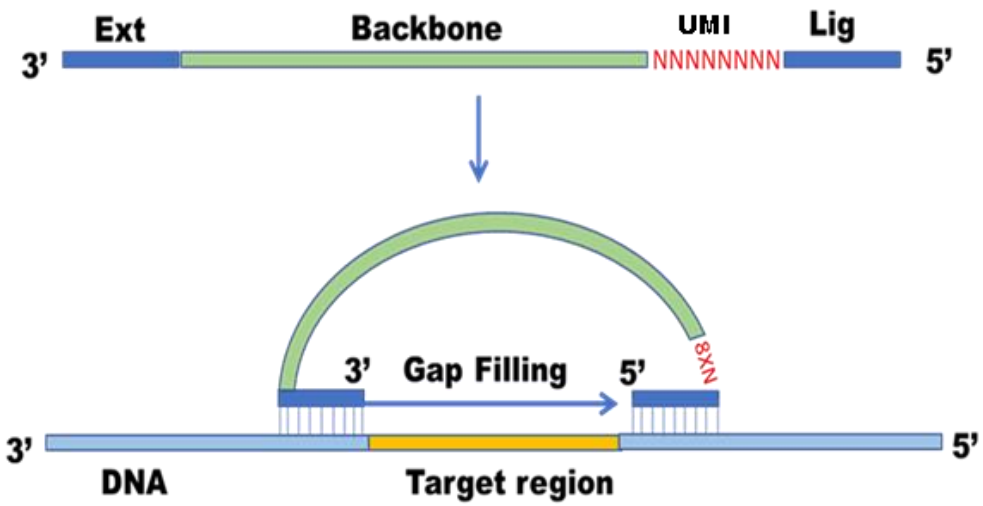


Figure 1. 10 Features of a smMIP probe.

The schematic illustrates an smMIP probe consisting of a common DNA backbone flanked by gene-specific sequences at both ends; these are termed extension and ligation arms. A short stretch of random bases (UMIs) are incorporated in the backbone to capture each single molecular event. Complementary sequences for universal primers are located in the probe backbone.

1.3.3 Third generation sequencing

The low-cost, high-throughput, capability of NGS has resulted in its universal adoption. However, one cited limitation is the production of relatively short sequence reads, typically 150bp in length. This limits the scope of analysis for a number of applications. For example, assembling a complete genome or transcriptome, through complex or repetitive regions remains challenging using short-read data. The assembly approach considers highly similar sequences to have originated from the same genomic location, and therefore often fails to map the repetitive regions that are scattered throughout the genome sequence correctly. Instead these are algorithmically collapsed into only a few sequences due to their similarity, producing an incorrect genome assembly (Baptista et al., 2018). Resolving large, complex structural variants, including insertions, deletions, duplications, inversions and translocations, is also difficult using short read sequence data because the read lengths are insufficient to detect the breakpoints of these changes (Snyder et al., 2015). Haplotype analysis, detecting inheritance patterns of genetic variants and assigning parental origin of *de novo* variants are also not within the scope of short-read sequencing analysis because it represents the sample as a haploid genome, introducing assembly errors in regions that diverge between haplotypes (Snyder et al., 2015; Vinson et al., 2005).

Long-read sequencing technologies address many of these limitations by enabling the sequencing of DNA fragments thousands, to tens of thousands of bases in length, providing a more comprehensive view of the genome. Long-read sequencing can produce megabase-scale phase blocks at chromosome scale, enabling the study of complex genomic features, repetitive regions and structural variations with reduced number of gaps or biases. Also, it is

possible to get DNA methylation information by analysing native genomic DNA (Lin et al., 2022). It also facilitates the identification and characterization of full-length transcripts and alternative splicing events by direct sequencing of the single-molecule native RNA without the need for amplification and reverse transcription (Loman & Watson, 2015). This has enabled the detection of isoform diversity, providing a more comprehensive understanding of the transcriptome and its functional implications (Logsdon et al., 2020). These benefits have broad implications in fields such as genomics, genetics, evolutionary biology, and precision medicine. Currently, Oxford Nanopore Technology (ONT) and Pacific Biosciences (PacBio) are the two most popular platforms, providing efficient long-read sequencing (Lu et al., 2016; Wenger et al., 2019). The workflows for these technologies are provided below.

1.3.3.1 Nanopore sequencing

Developed by Oxford Nanopore Technologies (ONT), nanopore sequencing works by passing single-stranded DNA or RNA molecules through a tiny protein pore. A sequencing flow-cell contains an array of nanopores, each with its own corresponding electrode connected to a channel and sensor chip, which measures electric current that flows through the individual nanopore. When the DNA or RNA molecule passes through the nanopore, a series of bases within the pore disrupt the flow of hydrogen ion producing a characteristic decrease in the current amplitude, which gets recorded by a detector. The change in the current provides information about the nucleotide sequence of the molecule, in real-time (Clarke et al., 2009). ONT enables the attainment of a sequence read length up to 100 kb. ONT provides various types of flow cells, each offering variable data output capacities. For instance, a PromethION flow cell can generate up to 100 Gb (gigabase) of data. A graphic illustration of the nanopore sequencing process is shown in Figure 1. 11.

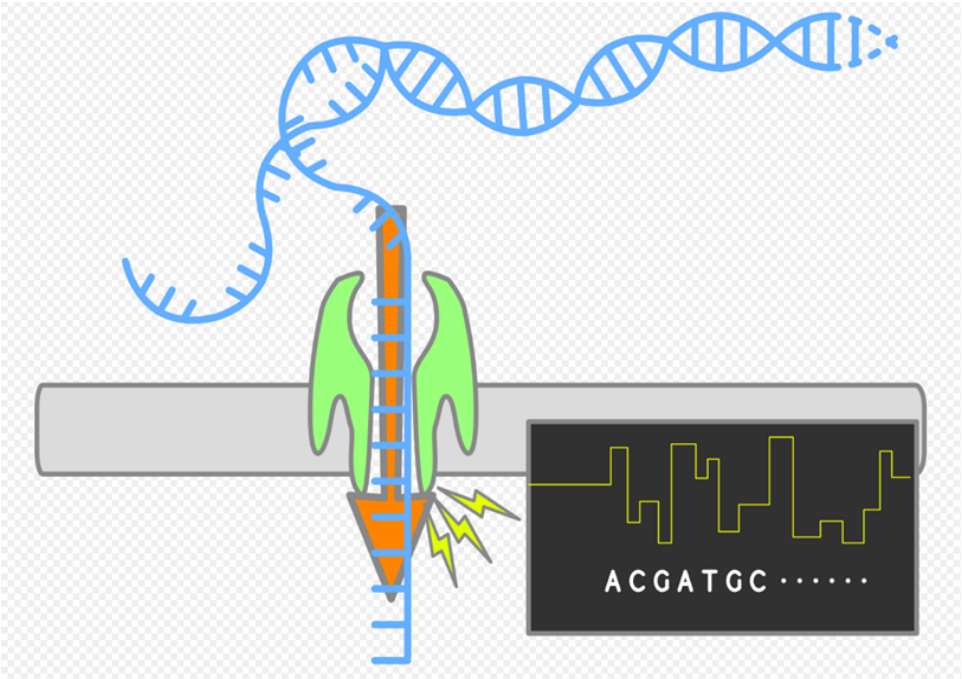


Figure 1. 11 Schematic representation of nanopore sequencing.

A schematic illustrating the process of nanopore sequencing. A single-stranded DNA molecule passes through a nanopore disrupts a continuous electrical current. The change in the current is recorded, providing information about the nucleotide sequence of the DNA molecule. The image was adapted with permission from DataBase Center for Life Science (DBCLS), https://commons.wikimedia.org/w/index.php?curid=86372818.

1.3.3.2 PacBio/SMRT sequencing

PacBio uses SMRT (Single-Molecule, Real-Time) technology to sequence read lengths up to 25 kb with 99.99% accuracy and can generate data in the range of several Gb to over 100 Gb per run, depending on the specific instrument and chemistry version (John Eid et al., 2009). Below is an overview and description of the key features of SMRT technology.

Principle of SMRT Technology

SMRT bell library preparation begins with the generation of blunt-end high quality, high molecular weight double stranded DNA (250 bp to greater than 25 kb). The double stranded template DNA is then circularized by adding SMRT bell hairpin adaptors. SMRT Bell adapters provide a binding site for a sequencing primer from which DNA sequencing is primed. The circular DNA template allows sequencing of both forward and reverse DNA strands simultaneously. Sequencing primers and DNA polymerases are added to the library and placed on the PacBio instrument for sequencing. A schematic representation of SMRT bell library preparation and sequencing is displayed in Figure 1. 12.

The polymerase used in PacBio sequencing is highly processive, meaning it can remain attached to the DNA template for an extended period, allowing sequencing of long DNA fragments in a single pass (J. Eid et al., 2009). Also, it is designed to have high fidelity by reading the molecule multiple times, which is crucial for generating accurate sequencing data. The core of SMRT sequencing is the SMRT cell, which is a small, disposable, and microfabricated device that serves as a sequencing chamber. Within the SMRT cell, there are millions of tiny wells called zero mode waveguides (ZMW). Each ZMW is a nanoscale well. A DNA template-polymerase complex is immobilized at the bottom of the ZMW.

Phospholinked dNTPs each labelled with a different coloured fluorophore are introduced into the ZMWs. As the DNA polymerase incorporates nucleotides during sequencing, the fluorescent signal generated by each nucleotide is detected by sensitive detectors within the ZMW (Korlach et al., 2008). The signal is detected in real-time using specialized optical systems, and the dynamics of the fluorescence signals determines the DNA sequences.

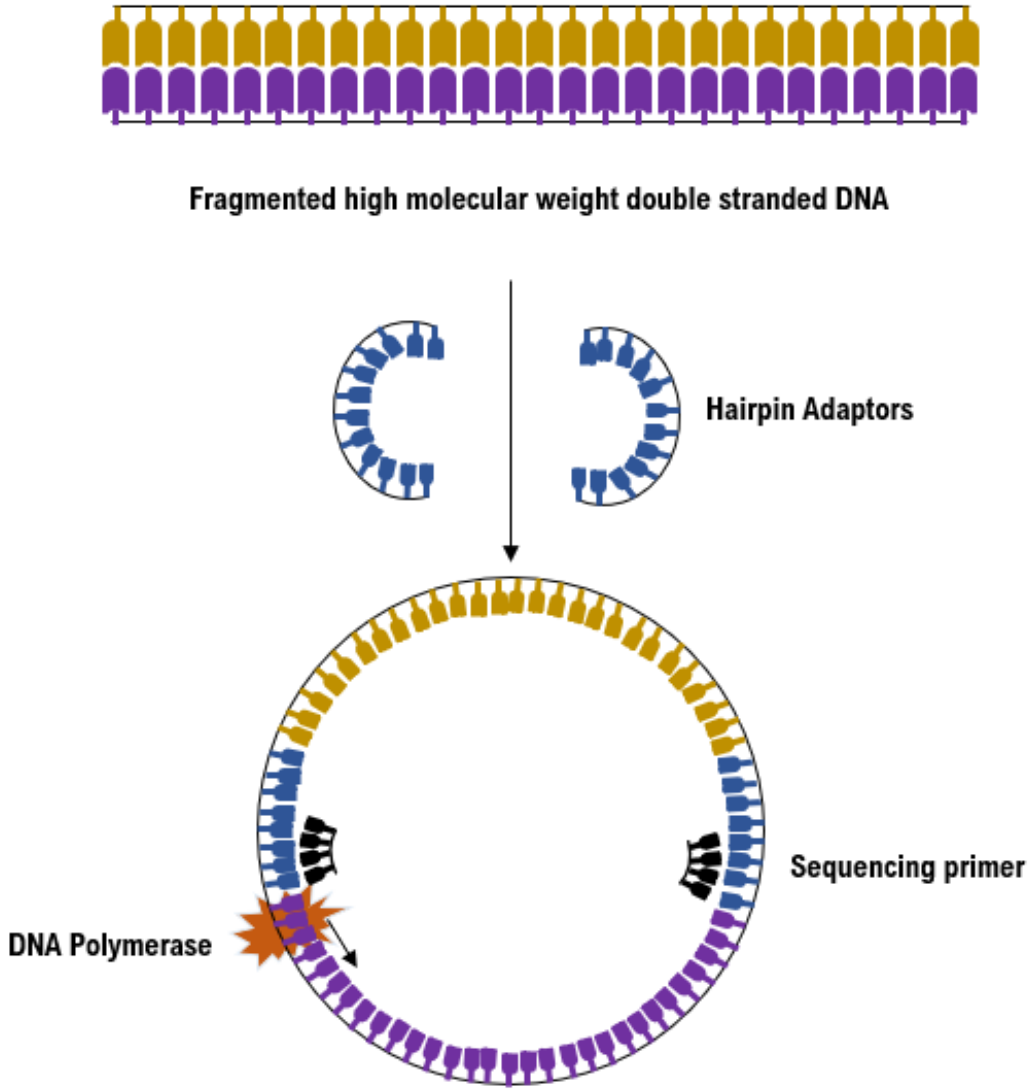


Figure 1. 12 Schematic illustration of SMRT bell sequencing.

A schematic illustration of SMRT-bell sequencing. The circular SMRT bell library is prepared by adding hairpin adaptors to the double stranded template DNA. As the DNA polymerase synthesizes the complementary strand, the incorporation of nucleotides emits light pulses that get detected by the SMRT cell's optical system in real-time for sequence determination.

1.4 Aims

Following the ubiquitous adoption of NGS, WES had become the preferred approach to AI research over the past decade. As a result, 20 novel genes were discovered to be associated with NS AI and 95 syndromes have been reported to have an enamel phenotype among their disease description in OMIM (Wright, 2023). Identifying disease associated genes and further studying the molecular pathways implicated in AI has increased our understanding of disease pathogenesis, benefitting AI patients and dentists around the world. However, there are still many AI families for whom it was not possible to determine a genetic diagnosis. The primary objective of this project was to develop a custom DNA sequencing reagent using smMIPs, for rapid, relatively low cost first-pass sequencing of genomic DNA from AI patients, that would enable sequencing of the coding and splicing site regions of the genes known to be associated with NS AI. The method aimed to provide flexibility in including or excluding further target regions later point, and to be cheaper than other conventional targeted DNA sequencing approaches. Consequently, this project began with an attempt to develop a low-cost rapid screening method applicable to a cohort of 181 AI patients. We subsequently directed our limited WES resources towards families negative for coding or splice site variants in known genes, to identify novel disease-associated genes underlying AI.

The aims of this project were to:

1. Develop and validate a targeted DNA sequencing method, single molecule molecular inversion probes (smMIPs), for the sequencing of coding and splice-site regions of nineteen genes known to cause NS AI.

2. Use the smMIPs reagent to screen all the unsolved cases in the Leeds AI cohort, to determine what proportion of cases are accounted for by mutations in the known genes.
3. Test the hypothesis that more genes underlying AI remain to be identified, by studying AI cases/families negative for smMIPs-AI gene panel analysis by whole exome sequencing (WES).

1.5 References

Aldred, M. J., Crawford, P. J., Roberts, E., & Thomas, N. S. (1992). Identification of a nonsense mutation in the amelogenin gene (AMELX) in a family with X-linked amelogenesis imperfecta (AIH1). *Hum Genet*, *90*(4), 413-416. https://doi.org/10.1007/bf00220469

Altug-Atac, A. T., & Erdem, D. (2007). Prevalence and distribution of dental anomalies in orthodontic patients. *Am J Orthod Dentofacial Orthop*, *131*(4), 510-514. https://doi.org/10.1016/j.ajodo.2005.06.027

Appelstrand, S. B., Robertson, A., & Sabel, N. (2022). Patient-reported outcome measures in individuals with amelogenesis imperfecta: a systematic review. *Eur Arch Paediatr Dent*, *23*(6), 885-895. https://doi.org/10.1007/s40368-022-00737-3

Ashikov, A., Abu Bakar, N., Wen, X.-Y., Niemeijer, M., Rodrigues Pinto Osorio, G., Brand-Arzamendi, K., Hasadsri, L., Hansikova, H., Raymond, K., Vicogne, D., Ondruskova, N., Simon, M. E. H., Pfundt, R., Timal, S., Beumers, R., Biot, C., Smeets, R., Kersten, M., Huijben, K., . . . Johann, t. W. N. (2018). Integrating glycomics and genomics uncovers SLC10A7 as essential factor for bone mineralization by regulating post-Golgi protein transport and glycosylation. *Human Molecular Genetics*, *27*(17), 3029-3045. https://doi.org/10.1093/hmg/ddy213

Aulestia, F. J., Groeling, J., Bomfim, G. H. S., Costiniti, V., Manikandan, V., Chaloemtoem, A., Concepcion, A. R., Li, Y., Wagner, L. E., 2nd, Idaghdour, Y., Yule, D. I., & Lacruz, R. S. (2020). Fluoride exposure alters Ca(2+) signaling and mitochondrial function in enamel cells. *Sci Signal*, *13*(619). https://doi.org/10.1126/scisignal.aay0086

Ayers, K. M., Drummond, B. K., Harding, W. J., Salis, S. G., & Liston, P. N. (2004). Amelogenesis imperfecta--multidisciplinary management from eruption to adulthood. Review and case report. *N Z Dent J*, *100*(4), 101-104.

Bäckman, B., & Holm, A. K. (1986). Amelogenesis imperfecta: prevalence and incidence in a northern Swedish county. *Community Dent Oral Epidemiol*, *14*(1), 43-47. https://doi.org/10.1111/j.1600-0528.1986.tb01493.x

Bäckman, B., & Holmgren, G. (1988). Amelogenesis imperfecta: a genetic study. *Hum Hered*, *38*(4), 189-206. https://doi.org/10.1159/000153785

Baptista, R. P., Reis-Cunha, J. L., DeBarry, J. D., Chiari, E., Kissinger, J. C., Bartholomeu, D. C., & Macedo, A. M. (2018). Assembly of highly repetitive genomes using short reads: the genome of discrete typing unit III Trypanosoma cruzi strain 231. *Microb Genom*, *4*(4). https://doi.org/10.1099/mgen.0.000156

Bartlett, J. D. (2013). Dental enamel development: proteinases and their enamel matrix substrates. *ISRN Dent*, *2013*, 684607. https://doi.org/10.1155/2013/684607

Bartold, P. M., Walsh, L. J., & Narayanan, A. S. (2000). Molecular and cell biology of the gingiva. *Periodontol 2000*, *24*, 28-55. https://doi.org/10.1034/j.1600-0757.2000.2240103.x

Beadle, G. W., & Tatum, E. L. (1941). Genetic Control of Biochemical Reactions in Neurospora. *Proc Natl Acad Sci U S A*, *27*(11), 499-506. https://doi.org/10.1073/pnas.27.11.499

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., . . . Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53-59. https://doi.org/10.1038/nature07517

Berkovitz, B. K. (2004). Periodontal ligament: structural and clinical correlates. *Dent Update*, *31*(1), 46-50, 52, 54. https://doi.org/10.12968/denu.2004.31.1.46

Bertola, D. R., Antequera, R., Rodovalho, M. J., Honjo, R. S., Albano, L. M. J., Furquim, I. M., Oliveira, L. A., & Kim, C. A. (2009). Brachyolmia with amelogenesis imperfecta: Further evidence of a distinct entity. *American Journal of Medical Genetics Part A*, *149A*(3), 532-534. https://doi.org/https://doi.org/10.1002/ajmg.a.32661

Bori, E., Guo, J., Rácz, R., Burghardt, B., Földes, A., Kerémi, B., Harada, H., Steward, M. C., Den Besten, P., Bronckers, A. L. J. J., & Varga, G. (2016). Evidence for Bicarbonate Secretion by Ameloblasts in a Novel Cellular Model. *Journal of Dental Research*, *95*(5), 588-596. https://doi.org/10.1177/0022034515625939

Bronckers, A. L. (2017). Ion Transport by Ameloblasts during Amelogenesis. *J Dent Res*, *96*(3), 243-253. https://doi.org/10.1177/0022034516681768

Brookes, S. J., Barron, M. J., Boot-Handford, R., Kirkham, J., & Dixon, M. J. (2014). Endoplasmic reticulum stress in amelogenesis imperfecta and phenotypic rescue using 4-phenylbutyrate. *Hum Mol Genet*, *23*(9), 2468-2480. https://doi.org/10.1093/hmg/ddt642

Brunton, P. A., Davies, R. P., Burke, J. L., Smith, A., Aggeli, A., Brookes, S. J., & Kirkham, J. (2013). Treatment of early caries lesions using biomimetic self-assembling peptides--a clinical safety trial. *Br Dent J*, *215*(4), E6. https://doi.org/10.1038/sj.bdj.2013.741

Bussaneli, D. G., Restrepo, M., Fragelli, C. M. B., Santos-Pinto, L., Jeremias, F., Cordeiro, R. C. L., Bezamat, M., Vieira, A. R., & Scarel-Caminaga, R. M. (2019). Genes Regulating Immune Response and Amelogenesis Interact in Increasing the Susceptibility to Molar-Incisor Hypomineralization. *Caries Res*, *53*(2), 217-227. https://doi.org/10.1159/000491644

Cantsilieris, S., Stessman, H. A., Shendure, J., & Eichler, E. E. (2017). Targeted Capture and High-Throughput Sequencing Using Molecular Inversion Probes (MIPs). *Methods Mol Biol*, *1492*, 95-106. https://doi.org/10.1007/978-1-4939-6442-0_6

Chen, S., Xie, H., Zhao, S., Wang, S., Wei, X., & Liu, S. (2022). The Genes Involved in Dentinogenesis. *Organogenesis*, *18*(1), 1-19. https://doi.org/10.1080/15476278.2021.2022373

Cho, S. H., Seymen, F., Lee, K. E., Lee, S. K., Kweon, Y. S., Kim, K. J., Jung, S. E., Song, S. J., Yildirim, M., Bayram, M., Tuna, E. B., Gencay, K., & Kim, J. W. (2012). Novel FAM20A mutations in hypoplastic amelogenesis imperfecta. *Hum Mutat*, *33*(1), 91-94. https://doi.org/10.1002/humu.21621

Chosack, A., Eidelman, E., Wisotski, I., & Cohen, T. (1979). Amelogenesis imperfecta among Israeli Jews and the description of a new type of local hypoplastic autosomal recessive amelogenesis imperfecta. *Oral Surg Oral Med Oral Pathol*, *47*(2), 148-156. https://doi.org/10.1016/0030-4220(79)90170-1

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(4), 265-270. https://doi.org/10.1038/nnano.2009.12

Collignon, A.-M., Vergnes, J.-N., Germa, A., Azogui, S., Breinig, S., Hollande, C., Bonnet, A.-L., & Nabet, C. (2022). Factors and Mechanisms Involved in Acquired Developmental Defects of Enamel: A Scoping Review [Review]. *Frontiers in Pediatrics*, *10*. https://doi.org/10.3389/fped.2022.836708

Colognato, H., & Yurchenco, P. D. (2000). Form and function: the laminin family of heterotrimers. *Developmental dynamics: an official publication of the American Association of Anatomists*, *218*(2), 213-234.

Crawford, P. J., Aldred, M., & Bloch-Zupan, A. (2007). Amelogenesis imperfecta. *Orphanet J Rare Dis*, *2*, 17. https://doi.org/10.1186/1750-1172-2-17

Crawford, P. J., Evans, R. D., & Aldred, M. J. (1988). Amelogenesis imperfecta: autosomal dominant hypomaturation-hypoplasia type with taurodontism. *British Dental Journal*, *164*(3), 71-73. https://doi.org/10.1038/sj.bdj.4806360

Crick, F. H., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, *192*, 1227-1232. https://doi.org/10.1038/1921227a0

Daniels, R. S., Harvey, R., Ermetal, B., Xiang, Z., Galiano, M., Adams, L., & McCauley, J. W. (2021). A Sanger sequencing protocol for SARS-CoV-2 S-gene. *Influenza Other Respir Viruses*, *15*(6), 707-710. https://doi.org/10.1111/irv.12892

Darling, A. I. (1959). The structure of the human tooth. *Proc Nutr Soc*, *18*(1), 70-75. https://doi.org/10.1079/pns19590018

de la Dure-Molla, M., Quentric, M., Yamaguti, P. M., Acevedo, A. C., Mighell, A. J., Vikkula, M., Huckert, M., Berdal, A., & Bloch-Zupan, A. (2014). Pathognomonic oral profile of Enamel Renal Syndrome (ERS) caused by recessive FAM20A mutations. *Orphanet J Rare Dis*, *9*, 84. https://doi.org/10.1186/1750-1172-9-84

Diaz, L. A., Ratrie, H., 3rd, Saunders, W. S., Futamura, S., Squiquera, H. L., Anhalt, G. J., & Giudice, G. J. (1990). Isolation of a human epidermal cDNA corresponding to the 180-kD autoantigen recognized by bullous pemphigoid and herpes gestationis sera. Immunolocalization of this protein to the hemidesmosome. *J Clin Invest*, *86*(4), 1088-1094. https://doi.org/10.1172/jci114812

Diekwisch, T. G., Ware, J., Fincham, A. G., & Zeichner-David, M. (1997). Immunohistochemical similarities and differences between amelogenin and tuftelin gene products during tooth development. *J Histochem Cytochem*, *45*(6), 859-866. https://doi.org/10.1177/002215549704500610

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., . . . Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, *323*(5910), 133-138. https://doi.org/doi:10.1126/science.1162986

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., & Mighell, A. J. (2009). Mutations in the beta propeller WDR72 cause autosomal-recessive hypomaturation amelogenesis imperfecta. *Am J Hum Genet*, *85*(5), 699-705. https://doi.org/10.1016/j.ajhg.2009.09.014

Elzein, R., Abdel-Sater, F., Mehawej, C., Jalkh, N., Ayoub, F., & Chouery, E. (2022). Identification by whole-exome sequencing of new single-nucleotide polymorphisms associated with molar-incisor hypomineralisation among the Lebanese population. *Eur Arch Paediatr Dent*, *23*(6), 919-928. https://doi.org/10.1007/s40368-022-00738-2

Fraser, D., & Nikiforuk, G. (1982). The etiology of enamel hypoplasia in children--a unifying concept. *J Int Assoc Dent Child*, *13*(1), 1-11.

Gadhia, K., McDonald, S., Arkutu, N., & Malik, K. (2012). Amelogenesis imperfecta: an introduction. *Br Dent J*, *212*(8), 377-379. https://doi.org/10.1038/sj.bdj.2012.314

Garot, E., Rouas, P., Somani, C., Taylor, G. D., Wong, F., & Lygidakis, N. A. (2022). An update of the aetiological factors involved in molar incisor hypomineralisation (MIH): a systematic review and meta-analysis. *European Archives of Paediatric Dentistry*, *23*(1), 23-38. https://doi.org/10.1007/s40368-021-00646-x

Ghanim, A., Manton, D., Bailey, D., Mariño, R., & Morgan, M. (2013). Risk factors in the occurrence of molar-incisor hypomineralization amongst a group of Iraqi children. *Int J Paediatr Dent*, *23*(3), 197-206. https://doi.org/10.1111/j.1365-263X.2012.01244.x

Hardenbol, P., Banér, J., Jain, M., Nilsson, M., Namsaraev, E. A., Karlin-Neumann, G. A., Fakhrai-Rad, H., Ronaghi, M., Willis, T. D., Landegren, U., & Davis, R. W. (2003). Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol*, *21*(6), 673-678. https://doi.org/10.1038/nbt821

Hardies, K., de Kovel, C. G. F., Weckhuysen, S., Asselbergh, B., Geuens, T., Deconinck, T., Azmi, A., May, P., Brilstra, E., Becker, F., Barisic, N., Craiu, D., Braun, K. P. J., Lal, D., Thiele, H., Schubert, J., Weber, Y., van 't Slot, R., Nürnberg, P., . . . Consortium, o. b. o. t. a. r. w. g. o. t. E. R. (2015). Recessive mutations in SLC13A5 result in a loss of citrate transport and cause neonatal epilepsy, developmental delay and teeth hypoplasia. *Brain*, *138*(11), 3238-3250. https://doi.org/10.1093/brain/awv263

Hart, P. S., Michalec, M. D., Seow, W. K., Hart, T. C., & Wright, J. T. (2003). Identification of the enamelin (g.8344delG) mutation in a new kindred and presentation of a standardized ENAM nomenclature. *Arch Oral Biol*, *48*(8), 589-596. https://doi.org/10.1016/s0003-9969(03)00114-6

Heimler, A., Fox, J. E., Hershey, J. E., & Crespi, P. (1991). Sensorineural hearing loss, enamel hypoplasia, and nail abnormalities in sibs. *Am J Med Genet*, *39*(2), 192-195. https://doi.org/10.1002/ajmg.1320390214

Hershey, A. D., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol*, *36*(1), 39-56. https://doi.org/10.1085/jgp.36.1.39

Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J., & Shendure, J. (2013). Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res*, *23*(5), 843-854. https://doi.org/10.1101/gr.147686.112

Hildebrand, C., Fried, K., Tuisku, F., & Johansson, C. S. (1995). Teeth and tooth nerves. *Prog Neurobiol*, *45*(3), 165-222. https://doi.org/10.1016/0301-0082(94)00045-j

Hou, J., Renigunta, A., Konrad, M., Gomes, A. S., Schneeberger, E. E., Paul, D. L., Waldegger, S., & Goodenough, D. A. (2008). Claudin-16 and claudin-19 interact and form a cation-selective tight junction complex. *J Clin Invest*, *118*(2), 619-628. https://doi.org/10.1172/jci33970

Iwasaki, K., Bajenova, E., Somogyi-Ganss, E., Miller, M., Nguyen, V., Nourkeyhani, H., Gao, Y., Wendel, M., & Ganss, B. (2005). Amelotin--a Novel Secreted, Ameloblast-specific Protein. *J Dent Res*, *84*(12), 1127-1132. https://doi.org/10.1177/154405910508401207

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., . . . Kleta, R. (2012). Nephrocalcinosis (enamel renal syndrome) caused by autosomal recessive FAM20A mutations. *Nephron Physiol*, *122*(1-2), 1-6. https://doi.org/10.1159/000349989

Jedeon, K., De la Dure-Molla, M., Brookes, S. J., Loiodice, S., Marciano, C., Kirkham, J., Canivenc-Lavier, M. C., Boudalia, S., Bergès, R., Harada, H., Berdal, A., & Babajko, S. (2013). Enamel defects reflect perinatal exposure to bisphenol A. *Am J Pathol*, *183*(1), 108-118. https://doi.org/10.1016/j.ajpath.2013.04.004

Jeremias, F., Pierri, R. A. G., Souza, J. F., Fragelli, C. M. B., Restrepo, M., Finoti, L. S., Bussaneli, D. G., Cordeiro, R. C. L., Secolin, R., Maurer-Morelli, C. V., Scarel-Caminaga, R. M., & Santos-Pinto, L. (2016). Family-Based Genetic Association for Molar-Incisor Hypomineralization. *Caries Research*, *50*(3), 310-318. https://doi.org/10.1159/000445726

Jernvall, J., & Thesleff, I. (2012). Tooth shape formation and tooth renewal: evolving with the same signals. *Development*, *139*(19), 3487-3497. https://doi.org/10.1242/dev.085084

Ji, Y., Li, C., Tian, Y., Gao, Y., Dong, Z., Xiang, L., Xu, Z., Gao, Y., & Zhang, L. (2021). Maturation stage enamel defects in Odontogenesis-associated phosphoprotein (Odaph) deficient mice. *Developmental Dynamics*, *250*(10), 1505-1517. https://doi.org/https://doi.org/10.1002/dvdy.336

Jiang, W., Chu, X., Wang, B., Pan, H., Xu, X., & Tang, R. (2009). Biomimetically triggered inorganic crystal transformation by biomolecules: a new understanding of biomineralization. *J Phys Chem B*, *113*(31), 10838-10844. https://doi.org/10.1021/jp904633f

Josephsen, K., Takano, Y., Frische, S., Praetorius, J., Nielsen, S., Aoba, T., & Fejerskov, O. (2010). Ion transporters in secretory and cyclically modulating ameloblasts: a new hypothesis for cellular control of preeruptive enamel maturation. *Am J Physiol Cell Physiol*, *299*(6), C1299-1307. https://doi.org/10.1152/ajpcell.00218.2010

Katsura, K., Nakano, Y., Zhang, Y., Shemirani, R., Li, W., & Den Besten, P. (2022). WDR72 regulates vesicle trafficking in ameloblasts. *Scientific Reports*, *12*(1), 2820. https://doi.org/10.1038/s41598-022-06751-1

Kiani, A. K., Pheby, D., Henehan, G., Brown, R., Sieving, P., Sykora, P., Marks, R., Falsini, B., Capodicasa, N., Miertus, S., Lorusso, L., Dondossola, D., Tartaglia, G. M., Ergoren, M. C., Dundar, M., Michelini, S., Malacarne, D., Bonetti, G., Dautaj, A., . . . Bertelli, M. (2022). Ethical considerations regarding animal experimentation. *J Prev Med Hyg*, *63*(2 Suppl 3), E255-e266. https://doi.org/10.15167/2421-4248/jpmh2022.63.2S3.2768

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., & Simmer, J. P. (2008). FAM83H mutations in families with autosomal-dominant hypocalcified amelogenesis imperfecta. *The American Journal of Human Genetics*, *82*(2), 489-494.

Kim, J.-W., Simmer, J. P., Hart, T. C., Hart, P. S., Ramaswami, M. D., Bartlett, J. D., & Hu, J. C.-C. (2005). MMP-20 mutation in autosomal recessive pigmented hypomaturation amelogenesis imperfecta. *Journal of Medical Genetics*, *42*(3), 271-275. https://doi.org/10.1136/jmg.2004.024505

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., & Hu, J. C. (2019). Mutations in RELT cause autosomal recessive amelogenesis imperfecta. *Clin Genet*, *95*(3), 375-383. https://doi.org/10.1111/cge.13487

Kim, K. H., Kim, E. J., Kim, H. Y., Li, S., & Jung, H. S. (2023). Fabrication of functional ameloblasts from hiPSCs for dental application. *Front Cell Dev Biol*, *11*, 1164811. https://doi.org/10.3389/fcell.2023.1164811

Konrad, M., Schaller, A., Seelow, D., Pandey, A. V., Waldegger, S., Lesslauer, A., Vitzthum, H., Suzuki, Y., Luk, J. M., Becker, C., Schlingmann, K. P., Schmid, M., Rodriguez-Soriano, J., Ariceta, G., Cano, F., Enriquez, R., Juppner, H., Bakkaloglu, S. A., Hediger, M. A., . . . Weber, S. (2006). Mutations in the tight-junction gene claudin 19 (CLDN19) are associated with renal magnesium wasting, renal failure, and severe ocular involvement. *Am J Hum Genet*, *79*(5), 949-957. https://doi.org/10.1086/508617

Kubota, K., Lee, D. H., Tsuchiya, M., Young, C. S., Everett, E. T., Martinez-Mier, E. A., Snead, M. L., Nguyen, L., Urano, F., & Bartlett, J. D. (2005). Fluoride Induces Endoplasmic Reticulum Stress in Ameloblasts Responsible for Dental Enamel Formation *. *Journal of Biological Chemistry*, *280*(24), 23194-23202. https://doi.org/10.1074/jbc.M503288200

Kuechler, A., Hentschel, J., Kurth, I., Stephan, B., Prott, E. C., Schweiger, B., Schuster, A., Wieczorek, D., & Lüdecke, H. J. (2012). A Novel Homozygous WDR72 Mutation in Two Siblings with Amelogenesis Imperfecta and Mild Short Stature. *Mol Syndromol*, *3*(5), 223-229. https://doi.org/10.1159/000343746

Lacruz, R., Smith, C., Kurtz, I., Hubbard, M., & Paine, M. (2013). New paradigms on the transport functions of maturation-stage ameloblasts. *Journal of Dental Research*, *92*(2), 122-129.

Lacruz, R. S. (2017). Enamel: Molecular identity of its transepithelial ion transport system. *Cell Calcium*, *65*, 1-7. https://doi.org/https://doi.org/10.1016/j.ceca.2017.03.006

Lacruz, R. S., Habelitz, S., Wright, J. T., & Paine, M. L. (2017). DENTAL ENAMEL FORMATION AND IMPLICATIONS FOR ORAL HEALTH AND DISEASE. *Physiol Rev*, *97*(3), 939-993. https://doi.org/10.1152/physrev.00030.2016

Lacruz, R. S., Smith, C. E., Bringas Jr, P., Chen, Y.-B., Smith, S. M., Snead, M. L., Kurtz, I., Hacia, J. G., Hubbard, M. J., & Paine, M. L. (2012). Identification of novel candidate genes involved in mineralization of dental enamel by genome-wide transcript profiling. *Journal of Cellular Physiology*, *227*(5), 2264-2275. https://doi.org/https://doi.org/10.1002/jcp.22965

Lagerström-Fermér, M., Nilsson, M., Bäckman, B., Salido, E., Shapiro, L., Pettersson, U., & Landegren, U. (1995). Amelogenin signal peptide mutation: correlation between mutations in the amelogenin gene (AMGX) and manifestations of X-linked amelogenesis imperfecta. *Genomics*, *26*(1), 159-162. https://doi.org/10.1016/0888-7543(95)80097-6

Lagerström, M., Dahl, N., Iselius, L., Bäckman, B., & Pettersson, U. (1990). Mapping of the gene for X-linked amelogenesis imperfecta by linkage analysis. *Am J Hum Genet*, *46*(1), 120-125.

Lagerström, M., Dahl, N., Nakahori, Y., Nakagome, Y., Bäckman, B., Landegren, U., & Pettersson, U. (1991). A deletion in the amelogenin gene (AMG) causes X-linked amelogenesis imperfecta (AIH1). *Genomics*, *10*(4), 971-975. https://doi.org/10.1016/0888-7543(91)90187-j

Lakhani, R. (2021). Amelogenesis imperfecta: the inside story. *British Dental Journal*, *231*(9), 564-565. https://doi.org/10.1038/s41415-021-3614-7

Law, K. B., Bronte-Tinkew, D., Di Pietro, E., Snowden, A., Jones, R. O., Moser, A., Brumell, J. H., Braverman, N., & Kim, P. K. (2017). The peroxisomal AAA ATPase complex prevents pexophagy and development of peroxisome biogenesis disorders. *Autophagy*, *13*(5), 868-884. https://doi.org/10.1080/15548627.2017.1291470

Lee, S. K., Hu, J. C., Bartlett, J. D., Lee, K. E., Lin, B. P., Simmer, J. P., & Kim, J. W. (2008). Mutational spectrum of FAM83H: the C-terminal portion is required for tooth enamel calcification. *Hum Mutat*, *29*(8), E95-99. https://doi.org/10.1002/humu.20789

Li, X. F., Kraev, A. S., & Lytton, J. (2002). Molecular cloning of a fourth member of the potassium-dependent sodium-calcium exchanger gene family, NCKX4. *J Biol Chem*, *277*(50), 48410-48417. https://doi.org/10.1074/jbc.M210011200

Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., Albrechtsen, A., Andersen, G., Cao, H., Korneliussen, T., Grarup, N., Guo, Y., Hellman, I., Jin, X., Li, Q., Liu, J., Liu, X., Sparsø, T., Tang, M., . . . Wang, J. (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet*, *42*(11), 969-972. https://doi.org/10.1038/ng.680

Liang, T., Hu, Y., Smith, C. E., Richardson, A. S., Zhang, H., Yang, J., Lin, B., Wang, S. K., Kim, J. W., Chun, Y. H., Simmer, J. P., & Hu, J. C. (2019). AMBN mutations causing hypoplastic amelogenesis imperfecta and Ambn knockout-NLS-lacZ knockin mice exhibiting failed amelogenesis and Ambn tissue-specificity. *Mol Genet Genomic Med*, *7*(9), e929. https://doi.org/10.1002/mgg3.929

Lima, L. H., Barbazetto, I. A., Chen, R., Yannuzzi, L. A., Tsang, S. H., & Spaide, R. F. (2011). Macular dystrophy in Heimler syndrome. *Ophthalmic Genet*, *32*(2), 97-100. https://doi.org/10.3109/13816810.2010.551797

Lin, J. H., Chen, L. C., Yu, S. C., & Huang, Y. T. (2022). LongPhase: an ultra-fast chromosome-scale phasing algorithm for small and large variants. *Bioinformatics*, *38*(7), 1816-1822. https://doi.org/10.1093/bioinformatics/btac058

Linde, A., & Goldberg, M. (1993). Dentinogenesis. *Crit Rev Oral Biol Med*, *4*(5), 679-728. https://doi.org/10.1177/10454411930040050301

Linnarsson, S. (2010). Recent advances in DNA sequencing methods - general principles of sample preparation. *Exp Cell Res*, *316*(8), 1339-1343. https://doi.org/10.1016/j.yexcr.2010.02.036

Liu, H., Yan, X., Pandya, M., Luan, X., & Diekwisch, T. G. H. (2016). Daughters of the Enamel Organ: Development, Fate, and Function of the Stratum Intermedium, Stellate Reticulum, and Outer Enamel Epithelium. *Stem Cells and Development*, *25*(20), 1580-1590. https://doi.org/10.1089/scd.2016.0267

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nat Rev Genet*, *21*(10), 597-614. https://doi.org/10.1038/s41576-020-0236-x

Loman, N. J., & Watson, M. (2015). Successful test launch for nanopore sequencing. *Nat Methods*, *12*(4), 303-304. https://doi.org/10.1038/nmeth.3327

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics*, *14*(5), 265-279. https://doi.org/10.1016/j.gpb.2016.05.004

Lu, T., Li, M., Xu, X., Xiong, J., Huang, C., Zhang, X., Hu, A., Peng, L., Cai, D., Zhang, L., Wu, B., & Xiong, F. (2018). Whole exome sequencing identifies an AMBN missense mutation causing severe autosomal-dominant amelogenesis imperfecta and dentin disorders. *Int J Oral Sci*, *10*(3), 26. https://doi.org/10.1038/s41368-018-0027-9

Lyne, A., Parekh, S., Patel, N., Lafferty, F., Brown, C., Rodd, H., & Monteiro, J. (2021). Patient-reported outcome measure for children and young people with amelogenesis imperfecta. *Br Dent J*, 1-6. https://doi.org/10.1038/s41415-021-3329-9

MacGibbon, D. (1972). Generalized enamel hypoplasia and renal dysfunction. *Aust Dent J*, *17*(1), 61-63. https://doi.org/10.1111/j.1834-7819.1972.tb02747.x

Markovich, D., & Murer, H. (2004). The SLC13 gene family of sodium sulphate/carboxylate cotransporters. *Pflügers Archiv*, *447*(5), 594-602. https://doi.org/10.1007/s00424-003-1128-6

Martelli-Júnior, H., Bonan, P. R., Dos Santos, L. A., Santos, S. M., Cavalcanti, M. G., & Coletta, R. D. (2008). Case reports of a new syndrome associating gingival fibromatosis and dental abnormalities in a consanguineous family. *J Periodontol*, *79*(7), 1287-1296. https://doi.org/10.1902/jop.2008.070520

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, *74*(2), 560-564. https://doi.org/10.1073/pnas.74.2.560

McInerney-Leo, A. M., Le Goff, C., Leo, P. J., Kenna, T. J., Keith, P., Harris, J. E., Steer, R., Bole-Feysot, C., Nitschke, P., Kielty, C., Brown, M. A., Zankl, A., Duncan, E. L., & Cormier-Daire, V. (2016). Mutations in LTBP3 cause acromicric dysplasia and geleophysic dysplasia. *J Med Genet*, *53*(7), 457-464. https://doi.org/10.1136/jmedgenet-2015-103647

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, *11*(1), 31-46. https://doi.org/10.1038/nrg2626

Moffatt, P., Smith, C. E., St-Arnaud, R., Simmons, D., Wright, J. T., & Nanci, A. (2006). Cloning of rat amelotin and localization of the protein to the basal lamina of maturation stage ameloblasts and junctional epithelium. *Biochem J*, *399*(1), 37-46. https://doi.org/10.1042/bj20060662

Morgan, T. H. (1910). Sex Limited Inheritance in Drosophila. *Science*, *32*(812), 120-122. https://doi.org/doi:10.1126/science.32.812.120

Mory, A., Dagan, E., Illi, B., Duquesnoy, P., Mordechai, S., Shahor, I., Romani, S., Hawash-Moustafa, N., Mandel, H., Valente, E. M., Amselem, S., & Gershoni-Baruch, R. (2012). A nonsense mutation in the human homolog of Drosophila rogdi causes Kohlschutter-Tonz syndrome. *Am J Hum Genet*, *90*(4), 708-714. https://doi.org/10.1016/j.ajhg.2012.03.005

Mukherjee, K., Ruan, Q., Nutt, S., Tao, J., De Yoreo, J. J., & Moradian-Oldak, J. (2018). Peptide-Based Bioinspired Approach to Regrowing Multilayered Aprismatic Enamel. *ACS Omega*, *3*(3), 2546-2557. https://doi.org/10.1021/acsomega.7b02004

Nakano, Y., Le, M. H., Abduweli, D., Ho, S. P., Ryazanova, L. V., Hu, Z., Ryazanov, A. G., Den Besten, P. K., & Zhang, Y. (2016). A critical role of TRPM7 as an ion channel protein in mediating the mineralization of the craniofacial hard tissues. *Frontiers in physiology*, *7*, 258.

Nanci, A. (2008). Enamel: composition, formation, and structure. *Ten Cate's oral histology: development, structure, and function*, 183-184.

Nguyen, T., Phillips, C., Frazier-Bower, S., & Wright, T. (2013). Craniofacial variations in the tricho-dento-osseous syndrome. *Clin Genet*, *83*(4), 375-379. https://doi.org/10.1111/j.1399-0004.2012.01907.x

Nurbaeva, M. K., Eckstein, M., Concepcion, A. R., Smith, C. E., Srikanth, S., Paine, M. L., Gwack, Y., Hubbard, M. J., Feske, S., & Lacruz, R. S. (2015). Dental enamel cells express functional SOCE channels. *Scientific Reports*, *5*(1), 15803.

Nurbaeva, M. K., Eckstein, M., Snead, M. L., Feske, S., & Lacruz, R. S. (2015). Store-operated Ca2+ Entry Modulates the Expression of Enamel Genes. *J Dent Res*, *94*(10), 1471-1477. https://doi.org/10.1177/0022034515598144

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*, *15*(2), 256-278. https://doi.org/10.1093/bib/bbs086

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, *15*(2), 256-278.

Park, J. C., Park, J. T., Son, H. H., Kim, H. J., Jeong, M. J., Lee, C. S., Dey, R., & Cho, M. I. (2007). The amyloid protein APin is highly expressed during enamel mineralization and maturation in rat incisors. *Eur J Oral Sci*, *115*(2), 153-160. https://doi.org/10.1111/j.1600-0722.2007.00435.x

Parry, D. A., Brookes, S. J., Logan, C. V., Poulter, J. A., El-Sayed, W., Al-Bahlani, S., Al Harasi, S., Sayed, J., Raïf el, 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., . . . Mighell, A. J. (2012). Mutations in C4orf26, encoding a peptide with in vitro hydroxyapatite crystal nucleation and growth activity, cause amelogenesis imperfecta. *Am J Hum Genet*, *91*(3), 565-571. https://doi.org/10.1016/j.ajhg.2012.07.020

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., & Mighell, A. J. (2016). A homozygous STIM1 mutation impairs store-operated calcium entry and natural killer cell effector function without clinical immunodeficiency. *J Allergy Clin Immunol*, *137*(3), 955-957.e958. https://doi.org/10.1016/j.jaci.2015.08.051

Parry, D. A., Mighell, A. J., El-Sayed, W., Shore, R. C., Jalili, I. K., Dollfus, H., Bloch-Zupan, A., Carlos, R., Carr, I. M., Downey, L. M., Blain, K. M., Mansfield, D. C., Shahrabi, M., Heidari, M., Aref, P., Abbasi, M., Michaelides, M., Moore, A. T., Kirkham, J., & Inglehearn, C. F. (2009). Mutations in CNNM4 cause Jalili syndrome, consisting of autosomal-recessive cone-rod dystrophy and amelogenesis imperfecta. *Am J Hum Genet*, *84*(2), 266-273. https://doi.org/10.1016/j.ajhg.2009.01.009

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., & Mighell, A. J. (2013). Identification of mutations in SLC24A4, encoding a potassium-dependent sodium/calcium exchanger, as a cause of amelogenesis imperfecta. *Am J Hum Genet*, *92*(2), 307-312. https://doi.org/10.1016/j.ajhg.2013.01.003

Parry, D. A., Smith, C. E., 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., Simmer, J. P., Ahmed, M., Jafri, H., Johnson, C. A., Inglehearn, C. F., & Mighell, A. J. (2016). Mutations in the pH-Sensing G-protein-Coupled Receptor GPR68 Cause Amelogenesis Imperfecta. *Am J Hum Genet*, *99*(4), 984-990. https://doi.org/10.1016/j.ajhg.2016.08.020

Pashley, D. H. (1989). Dentin: a dynamic substrate--a review. *Scanning Microsc*, *3*(1), 161-174; discussion 174-166.

Patel, P. R., & Brown, W. E. (1975). Thermodynamic Solubility Product of Human Tooth Enamel: Powdered Sample. *Journal of Dental Research*, *54*(4), 728-736. https://doi.org/10.1177/00220345750540040601

Pitiphat, W., Luangchaichaweng, S., Pungchanchaikul, P., Angwaravong, O., & Chansamak, N. (2014). Factors associated with molar incisor hypomineralization in Thai children. *Eur J Oral Sci*, *122*(4), 265-270. https://doi.org/10.1111/eos.12136

Poulsen, S., Gjørup, H., Haubek, D., Haukali, G., Hintze, H., Løvschall, H., & Errboe, M. (2008). Amelogenesis imperfecta - a systematic literature review of associated dental and oro-facial abnormalities and their impact on patients. *Acta Odontol Scand*, *66*(4), 193-199. https://doi.org/10.1080/00016350802192071

Poulter, J. A., Brookes, S. J., Shore, R. C., Smith, C. E., Abi Farraj, L., Kirkham, J., Inglehearn, C. F., & Mighell, A. J. (2014). A missense mutation in ITGB6 causes pitted hypomineralized amelogenesis imperfecta. *Hum Mol Genet*, *23*(8), 2189-2197. https://doi.org/10.1093/hmg/ddt616

Poulter, J. A., El-Sayed, W., Shore, R. C., Kirkham, J., Inglehearn, C. F., & Mighell, A. J. (2014). Whole-exome sequencing, without prior linkage, identifies a mutation in LAMB3 as a cause of dominant hypoplastic amelogenesis imperfecta. *Eur J Hum Genet*, *22*(1), 132-135. https://doi.org/10.1038/ejhg.2013.76

Poulter, J. A., Murillo, G., Brookes, S. J., Smith, C. E., Parry, D. A., Silva, S., Kirkham, J., Inglehearn, C. F., & Mighell, A. J. (2014). Deletion of ameloblastin exon 6 is associated with amelogenesis imperfecta. *Hum Mol Genet*, *23*(20), 5317-5324. https://doi.org/10.1093/hmg/ddu247

Pousette Lundgren, G., Wickström, A., Hasselblad, T., & Dahllöf, G. (2016). Amelogenesis Imperfecta and Early Restorative Crown Therapy: An Interview Study with Adolescents and Young Adults on Their Experiences. *PLOS ONE*, *11*(6), e0156879. https://doi.org/10.1371/journal.pone.0156879

Prakriya, M., & Lewis, R. S. (2015). Store-operated calcium channels. *Physiological reviews*.

Pugach, M. K., & Gibson, C. W. (2014). Analysis of enamel development using murine model systems: approaches and limitations. *Front Physiol*, *5*, 313. https://doi.org/10.3389/fphys.2014.00313

Purvis, R. J., Barrie, W. J., MacKay, G. S., Wilkinson, E. M., Cockburn, F., & Belton, N. R. (1973). Enamel hypoplasia of the teeth associated with neonatal tetany: a manifestation of maternal vitamin-D deficiency. *Lancet*, *2*(7833), 811-814. https://doi.org/10.1016/s0140-6736(73)90857-x

Qiao, L., Liu, X., He, Y., Zhang, J., Huang, H., Bian, W., Chilufya, M. M., Zhao, Y., & Han, J. (2021). Progress of Signaling Pathways, Stress Pathways and Epigenetics in the Pathogenesis of Skeletal Fluorosis. *International Journal of Molecular Sciences*, *22*(21).

Rajpar, M. H., Harley, K., Laing, C., Davies, R. M., & Dixon, M. J. (2001). Mutation of the gene encoding the enamel-specific protein, enamelin, causes autosomal-dominant amelogenesis imperfecta. *Hum Mol Genet*, *10*(16), 1673-1677. https://doi.org/10.1093/hmg/10.16.1673

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., . . . Van Camp, G. (2015). Heimler Syndrome Is Caused by Hypomorphic Mutations in the Peroxisome-Biogenesis Genes PEX1 and PEX6. *Am J Hum Genet*, *97*(4), 535-545. https://doi.org/10.1016/j.ajhg.2015.08.011

Rauth, R. J., Potter, K. S., Ngan, A. Y., Saad, D. M., Mehr, R., Luong, V. Q., Schuetter, V. L., Miklus, V. G., Chang, P., Paine, M. L., Lacruz, R. S., Snead, M. L., & White, S. N. (2009). Dental enamel: genes define biomechanics. *J Calif Dent Assoc*, *37*(12), 863-868.

Riemann, D., Wallrafen, R., & Dresbach, T. (2017). The Kohlschütter-Tönz syndrome associated gene Rogdi encodes a novel presynaptic protein. *Scientific Reports*, *7*(1), 15791. https://doi.org/10.1038/s41598-017-16004-1

Robinson, J. T., Thorvaldsdottir, H., Turner, D., & Mesirov, J. P. (2022). igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics*, *39*(1). https://doi.org/10.1093/bioinformatics/btac830

Rugg-Gunn, A. J., Al-Mohammadi, S. M., & Butler, T. J. (1998). Malnutrition and developmental defects of enamel in 2- to 6-year-old Saudi boys. *Caries Res*, *32*(3), 181-192. https://doi.org/10.1159/000016451

Rungroj, N., Nettuwakul, C., Sawasdee, N., Sangnual, S., Deejai, N., Misgar, R. A., Pasena, A., Khositseth, S., Kirdpon, S., Sritippayawan, S., Vasuvattakul, S., & Yenchitsomanus, P. (2018). Distal renal tubular acidosis caused by tryptophan-aspartate repeat domain 72 (WDR72) mutations. *Clinical Genetics*, *94*(5), 409-418. https://doi.org/https://doi.org/10.1111/cge.13418

Salido, E. C., Yen, P. H., Koprivnikar, K., Yu, L. C., & Shapiro, L. J. (1992). The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes. *Am J Hum Genet*, *50*(2), 303-316.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977a). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463-5467. https://doi.org/doi:10.1073/pnas.74.12.5463

Sanger, F., Nicklen, S., & Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, *74*(12), 5463-5467. https://doi.org/10.1073/pnas.74.12.5463

Sato, K., Mogi, C., Mighell, A. J., & Okajima, F. (2020). A missense mutation of Leu74Pro of OGR1 found in familial amelogenesis imperfecta actually causes the loss of the pH-sensing

mechanism. *Biochem Biophys Res Commun*, *526*(4), 920-926. https://doi.org/10.1016/j.bbrc.2020.04.005

Schossig, A., Bloch-Zupan, A., Lussi, A., Wolf, N. I., Raskin, S., Cohen, M., Giuliano, F., Jurgens, J., Krabichler, B., Koolen, D. A., de Macena Sobreira, N. L., Maurer, E., Muller-Bolla, M., Penzien, J., Zschocke, J., & Kapferer-Seebacher, I. (2017). SLC13A5 is the second gene associated with Kohlschütter-Tönz syndrome. *J Med Genet*, *54*(1), 54-62. https://doi.org/10.1136/jmedgenet-2016-103988

Schour, I. (1948). Development and growth of teeth. *Oral Surg Oral Med Oral Pathol*, *1*(4), 346-354. https://doi.org/10.1016/0030-4220(48)90258-8

Sedano, H. O. (1975). Congenital oral anomalies in argentinian children. *Community Dent Oral Epidemiol*, *3*(2), 61-63. https://doi.org/10.1111/j.1600-0528.1975.tb00281.x

Seymen, F., Kim, Y. J., 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, Y. J., Lee, S. H., Lee, Z. H., Zhang, H., Hu, J. C., Simmer, J. P., Cho, E. S., & Kim, J. W. (2016). Recessive Mutations in ACPT, Encoding Testicular Acid Phosphatase, Cause Hypoplastic Amelogenesis Imperfecta. *Am J Hum Genet*, *99*(5), 1199-1205. https://doi.org/10.1016/j.ajhg.2016.09.018

Silva, M., Arnaud, M. A., Lyra, M. C. A., Alencar Filho, A. V., Rocha MÂ, W., Ramos, R. C. F., Van Der Linden, V., Caldas, A. F. J., Heimer, M. V., & Rosenblatt, A. (2020). Dental development in children born to Zikv-infected mothers: a case-based study. *Arch Oral Biol*, *110*, 104598. https://doi.org/10.1016/j.archoralbio.2019.104598

Smith, C. E. (1998). Cellular and chemical events during enamel maturation. *Crit Rev Oral Biol Med*, *9*(2), 128-161. https://doi.org/10.1177/10454411980090020101

Smith, C. E. (1998). Cellular and Chemical Events During Enamel Maturation. *Critical Reviews in Oral Biology & Medicine*, *9*(2), 128-161. https://doi.org/10.1177/10454411980090020101

Smith, C. E., Chong, D. L., Bartlett, J. D., & Margolis, H. C. (2005). Mineral acquisition rates in developing enamel on maxillary and mandibular incisors of rats and mice: implications to extracellular acid loading as apatite crystals mature. *Journal of bone and mineral research*, *20*(2), 240-249.

Smith, C. E., Poulter, J. A., Levin, A. V., Capasso, J. E., Price, S., Ben-Yosef, T., Sharony, R., Newman, W. G., Shore, R. C., Brookes, S. J., Mighell, A. J., & Inglehearn, C. F. (2016). Spectrum of PEX1 and PEX6 variants in Heimler syndrome. *Eur J Hum Genet*, *24*(11), 1565-1571. https://doi.org/10.1038/ejhg.2016.62

Smith, C. E. L., Murillo, G., Brookes, S. J., Poulter, J. A., Silva, S., Kirkham, J., Inglehearn, C. F., & Mighell, A. J. (2016). Deletion of amelotin exons 3–6 is associated with amelogenesis imperfecta. *Human Molecular Genetics*, *25*(16), 3578-3587. https://doi.org/10.1093/hmg/ddw203

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., & Mighell, A. J. (2020). A missense variant in specificity protein 6 (SP6) is associated with amelogenesis imperfecta. *Human Molecular Genetics*, *29*(9), 1417-1425. https://doi.org/10.1093/hmg/ddaa041

Sneller, J., Buchanan, H., & Parekh, S. (2014). The impact of amelogenesis imperfecta and support needs of adolescents with AI and their parents: an exploratory study. *Int J Paediatr Dent*, *24*(6), 409-416. https://doi.org/10.1111/ipd.12086

Snyder, M. W., Adey, A., Kitzman, J. O., & Shendure, J. (2015). Haplotype-resolved genome sequencing: experimental methods and applications. *Nat Rev Genet*, *16*(6), 344-358. https://doi.org/10.1038/nrg3903

Spokes, S. (1890). Calcification. *Am J Dent Sci*, *24*(4), 155-162.

Strauch, S., & Hahnel, S. (2018). Restorative Treatment in Patients with Amelogenesis Imperfecta: A Review. *J Prosthodont*, *27*(7), 618-623. https://doi.org/10.1111/jopr.12736

Sundell, S., & Koch, G. (1985). Hereditary amelogenesis imperfecta. I. Epidemiology and clinical classification in a Swedish child population. *Swed Dent J*, *9*(4), 157-169.

Swaters, D., van Veen, A., van Meurs, W., Turner, J. E., & Ritskes-Hoitinga, M. (2022). A History of Regulatory Animal Testing: What Can We Learn? *Alternatives to Laboratory Animals*, *50*(5), 322-329. https://doi.org/10.1177/02611929221118001

Taft, J. (1881). Tooth Structure. *Dent Regist*, *35*(3), 89-93.

Tantibhaedhyangkul, W., Tantrapornpong, J., Yutchawit, N., Theerapanon, T., Intarak, N., Thaweesapphithak, S., Porntaveetus, T., & Shotelersuk, V. (2023). Dental characteristics of patients with four different types of skeletal dysplasias. *Clinical Oral Investigations*. https://doi.org/10.1007/s00784-023-05194-w

Ten Cate, A. R. (1996). The role of epithelium in the development, structure and function of the tissues of tooth support. *Oral Dis*, *2*(1), 55-62. https://doi.org/10.1111/j.1601-0825.1996.tb00204.x

Thesleff, I. (2006). The genetic basis of tooth development and dental defects. *American Journal of Medical Genetics Part A*, *140A*(23), 2530-2535. https://doi.org/https://doi.org/10.1002/ajmg.a.31360

Thesleff, I., Vaahtokari, A., & Vainio, S. (1990). Molecular changes during determination and differentiation of the dental mesenchymal cell lineage. *J Biol Buccale*, *18*(3), 179-188.

Thomaz É, B., Alves, C. M., Ribeiro, C. C., Batista, R. F., Simões, V. M., Cavalli, R., Saraiva Mda, C., Cardoso, V. C., Bettiol, H., Barbieri, M. A., & da Silva, A. A. (2015). Perinatal outcomes and changes in the oral cavity: Brazilian cohorts of Ribeirão Preto and São Luís. *Rev Bras Epidemiol*, *18*(4), 966-970. https://doi.org/10.1590/1980-5497201500040023

Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A., & Shendure, J. (2009). Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods*, *6*(5), 315-316. https://doi.org/10.1038/nmeth.f.248

Turner, J. G. (1912). Two Cases of Hypoplasia of Enamel. *Proc R Soc Med*, *5*(Odontol Sect), 73-76.

Utami, T. W., Miyoshi, K., Hagita, H., Yanuaryska, R. D., Horiguchi, T., & Noma, T. (2011). Possible linkage of SP6 transcriptional activity with amelogenesis by protein stabilization. *J Biomed Biotechnol*, *2011*, 320987. https://doi.org/10.1155/2011/320987

Via, W. F., Jr., & Churchill, J. A. (1959). Relationship of enamel hypoplasia to abnormal events of gestation and birth. *J Am Dent Assoc*, *59*, 702-707. https://doi.org/10.14219/jada.archive.1959.0209

Vinson, J. P., Jaffe, D. B., O'Neill, K., Karlsson, E. K., Stange-Thomann, N., Anderson, S., Mesirov, J. P., Satoh, N., Satou, Y., Nusbaum, C., Birren, B., Galagan, J. E., & Lander, E. S. (2005). Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome Res*, *15*(8), 1127-1135. https://doi.org/10.1101/gr.3722605

Wang, B., Wan, L., & Li, L. M. (2015). Achieving Accurate and Fast Base-calling by a Block Model of the Illumina Sequencing Data. *IFAC-PapersOnLine*, *48*(28), 1462-1465. https://doi.org/https://doi.org/10.1016/j.ifacol.2015.12.339

Wang, S., Choi, M., Richardson, A. S., Reid, B. M., Seymen, F., Yildirim, M., Tuna, E., Gençay, K., Simmer, J. P., & Hu, J. C. (2014). STIM1 and SLC24A4 Are Critical for Enamel Maturation. *J Dent Res*, *93*(7 Suppl), 94s-100s. https://doi.org/10.1177/0022034514527971

Wang, S. K., Choi, M., Richardson, A. S., Reid, B. M., Lin, B. P., Wang, S. J., Kim, J. W., Simmer, J. P., & Hu, J. C. (2014). ITGB6 loss-of-function mutations cause autosomal recessive amelogenesis imperfecta. *Hum Mol Genet*, *23*(8), 2157-2163. https://doi.org/10.1093/hmg/ddt611

Wang, S. K., Zhang, H., Wang, Y. L., Lin, H. Y., Seymen, F., Koruyucu, M., Wright, J. T., Kim, J. W., Simmer, J. P., & Hu, J. C. (2023). FAM20A mutations and transcriptome analyses of dental pulp tissues of enamel renal syndrome. *Int Endod J*, *56*(8), 943-954. https://doi.org/10.1111/iej.13928

Wang, S. K., Zhang, H., Wang, Y. L., Seymen, F., Koruyucu, M., Simmer, J. P., & Hu, J. C. (2022). Phenotypic variability in LAMA3-associated amelogenesis imperfecta. *Oral Dis*. https://doi.org/10.1111/odi.14425

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737-738. https://doi.org/10.1038/171737a0

Weinmann, J. P., Svoboda, J. F., & Woods, R. W. (1945). Hereditary disturbances of enamel formation and calcification. *The Journal of the American Dental Association*, *32*(7), 397-418.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., . . . Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*, *37*(10), 1155-1162. https://doi.org/10.1038/s41587-019-0217-9

Wiener, R. C., Shen, C., Findley, P., Tan, X., & Sambamoorthi, U. (2018). Dental Fluorosis over Time: A comparison of National Health and Nutrition Examination Survey data from 2001-2002 and 2011-2012. *J Dent Hyg*, *92*(1), 23-29.

Witkop, C. J., Jr. (1988). Amelogenesis imperfecta, dentinogenesis imperfecta and dentin dysplasia revisited: problems in classification. *J Oral Pathol*, *17*(9-10), 547-553. https://doi.org/10.1111/j.1600-0714.1988.tb01332.x

Wright, J. T. (2023). Enamel Phenotypes: Genetic and Environmental Determinants. *Genes (Basel)*, *14*(3). https://doi.org/10.3390/genes14030545

Yamaguti, P. M., Neves, F. d. A. R., Hotton, D., Bardet, C., Dure-Molla, M. d. L., Castro, L. C., Scher, M. d. C., Barbosa, M. E., Ditsch, C., Fricain, J.-C., Faille, R. d. L., Figueres, M.-L., Vargas-Poussou, R., Houiller, P., Chaussain, C., Babajko, S., Berdal, A., & Acevedo, A. C. (2017). Amelogenesis imperfecta in familial hypomagnesaemia and hypercalciuria with nephrocalcinosis caused by CLDN19 gene mutations. *Journal of Medical Genetics*, *54*(1), 26-37. https://doi.org/10.1136/jmedgenet-2016-103956

Yik, W. Y., Steinberg, S. J., Moser, A. B., Moser, H. W., & Hacia, J. G. (2009). Identification of novel mutations and sequence variation in the Zellweger syndrome spectrum of peroxisome biogenesis disorders. *Hum Mutat*, *30*(3), E467-480. https://doi.org/10.1002/humu.20932

Yoshinaga, Y., Daum, C., He, G., & O'Malley, R. (2018). Genome Sequencing. In R. P. de Vries, A. Tsang, & I. V. Grigoriev (Eds.), *Fungal Genomics: Methods and Protocols* (pp. 37-52). Springer New York. https://doi.org/10.1007/978-1-4939-7804-5_4

Zhang, Z., Tian, H., Lv, P., Wang, W., Jia, Z., Wang, S., Zhou, C., & Gao, X. (2015). Transcriptional factor DLX3 promotes the gene expression of enamel matrix proteins during amelogenesis. *PLOS ONE*, *10*(3), e0121288. https://doi.org/10.1371/journal.pone.0121288

Zheng, X., Huang, W., He, Z., Li, Y., Li, S., & Song, Y. (2023). Effects of Fam83h truncation mutation on enamel developmental defects in male C57/BL6J mice. *Bone*, *166*, 116595. https://doi.org/10.1016/j.bone.2022.116595

CHAPTER 2 A targeted smMIPs screen of non-syndromic Amelogenesis Imperfecta

Hany U, Watson CM, Liu L, Nikolopoulos G, Smith CEL, Poulter JA, Antanaviciute A, Rigby A, Balmer R, Brown CJ, Patel A, De Camargo MGA, Rodd HD, Moffat M, Murillo G, Mudawi A, Jafri H, Inglehearn CF, Mighell AJ.
Status: Final draft stage, to submit to Human Mutation.

This chapter introduces the subsequent paper, providing crucial background information, outlining the rationale for the research, detailing methodologies employed, and acknowledging the significant contributions made by other collaborators to enrich the overall work.

2.1 Research Rationale

This paper describes the development of a targeted screening reagent using single molecule Molecular Inversion Probe (smMIPs) technology, for the genetic analysis of patients with AI and details how it was successfully deployed to provide a molecular diagnosis for unsolved AI cases. The primary goal was to develop a sequencing workflow for targeted sequencing of the 19 genes known to harbour pathogenic variants causing NS AI, which could then be used routinely as a first screen for AI patients. These genes were selected on the basis that, in prior screening of cases from the Leeds AI cohort, they had comprised a significant proportion of the mutation spectrum associated with AI, and they were composed of a relatively small number of exons, requiring fewer probes to cover the target regions. The genes *LAMA3, LAMC2* and *COL7A1* were not included in the panel because they are large, and the contribution of these genes to the development of the AI phenotype was not established in the literature.

In the past, WES had been used by the Leeds group and others to identify genetic variants in AI patients. WES focuses on sequencing all approximately 20,000 (Frankish et al., 2019) protein-coding genes of the human genome, which is both time- and money-intensive. By contrast, the smMIPs method that we report has been devised to target only the protein coding and splice-site regions of 19 genes known to cause NS AI. The workflow requires less sequence yield than WES, is less challenging to analyse, constitutes less of a data storage issue thereafter and is quicker and cheaper. NHS England has commissioned genetic testing for AI patients. The eligibility criteria that are detailed in the national genomic test directory indicate that a patient should be referred by a specialist dentist, and the proband is eligible for a WES/medium panel genetic screen/CNV detection by MLPA (multiplex ligation-

dependent probe amplification). Since the AI research community does not have access to this service, smMIPs screening could provide a quick pre-screen for AI patients, or additional family members, for research purposes.

A further rationale for developing a targeted genetic screening method was to keep the analysis focused on genes already implicated in AI. Traditional genetic testing risks the discovery of incidental/secondary findings that raise ethical questions about how and when to communicate these findings to the individual who underwent testing or participated in research. Incidental genetic findings refer to unexpected genetic information that is discovered during a genetic test or research but is unrelated to the primary purpose of the testing or study. In contrast secondary (or pertinent) genetic findings refer to genetic information that is deliberately sought out or analysed as part of a genetic test or research study but is separate from the primary reason for the testing or research. The ACMG (American College of Medical Genetics and Genomics) has published a position statement, which recommends that pathogenic and likely pathogenic variants in 73 genes, associated mainly with hereditary cancer and cardiac disease, should be reported for patients undergoing diagnostic genetic testing (Miller et al., 2021). The question of whether incidental findings should be communicated to patients, particularly those who represent paediatric groups, remains contentious. It is important to note that the utility of these findings is context-specific and can vary based on the specific genetic variants, the healthcare setting, and the individual's values and preferences. A thoughtful approach to informed consent, genetic counselling, and clear communication of findings is essential to maximize their utility while respecting individual choices and autonomy. A focused genetic screening approach like smMIPs side-steps this issue by reducing the likelihood of

encountering incidental findings, as it intentionally avoids sequencing regions of the genome that are unrelated to the primary research or clinical question.

2.2 Research Contribution

Patients were diagnosed and recruited either by my supervisor Dr Alan Mighell or by other dental colleagues who collaborate with him. With oversight from my supervisory team, I designed the research study, planned, and undertook all the wet laboratory work, carried out all the data analysis and drafted the manuscript. All of the variants detected by smMIPs screening were further confirmed and segregated in available family members by Sanger sequencing. I designed PCR primers targeting all these variants, and under my supervision, Lu Liu, a technician funded by the Rosetrees Trust, carried out the wet laboratory work required to complete this Sanger sequencing. Some key steps of method development are described below.

2.2.1 Designing smMIPs probes using MIPGEN

I used the MIPGEN package available at https://github.com/shendurelab/MIPGEN (Hiatt et al., 2013) to design smMIPs probes targeting coding and splice-site regions of 19 genes causing non-syndromic AI. MIPGEN needs a C + + compiler and was used in conjunction with the human reference genome build hg19. Dr James Poulter installed the software on a local server. I then designed smMIPs probes for the AI gene panel using the Unix command line and Python. Details of the MIPGEN workflow are included in Appendix 1.

2.2.2 Sample preparation for smMIPs screening

I have used the smMIPs probe-based approach for preparing sequencing libraries for patient DNA and carried out the laboratory work required for library preparation for smMIPs sequencing. In brief, the steps involved in smMIPs library preparation are probe pooling, probe phosphorylation, target capture and ligation, exonuclease treatment, and post-capture PCR amplification (Cantsilieris et al., 2017). Sequencing was carried out using SBS chemistry on platforms manufactured by Illumina. Depending on the number of patients included in the experiment, this was analysed on either the MiSeq or NextSeq500. To sequence smMIPs libraries using an Illumina sequencer, custom sequencing primers were required; details of these primers are included in Appendix 2. The protocol for sequencing smMIPs libraries using MiSeq and NextSeq platforms are provided in Appendix 3.

2.2.3 Data analysis

To process the smMIPs data, Dr Christopher Watson deployed an in-house data processing pipeline that was based on the MIPVAR v.0.7.17 (https://sourceforge.net/projects/mipvar/) framework. MIPVAR is a Java application that uses the bioinformatics tools BWA (v.0.7.12), GATK (v.3.2.2) and BEDTools (v.2.24.0). I received training from Dr Watson on installing required software using Bioconda and on how to perform analysis using the pipeline. I then carried out data analysis for variant identification and analysed results for all the patients' samples included in this study. Details of the commands used to process the smMIPs data are provided in Appendix 4.

2.2.4 Additional methodology

Details of the probe design, library preparation, sequencing and data analysis are included in the manuscript.

A targeted smMIPs screen of non-syndromic Amelogenesis Imperfecta

Ummey Hany[1]. u.hany@leeds.ac.uk; Orcid ID 0000-0002-4486-1625.
Christopher M. Watson[1,2]. c.m.watson@leeds.ac.uk; Orcid ID 0000-0003-2371-1844.
Lu Liu[1,3]. L.Liu3@leeds.ac.uk; Orcid ID 0009-0008-3593-8409.
Georgios Nikolopoulos[4]. georknikolopoulos@gmail.com; Orcid ID 0000-0003-3166-8372.
Claire E.L. Smith[1]. c.e.l.Smith@leeds.ac.uk; Orcid ID 0000-0001-8320-5105.
James A. Poulter[1]. J.A.Poulter@leeds.ac.uk; Orcid ID 0000-0003-2048-5693.
Agne Antanaviciute[5]. agne.antanaviciute@ndm.ox.ac.uk; Orcid ID 0000-0002-9019-2215.
Alice Rigby[1,3]. A.L.Rigby@leeds.ac.uk; Orcid ID 0000-0001-6888-2255.
Richard Balmer[3]. R.C.Balmer@leeds.ac.uk; Orcid ID 0009-0004-5314-2157.
Catriona J. Brown[6]. Catriona.Brown@bhamcommunity.nhs.uk;
Anesha Patel[7]. Anesha.Patel1@nihr.ac.uk;
María Gabriela Acosta de Camargo[8]. gabyacosta647@gmail.com; Orcid ID 0000-0001-7615-918X.
Helen D Rodd[9]. H.D.Rodd@sheffield.ac.uk; Orcid ID 0000-0003-2973-2558.
Michelle Moffat[10]. michelle.moffat3@nhs.net;
Gina Murillo[11]. twingina24@hotmail.com; Orcid ID 0000-0003-2976-6301.
Amal Mudawi[1,12]. dnamym@leeds.ac.uk; Orcid ID 0000-0002-2265-5898.
Hussain Jafri[13]. hussain112345@hotmail.com;
Alan J. Mighell[3,14]. A.J.Mighell@leeds.ac.uk; Orcid ID 0000-0002-9624-6923.
Chris F. Inglehearn[1,14]. c.inglehearn@leeds.ac.uk; Orcid ID 0000-0002-5143-2562.

Affiliations

1: Leeds Institute of Medical Research, University of Leeds, St. James's University Hospital, Leeds, UK.
2: North East and Yorkshire Genomic Laboratory Hub, Central Lab, St. James's University Hospital, Leeds, UK.
3: School of Dentistry, Clarendon Way, University of Leeds, Worsley Building, Leeds, UK.
4: Institute for Fundamental Biomedical Research, B.S.R.C. 'Alexander Fleming', 16672 Vari, Attica, Greece.
5. MRC Human Immunology Unit, University of Oxford, Oxford, UK.
6: Birmingham Dental Hospital, Mill Pool Way, Edgbaston, Birmingham, UK.

7: LCRN West Midlands Core Team, NIHR Clinical Research Network (CRN), Research Park (West Wing), Vincent Drive, Edgbaston, Birmingham, UK.
8: Department of Dentistry of the Child and Adolescent, Universidad de Carabobo, Venezuela.
9: Academic Unit of Oral Health Dentistry and Society, School of Clinical Dentistry, University of Sheffield, Sheffield, UK.
10: Paediatric Dentistry, The Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK.
11: School of Dentistry, Universidad de Costa Rica, Ciudad Universitaria Rodrigo Facio, San Pedro Montes De Oca, Costa Rica.
12: University of Khartoum, Khartoum, Sudan.
13: Fatima Jinnah Medical University, Lahore, Pakistan.
14: Joint senior authors. Correspondence to C.Inglehearn@leeds.ac.uk

Acknowledgements

The authors thank the families involved for their support for this study.

Author contributions

Authors UH, CMW, GN, CELS, JAP, CFI and AJM contributed to the conception, design, data acquisition and interpretation, drafting and critical review of the manuscript. Authors LL, AA, AR, RB, CJB, AP, MGAC, HDR, MM, GM, AM, and HJ contributed to data acquisition, interpretation, and critical review of the manuscript. All authors approved the final version of the manuscript.

Funding

This work was supported by Rosetrees Trust Grant PGS19-2/10111, Wellcome Trust grant number WT093113MA and by a Leeds Doctoral Scholarship awarded to UH.

Conflict of Interest statement

None.

Keywords

Amelogenesis, tooth disease, enamel development.

Abstract

Amelogenesis is the process of tooth enamel formation. Disruption of amelogenesis causes the Mendelian inherited disorder amelogenesis imperfecta (AI). AI patients have weak, discoloured or brittle enamel, caused by reduced enamel quantity or mineralization, and AI can occur in isolation or, less commonly, as part of a syndromic condition. Pathogenic variants in at least 38 genes have been shown to cause the condition. Current genetic screening studies generally use exome sequencing, but this is expensive and requires complex analysis and a large data storage capacity. smMIPs (single molecule molecular inversion probes) provide a flexible alternative, allowing the creation of a disease-specific, targeted sequencing reagent for low cost, robust, high-throughput screening. Here we describe the development of an smMIP reagent targeting 19 genes implicated in isolated AI, and its use in screening a cohort of 181 AI probands. Whilst this analysis was intended only as a pre-screen to prioritise exome sequencing more efficiently, its use nevertheless led to molecular diagnoses for 66 probands (36%), at a cost per sample screened of ~ £43. Variants in three genes, *COL17A1*, *FAM83H* (both dominant) and *MMP20* (Hardies et al.)(Hardies et al.) accounted for approximately half of cases solved. There is scope for further improvement of the smMIP reagent by adding additional probes targeting gaps in the existing sequence or regions of low coverage, as well as ongoing improvements to take account of new information about the genetic basis of AI. The smMIPs reagent therefore provides a robust, flexible, high-throughput and low-cost approach to AI screening, and is now a resource available to other AI researchers around the world.

Introduction

Amelogenesis is the name given to the process of tooth enamel formation. Ameloblasts, which are derived from the oral epithelium, form a monolayer around the developing enamel. These cells express over 10,000 genes during amelogenesis (Hu et al., 2015), and the highly coordinated sequence of expression of key genes by ameloblasts is essential for the formation and mineralization of enamel during tooth development. Any disruption in amelogenesis can lead to Amelogenesis Imperfecta (AI), a Mendelian inherited disorder affecting the enamel appearance, quantity, quality and function of all teeth of both dentitions. AI results in weak, discoloured enamel that easily breaks down or a reduced enamel volume, with no enamel formed in the most extreme instances. It can occur in isolation or as a component of a series of syndromic conditions. It can be difficult to distinguish clinically between syndromic and non-syndromic AI (NSAI), reflecting the fact that additional clinical features can be subtle or of variable severity or timing in their clinical presentation. It can also be challenging to distinguish AI from other developmental defects of enamel. Within these limitations, reported prevalence of NSAI ranges from 1 in 233 in Turkey (Altug-Atac & Erdem, 2007) through 1 in 700 in Sweden (Bäckman & Holm, 1986), 1 in 1000 in Argentina (Sedano, 1975) and 1 in 8000 in Israel (Chosack et al., 1979) to 1 in 14,000 in USA (Witkop, 1988). Since it was first discovered that mutations in the X-linked gene *AMELX* (amelogenin, X) cause NSAI (Lagerström et al., 1991), a further 20 autosomal genes have been reported to be associated with NSAI (Simmer et al., 2021; Smith et al., 2017). Of these 21 genes, pathogenic variants in one cause X-linked AI, in ten cause autosomal recessive (AR) NSAI, in eight cause autosomal dominant (AD) NSAI, and in two can cause both dominant and recessive forms of NSAI.

Knowing which gene and variant(s) cause AI in patients gives a clear prognosis, informs management that may include genetic counselling for patients and relatives, and increases our understanding of the underlying causes, supporting future research. Next generation sequencing, also known as massively parallel sequencing, has revolutionised the speed and accuracy of diagnostic screening for pathogenic variants causing inherited conditions in the last fifteen years (Metzker, 2010). Although whole exome sequencing (WES) is less expensive than whole genome sequencing (WGS), it is still relatively costly compared to a customised "targeted" approach. Furthermore, WES and WGS analysis pipelines are

computationally demanding, data storage has governance, operational and cost implications, and these approaches do generate coincidental findings.

The single-molecule Molecular Inversion Probes (smMIPs) approach (Hiatt et al., 2013) presents an attractive alternative to WES/WGS, allowing selective screening of specific target genes or loci in large patient cohorts. Originally developed for targeted genotyping of SNPs (Single Nucleotide Polymorphism) in patients with immunoglobin nephropathy or Berger's disease (Hardenbol et al., 2003), its applications now span a wide range of fields, from clinical genetics to evolutionary biology. It has also been used successfully for diagnosing a wide range of diseases and conditions, including patients with *ABCA4*-associated Stargardt disease, macular dystrophy and male infertility (Hitti-Malin et al., 2022; Khan et al., 2020; Mc Clinton et al., 2023; Oud et al., 2017).

smMIPs are oligonucleotide probes consisting of a common DNA backbone flanked by target-specific sequences, ligation, and extension arms. Probes hybridize to the complementary target genomic sequence of the arms. These extension arms act as primers, allowing a DNA polymerase to fill the gap between them. The product is then circularized by DNA ligase. The circular DNA is linearized and amplified by PCR using universal primers complementary to the probe backbone. Thousands of probes can be mixed in a single reaction to amplify multiple target regions from a single DNA sample. Samples can then be multiplexed through the addition of short unique index sequences to the primer, allowing identification of sample-specific reads (Cantsilieris et al., 2017). The technique has been further improved by incorporating unique molecular identifiers (UMIs) with the potential to capture single molecular events (Hiatt et al., 2013). One recent study used smMIPs to screen exons amounting to over 450 kb of sequence, at loci distributed throughout the genome, in 300 patients in a single sequencing run (Panneman et al., 2023). This approach gives a far lower cost per-sample than other comparable targeted analysis workflows (Hiatt et al., 2013).

Here, we describe the development of a custom smMIPs reagent for screening AI and demonstrate its utility in providing molecular diagnoses in previously undiagnosed cases of AI. The smMIPs reagent was designed to capture the coding sequences and splice donor and acceptor sites of 19 genes associated with NSAI. The assay was validated using control samples that had previously been analysed by WES, then used to screen 181 unsolved cases from an AI cohort.

Materials and Methods

Participant recruitment

Participants with a clinical diagnosis of NSAI were recruited with informed written consent and ethical approval (REC 13/YH/0028), in accordance with the principles of the Declaration of Helsinki. Genomic DNA was obtained from saliva using Oragene® DNA Sample Collection kits (DNA Genotek).

Gene selection and smMIPs probe design

Probes were designed using MIPGEN (https://github.com/shendurelab/MIPGEN) (Boyle et al., 2014). Each probe was designed to have an approximately 82 nucleotide backbone (including the extension and ligation arms located at each end) and a 110 nucleotide region-specific target sequence. Extension and ligation arms together were 45 nucleotides long and were complementary to sequence adjacent to the region-specific target sequence. The common linker sequence joining the two arms contained universal PCR primer complementary sites, followed by an 8 nucleotide stretch of random bases. The latter provided a degenerate molecular index with 4^8 possible unique combinations for each amplicon (Eijkelenboom et al., 2016). Probe selection was performed by uploading all the probes on Integrative Genome Viewer (IGV) (Robinson et al., 2011). For each target region, a single probe was selected, targeting either the plus or minus strand of the DNA with a minimum of 10 bp overlap with adjacent probes wherever possible. Probes with high logistics scores, as calculated by MIPGEN, were selected where possible. 517 of the 609 probes used had logistic scores >=0.5. Probes with scores of less than 0.5 were used in challenging regions such as the *FAM83H* and *ACP4* gene loci (Hiatt et al., 2013).

Probe preparation

Each smMIPs probe was synthesized at 100 nanomole scale (in 96-well plate format) without modifications (Integrated DNA Technologies). Probes were then pooled at equimolar concentrations to create a "megapool" for hybridization to genomic DNA. A 25 µL aliquot of the megapool was 5'-phosphorylated in a reaction that comprised 1 µL of 10 units of T4 Polynucleotide Kinase (New England Biolabs, NEB), and 3 µL of 10x T4 DNA ligase reaction buffer with 10 mM ATP in a 30 µL total reaction volume. The reaction was incubated at 37°C for 45 minutes followed by 65°C for 20 minutes (Cantsilieris et al., 2017).

Library preparation and sequencing

For one affected participant from each of the 181 families investigated, 100 ng genomic DNA in 10 µL nuclease free water (NEB) was subjected to targeted hybridisation and ligation using the phosphorylated probe megapool. The probe megapool was diluted to obtain a ratio of 800 probe copies per single DNA molecule in the final capture reaction. A 15 µL hybridisation capture mastermix was prepared on ice for each sample, comprising 2.5 µL of Ampligase™, 10× reaction buffer (Epicenter), 0.32 µL of dNTP mix (0.025mM), 0.32 µL HemoKlenTaq (10U/µL) (NEB)), 0.2 µL of Ampligase™ (5 U/µL) (Epicenter), 3.29 µL of the smMIPs megapool (10^5 x diluted), and nuclease-free water to a total volume of 15 µL. 15 µL of hybridisation mastermix was then added to 10 µL of genomic DNA (100 ng). The reaction was incubated on a benchtop thermocycler at 95 ºC for 3 minutes followed by 22 hours at 65 ºC. The reaction was then exonuclease treated using a mastermix that contained 0.5 µL of Exonuclease I (NEB), 0.5 µL of Exonuclease III , 0.2 µL Ampligase™ 10x reaction buffer (Epicenter) and 0.8 µL nuclease-free water. The reaction was incubated at 37 ºC for 45 minutes then at 95 ºC for 2 minutes. Following exonuclease treatment, 10 µL of the treated sample was added to 15 µL of a post-capture PCR mastermix, which was prepared by combining 12 µL of Q5 Hot Start HiFi 2× mastermix , 1.25 µL of 10µM forward primer, 1.25 µL of 10µM barcoded reverse primer and 0.5 µL of nuclease-free water. Thermocycling conditions were an initial cycle of 98 ºC for 30 seconds followed by 23 cycles of 98 ºC for 10 seconds, 60 ºC for 30 seconds and 72 ºC for 30 seconds, before a final extension step at 72 ºC for 2 minutes. The reactions were then purified using a 0.8× Axygen® AxyPrep MAG PCR clean-up kit and the fragment distribution of the resulting library was visualized on a Tapestation using a DNA 1000 HS assay (Agilent Technologies, Wokingham, UK). Each library was individually quantified using a Qubit 2.0 fluorometer (Invitrogen) and HS DNA reagents. Libraries were then pooled in equimolar concentration for sequencing (Cantsilieris et al., 2017). Depending on the number of samples processed in an individual batch, sequencing was carried out using either a MiSeq (Illumina Inc.) or a NextSeq 500 (Illumina Inc.) to generate paired-end 150 bp reads. Manufacturer's instructions were followed throughout.

smMIPs probe rebalancing

Average read depth was calculated for each smMIPs probe from the initial test run. This value was used to adjust the volume of each underperforming or overperforming probe, either

increasing or decreasing their concentration in the probe megapool. A small number of probes failed. Alternative probes were designed to replace them in subsequent runs.

Data processing pipeline

An in-house bioinformatics pipeline was developed to process the raw sequence data. For each participant, MIPVAR https://sourceforge.net/projects/mipvar/ processed consecutive read-pairs by removing the unique molecular identifier (UMIs) then aligning the sequence read to the human reference genome (hg19) using MIPVAR v.0.7.17. The ligation and extension arms were trimmed to eliminate erroneous variant calls caused by hybridization bias. Read pairs containing identical UMIs that aligned to the same genomic position, were marked as PCR duplicates using Picard v.1.119 (https://broadinstitute.github.io/picard/). Non-reference bases were identified and collated in variant call format (VCF) using the Genome Analysis Toolkit's HaplotypeCaller v.3.7-0 (DePristo et al., 2011). Each per-participant VCF file was annotated with functionally relevant biological information and observed population frequency data using Annovar (Wang et al., 2010). ExomeDepth software (Plagnol et al., 2012) was used for CNV analysis. An in-house batch analysis script was used to process samples that were sequenced at the same time on the same machine following the same bioinformatics procedures.

Variant classification

The pathogenicity status of identified variants was classified according to American College of Medical Genetics and Genomics (ACMG) criteria using the web-based platform Franklin by Genoox (https://franklin.genoox.com/clinical-db/home) (Richards et al., 2015). Allele frequencies were obtained from the Genome Aggregation Database v.2.1.1 (https://gnomad.broadinstitute.org/) (Karczewski et al., 2020). Splicing predictions were obtained using SpliceAI (https://spliceailookup.broadinstitute.org) (Jaganathan et al., 2019).

Variant verification and segregation analysis

Primers were designed using AutoPrimer3 (https://github.com/david-a-parry/autoprimer3) and synthesised by IDT (Leuven, Belgium). 25 ng genomic DNA was amplified using Q5 High-Fidelity 2× Master Mix (NEB) according to manufacturer's instructions. 2.5 µL PCR products were purified using 1 µL ExoSAP-IT (Applied Biosystems). The sequencing reaction mix was prepared by adding 1 µL of ExoSAP-IT treated DNA to a mastermix containing 6 µL of nuclease free water, 0.5 µL of BigDye® Terminator v3.1 (Applied

Biosystems), 1.5 µL of BigDye® Terminator v3.1 Sequencing Buffer (Applied Biosystems) and 1 µL of primer (1.6 µM). After an initial denaturation step at 96 °C for 1 minute, the samples underwent 25 cycles of 96 °C for 10 seconds, 50 °C for 5 seconds and 60 °C for 4 minutes. All temperatures were ramped at 1 °C/second. Sequencing template were precipitated using 125 mM EDTA and 100% ethanol followed by centrifugation at 3900 rpm for 30 minutes at 4 °C. DNA was washed with 70% ethanol and dried at 37 °C for 1 minute. Precipitates were dissolved in 10 µL Hi-Di Formamide (Applied Biosystems) ready for sequencing. Sequencing was carried on an ABI3130xl Genetic Analyser (Applied Biosystems) following manufacturer's instructions. Electropherograms were analysed using SeqScape v.2.5 (Applied Biosystems).

Results

Nineteen genes associated with NSAI (listed in Table 1) were targeted with 609 smMIPs covering their coding exons and adjacent splice sites. After probe rebalancing, a mean capture efficiency of 97% at greater than 20 reads was achieved across the nineteen genes targeted in control DNA. The optimised reagent was then used to screen genomic DNA from eight validation control samples and 181 probands from unrelated families with NSAI. The variants identified per participant were filtered to exclude those with a CADD score <15 or a minor allele frequency (MAF) >0.01 for homozygous and >0.001 for heterozygous variants (Karczewski et al., 2020; Kircher et al., 2014). The variant list for each case was then filtered further to include only variants classified pathogenic, likely pathogenic or a variant of unknown significance (VUS) according to ACMG criteria. All potentially pathogenic genotypes were re-sequenced by Sanger sequencing and their segregation with disease was checked in all available family members.

Validation Samples

To validate the smMIPs library preparation method and bioinformatics pipeline, eight control DNAs with known pathogenic variants were analysed. These were from three individuals with AI due to homozygous variants (*MMP20*:NM_004771.4:c.955A>T;p.Ile319Phe, *KLK4*:NM_004917.4:c.632delT;p.Leu211Argfs37 and *RELT*:NM_152222.2:c.164C>T;p.Thr55Ile) and five from individuals with dominant AI due to heterozygous variants (*ENAM*: NM_031889.3:c.92T>G;p.Leu31Arg, *LAMB3*: NM_000228.3:c.2660G>A;p.Arg887His, *COL17A1*: NM_000494.4:c.3595G>C; p.Glu1199Gln, *FAM83H*: NM_198488.5:c.1354C>T;p.Gln452* and *FAM83H*: NM_198488.5:c.1192C>T;p.Gln398*). The smMIPs variant calling pipeline correctly identified all these variants in the corresponding samples, showing the reagent and protocol are effective in screening for a wide range of variants causing AI.

Participant Screening Results

Once validated, the reagent was then used to screen a cohort of 181 probands with NSAI. These included 25 families presenting with dominantly inherited AI, 48 with confirmed, and a further 29 with suspected, recessive AI, and 79 cases either with no family history or where family history was unknown. No families presented with an unambiguous X-linked family history.

A total of 58 variants considered likely to cause disease were detected and confirmed by Sanger sequencing in probands and additional family members where available. These consisted of 17 missense, 17 premature termination, 17 frameshift and 6 splice site variants and one large deletion. By ACMG criteria, 29 of these were classified as pathogenic, 20 as likely pathogenic and 9 as VUSs (Tables 2, 3 and 4 and Figure 1A and 1B). These data resulted in possible or probable molecular diagnoses explaining the condition in 66 probands (36%), by identifying potentially pathogenic genotype combinations in the known AI-associated genes.

Of these 66 probands, 7 (11%) were solved as X-linked AI (Table 2), 32 (48%) as dominant AI (Table 3) and 27 (41%) as recessive AI (Table 4). The 7 cases solved as X-linked AI had variants in *AMELX*, the only known gene causing AI on the X chromosome. Of the 32 dominant families, *COL17A1* (12 probands) and *FAM83H* (10 probands) gene variants accounted for the majority of cases, with smaller numbers having pathogenic heterozygous variants in *DLX3*, *ENAM* and *AMBN*. No likely dominant disease-causing variants were identified in *LAMB3*, *SP6*, or *AMTN* in this cohort. Among probands solved as recessive AI, only *MMP20* variants accounted for a relatively large proportion (a third) of cases, in part because of previously reported common founder variants in the local community (G. Nikolopoulos et al., 2021). Smaller numbers of variants were also found in *WDR72*, *ACP4*, *FAM20A*, *AMBN*, *SLC24A4*, *RELT* and *ENAM*. No variants were identified in three other genes implicated in ARAI, *ITGB6*, *ODAPH*, *GPR68* and *KLK4*. Variants in *ENAM* and *AMBN* were implicated in both dominant and recessive AI in this cohort. Proportions of families solved by variants in each gene are shown in Figure 1C.

Possible Digenic Cases

Interestingly, four families presented with potentially pathogenic variants in both *COL17A1* and *MMP20* (Table 5). In each case families were considered to be solved based on the presence of a genetic variant or variants within one of these genes, but variant(s) in the other were also present and segregated with AI where this information was available. Pedigrees of these families and Sanger sequencing chromatograms from the each proband are shown in Figure 2.

Discussion

This study describes the development and validation of a custom smMIPs sequencing reagent targeting the coding exons and splice sites of 19 genes known to harbour pathogenic variants presenting as NSAI, and its use in screening a single affected participant from each of a cohort of 181 unsolved NSAI families. Data analysis and storage are simplified with this targeted screening approach, and ethical issues posed by coincidental findings are reduced. The targeting reagent and optimised method used here proved rapid and robust, detecting all validation control variants and solving a third of AI probands screened in a single sequencing run. It was also cost-effective, at a per-sample cost of around £43, with further economies of scale possible through subsequent rounds of optimisation and greater multiplexing of samples.

Previous AI cohort studies have suggested that screening all the currently known AI-associated genes solves 50-60% of cases (Bloch-Zupan et al., 2023; Chan et al., 2011). The smMIPs screen described here identified likely causative variants in 36% of cases and families. However, the smMIPs screen was not intended to provide a comprehensive screen, but rather to act as a pre-screen for AI cases for known genes, allowing targeting of more comprehensive but costly and labour-intensive WES or WGS to cases unsolved in the initial screen. It is likely that the diagnostic success achieved with the approach described could be improved. The use of smMIPs gives flexibility, allowing further optimisation as knowledge advances. Small gaps were noted in the coverage of several genes included in the study (notably *MMP20*, *ENAM* and *AMBN*), meaning that some variants in these genes may have been missed. WES in cases not solved by smMIPs may reveal further blind-spots, either in the capture, sequencing or analysis pipeline used in this study, that lead to variants in the targeted genes being missed. Additional probes targeting these deficiencies could be added to the existing reagent, as could additional probes targeting new genes and variants implicated in AI in the literature, making smMIPs a flexible diagnostics tool for AI research.

The findings of this study can be compared to those of another AI cohort study published recently (Bloch-Zupan et al., 2023). Both studies identify dominant AI as accounting for nearly half of NSAI cases, with recessive disease at approximately 40% and the remainder being solved as X-linked disease. However, screening results may reflect biases in current knowledge or differences in the populations screened. Previous cohort studies have suggested

that dominant AI is much more common than recessive disease (Bäckman & Holmgren, 1988; Chan et al., 2011). The relatively high frequency of recessive AI in the cohort studied here may reflect the inclusion of families from the Yorkshire Pakistani community, which has a high level of first cousin marriage and consequent increased risk of recessive disease (Arciero et al., 2021). The spectrum of variants and frequencies of the different forms of AI revealed by this study and that of Bloch-Zupan and colleagues (Bloch-Zupan et al., 2023) are broadly similar, but include some notable differences. This study found dominant *COL17A1* variants to be the most common cause of NSAI, accounting for 18% of cases, whereas the other study found recessive *MMP20* variants to be the leading cause. Data on *COL17A1* AI from this study is included in a more detailed study reported elsewhere (Hany et al., 2023). Furthermore, our findings reveal variants in *AMBN* accounting for 8% of solved cases, compared to only 3% in the other study, and causing both recessive and dominant NSAI. These findings are also described in more detailed elsewhere (Hany et al., 2023).

Another notable finding of this study is the identification of potentially disease-causing variants in both *COL17A1* and *MMP20* in five individuals from four families screened. Digenic AI has been suggested in three previous studies. One showed co-segregation of *ENAM* and *LAMA3* variants with AI through six meioses in a family (Zhang et al., 2019). The second reported a single case with *COL17A1* and *LAMA3* variants (Prasad et al., 2016). The third consisted of a father with AI thought to be due to a *LAMA3* VUS and a son with more severe AI, who carried the same *LAMA3* variant but was also compound heterozygote for two likely pathogenic *MMP20* variants (Wang et al., 2023). There is also published evidence that specific heterozygous variants in each of *Mmp20* and *Klk4* caused an enamel phenotype in mice but a single heterozygous variant in either gene did not (Hu et al., 2016).

Though speculative, these reports suggest that some AI might in fact be polygenic rather than Mendelian in origin, which may explain the variation in AI phenotype sometimes seen between individuals in the same family. This study provides further circumstantial evidence for such an effect, and interestingly, identifies four probands with allele combinations involving the same gene pair, *MMP20* and *COL17A1*. However, there is no evidence of a direct functional link between these two proteins. Furthermore, AI in each case is fully explained by one genotype (the *COL17A1* genotype in families 4 and 18, the *MMP20* genotype in families 25 and 62), without the need to invoke any contribution from the other gene, and the phenotypes observed in these families are consistent with previously

documented phenotypes for variants in these genes. In the absence of functional evidence or further cases with significant co-segregation, the case for polygenic AI therefore remains unproven.

Conclusions

In summary, we have developed and validated a flexible smMIPs reagent for rapid, high-throughput, cost-effective screening for variants in nineteen genes known to be implicated in NSAI, with further optimisation possible. Intended as a pre-screen for AI cases, its use in a cohort of individuals with AI nevertheless resulted in molecular diagnoses for 66 probands and their families. This analysis confirmed dominant inheritance as the most common mode of inheritance in AI, with *COL17A1* and *FAM83H* variants as the most common underlying cause. The proven success of this approach demonstrates the power of smMIPs, gives insights into the epidemiology of NSAI and provides a reagent that is now available to AI research groups around the world and can be adapted to account for future developments in the field.

References

Aldred, M. J., Crawford, P. J., Roberts, E., & Thomas, N. S. (1992). Identification of a nonsense mutation in the amelogenin gene (AMELX) in a family with X-linked amelogenesis imperfecta (AIH1). *Hum Genet*, *90*(4), 413-416. https://doi.org/10.1007/bf00220469

Arciero, E., Dogra, S. A., Malawsky, D. S., Mezzavilla, M., Tsismentzoglou, T., Huang, Q. Q., Hunt, K. A., Mason, D., Sharif, S. M., van Heel, D. A., Sheridan, E., Wright, J., Small, N., Carmi, S., Iles, M. M., & Martin, H. C. (2021). Fine-scale population structure and demographic history of British Pakistanis. *Nat Commun*, *12*(1), 7189. https://doi.org/10.1038/s41467-021-27394-2

Bäckman, B., & Holm, A. K. (1986). Amelogenesis imperfecta: prevalence and incidence in a northern Swedish county. *Community Dent Oral Epidemiol*, *14*(1), 43-47. https://doi.org/10.1111/j.1600-0528.1986.tb01493.x

Bäckman, B., & Holmgren, G. (1988). Amelogenesis imperfecta: a genetic study. *Hum Hered*, *38*(4), 189-206. https://doi.org/10.1159/000153785

Bloch-Zupan, A., Rey, T., Jimenez-Armijo, A., Kawczynski, M., Kharouf, N., , O.-R. c., Dure-Molla, M. d. L., Noirrit, E., Hernandez, M., Joseph-Beaudin, C., Lopez, S., Tardieu, C., Thivichon-Prince, B., , E. C. C., Dostalova, T., Macek, M., , I. C., Alloussi, M. E., Qebibo, L., . . . Urzúa Orellana, B. (2023). Amelogenesis imperfecta: Next-generation sequencing sheds light on Witkop's classification [Original Research]. *Frontiers in physiology*, *14*. https://doi.org/10.3389/fphys.2023.1130175

Boyle, E. A., O'Roak, B. J., Martin, B. K., Kumar, A., & Shendure, J. (2014). MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics*, *30*(18), 2670-2672. https://doi.org/10.1093/bioinformatics/btu353

Cantsilieris, S., Stessman, H. A., Shendure, J., & Eichler, E. E. (2017). Targeted Capture and High-Throughput Sequencing Using Molecular Inversion Probes (MIPs). *Methods Mol Biol*, *1492*, 95-106. https://doi.org/10.1007/978-1-4939-6442-0_6

Chan, H. C., Estrella, N. M., Milkovich, R. N., Kim, J. W., Simmer, J. P., & Hu, J. C. (2011). Target gene analyses of 39 amelogenesis imperfecta kindreds. *Eur J Oral Sci*, *119 Suppl 1*(Suppl 1), 311-323. https://doi.org/10.1111/j.1600-0722.2011.00857.x

Chosack, A., Eidelman, E., Wisotski, I., & Cohen, T. (1979). Amelogenesis imperfecta among Israeli Jews and the description of a new type of local hypoplastic autosomal recessive amelogenesis imperfecta. *Oral Surg Oral Med Oral Pathol*, *47*(2), 148-156. https://doi.org/10.1016/0030-4220(79)90170-1

Condrat, I., He, Y., Cosgarea, R., & Has, C. (2018). Junctional Epidermolysis Bullosa: Allelic Heterogeneity and Mutation Stratification for Precision Medicine. *Front Med (Lausanne)*, *5*, 363. https://doi.org/10.3389/fmed.2018.00363

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, *43*(5), 491-498. https://doi.org/10.1038/ng.806

Eijkelenboom, A., Kamping, E. J., Kastner-van Raaij, A. W., Hendriks-Cornelissen, S. J., Neveling, K., Kuiper, R. P., Hoischen, A., Nelen, M. R., Ligtenberg, M. J., & Tops, B. B. (2016). Reliable Next-Generation Sequencing of Formalin-Fixed, Paraffin-Embedded Tissue Using Single Molecule Tags. *J Mol Diagn*, *18*(6), 851-863. https://doi.org/10.1016/j.jmoldx.2016.06.010

El-Sayed, W., Shore, R. C., Parry, D. A., Inglehearn, C. F., & 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), 235-239. https://doi.org/10.1159/000252801

Hany, U., Watson, C. M., Liu, L., Smith, C. E. L., Harfoush, A., Poulter, J. A., Nikolopoulos, G., Balmer, R., Brown, C. J., Patel, A., Simmonds, J., Charlton, R., Acosta de Camargo, M. G., Rodd, H. D., Jafri, H., Antanaviciute, A., Moffat, M., Al-Jawad, M., Inglehearn, C. F., & Mighell, A. J. (2023).

Heterozygous COL17A1 variants are a frequent cause of amelogenesis imperfecta. *J Med Genet*. https://doi.org/10.1136/jmg-2023-109510

Hardenbol, P., Banér, J., Jain, M., Nilsson, M., Namsaraev, E. A., Karlin-Neumann, G. A., Fakhrai-Rad, H., Ronaghi, M., Willis, T. D., Landegren, U., & Davis, R. W. (2003). Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol*, *21*(6), 673-678. https://doi.org/10.1038/nbt821

Hardies, K., de Kovel, C. G. F., Weckhuysen, S., Asselbergh, B., Geuens, T., Deconinck, T., Azmi, A., May, P., Brilstra, E., Becker, F., Barisic, N., Craiu, D., Braun, K. P. J., Lal, D., Thiele, H., Schubert, J., Weber, Y., van 't Slot, R., Nürnberg, P., . . . Consortium, o. b. o. t. a. r. w. g. o. t. E. R. (2015). Recessive mutations in SLC13A5 result in a loss of citrate transport and cause neonatal epilepsy, developmental delay and teeth hypoplasia. *Brain*, *138*(11), 3238-3250. https://doi.org/10.1093/brain/awv263

Hart, P. S., Becerik, S., Cogulu, D., Emingil, G., Ozdemir-Ozenen, D., Han, S. T., Sulima, P. P., Firatli, E., & Hart, T. C. (2009). Novel FAM83H mutations in Turkish families with autosomal dominant hypocalcified amelogenesis imperfecta. *Clin Genet*, *75*(4), 401-404. https://doi.org/10.1111/j.1399-0004.2008.01112.x

Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J., & Shendure, J. (2013). Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res*, *23*(5), 843-854. https://doi.org/10.1101/gr.147686.112

Hitti-Malin, R. J., Dhaenens, C. M., Panneman, D. M., Corradi, Z., Khan, M., den Hollander, A. I., Farrar, G. J., Gilissen, C., Hoischen, A., van de Vorst, M., Bults, F., Boonen, E. G. M., Saunders, P., Roosing, S., & Cremers, F. P. M. (2022). Using single molecule Molecular Inversion Probes as a cost-effective, high-throughput sequencing approach to target all genes and loci associated with macular diseases. *Hum Mutat*, *43*(12), 2234-2250. https://doi.org/10.1002/humu.24489

Hu, S., Parker, J., & Wright, J. T. (2015). Towards Unraveling the Human Tooth Transcriptome: The Dentome. *PLOS ONE*, *10*(4), e0124801. https://doi.org/10.1371/journal.pone.0124801

Hu, Y., Smith, C. E., Richardson, A. S., Bartlett, J. D., Hu, J. C., & Simmer, J. P. (2016). MMP20, KLK4, and MMP20/KLK4 double null mice define roles for matrix proteases during dental enamel formation. *Mol Genet Genomic Med*, *4*(2), 178-196. https://doi.org/10.1002/mgg3.194

Hyun, H. K., Lee, S. K., Lee, K. E., Kang, H. Y., Kim, E. J., Choung, P. H., & Kim, J. W. (2009). Identification of a novel FAM83H mutation and microhardness of an affected molar in autosomal dominant hypocalcified amelogenesis imperfecta. *Int Endod J*, *42*(11), 1039-1043. https://doi.org/10.1111/j.1365-2591.2009.01617.x

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., & Farh, K. K. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, *176*(3), 535-548.e524. https://doi.org/10.1016/j.cell.2018.12.015

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., . . . Genome Aggregation Database, C. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434-443. https://doi.org/10.1038/s41586-020-2308-7

Khan, M., Cornelis, S. S., Pozo-Valero, M. D., Whelan, L., Runhart, E. H., Mishra, K., Bults, F., AlSwaiti, Y., AlTalbishi, A., De Baere, E., Banfi, S., Banin, E., Bauwens, M., Ben-Yosef, T., Boon, C. J. F., van den Born, L. I., Defoort, S., Devos, A., Dockery, A., . . . Cremers, F. P. M. (2020). Resolving the dark matter of ABCA4 for 1054 Stargardt disease probands through integrated genomics and transcriptomics. *Genet Med*, *22*(7), 1235-1246. https://doi.org/10.1038/s41436-020-0787-4

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., & Simmer, J. P. (2008). FAM83H Mutations in Families with Autosomal-Dominant Hypocalcified Amelogenesis Imperfecta. *The American Journal of Human Genetics*, *82*(2), 489-494. https://doi.org/https://doi.org/10.1016/j.ajhg.2007.09.020

Kim, J. W., Simmer, J. P., Hart, T. C., Hart, P. S., Ramaswami, M. D., Bartlett, J. D., & Hu, J. C. C. (2005). MMP-20 mutation in autosomal recessive pigmented hypomaturation amelogenesis imperfecta. *Journal of Medical Genetics*, *42*(3), 271. https://doi.org/10.1136/jmg.2004.024505

Kim, Y. J., Zhang, H., Lee, Y., Seymen, F., Koruyucu, M., Kasimoglu, Y., Simmer, J. P., Hu, J. C., & Kim, J. W. (2023). Novel WDR72 Mutations Causing Hypomaturation Amelogenesis Imperfecta. *J Pers Med*, *13*(2). https://doi.org/10.3390/jpm13020326

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, *46*(3), 310-315. https://doi.org/10.1038/ng.2892

Lagerström, M., Dahl, N., Nakahori, Y., Nakagome, Y., Bäckman, B., Landegren, U., & Pettersson, U. (1991). A deletion in the amelogenin gene (AMG) causes X-linked amelogenesis imperfecta (AIH1). *Genomics*, *10*(4), 971-975. https://doi.org/10.1016/0888-7543(91)90187-j

Leban, T., Trebušak Podkrajšek, K., Kovač, J., Fidler, A., & Pavlič, A. (2022). An Intron c.103-3T>C Variant of the AMELX Gene Causes Combined Hypomineralized and Hypoplastic Type of Amelogenesis Imperfecta: Case Series and Review of the Literature. *Genes (Basel)*, *13*(7). https://doi.org/10.3390/genes13071272

Lee, S. K., Seymen, F., Lee, K. E., Kang, H. Y., Yildirim, M., Tuna, E. B., Gencay, K., Hwang, Y. H., Nam, K. H., De La Garza, R. J., Hu, J. C., Simmer, J. P., & Kim, J. W. (2010). Novel WDR72 mutation and cytoplasmic localization. *J Dent Res*, *89*(12), 1378-1382. https://doi.org/10.1177/0022034510382117

Lench, N. J., & Winter, G. B. (1995). Characterisation of molecular defects in X-linked amelogenesis imperfecta (AIH1). *Hum Mutat*, *5*(3), 251-259. https://doi.org/10.1002/humu.1380050310

Mc Clinton, B., Corradi, Z., McKibbin, M., Panneman, D. M., Roosing, S., Boonen, E. G. M., Ali, M., Watson, C. M., Steel, D. H., Cremers, F. P. M., Inglehearn, C. F., Hitti-Malin, R. J., & Toomes, C. (2023). Effective smMIPs-Based Sequencing of Maculopathy-Associated Genes in Stargardt Disease Cases and Allied Maculopathies from the UK. *Genes*, *14*(1), 191. https://www.mdpi.com/2073-4425/14/1/191

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, *11*(1), 31-46. https://doi.org/10.1038/nrg2626

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., & Mighell, A. J. (2021). Spectrum of pathogenic variants and founder effects in amelogenesis imperfecta associated with MMP20. *Human Mutation*, *42*(5), 567-576. https://doi.org/https://doi.org/10.1002/humu.24187

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., & Mighell, A. J. (2021). Spectrum of pathogenic variants and founder effects in amelogenesis imperfecta associated with MMP20. *Hum Mutat*, *42*(5), 567-576. https://doi.org/10.1002/humu.24187

O'Sullivan, J., Bitu, Carolina C., Daly, Sarah B., Urquhart, Jill E., Barron, Martin J., Bhaskar, Sanjeev S., Martelli-Júnior, H., dos Santos Neto, Pedro E., Mansilla, Maria A., Murray, Jeffrey C., Coletta, Ricardo D., Black, Graeme C. M., & Dixon, Michael J. (2011). Whole-Exome Sequencing Identifies FAM20A Mutations as a Cause of Amelogenesis Imperfecta and Gingival Hyperplasia Syndrome. *The American Journal of Human Genetics*, *88*(5), 616-620. https://doi.org/https://doi.org/10.1016/j.ajhg.2011.04.005

Oud, M. S., Ramos, L., O'Bryan, M. K., McLachlan, R. I., Okutman, Ö., Viville, S., de Vries, P. F., Smeets, D., Lugtenberg, D., Hehir-Kwa, J. Y., Gilissen, C., van de Vorst, M., Vissers, L., Hoischen, A., Meijerink, A. M., Fleischer, K., Veltman, J. A., & Noordam, M. J. (2017). Validation and application of a novel integrated genetic screening method to a cohort of 1,112 men with idiopathic azoospermia or severe oligozoospermia. *Hum Mutat*, *38*(11), 1592-1605. https://doi.org/10.1002/humu.23312

Panneman, D. M., Hitti-Malin, R. J., Holtes, L. K., de Bruijn, S. E., Reurink, J., Boonen, E. G. M., Khan, M. I., Ali, M., Andréasson, S., De Baere, E., Banfi, S., Bauwens, M., Ben-Yosef, T., Bocquet, B., De Bruyne, M., de la Cerda, B., Coppieters, F., Farinelli, P., Guignard, T., . . . Roosing, S. (2023). Cost-effective sequence analysis of 113 genes in 1,192 probands with retinitis pigmentosa and Leber congenital amaurosis. *Front Cell Dev Biol*, *11*, 1112270. https://doi.org/10.3389/fcell.2023.1112270

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., & Nejentsev, S. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, *28*(21), 2747-2754. https://doi.org/10.1093/bioinformatics/bts526

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., . . . Bloch-Zupan, A. (2016). A targeted next-generation sequencing assay for the molecular diagnosis of genetic disorders with orodental involvement. *J Med Genet*, *53*(2), 98-110. https://doi.org/10.1136/jmedgenet-2015-103302

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*, *17*(5), 405-424. https://doi.org/10.1038/gim.2015.30

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol*, *29*(1), 24-26. https://doi.org/10.1038/nbt.1754

Sedano, H. O. (1975). Congenital oral anomalies in argentinian children. *Community Dent Oral Epidemiol*, *3*(2), 61-63. https://doi.org/10.1111/j.1600-0528.1975.tb00281.x

Seymen, F., Kim, Y. J., 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, Y. J., Lee, S. H., Lee, Z. H., Zhang, H., Hu, J. C., Simmer, J. P., Cho, E. S., & Kim, J. W. (2016). Recessive Mutations in ACPT, Encoding Testicular Acid Phosphatase, Cause Hypoplastic Amelogenesis Imperfecta. *Am J Hum Genet*, *99*(5), 1199-1205. https://doi.org/10.1016/j.ajhg.2016.09.018

Seymen, F., Lee, K. E., Tran Le, C. G., Yildirim, M., Gencay, K., Lee, Z. H., & Kim, J. W. (2014). Exonal deletion of SLC24A4 causes hypomaturation amelogenesis imperfecta. *J Dent Res*, *93*(4), 366-370. https://doi.org/10.1177/0022034514523786

Simmer, J. P., Hu, J. C. C., Hu, Y., Zhang, S., Liang, T., Wang, S.-K., Kim, J.-W., Yamakoshi, Y., Chun, Y.-H., Bartlett, J. D., & Smith, C. E. (2021). A genetic model for the secretory stage of dental enamel formation. *Journal of Structural Biology*, *213*(4), 107805. https://doi.org/https://doi.org/10.1016/j.jsb.2021.107805

Simmer, S. G., Estrella, N. M., Milkovich, R. N., & Hu, J. C. (2013). Autosomal dominant amelogenesis imperfecta associated with ENAM frameshift mutation p.Asn36Ilefs56. *Clin Genet*, *83*(2), 195-197. https://doi.org/10.1111/j.1399-0004.2012.01887.x

Smith, C. E. L., Poulter, J. A., Antanaviciute, A., Kirkham, J., Brookes, S. J., Inglehearn, C. F., & Mighell, A. J. (2017). Amelogenesis Imperfecta; Genes, Proteins, and Pathways. *Front Physiol*, *8*, 435. https://doi.org/10.3389/fphys.2017.00435

Song, J. S., Lee, Y., Shin, T. J., Hyun, H. K., Kim, Y. J., & Kim, J. W. (2022). Identification of a Novel FAM83H Mutation and Management of Hypocalcified Amelogenesis Imperfecta in Early Childhood. *Children (Basel)*, *9*(3). https://doi.org/10.3390/children9030429

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Witkop, C. J., Jr. (1988). Amelogenesis imperfecta, dentinogenesis imperfecta and dentin dysplasia revisited: problems in classification. *J Oral Pathol*, *17*(9-10), 547-553. https://doi.org/10.1111/j.1600-0714.1988.tb01332.x

Wright, J. T., Frazier-Bowers, S., Simmons, D., Alexander, K., Crawford, P., Han, S. T., Hart, P. S., & Hart, T. C. (2009). Phenotypic variation in FAM83H-associated amelogenesis imperfecta. *J Dent Res*, *88*(4), 356-360. https://doi.org/10.1177/0022034509333822

Wright, J. T., Hong, S. P., Simmons, D., Daly, B., Uebelhart, D., & Luder, H. U. (2008). DLX3 c.561_562delCT mutation causes attenuated phenotype of tricho-dento-osseous syndrome. *Am J Med Genet A*, *146a*(3), 343-349. https://doi.org/10.1002/ajmg.a.32132

Wright, J. T., Torain, M., Long, K., Seow, K., Crawford, P., Aldred, M. J., Hart, P. S., & Hart, T. C. (2011). Amelogenesis imperfecta: genotype-phenotype studies in 71 families. *Cells Tissues Organs*, *194*(2-4), 279-283. https://doi.org/10.1159/000324339

Zhang, H., Hu, Y., Seymen, F., Koruyucu, M., Kasimoglu, Y., Wang, S. K., Wright, J. T., Havel, M. W., Zhang, C., Kim, J. W., Simmer, J. P., & Hu, J. C. (2019). ENAM mutations and digenic inheritance. *Mol Genet Genomic Med*, *7*(10), e00928. https://doi.org/10.1002/mgg3.928

Table 1: Genes selected for inclusion in the smMIPs reagent. Genomic coordinates are provided according to the human reference genome build hg19. OMIM: Online Mendelian Inheritance in Man.

Gene name	Gene symbol	Strand	OMIM	Genomic coordinates	Cytoband
LAMININ, BETA-3	*LAMB3:NM_000228.3*	-	150310	chr1:209,788,218-209,825,820	1q32.2
INTEGRIN, BETA-6	*ITGB6:NM_000888.5*	-	147558	chr2:160,958,233-161,056,589	2q24.2
AMELOTIN	*AMTN:NM001286731.2*	+	610912	chr4:71,384,298-71,398,459	4q13.3
AMELOBLASTIN	*AMBN:NM_016519*	+	601259	chr4:71,457,975-71,473,004	4q13.3
ENAMELIN	*ENAM:NM_031889.3*	+	606585	chr4:71,494,461-71,512,536	4q13.3
ODONTOGENESIS-ASSOCIATED PHOSPHOPROTEIN	*ODAPH:NM_001206981*	+	614829	chr4:76,481,258-76,491,103	4q21.1
FAMILY WITH SEQUENCE SIMILARITY 83	*FAM83H:NM_198488.5*	-	611927	chr8:144,806,103-144,815,914	8q24.3
COLLAGEN, TYPE XVII, ALPHA-1	*COL17A1:NM_000494.4*	-	113811	chr10:105,791,046-105,845,638	10q25.1
RECEPTOR EXPRESSED IN LYMPHOID TISSUES	*RELT:NM_152222.2*	+	611211	chr11:73,087,405-73,108,519	11q13.4
MATRIX METALLOPROTEINASE 20	*MMP20:NM_004771.4*	-	604629	chr11:102,447,566-102,496,063	11q22.2
G PROTEIN-COUPLED RECEPTOR 68	*GPR68:NM001177676.2*	-	601404	chr14:91,698,876-91,710,852	14q32.11
SOLUTE CARRIER FAMILY 24 (SODIUM/POTASSIUM/CALCIUM EXCHANGER), MEMBER 4	*SLC24A4:NM_153646*	+	609840	chr14:92,790,152-92,967,825	14q32.12
WD REPEAT-CONTAINING PROTEIN 72	*WDR72:NM_182758.4*	-	613214	chr15:53,805,938-54,051,859	15q21.3
TRANSCRIPTION FACTOR Sp6	*SP6:NM_199262*	-	608613	chr17:45,922,280-45,928,516	17q21.32
DISTAL-LESS HOMEOBOX 3	*DLX3:NM_005220.3*	-	600525	chr17:48,067,369-48,072,588	17q21.33
FAMILY WITH SEQUENCE SIMILARITY 20, MEMBER A	*FAM20A:NM_017565.4*	-	611062	chr17:66,531,257-66,597,095	17q24.2
ACID PHOSPHATASE 4	*ACP4:NM_033068.3*	+	606362	chr19:51,293,672-51,298,481	19q13.33
KALLIKREIN-RELATED PEPTIDASE 4	*KLK4: NM_004917.4*	-	603767	chr19:51,409,608-51,413,994	19q13.41
AMELOGENIN	*AMELX:NM_182680.1*	+	300391	chrX:11,311,533-11,318,881	Xp22.2

Table 2: Details of the variants identified in the X-linked families. Variants are reported as hemizygous (hemi) in males and heterozygous (het) in females. Nomen: Nomenclature P: Pathogenic. VUS: Variant of uncertain significance. LP: Likely Pathogenic. CADD v.1.3: Combined annotation-dependent depletion. gnomAD v.2.1.1: Genome aggregation database. Nomenclature is reported according to the human reference genome build hg19.

ID	Gene	Genomic nomen	Transcript nomen	Predicted protein nomen	Zygosity	CADD	gnomAD	Pathogenicity	Reference
Fam 8	*AMELX*	X-11314944G>A	c.100G>A	p.(Glu34Lys)	Het, female	33	Absent	VUS	
Fam 9	*AMELX*	X-11316220T>C	c.103-3T>C	p.?	Hemi, male	15	Absent	LP	(Leban et al., 2022)
Fam 21	*AMELX*	X-11316953GC>G	c.472del	p.(Pro158Hisfs*31)	Hemi, male		Absent	P	(Lench & Winter, 1995)
Fam 27	*AMELX*	X-11316363AC>A	c.152del	p.(Pro52Leufs*2)	Het, female		Absent	P	(Aldred et al., 1992)
Fam 32	*AMELX*	X-11316220T>C	c.103-3T>C	p.?	Het, female	15	Absent	LP	(Leban et al., 2022)
Fam 38	*AMELX*	X-11316363	c.152del	p.(Pro52Leufs*2)	Hemi, male		Absent	P	(Aldred et al., 1992)
Fam 65	*AMELX*	X-11316927	c.446A>C	p.(Gln149Pro)	Hemi, male	19	0.00001	VUS	

Table 3: Details of the variants identified in the dominant families. All variants listed below are heterozygous. Nomen: Noemclature. P: Pathogenic. VUS: Variant of uncertain significance. LP: Likely Pathogenic. CADD v.1.3: Combined annotation-dependent depletion. gnomAD v.2.1.1: Genome aggregation database. Nomenclature is reported according to the human reference genome build hg19.

ID	Gene	Genomic nomen	Transcript nomen	Predicted protein nomen	CADD	gnomAD	Pathogenicity	Reference
Fam 1	*COL17A1*	10-105811247C>T	c.2030G>A	p.(Gly677Asp)	26	Absent	LP	
Fam 4	*COL17A1*	10- 105798865del	c.2912del	p.(Pro971Glnfs*95)	33	Absent		
Fam 5	*DLX3*	17- 48069185_48069186del	c.561_562del	p.(Tyr188Glnfs*13)		Absent	P	(Wright et al., 2008)
Fam 7	*COL17A1*	10-105796271G>A	c.3397C>T	p.(Arg1133Cys)	33	Absent	VUS	
Fam 10	*FAM83H*	8-144811340G>A	c.601C>T	p.(Gln201*)	40	0.00168	VUS	
Fam 11	*COL17A1*	10-105796802C>T	c.3277+1G>A	p.?	35	Absent	P	
Fam 13	*COL17A1*	10-105811266C>T	c.2011G>A	p.(Gly671Ser)	25	Absent	LP	
Fam 14	*ENAM*	4-71503505A>T	c.535-2A>T	p.?	33	Absent	LP	(Wright et al., 2011)
Fam 16	*AMBN*	4-71465278C>G	c.209C>G	p.(Ser70*)	36	0.00010	P	
Fam 18	*COL17A1*	10-105830245_105830254del	c.541_550del	p.(Asn181Profs*13)	28	Absent	P	
Fam 22	*FAM83H*	8-144810257	c.1374C>A	p.(Tyr458*)	38	Absent	P	(El-Sayed et al., 2010)
Fam 24	*COL17A1*	10-105795287del	c.3456del	p.(Pro1154Leufs*97)	21	0.00002	P	
Fam 26	*COL17A1*	105793715_105793716del	c.4147_4148del	p.(Ser1383Hisfs*71)	34	Absent	P	
Fam 29	*FAM83H*	8-144811340	c.601C>T	p.(Gln201*)	36	0.00168	VUS	
Fam 30	*ENAM*	4-71497385AG>A	c.55-1del	p.?	33	Absent	LP	
Fam 31	*FAM83H*	8-144811340G>A	c.601C>T	p.(Gln201*)	36	0.00168	VUS	
Fam 33	*ENAM*	4-71497438TA>T	c.106del	p.(Asn36Ilefs*22)		Absent	P	(Simmer et al., 2013)
Fam 34	*ENAM*	4-71501548G>A	c.472-1G>A	p.?	25	Absent	LP	
Fam 35	*ENAM*	4-71497438TA>T	c.106del	p.(Asn36Ilefs*22)		Absent	P	(Simmer et al., 2013)
Fam 36	*COL17A1*	10-105816859C>A	c.1339G>T	p.(Gly447Cys)	24	0.00070	VUS	
Fam 39	*FAM83H*	8-144809494	c.2137C>T	p.(Gln713*)	36	Absent	LP	
Fam 40	*AMBN*	4-71465278	c.209C>G	p.(Ser70*)	36	0.00010	P	
Fam 41	*DLX3*	17-48072078G>C	c.285C>G	p.(Tyr95*)	36	Absent	P	
Fam 43	*FAM83H*	8-144810439	c.1192C>T	p.(Gln398*)	36	Absent	LP	(Hart et al., 2009)
Fam 44	*AMBN*	4-71459104	c.76G>A	p.(Ala26Thr)	26	Absent	LP	
Fam 45	*FAM83H*	8-144810277	c.1354C>T	p.(Gln452*)	37	Absent	LP	(Hyun et al., 2009)
Fam 48	*COL17A1*	10-105831793G>A	c.460C>T	p.(Arg154*)	36	Absent	P	(Condrat et al., 2018)
Fam 49	*COL17A1*	10-105795287del	c.3456del	p.(Pro1154Leufs*97)	21	0.00002	P	
Fam 55	*FAM83H*	8-144810268	c.1363C>T	p.(Gln455*)	37	Absent	P	(Song et al., 2022)
Fam 56	*COL17A1*	10-105795035G>A	c.3605C>T	p.(Ser1202Leu)	27	0.00001	VUS	
Fam 57	*FAM83H*	8-144810658	c.973C>T	p.(Arg325*)	36	Absent	P	(Kim et al., 2008)
Fam 63	*FAM83H*	8-144810707	c.923_924del	p.(Leu308Argfs*16)		Absent	LP	(Wright et al., 2009)

Table 4: Details of the variants identified in the recessive families. Families reported with single variant are homozygous, while families reported with two variants are compound heterozygous. Nomen: nomenclature, P: Pathogenic, VUS: Variant of unknown significance, LP: Likely Pathogenic, CADD v.1.3: Combined annotation-dependent depletion, gnomAD v.2.1.1: Genome aggregation database. Nomenclature is reported according to human reference genome build hg19.

ID	Gene	Genomic nomen	Transcript nomen	Predicted protein nomen	CADD	gnomAD	Pathogenicity	Reference
Fam 2	*ACP4*	19-51294940	c.331C>T	p.(Arg111Cys)	26	0.00012	LP	(Seymen et al., 2016)
Fam 3	*FAM20A*	17-66538244	c.987_990del	p.(Cys330Alafs*51)	33	Absent	LP	
Fam 6	*MMP20*	11-102480660C>G	c.625G>C	p.(Glu209Gln)	30	0.00006	LP	(Nikolopoulos et al., 2021)
Fam 12	*WDR72*	15-53908070	c.2332dupA	p.(Met778Asnfs*4)	24	0.00001	P	(Kim et al., 2023)
	WDR72	15-54025229	c.118C>T	p.(Gln40*)	40	0.00002	P	
Fam 15	*MMP20*	11- 102465490T>A	c.954-2A>T	p.?	25	0.00110	P	(Kim et al., 2005)
Fam 17	*MMP20*	11-102480660C>G	c.625G>C	p.(Glu209Gln)	27	0.00007	P	(Nikolopoulos et al., 2021)
Fam 19	*AMBN*	4-71465278	c.209C>G	p.(Ser70*)	36	0.00010	P	
Fam 20	*MMP20*	11-102480660C>G	c.625G>C	p.(Glu209Gln)	27	0.00007	P	(Nikolopoulos et al., 2021)
Fam 23	*SLC24A4*	Exon 15, 16, 17 deletion						(Seymen et al., 2014)
Fam 25	*MMP20*	11:102479824T>A	c.655A>T	p.(Asn219Tyr)	27	Absent	VUS	
Fam 28	*AMBN*	4-71465278	c.209C>G	p.(Ser70*)	36	0.00010	P	
	AMBN	4-71467135	c.295T>C	p.(Tyr99His)	26	0.00008	LP	
Fam 37	*ITGB6*	2-161052847C>A	c.226G>T	p.(Glu76*)	39	0.00001	P	
Fam 42	*RELT*	11-73101947	c.268T>C	p.(Cys90Arg)	26	Absent	LP	
Fam 46	*MMP20*	11-102479803G>A	c.676C>T	p.(His226Tyr)	28	Absent	LP	
Fam 47	*MMP20*	11- 102465490T>A	c.954-2A>T	p.?	25	0.00110	P	(Kim et al., 2005)
Fam 50	*ACP4*	19- 51294940	c.331C>T	p.(Arg111Cys)	33	0.00012	P	(Seymen et al., 2016)
	ACP4	19- 51295042	c.433delC	p.(Val146Trpfs*7)	32	Absent	P	
Fam 51	*SLC24A4*	14-92920382	c.1019T>C	p.(Leu340Pro)	25	Absent	VUS	
Fam 52	*WDR72*	15-53907717	c.2686C>T	p.(Arg896*)	36	0.00003	P	
Fam 53	*FAM20A*	17-66535488	c.1351C>T	p.(Gln451*)	47	Absent	LP	
Fam 54	*WDR72*	15-53994432 CAT>C	c.1467_1468del	p.(Val491Aspfs*8)		0.00007	P	(Lee et al., 2010)
Fam 58	*MMP20*	11- 102465490T>A	c.954-2A>T	p.?	25	0.00110	P	(Kim et al., 2005)
	MMP20	11-102477286del	c.933del	p.(Glu311Aspfs*59)	32	0.00001	P	
Fam 59	*WDR72*	15-53992111 GCA>G	c.1600_1601del	p.(Cys534Argfs*2)		Absent	P	
	WDR72	15-53907897	c.2506G>T	p.(Glu836*)	37	Absent	P	
Fam 60	*ACP4*	19-51297211	c.845T>C	p.(Met282Thr)	27	Absent	VUS	
Fam 61	*FAM20A*	17-66551883	c.406C>T	p.(Arg136*)	39	0.00004	P	(O'Sullivan et al., 2011)
	FAM20A	17-66538120	c.1109+6T>G	p.?	23	Absent	LP	
Fam 62	*MMP20*	11- 102465490T>A	c.954-2A>T	p.?	25	0.00110	P	(Kim et al., 2005)
Fam 64	*ENAM*	4-71508226G>A	c.1083G>A	p.(Trp361*)	36	Absent	LP	
Fam 66	*WDR72*	15-54003125	c.883G>A	p.(Ala295Thr)	20	Absent	LP	

Table 5: Details of the variants identified in the four participants presented with potentially relevant variants in 2 genes known to be associated with NSAI. Nomen: Nomenclature, P: Pathogenic, LP: Likely Pathogenic. VUS: Variant of unknown significance. LB: Likely benign. Het: Heterozygous. Hom: Homozygous. CADD v.1.3: Combined annotation-dependent depletion. gnomAD v.2.1.1: Genome aggregation database. ClinVar: Public archive of interpretations of clinically relevant variants. Nomenclature is reported according to human reference genome build hg19.

ID	Gene	Genomic nomen	Transcript nomen	Predicted protein nomen	Zygosity	CADD	gnomAD	Pathogenicity	ClinVar
Fam 25	*COL17A1*	10-105799724A>T	c.2788+7T>A	p.?	Het	17	0.00006	VUS	
	MMP20	11-102479824T>A	c.655A>T	p.(Asn219Tyr)	Hom	26	Absent	VUS	
Fam 62	*MMP20*	11- 102465490T>A	c.954-2A>T	p.?	Hom	25	0.00110	P	
	COL17A1	10-105830262G>A	c.529C>T	p.(Arg177Trp)	Het	27	0.00005	VUS	
Fam 4	*MMP20*	11-102477309 C>T	c.910G>A	p.(Ala304Thr)	Het	25	0.00156	VUS	VCV000301942.9
	MMP20	11-102495959 G>A	c.92C>T	p.(Pro31Leu)	Het	21	0.00460	LB	VCV000301957.9
	COL17A1	10- 105798865del	c.2912del	p.(Pro971Glnfs*95)	Het		Absent	LP	
Fam 18	*COL17A1*	10-105830245_105830254del	c.541_550del	p.(Asn181Profs*13)	Het	28	Absent	P	
	MMP20	11-102477309 C>T	c.910G>A	p.(Ala304Thr)	Het	25	0.00156	VUS	VCV000301942.9
	MMP20	11-102495959 G>A	c.92C>T	p.(Pro31Leu)	Het	21	0.00460	LB	VCV000301957.9

Figure 1

A. Variant classification for variants detected by smMIPs screening of the AI cohort. A total of 58 disease-causing variants were detected in 66 families. According to the ACMG classification, 29 of these are classified as pathogenic, 20 are likely pathogenic and 9 are VUS. Nomenclature is reported according to the human reference genome build hg19. **B.** The types of mutations detected in smMIPs-AI cohort screening. Of the 58 disease causing variants detected in 66 families, 17 are missense, 17 are frameshift, 17 are nonsense mutations predicted to lead to a premature termination codon, six are predicted to alter splice sites and one is a large deletion. Nomenclature is reported according to the human reference genome build hg19. **C.** Genetic diagnosis of the AI cohort by smMIPs screening. The most commonly identified genes with variants were *COL17A1* (12), *FAM83H* (10) and *MMP20* (9). The numbers of observations in each category are given in brackets.

Figure 1A

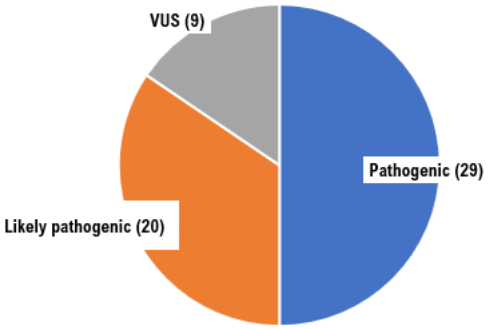


Figure 1B

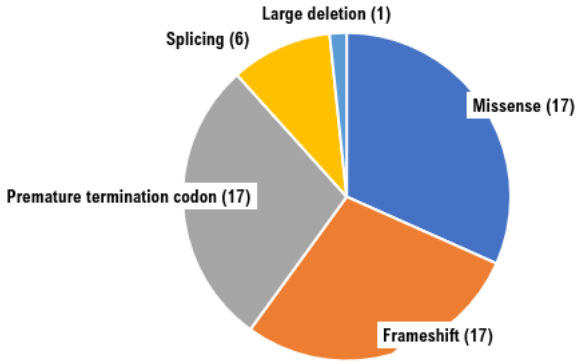


Figure 1C

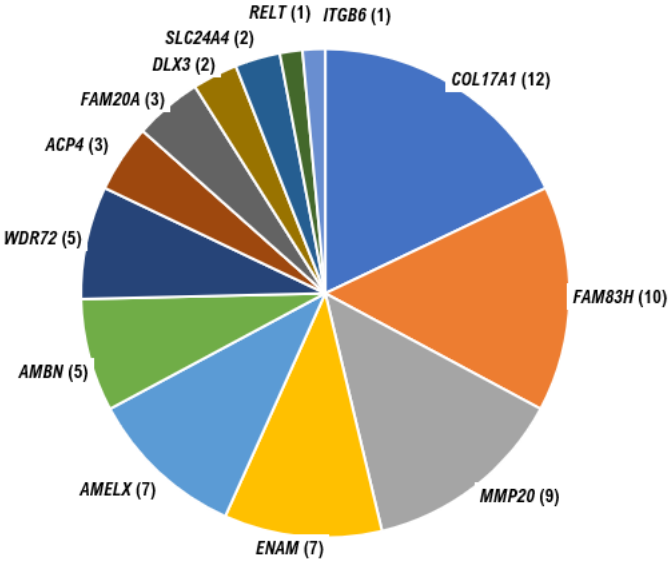
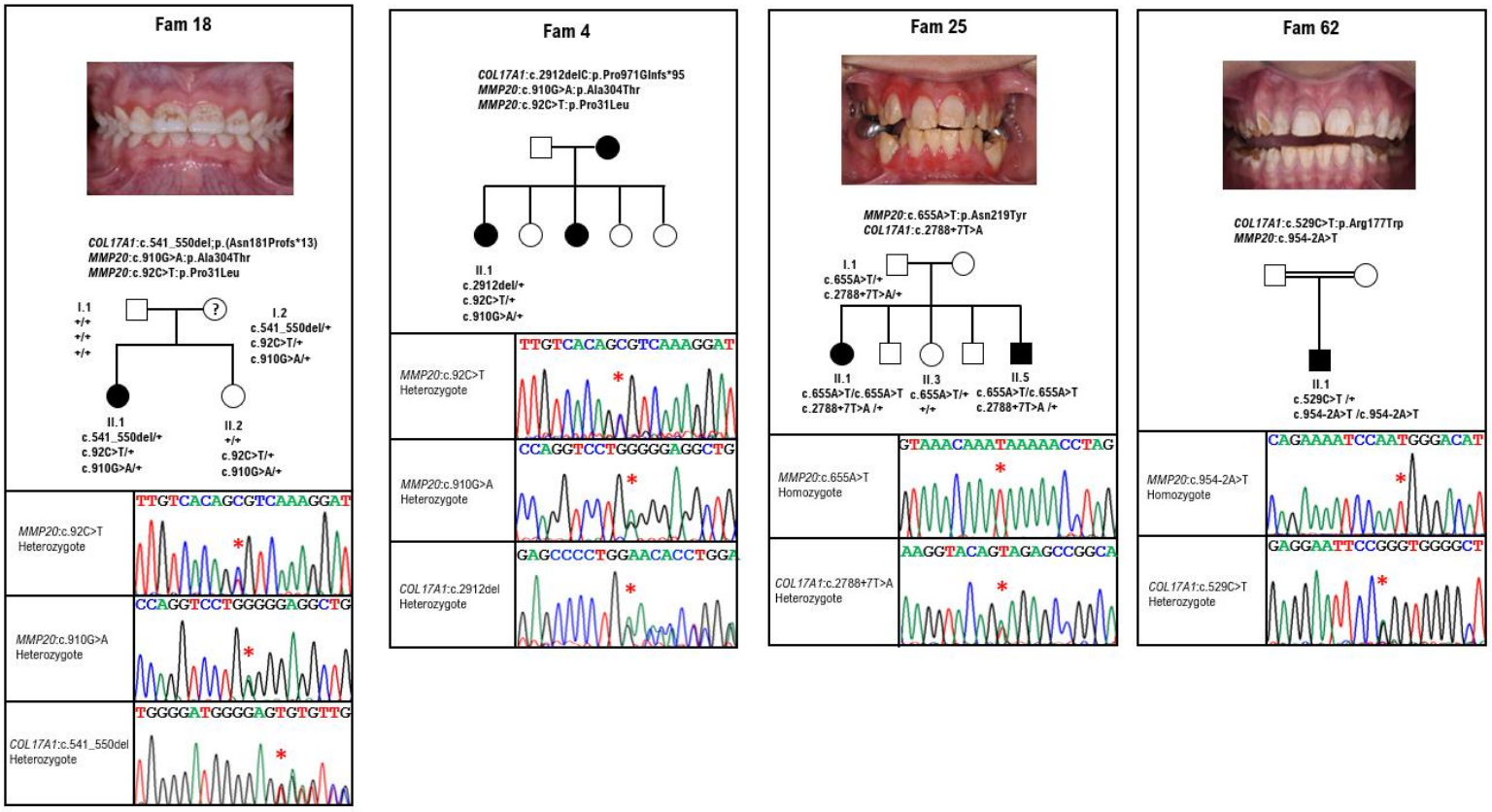


Figure 2: Pedigrees of four putative digenic families recruited for this study and, where available, intraoral images of the proband in each family. Sanger sequencing chromatograms from the proband for each family are displayed beneath each pedigree. A question mark in the pedigree denotes an individual with possible AI who has not been clinically assessed. Probands from families 4 and 18 were diagnosed as having hypoplastic AI (no image is available from family 4), consistent with AI due to a dominant heterozygous *COL17A1* variant, and are reported elsewhere (Hany et al., 2023). Probands from families 25 and 62 were diagnosed as having hypomineralised AI, consistent with AI due to recessive *MMP20* variants.



CHAPTER 3 Novel ameloblastin variants, contrasting Amelogenesis Imperfecta phenotypes

Hany U, Watson CM, Liu L, Nikolopoulos G, Smith CEL, Poulter JA, Brown CJ, Patel A, Rodd HD, Balmer R, Harfoush A, Al-Jawad M, Inglehearn CF, Mighell AJ. Novel Ameloblastin Variants, Contrasting Amelogenesis Imperfecta Phenotypes. J Dent Res. 2024 Jan;103(1):22-30. doi: 10.1177/00220345231203694. Epub 2023 Dec 6. PMID: 38058155; PMCID: PMC10734210.

This chapter introduces the subsequent paper, providing crucial background information, outlining the rationale for the research, detailing methodologies employed, and acknowledging the significant contributions made by other collaborators to enrich the overall work.

3.1 Research Rationale

This paper summarizes findings specifically related to AI caused by variants in the *AMBN* gene. Reported data were from both the smMIPs screen and subsequent WES screening of cases unsolved by smMIPs analysis. AMBN plays a critical role in amelogenesis, and biallelic variants in *AMBN* are known to cause AR AI (Poulter, Murillo, et al., 2014). However, the suggestion that dominant AI can be caused by monoallelic *AMBN* variants remains controversial. There was only one study that reported a heterozygous *AMBN* variant p.(Pro357Ser) in patients with a mixed AI and DI (dentinogenesis imperfecta) phenotype that segregated in a large family with a clear dominant pattern of inheritance (Lu et al., 2018). In our study, we detected five novel *AMBN* variants, in combinations that constituted both monoallelic and biallelic genotypes, in eleven families with contrasting AI phenotypes that correlated with their underlying genotypes. In this paper we discussed the potential molecular mechanisms behind these differing phenotypes and evaluated possible disease mechanisms for both dominant and recessive AI caused by *AMBN* mutation.

To validate the hypothesis that *AMBN* variants can cause dominantly inherited AI we further investigated monoallelic cases by sequencing the entire *AMBN* gene, using long-read nanopore reads, to look for second variants in the non-coding regions (introns and promoters) and determine the phase (whether or not they are arranged in *trans*) when multiple variants were found. Using these long-read data we were able to carry out haplotype analysis of three variants that were shared by multiple families, to confirm inheritance from a common ancestor.

3.2 Research Contribution

Patients were diagnosed and recruited either by my supervisor Dr Alan Mighell or by other dental colleagues who collaborate with him. I carried out the wet laboratory work including smMIPs, WES, and long-read sequencing for all the patients involved in this study. As previously and again under my supervision, Lu Liu performed the wet laboratory work for the Sanger sequencing. I analysed the data generated using the three sequencing techniques. I wrote the first draft of the paper and managed the master copy, entering feedback from others during the writing process.

3.2.1 Additional methodology

Further details of the methodology are reported in the manuscript.

Research Reports: Clinical

Novel Ameloblastin Variants, Contrasting Amelogenesis Imperfecta Phenotypes

Journal of Dental Research
2024, Vol. 103(1) 22–30

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00220345231203694
journals.sagepub.com/home/jdr

U. Hany[1], C.M. Watson[1,2], L. Liu[1,3], G. Nikolopoulos[1], C.E.L. Smith[1], J.A. Poulter[1], C.J. Brown[4], A. Patel[5], H.D. Rodd[6], R. Balmer[3], A. Harfoush[3], M. Al-Jawad[3], C.F. Inglehearn[1*], and A.J. Mighell[3*]

Abstract

Amelogenesis imperfecta (AI) comprises a group of rare, inherited disorders with abnormal enamel formation. Ameloblastin (AMBN), the second most abundant enamel matrix protein (EMP), plays a critical role in amelogenesis. Pathogenic biallelic loss-of-function *AMBN* variants are known to cause recessive hypoplastic AI. A report of a family with dominant hypoplastic AI attributed to AMBN missense change p.Pro357Ser, together with data from animal models, suggests that the consequences of *AMBN* variants in human AI remain incompletely characterized. Here we describe 5 new pathogenic *AMBN* variants in 11 individuals with AI. These fall within 3 groups by phenotype. Group 1, consisting of 6 families biallelic for combinations of 4 different variants, have yellow hypoplastic AI with poor-quality enamel, consistent with previous reports. Group 2, with 2 families, appears monoallelic for a variant shared with group 1 and has hypomaturation AI of near-normal enamel volume with pitting. Group 3 includes 3 families, all monoallelic for a fifth variant, which are affected by white hypoplastic AI with a thin intact enamel layer. Three variants, c.209C>G; p.(Ser70*) (groups 1 and 2), c.295T>C; p.(Tyr99His) (group 1), and c.76G>A; p.(Ala26Thr) (group 3) were identified in multiple families. Long-read *AMBN* locus sequencing revealed these variants are on the same conserved haplotype, implying they originate from a common ancestor. Data presented therefore provide further support for possible dominant as well as recessive inheritance for *AMBN*-related AI and for multiple contrasting phenotypes. In conclusion, our findings suggest pathogenic *AMBN* variants have a more complex impact on human AI than previously reported.

Keywords: AMBN, amelogenesis, dental enamel, hypoplastic AI, founder effect, X-ray microtomography

Introduction

Enamel is formed when ameloblasts secrete, then mineralize, an extracellular matrix (ECM) composed of enamel matrix proteins (EMPs), in a process known as amelogenesis. Throughout the secretory stage of amelogenesis, ameloblasts secrete the EMPs amelogenin (AMELX), ameloblastin (AMBN), enamelin (ENAM), and amelotin (AMTN) and the matrix modifier matrix metallopeptidase 20 (MMP20) (Lee et al. 1996). In the later maturation stage, a further matrix modifier, kallikrein-related peptidase 4 (KLK4), is also secreted (Smith et al. 2017; Pandya and Diekwisch 2021). EMPs play an essential role in the biomineralization and structural organization of enamel (Bartlett et al. 2006).

Amelogenesis imperfecta (AI) describes a heterogeneous group of Mendelian disorders causing abnormal amelogenesis, affecting all teeth of both dentitions (Smith et al. 2017). Reported prevalence ranges between 1 in 700 (Sweden) and 1 in 14,000 (United States) (Bäckman and Holm 1986; Witkop 1988). Poor aesthetics and early functional failure create considerable challenges for affected individuals and those providing care. AI can be isolated or part of syndromic conditions, with many genes implicated (Smith et al. 2017; Wright 2023).

AMBN, a phosphorylated glycoprotein, is the second most abundant EMP after AMELX. The *AMBN* gene encodes a 447–amino acid protein that is acidic and proline rich (15.2%), a characteristic shared with other EMPs (Krebsbach et al. 1996;

[1]Leeds Institute of Medical Research, University of Leeds, St. James's University Hospital, Leeds, UK
[2]North East and Yorkshire Genomic Laboratory Hub, Central Lab, St. James's University Hospital, Leeds, UK
[3]School of Dentistry, Clarendon Way, University of Leeds, Leeds, UK
[4]Birmingham Dental Hospital, Mill Pool Way, Edgbaston, Birmingham, UK
[5]LCRN West Midlands Core Team, NIHR Clinical Research Network (CRN), Birmingham Research Park (West Wing), Edgbaston, Birmingham, UK
[6]Academic Unit of Oral Health Dentistry and Society, School of Clinical Dentistry, University of Sheffield, Sheffield, S Yorks, UK
*Joint senior authors.

A supplemental appendix to this article is available online.

Corresponding Author:
A.J. Mighell, School of Dentistry, University of Leeds, Worsley Building, Leeds, LS2 9LU, UK.
Email a.j.mighell@leeds.ac.uk

Pandya and Diekwisch 2021). Proline residues in EMPs are understood to inhibit formation of secondary structures such as α-helices and β-sheets. This makes EMPs intrinsically disordered proteins that do not form stable 3-dimensional (3D) structures but instead exist in heterogeneous oligomeric states (Wald et al. 2013; Vetyskova et al. 2020). This characteristic is important for amelogenesis (Stakkestad et al. 2017). Recombinant human AMBN fails to oligomerize when exon 5 is deleted (Wald et al. 2013) and mice homozygous for a deletion of *Ambn* exons 5 and 6, expressing truncated AMBN protein, produce very thin enamel (Fukumoto et al. 2005).

AMBN is therefore a strong candidate gene for involvement in human AI. In 2014, we reported a family with recessive hypoplastic AI due to an in-frame homozygous biallelic deletion of *AMBN* exon 6 (Poulter et al. 2014). Two further recessive AI families with biallelic *AMBN* pathogenic variants have since been reported (Prasad et al. 2016; Liang et al. 2019). A report of a large dominant family where AI and dentinogenesis imperfecta (DI) segregated with a heterozygous AMBN missense variant p.(Pro357Ser) challenged our understanding of *AMBN*-associated disease (Lu et al. 2018). Liang and colleagues suggested that, given the mixed phenotype, there could also be a variant in *DSPP*, a gene linked to *AMBN* on chromosome 4, which contains a region poorly covered by whole-exome sequencing (Liang et al. 2019). No further dominant pathogenic *AMBN* variants have been reported. This raises the possibility that the consequences of *AMBN* variants in human AI remain incompletely characterized.

Here, we report 5 novel *AMBN* variants in 11 individuals with AI that can be divided into 3 clinical groups. One has a dominant family history spanning 4 generations, and the likely causative variant in this family was also identified as monoallelic/heterozygous in 2 other apparently unrelated individuals with isolated AI. These data provide further evidence suggesting *AMBN* variants can cause both dominant and recessive AI with variations in clinical phenotypes.

Materials and Methods

Patients were recruited though UK dental clinics in accordance with the principles of the Declaration of Helsinki (ethical approval REC 13/YH/0028). Genomic DNA was isolated from saliva or from peripheral blood by standard approaches as detailed in the Appendix methods.

Proband genomic DNA was analyzed by short-read next generation sequencing of either whole-exome sequencing (WES) or single-molecule molecular inversion probes (smMIPs) data generated on HiSeq 3000, NextSeq 500, or NextSeq 2000 sequencers (Illumina). Further details of methods used in library preparation and sequence analysis are given in the Appendix methods. The pathogenicity status of detected variants was classified according to American College of Medical Genetics and Genomics (ACMG) guidelines using Franklin (https://franklin.genoox.com) (Richards et al. 2015).

Long-read sequencing was carried out on a Flongle flowcell (R.9.4.1), using a MinION (ONT) device running MinKNOW, to analyze long-range polymerase chain reaction (PCR) products amplified by SequelPrep polymerase (ThermoFisher Scientific), following the manufacturer's guidelines. Methods used in sequence analysis are detailed in the Appendix methods. Haplotypes were defined after selection of reference and non-reference-matching nucleotides at the positions being examined, using the Jvarkit tool biostar214299 (http://lindenb.github.io/jvarkit/Biostar214299.html) (Lindenbaum 2015). Aligned sequence reads were visualized using the Integrated Genome Viewer (v.2.7.2) (Robinson et al. 2011).

Variants were confirmed and segregation tested by PCR amplification and Sanger sequencing on an ABI3130xl Genetic Analyser (Applied Biosystems). Electropherograms were analyzed using SeqScapeTM (v.2.5) (ThermoFisher Scientific).

Intact teeth were analyzed using a high-resolution micro–computed tomography (μCT) SkyScan 1172 (Bruker) scanner to quantify mineral density. Video showing the 3D internal and external features was created using CTVox (Bruker). Longitudinal mid-bucco slices of the teeth were imaged on an S-3400N scanning electron microscope (SEM) (Hitachi). Further details are in the Appendix methods.

Results

Probands in a large cohort of apparently unrelated patients/families with AI were subject to ongoing screening of AI-associated genes, either by targeted smMIPs or WES. Members of 11 families with likely pathogenic *AMBN* variants were identified to date, as shown in Figure 1A. All affected individuals were diagnosed with AI by experienced dental practitioners. No evidence was found of dentine changes.

Genetic Findings

The 11 families could be sorted into 3 groups according to their *AMBN* genotypes (Table). Group 1 includes 6 families, G1-1, G1-2, G1-3, G1-4, G1-5, and G1-6, all with an *AMBN* genotype and family history consistent with autosomal recessive AI. Affected individuals from G1-1, G1-2, and G1-3 are homozygous for pathogenic variant c.209C>G; p.(Ser70*), a stop-gain mutation that is likely to undergo nonsense-mediated decay (NMD) (Kurosaki and Maquat 2016). The affected individual from a fourth family, G1-4, is compound heterozygous for a single base duplication c.539dup; p.(Val181Serfs*5) and a splice site variant, c.571-1G>C; p.(?). The c.539dup variant gives rise to a frameshift variant (in exon 7 of 13) that is predicted to undergo NMD. The variant c.571-1G>C, with a splice-AI acceptor loss score of 0.83, donor loss of 0.20, and acceptor gain of 0.79, alters the splice acceptor site for *AMBN* exon 8 and is predicted to result in exon skipping. Probands from the remaining 2 group 1 families, G1-5 and G1-6, are compound heterozygotes for the stop-gain *AMBN* variant c.209C>G; p.(Ser70*) and a novel *AMBN* missense variant c.295T>C; p.(Tyr99His). The variant c.295T>C; p.(Tyr99His) has a combined annotation dependent depletion (CADD) score of 25.8 and is classified as likely pathogenic.

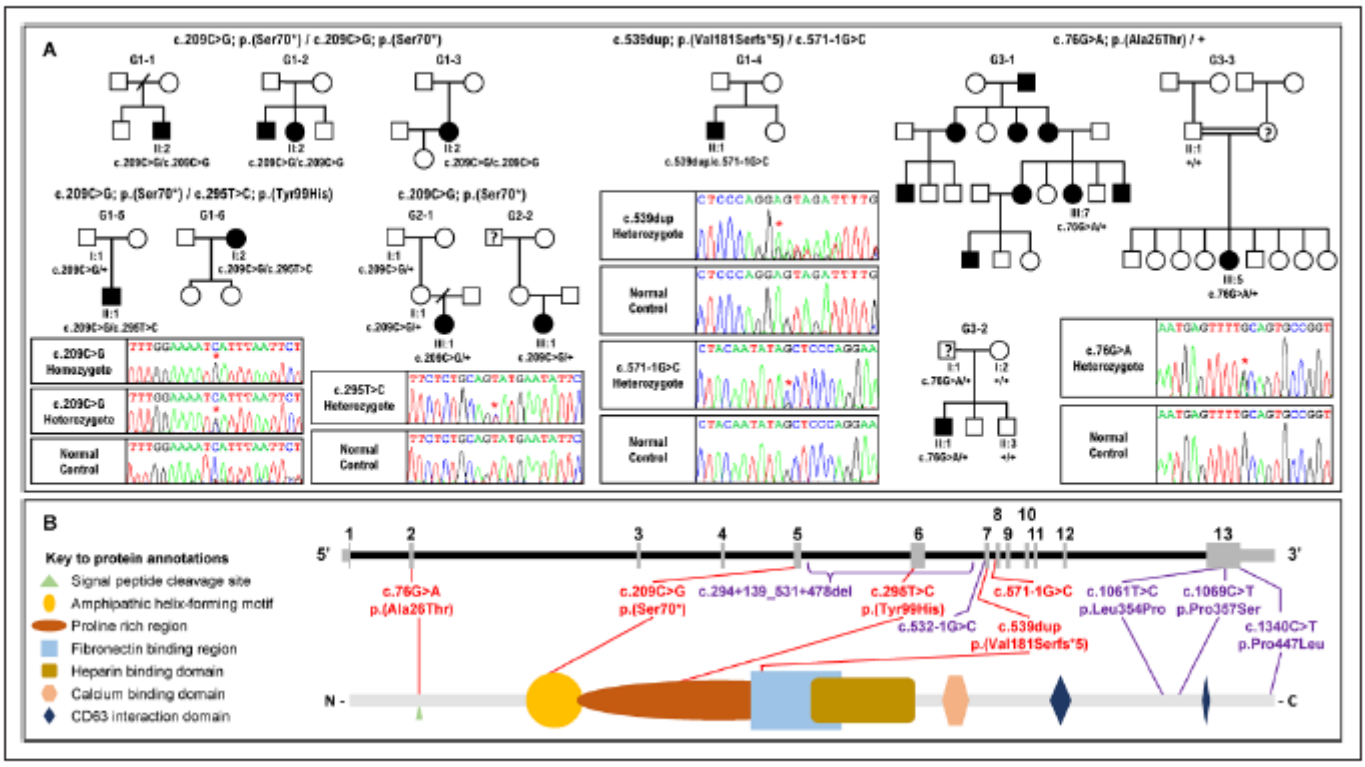


Figure 1. Pedigrees of the families recruited in this study, electropherograms of the variants identified and the location of the variants in the gene and protein. (**A**) Pedigrees of the 11 families described in this study with exemplar electropherograms of the variant(s) identified by whole-exome sequencing (WES) and single-molecule molecular inversion probes (smMIPs) analysis and then verified by Sanger sequencing. G1-1 and G1-3 present as cases of isolated amelogenesis imperfecta (AI) with no family history, while G1-2 consists of an affected sibling pair. Probands in these families are homozygous for the novel variant c.209C>G; p.(Ser70*) in exon 5. Families G1-5 and G1-6 also present as isolated cases, and affected individuals in each are compound heterozygotes for variants c.209C>G; p.(Ser70*) and c.295T>C; p.(Tyr99His) in exons 5 and 6, respectively. Probands from G2-1 and G2-2 present isolated AI with no apparent family history, and each carries the same heterozygous variant, c.209C>G; p.(Ser70*). G1-4 proband represents an isolated AI who is a compound heterozygote for the variants c.539dup; p.(Val181Serfs*5) and c.571-1G>C in exons 7 and 8, respectively. Three families, G3-1, G3-2, and G3-3, all carry the same heterozygous variant c.76G>A; p.(Ala26Thr) in exon 2. G3-1 proband is a single individual from a family with a clear history of dominant AI, while G3-2 and G3-3 were recruited as isolated patients with AI. G3-2, for which the father of the proband self-reported as unaffected on recruitment but was not clinically examined, carries the variant and the proband of G3-3 was reported as probable dominant AI, for which the mother was self-reporting possibly having AI. The "?" in the pedigree denotes possible AI in individuals not clinically assessed. Red stars (*) highlight nonreference nucleotides. Variant nomenclature is according to the transcript NM_016519.6. (**B**) Schematic representation of the *AMBN* gene (NM_016519.6) and its translated product showing the known domains of the 447–amino acid protein. Variants identified in this study are annotated in red; previously published variants are annotated in purple.

Group 2 includes 2 cases of isolated AI from families G2-1 and G2-2, without any history of AI in the family. The probands in each family are heterozygous for the *AMBN* variant c.209C>G; p.(Ser70*). No second *AMBN* variant was identified in *trans* in G2-1; however, the normal allele in the proband from G2-2 is a complex allele carrying 2 common *AMBN* coding variants: an in-frame deletion c.539_541del: p.(Gly180del) and the missense variant c.764C>T: p.(Ala255Val). The population allele frequencies of these variants are 0.082 and 0.086, respectively, and both are predicted to be benign. Investigation of other genes known to cause AI did not identify any potentially relevant pathogenic variants.

Group 3 includes family G3-1, with an extensive family history of dominant AI but only a single affected individual recruited. This group also includes 2 additional families with possible AD-AI: G3-2, for which the father of the proband self-reported as unaffected but was not clinically examined, and G3-3, for which the proband's mother is said to have AI but was not examined clinically. Probands from these families are all heterozygous carriers of the novel missense variant c.76G>A; p.(Ala26Thr). This variant was classified as likely pathogenic. No second *AMBN* variant was identified on the normal allele in these families.

Founder Effect Screening

The presence of variant c.209C>G; p.(Ser70*) in 7 families (groups 1 and 2), c.295T>C; p.(Tyr99His) in 2 families (group 1), and c.76G>A; p.(Ala26Thr) in 3 families (group 3) suggests these variants may have been inherited from a common ancestor. To test this hypothesis, we examined the haplotype backgrounds of these variants at the *AMBN* locus, using long-range PCR and third-generation nanopore sequencing.

A 9,681-bp DNA segment spanning exons 4 to 13 of *AMBN* and including amino acid residues 70 (exon 5) and 99 (exon 6) was PCR amplified and analyzed by long-read sequencing in probands from families G1-1, G1-2, and G1-3 (homozygous for c.209C>G; p.(Ser70*)), G1-5 and G1-6 (compound

Table. Details of AMBN Variants Detected in the Probands of 11 Recruited Families.

					Variant(s)						
Group	Family ID	Family History	Phen	Method	Transcript Change	Amino Acid Change	Zygosity	CADD	gnomAD Frequency	ACMG	ClinVar
1	G1-1	IC	HP	smMIP	c.209C>G	p.(Ser70*)	Hom	36	0.0001	Path	VCV001702585.3
	G1-2	SP	HP	WES	c.209C>G	p.(Ser70*)	Hom	36	0.0001	Path	VCV001702585.3
	G1-3	IC	HP	WES	c.209C>G	p.(Ser70*)	Hom	36	0.0001	Path	VCV001702585.3
	G1-4	IC	HP	WES	c.539dup	p.(Val181Serfs*5)	Het	22.8	Absent	Path	
					c.571-1G>C	p.?	Het	22.6	0.0001	Likely Path	VCV002444856.1
	G1-5	IC	HP	smMIP	c.209C>G	p.(Ser70*)	Het	36	0.0001	Path	VCV001702585.3
					c.295T>C	p.(Tyr99His)	Het	25.8	0.0001	Likely Path	VCV002233469.1
	G1-6	IC	HP	WES	c.209C>G	p.(Ser70*)	Het	36	0.0001	Path	VCV001702585.3
					c.295T>C	p.(Tyr99His)	Het	25.8	0.0001	Likely Path	
2	G2-1	IC	HM	smMIP	c.209C>G	p.(Ser70*)	Het	36	0.0001	Path	VCV001702585.3
	G2-2	IC	HM	smMIP	c.209C>G	p.(Ser70*)	Het	36	0.0001	Path	VCV001702585.3
3	G3-1	IC	HP	WES	c.76G>A	p.(Ala26Thr)	Het	26	Absent	Likely Path	
	G3-2	AD	HP	smMIP	c.76G>A	p.(Ala26Thr)	Het	26	Absent	Likely Path	
	G3-3	IC	HP	WES	c.76G>A	p.(Ala26Thr)	Het	26	Absent	Likely Path	

Variants are reported according to AMBN transcript NM_016519.6 and protein NP_057603.1, using human reference genome GRCh37/hg19.
ACMG criteria for p.(Ser70*) and p.(Val181Serfs*5) are Path: pathogenic (PP4, PVS1, PM2), for c.571-1G>C is likely Path: likely pathogenic (PP4, PM3, PM2, PVS1), for p.(Tyr99His) is likely pathogenic (PP4, PM3, PM2, PP3), for p.(Ala26Thr) is likely pathogenic (PP4, PS4, PM2). Scoring criteria: PP3, pathogenic supporting; PP4, pathogenic supporting; PS4, pathogenic strong; PVS1, pathogenic very strong; PM2, pathogenic moderate; PM3, pathogenic moderate.
ACMG, American College of Medical Genetics; AD, autosomal dominant; CADD, combined annotation dependent depletion; ClinVar, public archive of interpretations of clinically relevant variants; gnomAD, genome aggregation database (Karczewski et al. 2020); Het, heterozygous; HM, hypomaturation; Hom, homozygous; HP, hypoplastic; IC, isolated case; Phen, phenotype; smMIP, single-molecule molecular inversion probe; SP, sibling pair; WES, whole-exome sequencing.

heterozygotes for c.209C>G; p.(Ser70*) and c.295T>C; p.(Tyr99His)), and G2-1 and G2-2 (heterozygotes for c.209C>G; p.(Ser70*)). We observed a haplotype consisting of 11 nonreference nucleotides arranged in *cis* with c.209C>G; p.(Ser70*) in all 7 samples (Fig. 2A). The c.295T>C; p.(Tyr99His) variant was found on a different haplotype background characterized by 5 nonreference nucleotides in the 2 families carrying it (Fig. 2B).

An 8,520-bp amplicon, spanning exons 1 to 5 and including residue 26, was PCR amplified and analyzed by long-read nanopore sequencing in probands from families G3-1, G3-2, and G3-3, all heterozygous for the *AMBN* variant c.76G>A; p.(Ala26Thr). These data revealed a haplotype shared by all 3 families that comprised 3 nonreference nucleotides (Fig. 2C).

Phenotyping

Images of teeth and dental radiographs identified differences in the clinical phenotypes of the 3 groups recognized through genetic analyses (Fig. 3 and Appendix Fig. 1). Affected individuals in group 1 were characterized by hypoplastic AI with poor-quality enamel, with teeth having a yellow appearance following early posteruption loss of a thin layer of creamy, opaque mineralized tissue. Affected individuals in group 3 also had hypoplastic AI, but this differed from group 1 through the presence of a thin layer of more persistent enamel, which gives the teeth a whiter long-term appearance than group 1. Group 2 has a very different phenotype, characterized by hypomaturation AI with associated pits and minor morphological variations within a near-normal enamel volume that is more radio-dense than the supporting dentine on clinical radiography. No clear dentine abnormalities were evident on dental radiographs.

Teeth were available from a primary upper lateral incisor from the G2-1 proband and a permanent canine from the G2-2 proband for laboratory analyses. µCT of these teeth revealed normal enamel volume (Fig. 4i–iv). No significant differences were observed in average enamel mineral density (EMD) between affected and control teeth of the same type obtained from unrelated unaffected individuals. The EMD in G2-1 and its respective control were 2.561 $g.cm^{-3}$ and 2.546 $g.cm^{-3}$, and in G2-2 and its respective control, they were 2.569 $g.cm^{-3}$ and 2.721 $g.cm^{-3}$. An outer layer of particularly high mineral density seen in the control was missing in the G2-1, while the enamel of the G2-2 appeared pitted (Fig. 4i, iii) with pits extending through the enamel layer to the dentine–enamel junction (DEJ) (Fig. 4vii–ix, Appendix video). SEM analysis of these teeth showed disrupted, poorly formed prismatic microstructure with little demarcation between rod and interrod regions (Fig. 4xiii–xv). Hunter–Schreger banding was also absent in the affected tooth (Fig. 4xvi) as opposed to the control (Fig. 4xvii).

Discussion

The data presented support *AMBN* variants having a complex impact on human AI that highlights our incomplete understanding.

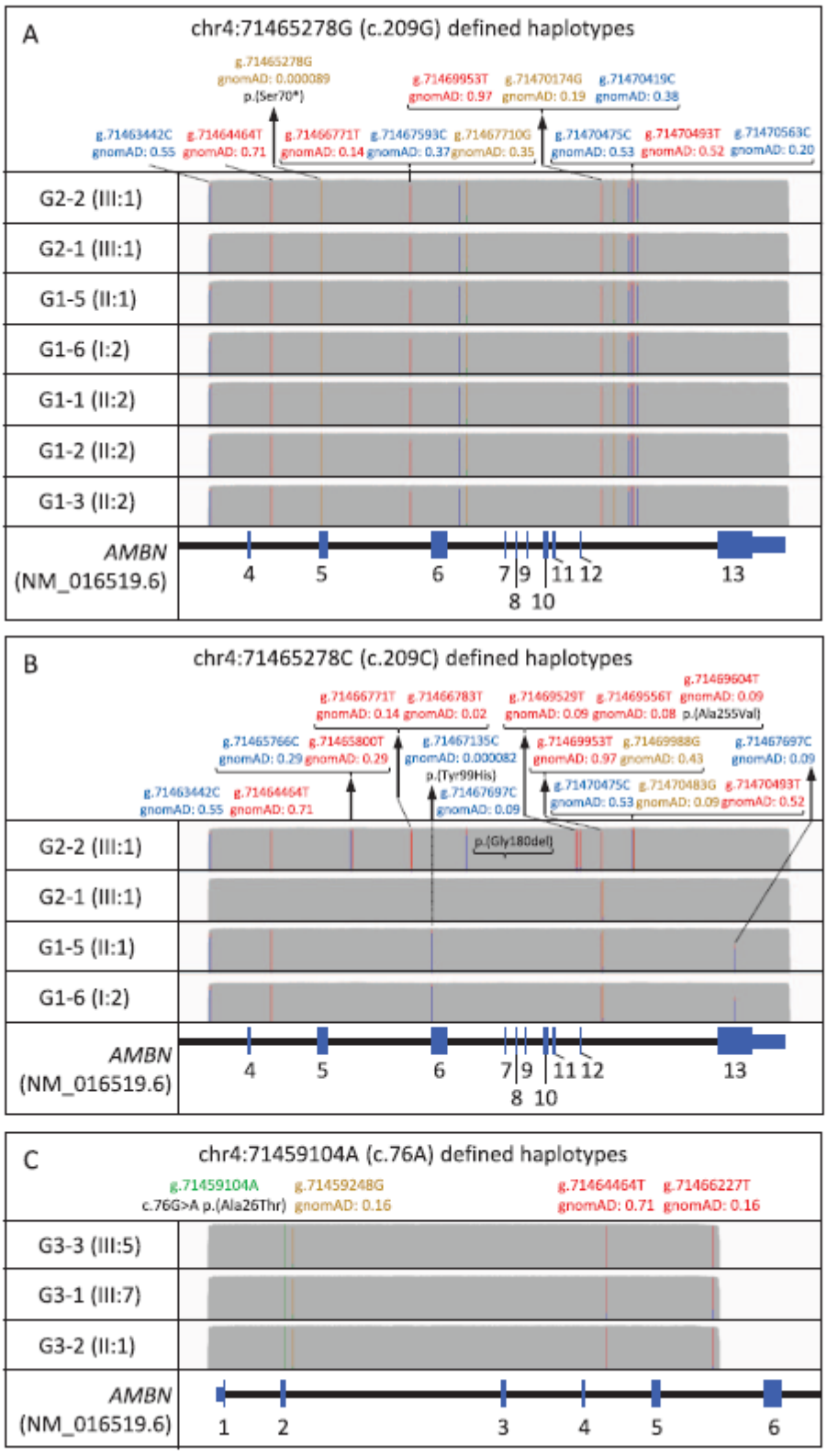


Figure 2. Long-read sequencing of a 9.7-kb amplification product from the *AMBN* locus spanning exons 4 to 12 of *AMBN*. Sequence analysis reveals the founder haplotype backgrounds on which variants c.209C>G; p.(Ser70*) and c.295T>C; p.(Tyr99His) have arisen, which are shared by all the families that carry them. Nucleotide positions are reported according to human genome build hg19. Allele frequencies are from the gnomAD database v.3.1.2 and are based on high-quality genotypes from a data set of 76,156 samples. The IGV allele frequency threshold is 0.6. Figure created using IGV version 2.12.2, with y-axis coverage tracks scaled to 2,300×. (**A**) Ten alleles identified from 3 homozygous (G1-1, G1-2, and G1-3), 2 heterozygous (G2-1 and G2-2), and 2 compound heterozygous (G1-5 and G1-6) individuals, all bearing the c.209C>G; p.(Ser70*) variant on a shared 11 nonreference single-nucleotide polymorphism (SNP) haplotype background. (**B**) The alleles not carrying the c.209C>G; p.(Ser70*) variant (wild type alleles) in families G2-1, G2-2, G1-5, and G1-6 tagged by the reference C nucleotide at position c.209 have haplotype backgrounds distinct from those bearing the pathogenic variants. Shared haplotype for the 2 alleles bearing the pathogenic variant c.295T>C; p.(Tyr99His) in families G1-5 and G1-6, each on the same background haplotype of 5 nonreference SNPs. (**C**) Haplotype background of the variant c.76G>A; p.(Ala26Thr) in families G3-1, G3-2, and G3-3 is identical, consisting of 3 other nonreference variants.

Figure 3. Clinical images and radiographs of the teeth capture the differences between the 3 groups. Group 1 (i–iii): Yellow hypoplastic amelogenesis imperfecta (AI) reflects the absence of any meaningful enamel on dental radiography (i and ii G1-2; iii G1-5 bitewing). Group 2 (iv–vi): Hypomaturation AI is characterized by variations in color with pits and other localized morphological changes that disrupt the normal clinical enamel surface. Dental radiography confirms near-normal enamel volumes with a clear difference between enamel and dentine radiodensity (iv and v G2-2; vi G2-2 detail from panoramic radiograph). Group 3 (vii–ix): White hypoplastic AI reflects the presence of a thin layer of enamel on dental radiography (vii and viii G3-3; ix G3-1 detail from panoramic radiograph). Further clinical images are included in Appendix Figure 1.

Six of the families described here have genotypes and inheritance patterns consistent with autosomal recessive hypoplastic AI with poor-quality enamel. Clinical images were consistent with rapid failure of a thin creamy mineralized tissue after eruption, leaving a predominantly yellow appearance. These group 1 families fit with previous reports of recessive AI due to biallelic pathogenic variants in *AMBN*, with no clinically significant enamel changes in heterozygous carriers (Poulter et al. 2014). Affected individuals from 5 of the 6 group 1 families were homozygous ($n=3$) or heterozygous ($n=2$) for the c.209C>G p.(Ser70*) variant, which occurred on the same haplotype consistent with a common UK founder allele. This variant is predicted to produce no AMBN protein. In the 2 families in whom heterozygous c.209C>G p.(Ser70*) was paired with the missense c.295T>C p.(Tyr99His) change, this also resulted in a poor enamel quality form of hypoplastic AI.

Intriguingly, group 2 included isolated AI cases heterozygous for *AMBN* c.209C>G; p.(Ser70*), with a very different AI phenotype from group 1 families. Heterozygosity for p.(Ser70*) alone being sufficient to cause AI is inconsistent with the existence of many apparently unaffected heterozygous carriers in group 1 families. Individuals G1-5 I-1 and G2-1 II-1, who are heterozygous carriers of p.(Ser70*), consented to give DNA but were not subject to clinical examination for possible subtle enamel developmental abnormalities, although neither was flagged as having AI. In contrast, probands in group 2 families were diagnosed clinically as having AI. Full gene screening of *AMBN* by long-read sequencing did not detect any second pathogenic variants in the coding sequence, introns, or 5′ or 3′ UTRs, and there was no evidence of allele dropout, which might imply the presence of structural variants missed by short-read sequencing in either of these 2 families (Appendix Fig. 2). The *trans* allele in G2-2 was found to carry 2 further coding variants that could act as hypomorphic alleles contributing to AI in G2-2 but are too common to be pathogenic themselves, even when homozygous. It is therefore necessary to consider alternative hypotheses, either that there may be other hypomorphic variants on the second *AMBN* alleles in

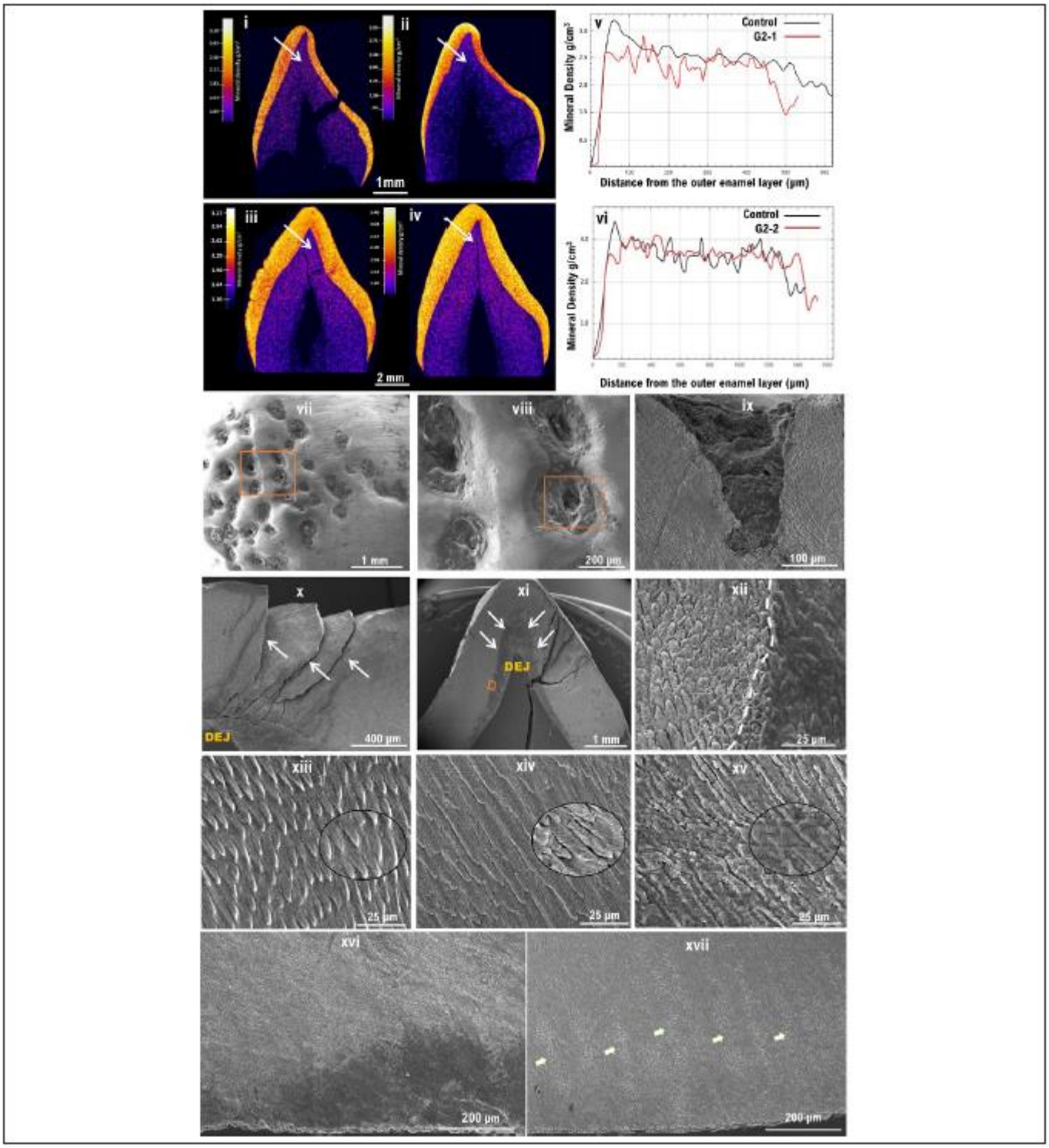


Figure 4. Phenotypic characterization of the teeth of group 2 families. (i–iv) Micro–computed tomography (µCT): False-colored calibrated heatmaps. (i) A primary upper lateral incisor from the affected individual from G2-1 and (iii) a permanent canine from G2-2. Teeth of the same types obtained from unrelated unaffected individuals were used as controls for comparison and are shown in (ii) and (iv), respectively. (i, iv) No significant differences in enamel volume was observed between affected and healthy controls, but the mineral density distribution was found to be disturbed. (v, vi) Mineral density line scans showing the distribution of enamel density form the enamel surface to the dentine–enamel junction (DEJ) in G2-1 and G2-2, respectively, as shown by the arrows in i–iv. In both controls, note the initial high peaks present at the surface and the gradual decrease of mineral density toward the DEJ, which are absent in the amelogenesis imperfecta–affected teeth confirming the disturbed mineral density observed in the µCT images. (iii) G2-2 tooth shows an uneven, pitted surface that is absent from the control tooth. (vii–xvii) Scanning electron microscope images: (vii–viii) Surface topography features in G2-2 showing numerous clear pitting of the labial enamel surface, extending through the enamel (ix) and to the DEJ (x). (xi) Labiolingual section of G2-2, a more disrupted inner enamel below the cuspal area is shown by the white arrows. (xii) Higher magnification of the inset in (xi) showing a clear demarcation (white dotted line) between a less dense enamel located near the DEJ and the rest of the enamel. (xiii–xv) Higher magnification of the enamel microstructure at ×1k, with ×3.5k insets to illustrate prismatic structure as well as crystallite orientation. (xiii) Healthy enamel. (xiv–xv) Affected teeth from the family G2-1 and G2-2, respectively. (xiv–xv) Note the poorly formed enamel rods that appear fused at many areas, making it difficult to distinguish the boundaries between rod and interrod regions. (xvi) Low magnification of G2-2 showing missing Hunter–Schreger banding as opposed to the healthy enamel (arrows) in (xvii).

G2-1 and G2-2, which have not been detected in genetic screening, or that genetic or environmental modifiers combine with the p.(Ser70*) variant to cause disease. Clinical evaluation of individuals presenting with whole-dentition abnormal enamel development considered possible environmental factors such as dental fluorosis or major systemic illness, before a clinical diagnosis of AI was made (Wright 2023). This does not preclude some attenuation by environmental factors of what is primarily a genetically driven enamel phenotype. No potentially pathogenic variants were identified in other known AI genes, but noncoding variants in known genes or variants in previously undiscovered amelogenesis-related genes could be present. However, assessing the relative contributions of genes and environment to severity in AI would require a large patient cohort and is beyond the scope of this study.

Group 3 includes 3 families with the same heterozygous variant, c.76G>A; p.(Ala26Thr), which is absent from gnomAD. Of these, G3-1 has a clear family history of dominantly inherited AI. However, only a single family member was recruited, so it was not possible to confirm cosegregation of the variant with AI. Observation of 2 other apparently unrelated patients with AI (G3-2 and G3-3) with the same variant on the same founder haplotype implies descent from a common ancestor, which is further evidence of cosegregation of this variant with AI as a dominantly inherited trait. No potentially pathogenic second *AMBN* variant was identified in smMIPs, WES, or long-read nanopore sequencing data in these individuals.

The spectrum of pathogenic variants in *AMBN* revealed by this and previous studies encompasses a premature termination codon (PTC), a 1-bp deletion leading to a frameshift, 2 splice acceptor site variants, an in-frame whole-exon deletion, and 5 missense variants. The consequences of these variants have not been determined experimentally, but it seems likely that the PTC, frameshift, splice variants, and exon deletion will act as null alleles, reducing the amount of functional AMBN available during amelogenesis. Teepe and coworkers reported that transgenic mice expressing Ambn at concentrations lower or higher than the wild-type level had enamel abnormalities (Teepe et al. 2014), suggesting that a specific AMBN concentration is crucial for amelogenesis. Furthermore, group 1 families G1-1, G1-2, and G1-3 are homozygous for likely null variants, providing additional support for the interpretation that lack of intact AMBN is the likely disease mechanism in these cases.

For missense variants, the disease mechanism is less clear, but they may also be functional knockouts. The p.(Tyr99His) substitution changes an aromatic tyrosine to a basic histidine in the first of 15 highly conserved amino acid residues in the proline-rich region of AMBN. This region is retained in AMBN isoform I (ISOI) but removed in isoform II (ISOII) due to alternative splicing, and both isoforms are highly conserved and are coexpressed in vitro, suggesting they might perform different functions during enamel development (MacDougall et al. 2000; Vetyskova et al. 2020). Probands in families G1-5 and G1-6, who are compound heterozygotes for p.(Ser70*) and p.(Tyr99His), had thin poor-quality enamel, similar to families homozygous for p.(Ser70*). It is therefore likely that p.(Tyr99His) is also a functional knockout.

On clinical examination, group 1 enamel is yellow, thin, and of poor quality without normal microstructure (Poulter et al. 2014). By contrast, group 3 enamel associated with c.76G>A; p.(Ala26Thr) is thin on radiography, yet the white appearance of the teeth is consistent with little posteruption breakdown. Without laboratory analyses, it is unknown if this has a normal enamel microstructure. This substitution changes a nonpolar hydrophobic alanine (conserved in all mammals except the toothless platypus) at the C-terminal amino acid of the AMBN secretory signal peptide, immediately adjacent to the cleavage site, to a polar hydrophilic threonine (Delsuc et al. 2015). Proteins such as AMBN that are destined for the extracellular environment are transported to the endoplasmic reticulum (ER) under the direction of the signal peptide, which is cleaved from the protein before secretion. Many human diseases, including AI, are caused by ER stress, resulting from the misfolding of newly synthesized proteins as they are trafficked through the ER (Brookes et al. 2017; Morikawa and Urano 2022). It is therefore plausible that the apparently dominant hypoplastic AI phenotype associated with the p.(Ala26Thr) signal peptide variant arises by a different disease mechanism compared to null variants causing recessive AI. A dominant negative effect may lead to impairment of the normal ameloblast secretory pathway, ER stress, and ultimately ameloblast apoptosis.

Accordingly, these data contribute to the debate as to whether mutations in *AMBN* cause dominant as well as recessive AI. A large dominant family with a combined AI/DI phenotype was segregated with the heterozygous *AMBN* missense variant p.(Pro357Ser) (Lu et al. 2018). This report was questioned by Liang and coworkers (2019), who cautioned that the diagnosis may be DI caused by a variation in *DSPP* (a gene linked to *AMBN* on chromosome 4 and with a repetitive region that is difficult to sequence by WES). However, a crossover below marker D4S2931, visible in individual IV7 in the microsatellite data presented but not discussed by Lu and colleagues (2018), clearly excludes the possibility that a variant in the *DSPP* gene, 17 Mb distal to *AMBN*, could cause the phenotype seen in this family. Furthermore, the families described in our study were diagnosed with AI in the absence of any clear dentine abnormalities, meaning that specific exclusion of *DSPP* or other genes involved in DI in these families is not required. The patterns of inheritance in these families, together with evidence that they share a common ancestor, therefore provide additional support for dominant inheritance of AI due to variants in *AMBN*.

These data provide new insight into how null *AMBN* variants contribute to human AI and emphasize the importance of AMBN concentration during amelogenesis. Data are also presented consistent with a very specific, rare heterozygous missense change causing a type of AI distinctive from that due to loss of function, consistent with dominant inheritance. It is plausible that *AMBN* variants will not only cause different types of AI but may also influence enamel formation in other

situations that are of relevance to enamel failure and how it is clinically managed.

Author Contributions

U. Hany, C.M. Watson, C.F. Inglehearn, A.J. Mighell, contributed to conception, design, data acquisition, analysis, and interpretation, drafted and critically revised the manuscript; L. Liu, G. Nikolopoulos, C.J. Brown, A. Patel, H.D. Rodd, R. Balmer, A. Harfoush, M. Al-Jawad, contributed to data acquisition and interpretation, critically revised the manuscript; C.E.L. Smith, J.A. Poulter, contributed to conception, design, data acquisition and interpretation, critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of the work.

Acknowledgments

The authors thank the families involved for their support for this study.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Rosetrees Trust Grant PGS19-2/10111, Wellcome Trust Grant WT093113MA, and a Leeds Doctoral Scholarship awarded to U. Hany.

ORCID iDs

U. Hany https://orcid.org/0000-0002-4486-1625
C.M. Watson https://orcid.org/0000-0003-2371-1844
C.F. Inglehearn https://orcid.org/0000-0002-5143-2562

References

Bäckman B, Holm AK. 1986. Amelogenesis imperfecta: prevalence and incidence in a northern Swedish county. Community Dent Oral Epidemiol. 14(1):43–47.

Bartlett JD, Ganss B, Goldberg M, Moradian-Oldak J, Paine ML, Snead ML, Wen X, White SN, Zhou YL. 2006. Protein-protein interactions of the developing enamel matrix. Curr Top Dev Biol. 74:57–115.

Brookes SJ, Barron MJ, Smith CEL, Poulter JA, Mighell AJ, Inglehearn CF, Brown CJ, Rodd H, Kirkham J, Dixon MJ. 2017. Amelogenesis imperfecta caused by N-terminal enamelin point mutations in mice and men is driven by endoplasmic reticulum stress. Hum Mol Genet. 26(10):1863–1876.

Delsuc F, Gasse B, Sire JY. 2015. Evolutionary analysis of selective constraints identifies ameloblastin (AMBN) as a potential candidate for amelogenesis imperfecta. BMC Evol Biol. 15:148.

Fukumoto S, Yamada A, Nonaka K, Yamada Y. 2005. Essential roles of ameloblastin in maintaining ameloblast differentiation and enamel formation. Cells Tissues Organs. 181(3–4):189–195.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Genome Aggregation Database, C. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581(7809), 434–443.

Krebsbach PH, Lee SK, Matsuki Y, Kozak CA, Yamada KM, Yamada Y. 1996. Full-length sequence, localization, and chromosomal mapping of ameloblastin. A novel tooth-specific gene. J Biol Chem. 271(8):4431–4435.

Kurosaki T, Maquat LE. 2016. Nonsense-mediated mRNA decay in humans at a glance. J Cell Sci. 129(3):461–467.

Lee SK, Krebsbach PH, Matsuki Y, Nanci A, Yamada KM, Yamada Y. 1996. Ameloblastin expression in rat incisors and human tooth germs. Int J Dev Biol. 40(6):1141–1150.

Liang T, Hu Y, Smith CE, Richardson AS, Zhang H, Yang J, Lin B, Wang SK, Kim JW, Chun YH, et al. 2019. AMBN mutations causing hypoplastic amelogenesis imperfecta and Ambn knockout-NLS-lacZ knockin mice exhibiting failed amelogenesis and ambn tissue-specificity. Mol Genet Genomic Med. 7(9):e929.

Lindenbaum P. 2015. JVarkit: java-based utilities for bioinformatics. doi:10.6084/m9.figshare.1425030.v1.

Lu T, Li M, Xu X, Xiong J, Huang C, Zhang X, Hu A, Peng L, Cai D, Zhang L, et al. 2018. Whole exome sequencing identifies an AMBN missense mutation causing severe autosomal-dominant amelogenesis imperfecta and dentin disorders. Int J Oral Sci. 10(3):26.

MacDougall M, Simmons D, Gu TT, Forsman-Semb K, Kärrman Mårdh C, Mesbah M, Forest N, Krebsbach PH, Yamada Y, Berdal A. 2000. Cloning, characterization and immunolocalization of human ameloblastin. Eur J Oral Sci. 108(4):303–310.

Morikawa S, Urano F. 2022. The role of ER stress in diabetes: exploring pathological mechanisms using wolfram syndrome. Int J Mol Sci. 24(1):230.

Pandya M, Diekwisch TGH. 2021. Amelogenesis: transformation of a protein-mineral matrix into tooth enamel. J Struct Biol. 213(4):107809.

Poulter JA, Murillo G, Brookes SJ, Smith CE, Parry DA, Silva S, Kirkham J, Inglehearn CF, Mighell AJ. 2014. Deletion of ameloblastin exon 6 is associated with amelogenesis imperfecta. Hum Mol Genet. 23(20):5317–5324.

Prasad MK, Geoffroy V, Vicaire S, Jost B, Dumas M, Le Gras S, Switala M, Gasse B, Laugel-Haushalter V, Paschaki M, et al. 2016. A targeted next-generation sequencing assay for the molecular diagnosis of genetic disorders with orodental involvement. J Med Genet. 53(2):98–110.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 17(5):405–424.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol. 29(1):24–26.

Smith CEL, Poulter JA, Antanaviciute A, Kirkham J, Brookes SJ, Inglehearn CF, Mighell AJ. 2017. Amelogenesis imperfecta; genes, proteins, and pathways. Front Physiol. 8:435.

Stakkestad O, Lyngstadaas SP, Thiede B, Vondrasek J, Skalhegg BS, Reseland JE. 2017. Phosphorylation modulates ameloblastin self-assembly and Ca^{2+} binding. Front Physiol. 8:531.

Teepe JD, Schmitz JE, Hu Y, Yamada Y, Fajardo RJ, Smith CE, Chun YH. 2014. Correlation of ameloblastin with enamel mineral content. Connect Tissue Res. 55(Suppl 1):38–42.

Vetyskova V, Zouharova M, Bednarova L, Vanek O, Sazelova P, Kasicka V, Vymetal J, Srp J, Rumlova M, Charnavets T, et al. 2020. Characterization of AMBN I and II isoforms and study of their Ca^{2+}-binding properties. Int J Mol Sci. 21(23):9293.

Wald T, Osickova A, Sulc M, Benada O, Semeradtova A, Rezabkova L, Veverka V, Bednarova L, Maly J, Macek P, et al. 2013. Intrinsically disordered enamel matrix protein ameloblastin forms ribbon-like supramolecular structures via an N-terminal segment encoded by exon 5. J Biol Chem. 288(31):22333–22345.

Witkop CJ Jr. 1988. Amelogenesis imperfecta, dentinogenesis imperfecta and dentin dysplasia revisited: problems in classification. J Oral Pathol. 17(9-10):547–553.

Wright JT. 2023. Enamel phenotypes: genetic and environmental determinants. Genes (Basel). 14(3):545.

1. Supplementary Methods

Patient recruitment

Patients were recruited through UK paediatric dental clinics. Written consent was obtained in accordance with the principles of the Declaration of Helsinki and local ethical approval was granted for this study (REC 13/YH/0028). Genomic DNA was isolated from saliva samples using Oragene® sample collection tubes (DNA Genotek Inc. Ontario, Canada) and following manufacturer's instructions, or from peripheral blood lymphocytes using either a Chemagic 360 (Perkin Elmer, Waltham, MA, USA) or standard salting-out techniques.

Whole exome sequencing

Two wet-laboratory workflows were used to perform whole exome sequencing (WES), these comprised either the SureSelect Human All Exon v6 kit (Agilent Technologies, Wokingham, UK) or the Human Comprehensive Exome kit (10-50 Mb) (Twist Bioscience, San Francisco, CA, USA) following manufacturer's protocols throughout. SureSelect libraries were sequenced on a HiSeq 3000 (Illumina Inc., San Diego, CA, USA) which generated paired-end 150 bp reads. For the Twist workflow, sequencing was carried out using a P3 flowcell on a NextSeq 2000 (Illumina Inc.) which also generated paired-end 150 bp reads. Confirmation of raw read quality was performed using FastQC (v.0.11.3) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), before sequence reads were aligned to an indexed human reference genome (build hg19) using BWA (v.0.7.12-r1.39) (https://bio-bwa.sourceforge.net/) (Li and Durbin 2009). PCR duplicates were removed using Picard (v.2.5.0) (https://broadinstitute.github.io/picard/), before non-reference bases were identified using the Genome Analysis Tool Kit (GATK) HaplotypeCaller (v.3.5) (https://gatk.broadinstitute.org) according to recommended best practice workflows (DePristo et al. 2011). Identified sequence variants were annotated with functional and population frequency data using the Variant Effect Predictor (VEP) (v.83) (McLaren et al. 2016). Allele frequencies are reported according to the Genome Aggregation Database (gnomAD; v.2.1.1) (https://gnomad.broadinstitute.org/) (Karczewski et al. 2020) and splicing predictions are according to a web-lookup implementation of Splice-AI (https://spliceailookup.broadinstitute.org) (Jaganathan et al. 2019).

Single-molecule molecular inversion probes (smMIPs) sequencing

smMIPs covering the coding sequences of 19 genes (Table S1) associated with non-syndromic AI were designed using MIPGEN (https://github.com/shendurelab/MIPGEN) (Boyle et al. 2014) and synthesized by Integrated DNA Technologies (IDT; Leuven, Belgium) before being pooled in equimolar ratios. The smMIPs pool was next 5'-phosphorylated and 100 ng of each genomic DNA was subjected to targeted capture and ligation using the phosphorylated probe pool that was diluted to a ratio of 800 smMIPs copies for each DNA molecule in the final capture reaction. Sequencing was carried out using a Nextseq 500 (Illumina Inc.) generating paired-end 150 bp reads.

Sample demultiplexing and the removal of unique molecular identifiers (UMIs) was performed using a local implementation of the MIPVAR pipeline (https://sourceforge.net/projects/mipvar/) which was modified for compatibility with local computing hardware. This enabled sample demultiplexing and the removal of unique molecular identifies prior to ligation- and extension-arm processing using standard tools (BWA- v.0.7.12 (human reference genome build hg19), Picard v.1.102.0 and the GATK HaplotypeCaller v.3.2-2). Variants were recorded in VCF format and annotated using VEP.

Variant interpretation

The pathogenicity status of detected variants was classified according to the American College of Medical Genetics and Genomics (ACMG) guidelines using Franklin (https://franklin.genoox.com) (genoox, Palo Alto, CA ,USA) (Richards et al. 2015).

Long-read sequencing

Primer3 (http://primer3.ut.ee/) was used to design PCR primers which were synthesized by IDT (Leuven, Belgium). Two different amplicons were sequenced to cover the full AMBN gene; first one spanning exons 1 to 5 including 100bp of the promoter region, targeted a region of 8520 bp and the second one covered exons 4 to 13 targeted a region of 9681 bp. Primers for the first amplicon were, forward /CAATGTCCCTGCACGCAATA and reverse /GCAAGGAAGTCTCGCAACAA. Primers for the second amplicon were: forward /AAGCTGGGGGCAGTCAATAC and reverse /AGCAAAGGTAGAGGATGAGTATGC. Long-range PCR was performed using the SequelPrepTM polymerase (ThermoFisher Scientific), following the manufacturer's guidelines. Sequencing libraries were prepared using the SQK-LSK109 ligation kit (Oxford Nanopore Technologies (ONT), Oxford, UK). A

24-hr sequencing run was initiated for each sample on a Flongle flowcell (R.9.4.1) using a MinION (ONT) device running MinKNOW.

Guppy (v.5.0.16) (https://nanoporetech.com/) was used to perform basecalling, which converted raw data from fast5 to FASTQ.gz format. Sequencing adaptors were removed using Porechop (v.0.2.4) (https://github.com/rrwick/Porechop) before NanoFilt (v.2.8.0) (De Coster et al., 2018) was used to select reads by length (± 500 bp surrounding the expected size of the amplified fragment) and quality (Q≥10) (https://github.com/wdecoster/nanofilt). Processed reads were aligned to the human reference genome (build hg19) using MiniMap2 (v.2.22) (https://github.com/lh3/minimap2). Haplotypes were defined following the selection of reference and non-reference matching nucleotides at position chr4:70599561 using the Jvarkit tool biostar214299 (http://lindenb.github.io/jvarkit/Biostar214299.html) (Lindenbaum 2015). Samtools (v.1.14) was used to downsample corresponding BAM files to 2,000×. Aligned sequence reads were visualised using the Integrated Genome Viewer (v.2.7.2) (Robinson et al. 2011).

Sanger sequencing verification

Primer pairs were designed using AutoPrimer3 (https://github.com/gantzgraf/autoprimer3) and synthesized by IDT (Leuven, Belgium). Q5® High-Fidelity 2X Master Mix (New England Biolabs, Ipswich, MA, USA) was used for PCR amplification, which was confirmed by agarose gel electrophoresis. PCR products were purified using ExoSAP-IT (ThermoFisher Scientific, Waltham, MA, USA) prior to Sanger sequencing using BigDye Terminator v.3.1 and resolved on an ABI3130xl Genetic Analyser (Applied Biosystems, Paisley, UK). Electropherograms were analysed using SeqScapeTM (v.2.5) (ThermoFisher Scientific).

X-ray microtomography (µCT)

Intact teeth were analysed using a high-resolution micro-computed tomography (µ-CT) SkyScan 1172 (Bruker, Belgium) scanner to quantify mineral density. Mineral density values were calculated relative to three hydroxyapatite standards, of 0.25 and 0.75 g/cm3 (Bruker, Belgium), and 2.9 g/cm3 (Himed, USA). Fiji/Image J was used to analyse enamel density using a pixel threshold above 2.0 g/cm3. Video showing the 3D internal and external features were created using CTVox (Bruker, Belgium).

Scanning electron microscopy (SEM)

Longitudinal mid-bucco slices of the teeth were obtained using an Accutom 10 cutting machine and diamond cutting wheel (Struers, Germany). After removing surface debris, slices were gold coated (Agar Scientific, Elektron Technology, UK). Imaging was performed by S-3400N (Hitachi, Japan) SEM. References

2. Supplementary Figures

Figure S1: Clinical images and dental radiographs available for Group 1 (G1-1, G1-2, G1-3, G1-4, G1-5 and G1-6), Group 2 (G2-1 and G2-2) and Group 3 (G3-1, G3-2 and G3-3) families.

Images and radiographs from group 1 families

Probands from group 1 families (G1-1, G1-2, G1-3, G1-4, G1-5 and G1-6) display yellow hypoplastic AI, reflecting an absence of meaningful enamel obvious in dental radiography in G1-1 and G1-5.

Group 1: G1-3 Homozygous c.209C>G; p.(Ser70*)

Group 1: G1-6 Compound Heterozygous c.209C>G; p.(Ser70*) and c.295T>C; p.(Tyr99His)

Group 1: G1-4 Compound Heterozygous c.539dup; p.(Val181Serfs*5) and c.571-1G>C; p.?

Group 1: G1-5 Compound Heterozygous c.209C>G; p.(Ser70*) and c.295T>C; p.(Tyr99His)

Images and radiographs from group 2 families

Probands from G2-1 and G2-2 demonstrate hypomaturation AI, characterised by variations in colour with pits and other localised morphological changes that disrupt the normal clinical enamel surface.

Group 2: G2-1 Heterozygous c.209C>G; p.(Ser70*)

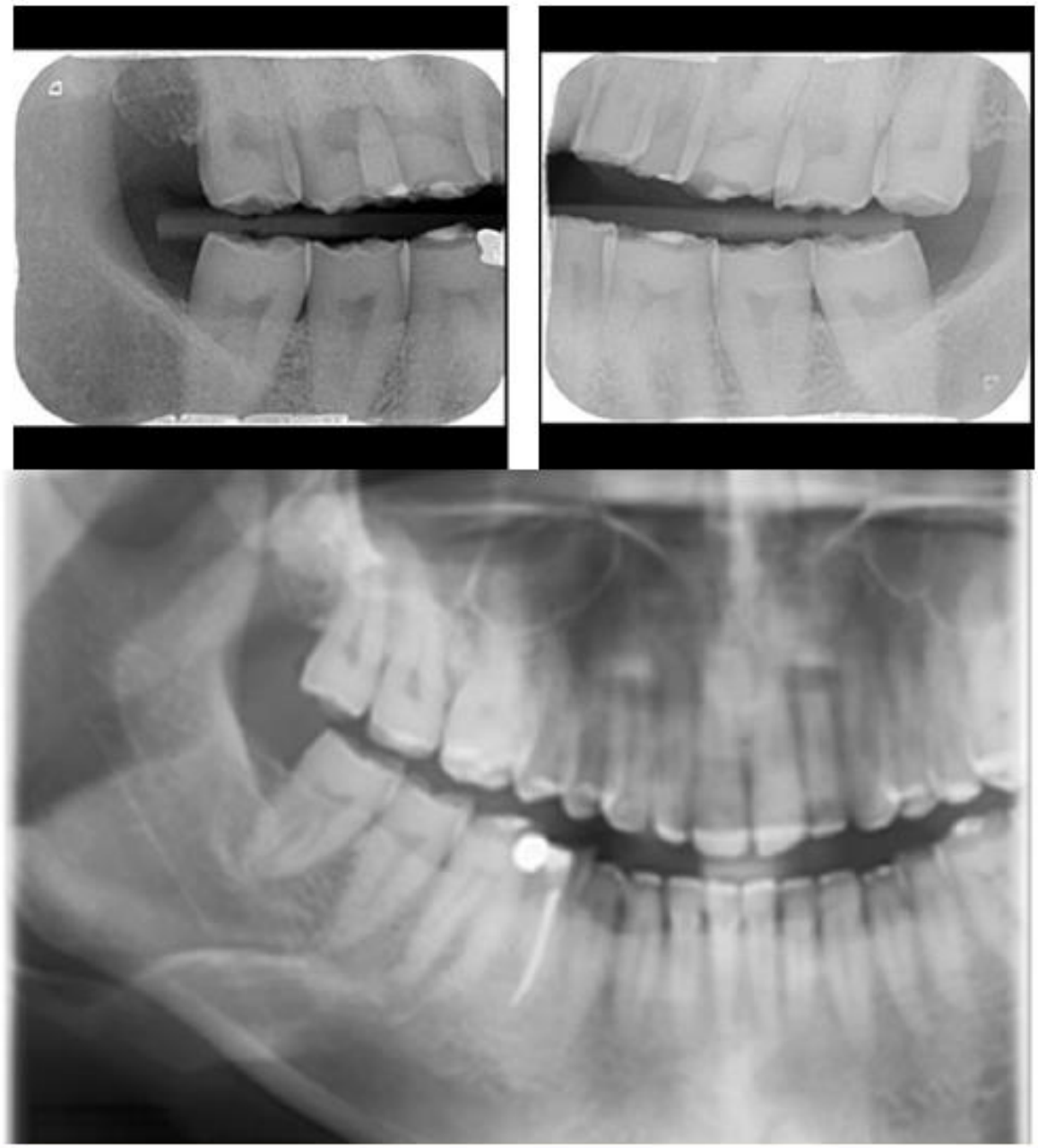
Group 2: G2-2 Heterozygous c.209C>G; p.(Ser70*)

Group 2: G2-2 Heterozygous c.209C>G; p.(Ser70*), continued

Images and radiographs from group 3 families

Group 3 (G3-1, G3-3) families are characterized by white hypoplastic AI reflecting the presence of a thin layer of enamel on dental radiography.

Group 3: G3-1 Heterozygous c.76G>A; p.(Ala26Thr)



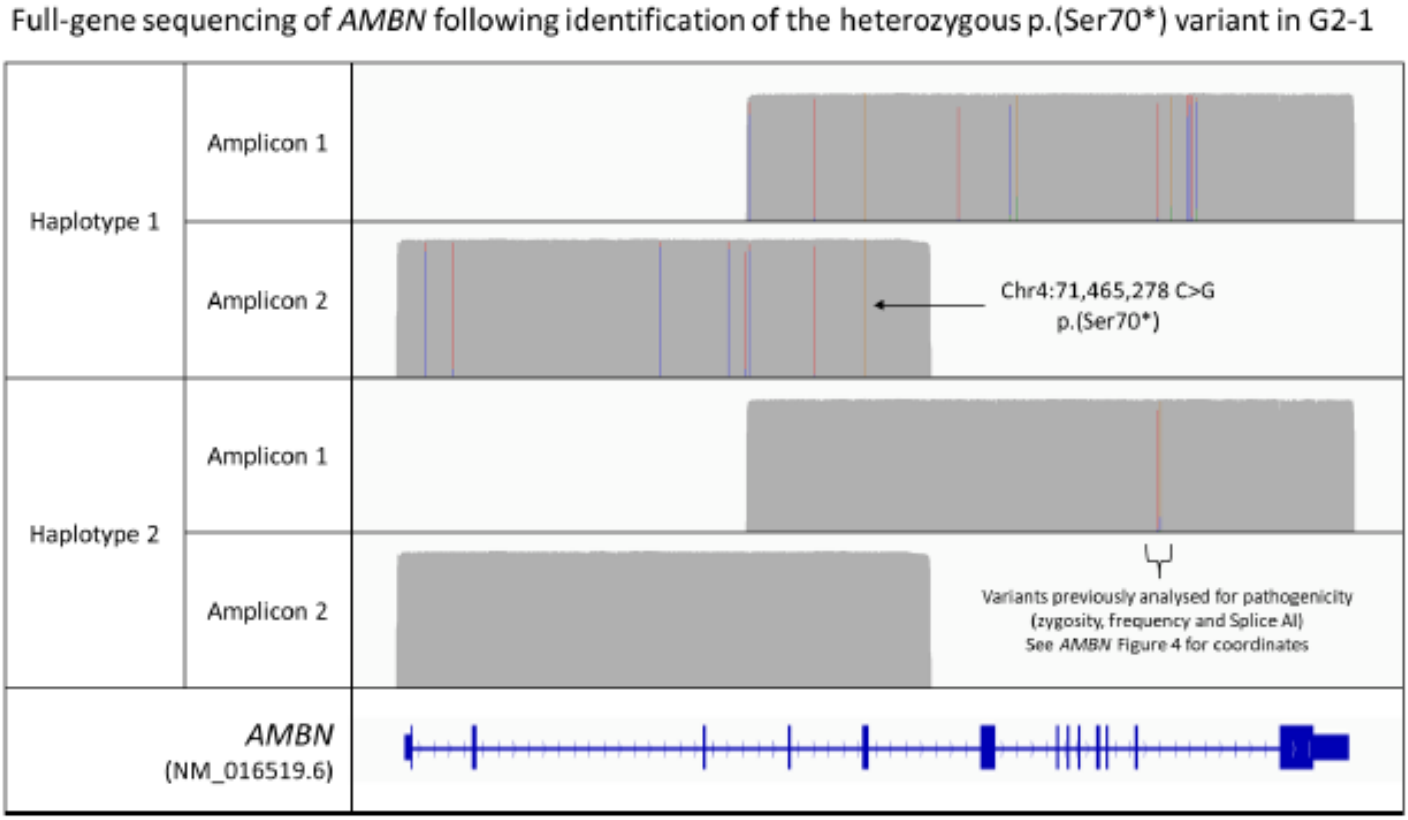
Group 3: G3-2 Heterozygous c.76G>A; p.(Ala26Thr)

No clinical images or dental X-rays available
The phenotype was described at recruitment as:

- Hypoplastic Amelogenesis Imperfecta

Group 3: G3-3 Heterozygous c.76G>A; p.(Ala26Thr)

Figure S2: Full-gene sequencing of *AMBN* following identification of the heterozygous p.(Ser70*) variant in G2-1 and G2-2. The data displays "full-gene screen" by two amplicons (18201 bp), the p.(Ser70*) is located roughly in the middle of the gene. The identified variants are on the "correct" haplotype in the overlapping area between amplicon 1 and 2 in both G2-1 and G2-2. Haplotype 1 is C defined and haplotype 2 is G defined. The haplotypes were set based on the reference and non-reference nucleotides that define the p.(Ser70*) mutation.



The image displays an 18201bp-long PCR product covering the whole AMBN gene in G2-1. No potential second pathogenic variants or large structural variants were detected other than p.(Ser70*) in the whole gene.

3. Supplementary Table

Table S1: Genes included in the smMIP reagent. Reference genome GRCh37/hg19.

Gene Name	Gene Symbol	OMIM	Genomic Coordinates	Cytoband
LAMININ, BETA-3	*LAMB3*	150310	chr1:209,788,218-209,825,820	1q32.2
INTEGRIN, BETA-6	*ITGB6*	147558	chr2:160,958,233-161,056,589	2q24.2
AMELOTIN	*AMTN*	610912	chr4:71,384,298-71,398,459	4q13.3
AMELOBLASTIN	*AMBN*	601259	chr4:71,457,975-71,473,004	4q13.3
ENAMELIN	*ENAM*	606585	chr4:71,494,461-71,512,536	4q13.3
ODONTOGENESIS-ASSOCIATED PHOSPHOPROTEIN	*ODAPH*	614829	chr4:76,481,258-76,491,103	4q21.1
FAMILY WITH SEQUENCE SIMILARITY 83	*FAM83H*	611927	chr8:144,806,103-144,815,914	8q24.3
COLLAGEN, TYPE XVII, ALPHA-1	*COL17A1*	113811	chr10:105,791,046-105,845,638	10q25.1
RECEPTOR EXPRESSED IN LYMPHOID TISSUES	*RELT*	611211	chr11:73,087,405-73,108,519	11q13.4
MATRIX METALLOPROTEINASE 20	*MMP20*	604629	chr11:102,447,566-102,496,063	11q22.2
G PROTEIN-COUPLED RECEPTOR 68	*GPR68*	601404	chr14:91,698,876-91,710,852	14q32.11
SOLUTE CARRIER FAMILY 24 (SODIUM/POTASSIUM/CALCIUM EXCHANGER), MEMBER 4	*SLC24A4*	609840	chr14:92,790,152-92,967,825	14q32.12
WD REPEAT-CONTAINING PROTEIN 72	*WDR72*	613214	chr15:53,805,938-54,051,859	15q21.3
TRANSCRIPTION FACTOR Sp6	*SP6*	608613	chr17:45,922,280-45,928,516	17q21.32
DISTAL-LESS HOMEOBOX 3	*DLX3*	600525	chr17:48,067,369-48,072,588	17q21.33
FAMILY WITH SEQUENCE SIMILARITY 20, MEMBER A	*FAM20A*	611062	chr17:66,531,257-66,597,095	17q24.2
ACID PHOSPHATASE 4	*ACP4*	606362	chr19:51,293,672-51,298,481	19q13.33
KALLIKREIN-RELATED PEPTIDASE 4	*KLK4*	603767	chr19:51,409,608-51,413,994	19q13.41
AMELOGENIN	*AMELX*	300391	chrX:11,311,533-11,318,881	Xp22.2

4. Supplementary references

Boyle EA, O'Roak BJ, Martin BK, Kumar A, Shendure J. 2014. Mipgen: Optimized modeling and design of molecular inversion probes for targeted resequencing. Bioinformatics. 30(18):2670-2672.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-498.

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB et al. 2019. Predicting splicing from primary sequence with deep learning. Cell. 176(Diaz et al.):535-548.e524.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 581(7809):434-443.

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 25(14):1754-1760.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. Genome Biol. 17(1):122.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E et al. 2015. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. Genet Med. 17(5):405-424.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol. 29(1):24-26.

CHAPTER 4 Heterozygous *COL17A1* variants are a frequent cause of Amelogenesis Imperfecta

Hany U, Watson CM, Liu L, Smith CEL, Harfoush A, Poulter JA, Nikolopoulos G, Balmer R, Brown CJ, Patel A, Simmonds J, Charlton R, Acosta de Camargo MG, Rodd HD, Jafri H, Antanaviciute A, Moffat M, Al-Jawad M, Inglehearn CF, Mighell AJ. Heterozygous COL17A1 variants are a frequent cause of amelogenesis imperfecta. J Med Genet. 2023 Nov 18:jmg-2023-109510. doi: 10.1136/jmg-2023-109510. Epub ahead of print. PMID: 37979963.

This chapter introduces the subsequent paper, providing crucial background information, outlining the rationale for the research, detailing methodologies employed, and acknowledging the significant contributions made by other collaborators to enrich the overall work.

4.1 Research Rationale

This paper describes findings from the smMIPs screen and subsequent WES analyses specifically relating to AI due to variants in *COL17A1*, and the potential links to skin and corneal disease. It is well documented in the literature that biallelic *COL17A1* variants are a cause of Junctional Epidermolysis Bullosa (JEB; OMIM: 619787) (Has et al., 2020), and that monoallelic *COL17A1* variants cause the corneal dystrophy Epithelial Recurrent Erosion Dystrophy (ERED; OMIM: 122400) (Jonsson et al., 2015). However, despite recognition that heterozygous carriers in JEB families can have AI, heterozygous *COL17A1* variants are not listed in the OMIM database (www.omim.org) as a cause of dominant non-syndromic AI. They are described in association with JEB, but only via the non-specific term 'enamel hypoplasia'. Interestingly, we found that *COL17A1* gene variants are the most frequent cause of dominant AI in our cohort. We reported seventeen heterozygous, potentially pathogenic *COL17A1* variants causing non-syndromic AI in nineteen unrelated families. All the families share a common phenotype in which enamel has near normal thickness but variable focal hypoplasia, with surface irregularities including pitting. We showed that the mutation spectra for JEB, ERED, and AI were comparable, and that there may be some mutations that are involved in causing two or even all three conditions. Therefore, we propose that patients with these three conditions, caused by *COL17A1* variants, require multidisciplinary care, and that people with AI and ERED caused by *COL17A1* variants should be treated as potential JEB carriers and given the appropriate counselling. A further important consideration is that carriers in JEB families may require additional dental and ophthalmological care.

4.2 Research Contribution

Patients were diagnosed and recruited either by my supervisor Dr Alan Mighell or by other dental colleagues who collaborate with him. For all of the patients and additional family members that participated in this study, I carried out the wet laboratory tasks relating to smMIPs and WES genetic screening. Lu Liu performed the wet laboratory work for Sanger sequencing under my supervision. I analysed all of the data generated by smMIPs and WES. During the writing process, I created the initial draft of the paper, edited the master copy, and added comments from reviewers.

4.2.1 Additional methodology

Further details of the methodology are available in the manuscript.

Original research

Heterozygous *COL17A1* variants are a frequent cause of amelogenesis imperfecta

Ummey Hany ,[1] Christopher M Watson ,[1,2] Lu Liu,[1,3] Claire E L Smith,[1] Asmaa Harfoush,[3] James A Poulter ,[1] Georgios Nikolopoulos,[4] Richard Balmer,[3] Catriona J Brown,[5] Anesha Patel,[6] Jenny Simmonds,[2] Ruth Charlton,[2] María Gabriela Acosta de Camargo,[7] Helen D Rodd,[8] Hussain Jafri,[9] Agne Antanaviciute,[10] Michelle Moffat,[11] Maisoon Al-Jawad,[3] Chris F Inglehearn,[1] Alan J Mighell[3]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/jmg-2023-109510).

For numbered affiliations see end of article.

Correspondence to
Alan J Mighell, School of Dentistry, Clarendon Way, University of Leeds, Leeds LS2 9NL, UK;
a.j.mighell@leeds.ac.uk

CFI and AJM are joint senior authors.

Received 13 July 2023
Accepted 17 October 2023

To cite: Hany U, Watson CM, Liu L, *et al*. *J Med Genet* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jmg-2023-109510

ABSTRACT

Background Collagen XVII is most typically associated with human disease when biallelic *COL17A1* variants (>230) cause junctional epidermolysis bullosa (JEB), a rare, genetically heterogeneous, mucocutaneous blistering disease with amelogenesis imperfecta (AI), a developmental enamel defect. Despite recognition that heterozygous carriers in JEB families can have AI, and that heterozygous *COL17A1* variants also cause dominant corneal epithelial recurrent erosion dystrophy (ERED), the importance of heterozygous *COL17A1* variants causing dominant non-syndromic AI is not widely recognised.

Methods Probands from an AI cohort were screened by single molecule molecular inversion probes or targeted hybridisation capture (both a custom panel and whole exome sequencing) for *COL17A1* variants. Patient phenotypes were assessed by clinical examination and analyses of affected teeth.

Results Nineteen unrelated probands with isolated AI (no co-segregating features) had 17 heterozygous, potentially pathogenic *COL17A1* variants, including missense, premature termination codons, frameshift and splice site variants in both the endo-domains and the ecto-domains of the protein. The AI phenotype was consistent with enamel of near normal thickness and variable focal hypoplasia with surface irregularities including pitting.

Conclusion These results indicate that *COL17A1* variants are a frequent cause of dominantly inherited non-syndromic AI. Comparison of variants implicated in AI and JEB identifies similarities in type and distribution, with five identified in both conditions, one of which may also cause ERED. Increased availability of genetic testing means that more individuals will receive reports of heterozygous *COL17A1* variants. We propose that patients with isolated AI or ERED, due to *COL17A1* variants, should be considered as potential carriers for JEB and counselled accordingly, reflecting the importance of multidisciplinary care.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ There is an established understanding that biallelic *COL17A1* variants are a cause of junctional epidermolysis bullosa (JEB).

WHAT THIS STUDY ADDS

⇒ Heterozygous *COL17A1* variants are a much more common cause than previously recognised of isolated autosomal dominant amelogenesis imperfecta (AI) (developmental enamel defects) in the absence of mucocutaneous disease.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ As genetic testing availability increases, including as part of AI and corneal dystrophy care, more individuals will receive reports of heterozygous *COL17A1* variants.

⇒ This study provides a reference point to inform how genetic counselling and clinical care are advanced.

⇒ Furthermore, there is a need to consider specialist dental and ophthalmic evaluation of carriers in JEB families to ensure that their care needs are also being met.

INTRODUCTION

Collagen type XVII alpha 1 chain (COL17A1), hereafter referred to as collagen XVII, is a hemidesmosomal transmembrane protein widely expressed in humans. It has diverse biological functions in cell adhesion, morphogenesis, neuromuscular signalling and host defence.[1] As a hemidesmosome component, it is present in the cutaneous basement membrane zone, which connects the skin epidermis and dermis.[2 3] It is also expressed during amelogenesis, the process by which dental enamel is formed, and contributes to the differentiation of ameloblasts.[4 5] Col17 knockout ($Col17^{-/-}$) mice exhibit distorted Tomes' processes, a reduced volume of enamel matrix during the secretory stage and prolonged calcification in the maturation stage of amelogenesis.[4]

The 56 exons of the *COL17A1* gene encode a 1497-amino acid protein which acts as a homotrimer composed of three alpha (α1) chains, each with a molecular mass of 180 kDa. These consist of a globular amino-terminal intracellular endodomain, a short transmembrane domain and

Genotype-phenotype correlations

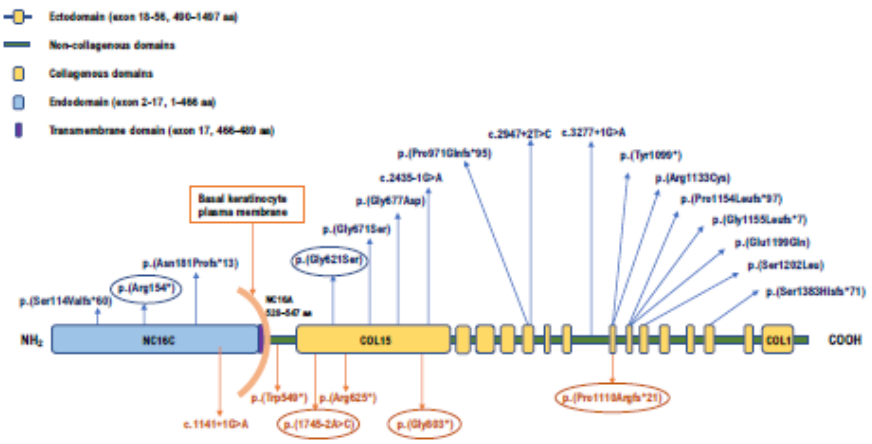


Figure 1 Schematic representation of the domain organisation of the collagen XVII protein. The extracellular domain or ectodomain is comprised of 15 collagenous (COL1–COL15, yellow vertical boxes) flanked by stretches of non-collagen sequence (NC1–NC16a, green horizontal lines). The non-collagen domain, NC16, spans from the extracellular matrix to cytoplasm and comprises a transmembrane domain NC16b adjoined by NC16a and NC16c to the C-terminal and N-terminal ends respectively. *COL17A1* variants identified in this study are denoted in blue text above the protein domains and variants published by others as causes of amelogenesis imperfecta are displayed in orange text below the protein domains.[25 50] The circled variants have been previously published in association with junctional epidermolysis bullosa.

a flexible rod-like carboxy-terminal extracellular ectodomain.[6] The ectodomain consists of 15 collagenous (COL1–COL15) sequences containing repeating Gly-X-Y tripeptides which, in the homotrimer, form the characteristic collagen triple helices. These are flanked by 16 non-collagenous sequences (NC1–NC16) (figure 1).[7] A notable characteristic of collagen XVII is the shedding of the ectodomain after cleavage at the cell surface by the sheddases ADAM 9, 10 and 17, to yield its soluble intracellular form; the biological significance of this remains to be determined.[8]

Biallelic variants in *COL17A1* (OMIM 113811) are a well-documented cause of the recessively inherited, genetically heterogeneous mucocutaneous blistering condition junctional epidermolysis bullosa (JEB).[9] JEB is genetically heterogeneous and characterised by erosions and blistering of the skin and mucous membranes, with cleavage at the basement membrane zone. There are a range of clinical presentations, but it is generally classified into one of two major subtypes: intermediate or severe.[10] JEB prevalence is estimated to be approximately 2 per million live births in the USA and 1 per million in England and Wales.[11–13]

Corneal epithelial erosions and enamel hypoplasia are distinct features in JEB families.[13] Corneal erosions are found in only a proportion of JEB cases,[14] while it has been reported that enamel hypoplasia is always associated with JEB.[15] Hintner and Wolff[16] first reported defective enamel in their patients with JEB, and since then, enamel hypoplasia in association with JEB has been further corroborated.[17 18] In comparison to healthy enamel, the enamel of patients with JEB has increased tissue porosity, reduced mineral content and contains serum albumin, with enamel hypoplasia.[19] These studies identified JEB on clinical features, without knowing which of the genes known to cause JEB was responsible. In papers primarily about JEB due to biallelic *COL17A1* changes, carrier parents or siblings of patients with JEB have been described as having developmental enamel defects, typically using the non-specific descriptive term enamel hypoplasia.[20 21]

Monoallelic *COL17A1* variants can also cause dominantly inherited epithelial recurrent erosion dystrophy (ERED, OMIM 122400), a corneal disease with the potential for lifelong progression and vision loss (three variants reported).[22 23] Furthermore, there are documented cases of monoallelic *COL17A1* variants causing dominantly inherited amelogenesis imperfecta (AI) in the absence of other co-segregating features or any family history of JEB. AI is a developmental failure of normal dental enamel formation affecting all teeth, which can be inherited as a dominant, recessive or X-linked trait, either in isolation or as a component of syndromic conditions.[24] Dominant isolated AI caused by heterozygous *COL17A1* variants has only been reported in two cohort studies, where *COL17A1* was only one of several genes implicated, and in one report of a genetically complex AI family (total six variants),[25–27] and *COL17A1* variants were not listed as a cause of AI in OMIM at the time of submission (13 July 2023). The distinction between AI and descriptive terms such as enamel hypoplasia is important. The latter term does not link to aetiology or inheritance, unlike AI.

Here, we describe 19 unrelated families with isolated AI in which probands are heterozygous for 17 different monoallelic *COL17A1* variants, consistent with this being a frequent cause of autosomal dominant AI presenting in the absence of other clinical features.

MATERIALS AND METHODS

Patient recruitment

Patients were recruited though UK dental clinics, with informed written consent and local ethical approval (REC 13/YH/0028), in accordance with the principles of the Declaration of Helsinki. Genomic DNA was obtained from venous blood using conventional extraction techniques, or from saliva using Oragene DNA

Sample Collection kits (DNA Genotek). Screening of families F2-F19 was carried out by the University of Leeds Amelogenesis Research Group, while family F1 was screened via the NHS testing service for AI (https://www.england.nhs.uk/publication/national-genomic-test-directories/).

Reference genome and transcript

The human reference genome used for this study was GRCh37/hg19, the transcript sequence of *COL17A1* used was NM_000494.4 and the collagen XVII protein sequence used was NP_000485.

Whole exome sequencing (WES)

Two different hybridisation capture reagents were used for WES library preparation: the SureSelect Human All Exon V6 kit (Agilent Technologies) and the Human Comprehensive Exome kit (10–50 Mb) (Twist Bioscience). For SureSelect, 3 µg genomic DNA was used to make libraries, following the manufacturer's protocol. These were sequenced on a HiSeq 3000 which generated paired-end 150 bp reads (Illumina). For the Twist kit, 50 ng genomic DNA was used to make libraries following the manufacturer's protocol. Sequencing was carried out on a NextSeq 2000 using a P3 kit to generate paired-end 150 bp reads (Illumina).

The quality of the raw sequence reads was reviewed using FastQC (V.0.11.3) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). These were then aligned to an indexed human reference genome using BWA (V.0.7.12) (https://bio-bwa.sourceforge.net/).[28] PCR duplicates were removed using Picard (V.2.5.0) (https://broadinstitute.github.io/picard/). Non-reference bases were identified and recorded in variant call format using the Genome Analysis Tool Kit (GATK) HaplotypeCaller (V.3.5) (https://gatk.broadinstitute.org).[29] Prior to filtering, identified variants were annotated with functionally relevant biological information and observed population allele frequencies using the Variant Effect Predictor (VEP) (V.83).[30]

Targeted sequencing of known AI genes (NHS)

Genes in the R340 (amelogenesis imperfecta) panel of the UK NHS National Genomic Test Directory were subject to hybridisation capture using a custom SureSelect reagent. Libraries were generated from 3 µg genomic DNA according to the manufacturer's instructions (Agilent Technologies). Libraries were sequenced using a NovaSeq 6000 (Illumina). The resulting FASTQ files were processed and aligned as described above.

Single molecule molecular inversion probes (smMIP) sequencing

smMIPs targeting the coding sequences of 19 genes (online supplemental table S4) implicated in non-syndromic AI were designed using MIPGEN (https://github.com/shendurelab/MIPGEN)[31] and synthesised by Integrated DNA Technologies (IDT, Leuven, Belgium) at 100 nmol scale. Then, 100 ng genomic DNA was subjected to targeted capture and ligation using the smMIPs probe pool diluted to reach a ratio of 800 smMIPs copies for every one DNA molecule in the final capture reaction.[32] Sequencing was carried out on a NextSeq 500 (Illumina) which generated paired-end 150 bp reads.

Data processing was performed using the MIPVAR pipeline (https://sourceforge.net/projects/mipvar/) which was modified for compatibility with local computing hardware. This enabled sample demultiplexing and the removal of unique molecular identifiers prior to ligation-arm and extension-arm processing using standard tools BWA (V.0.7.12), Picard (V.1.102.0) and the GATK HaplotypeCaller (V.3.2–2).

Variant classification

The pathogenicity of the variants was assessed according to the American College of Medical Genetics and Genomics (ACMG) criteria using Franklin by Genoox (https://franklin.genoox.com/clinical-db/home).[33] Allele frequencies were obtained from the Genome Aggregation Database V.2.1.1 (https://gnomad.broadinstitute.org/).[34] Splicing predictions were generated using SpliceAI (https://spliceailookup.broadinstitute.org).[35]

Sanger sequencing verification

Primers were designed using AutoPrimer3 (https://github.com/david-a-parry/autoprimer3) and synthesised by IDT. About 25 ng genomic DNA was amplified using Q5 High-Fidelity 2X Master Mix (NEB) according to manufacturer's instructions. PCR products were purified using ExoSAP-IT (Applied Biosystems) then sequenced using BigDye Terminator V.3.1 chemistry on an ABI3130xl Genetic Analyser (Applied Biosystems). Electropherograms were analysed using SeqScape software V.2.5 (Applied Biosystems).

Micro-computed tomography (µCT)

Intact teeth were analysed using a high resolution µ-CT SkyScan 1172 (Bruker, Belgium) scanner to quantify mineral density. Mineral density values were calculated relative to three hydroxyapatite standards of 0.25 and 0.75 g/cm^3 (Bruker, Belgium) and 2.9 g/cm^3 (Himed, USA). Fiji/ImageJ was used to analyse enamel density, with a pixel threshold above 2.0 g/cm^3.

Scanning electron microscopy (SEM)

Longitudinal mid buccal slices of teeth were obtained using an Accutom 10 cutting machine and diamond cutting wheel (Struers, Germany). After removing surface debris, slices were gold coated (Agar Scientific, Elektron Technology, UK). Imaging was performed by S-3400N (Hitachi, Japan) SEM.

RESULTS

Cohort screening

Genomic DNA from probands in a large cohort of apparently unrelated families with non-syndromic AI were investigated either by targeted smMIP screening, NHS diagnostic AI screening or WES. Variants identified were excluded if the CADD score was <15 or minor allele frequency was >0.001. The variant list for each case was then filtered further to include only candidate pathogenic variants in known or potential candidate AI genes. Where possible, additional filtering was performed based on family history. Probands from 19 families were identified as carrying potentially pathogenic variants in the *COL17A1* gene. In each case, this was the only variant in the known non-syndromic AI genes that met these criteria. Of these, inheritance in 12 families appeared dominant, while the mode of inheritance could not be determined in the remaining families due to incomplete clinical information about additional family members. Variant pathogenicity was assigned according to the ACMG criteria, with the outcome being pathogenic, likely pathogenic or a variant of unknown significance (VUS). All the suspected variants were re-sequenced by Sanger sequencing and their segregation with disease was checked in available family members (figure 2 and online supplemental figure S2).

Genotype-phenotype correlations

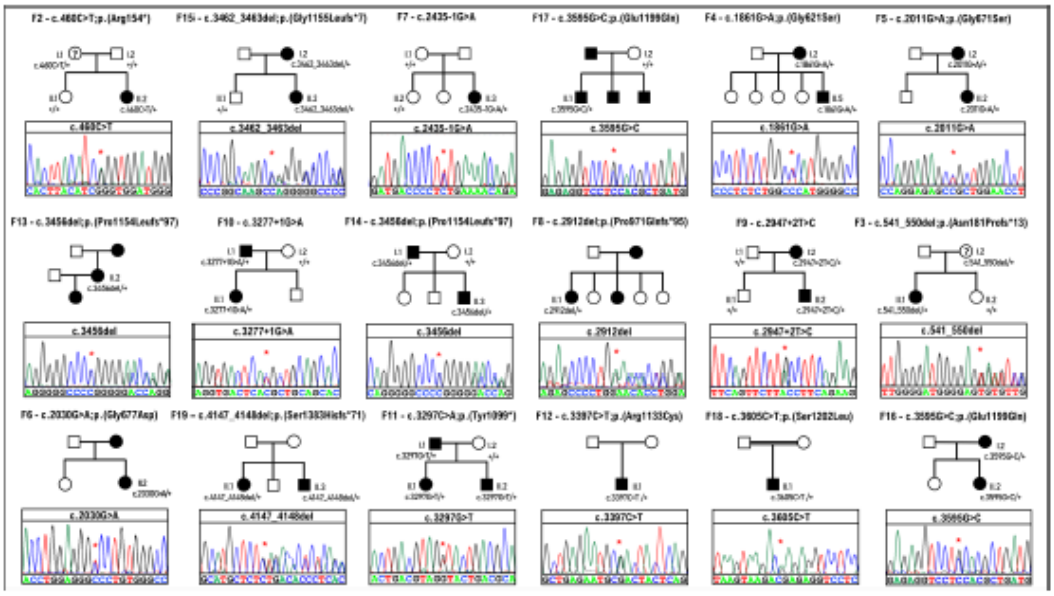


Figure 2 Pedigrees of 18 of the 19 families recruited for this study. The Sanger sequencing chromatogram from the proband from each family is displayed beneath each pedigree. Details of the family pedigree and variant identified in family F1 are displayed in online supplemental figure S2. A '?' mark in the pedigree means 'individuals with possible AI not clinically assessed'.

COL17A1 variants

Probands in the 19 non-syndromic AI families displayed in figure 2 and online supplemental figure S2 were found to carry heterozygous, potentially pathogenic variants in *COL17A1*. Two variants were present in two families, with the remainder each occurring in only one family. None of the 17 variants described in this study were previously associated with AI. Only seven are present in the gnomAD database (table 1). Of the 17 variants, 6 are missense: c.1861G>A; p.(Gly621Ser), c.2011G>A; p.(Gly671Ser), c.2030G>A; p.(Gly677Asp), c.3397C>T;

Table 1 Details of *COL17A1* variants reported in this study

Family ID	ACMG criteria	Variants			CADD score	gnomAD frequency
		Genomic nomenclature	Transcript nomenclature	Predicted protein nomenclature		
F1	P (PVS1, PM2, PP5)	g.105833981del	c.340del	p.(Ser114Valfs*60)	32.0	0.00001591
F2	P (PP1, PP4, PVS1, PM2, PP5)	g.105831793G>A	c.460C>T	p.(Arg154*)	36.0	0.000003977
F3	P (PP1, PP4, PVS1, PM2)	g.105830245_105830254del	c.541_550del	p.(Asn181Profs*13)	32.0	Absent
F4	LP (PP1, PP3, PP4, PM2)	g.105812867C>T	c.1861G>A	p.(Gly621Ser)	23.9	Absent
F5	LP (PP1, PP3, PP4, PM2)	g.105811266C>T	c.2011G>A	p.(Gly671Ser)	26.3	0.0001135
F6	LP (PP3, PP4, PM2)	g.105811247C>T	c.2030G>A	p.(Gly677Asp)	26.1	Absent
F7	LP (PP1, PP4, PVS1, PM2)	g.105803340C>T	c.2435–1G>A	p.?	34.0	Absent
F8	P (PP4, PVS1, PM2)	g.105798865del	c.2912del	p.(Pro971Glnfs*95)	33.0	Absent
F9	LP (PP1, PP4, PVS1, PM2)	g.105798827A>G	c.2947+2T>C	p.?	30.0	Absent
F10	P (PP1, PP4, PVS1, PM2, PP5)	g.105796802C>T	c.3277+1G>A	p.?	27.7	0.00006312
F11	P (PP1, PP4, PVS1, PM2)	g.105796371G>T	c.3297C>A	p.(Tyr1099*)	36.0	Absent
F12	VUS (PP4, PM2)	g.105796271G>A	c.3397C>T	p.(Arg1133Cys)	33.0	Absent
F13	P (PP4, PVS1, PM2)	g.105795287del	c.3456del	p.(Pro1154Leufs*97)	20.6	0.000008021
F14	P (PP1, PP4, PVS1, PM2)	g.105795287del	c.3456del	p.(Pro1154Leufs*97)	20.6	0.000008021
F15	P (PP1, PP4, PVS1, PM2)	g.105795277_105795278del	c.3462_3463del	p.(Gly1155Leufs*7)	33.0	Absent
F16	LP (PP1, PP4, PS4, PM2)	g.105795045C>G	c.3595G>C	p.(Glu1199Gln)	25.1	Absent
F17	LP (PP1, PP4, PS4, PM2)	g.105795045C>G	c.3595G>C	p.(Glu1199Gln)	25.1	Absent
F18	VUS (PP4, PM2)	g.105795035G>A	c.3605C>T	p.(Ser1202Leu)	27.0	0.00002800
F19	P (PP1, PP4, PVS1, PM2)	g.105793715_105793716del	c.4147_4148del	p.(Ser1383Hisfs*71)	34.0	Absent

ACMG scoring criteria: PP1: segregation data pathogenic supporting; PP4: phenotype pathogenic supporting; PS4: case control studies pathogenic strong; PVS1: effect on protein pathogenic very strong; PM2: population data pathogenic moderate; PP5: reputable source data pathogenic supporting.
Nomenclature is reported according to *COL17A1* transcript NM_000494.4 and chromosome 10 of human reference genome build hg19.
CADD V.1.3, combined annotation dependent depletion; gnomAD V.2.1.1, genome aggregation database; LP, likely pathogenic; P, pathogenic; VUS, variant of unknown significance.

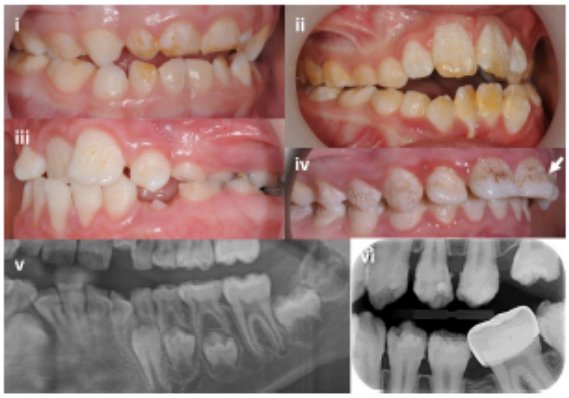


Figure 3 Intraoral images and dental radiographs illustrating the variation in enamel phenotypes associated with heterozygous *COL17A1* variants in primary and secondary teeth. (i) Primary tooth enamel changes can be minimal and easily missed and are primarily characterised by hypomaturation changes with subtle surface focal pitting (F10). (ii) A predominantly hypomaturation AI phenotype with some surface irregularities (F9). (iii) Surface pits and other irregularities are the clinically dominant feature, on a background of hypomaturation (F4). (iv) Hypomaturation enamel is combined with more exaggerated surface pits merging into grooves with mid-third crown regional hypoplasia (arrow) (F3). (v) Section of an orthopantomogram of a mixed dentition illustrating near normal enamel thickness with a normal difference in radiodensity between the enamel and the supporting dentine (F15). (vi) Intraoral radiograph illustrating near normal enamel thickness, but with enamel irregularities and a lesser difference in radiodensity between enamel and dentine than would be expected (F5). Further clinical images and dental radiographs are included in online supplemental figure S4.

p.(Arg1133Cys), c.3595G>C; p.(Glu1199Gln), c.3605C>T; p.(Ser1202Leu); six result in frameshifts: c.340del; p.(Ser114Valfs*60), c.541_550del; p.(Asn181Profs*13), c.2912del; p.(Pro971Glnfs*95), c.3456del; p.(Pro1154Leufs*97), c.3462_3463del; p.(Gly1155Leufs*7) and c.4147_4148del; p.(Ser1383Hisfs*71); 3 are predicted to affect splice sites: c.2435–1G>A (acceptor loss score 0.99, acceptor gain score 0.92), c.2947+2T>C (donor loss score 0.98) and c.3277+1G>A (donor loss score 0.97); and 2 create premature termination codons (PTC): c.460C>T; p.(Arg154*) and c.3297C>A; p.(Tyr1099*).

All the stop and frameshift variants identified are classified as pathogenic, while the three splice site variants are classified as pathogenic or likely pathogenic. Among the six missense variants three are glycine substitutions, and these are classified as likely pathogenic. Of the three remaining missense variants, one, p.(Glu1199Gln), was initially classed as a VUS, but is absent from gnomAD and was observed to co-segregate in two families reported here, leading to reclassification as likely pathogenic. The remaining two non-glycine missense variants are currently classified as variants of unknown significance (VUS). All missense variants identified are in the extracellular domain of the protein (figure 1).

Oral clinical phenotype

All families presented as isolated AI with no history of co-segregating health issues. Variability in the clinical AI phenotype was evident, with features that reflected enamel qualitative and quantitative changes (figure 3 and online supplemental figure S4).

Clinical enamel changes in the primary dentition were minimal and could be easily overlooked. Hypomaturation changes were more evident where there had been some post-eruptive enamel loss. Focal surface pitting was subtle.

In the secondary dentition, there was generalised, but clinically variable enamel hypomaturation characterised by white to yellow/brown colouration and greater enamel opacity than expected. Surface irregularities were also variable, with distinct, deep pits that in some instances were obvious due to extrinsic staining or formed linear, vertical defects in the most pronounced cases. Shallow surface irregularities were also present in some teeth. Regional enamel hypoplasia involving the middle third of the labial aspect of anterior teeth was observed in some cases. Dental radiographs confirmed that enamel thickness was for the most part within expected normal limits. A clear distinction between the radiodensity of enamel compared with the supporting dentine confirmed that any reduction in enamel mineralisation was at the mild end of the spectrum, consistent with the clinical hypomaturation phenotype. Post-eruption enamel loss was not obviously exaggerated.

Tooth root morphology including pulp spaces was within expected normal limits with no taurodontism. No oral mucosal or other oral cavity changes were evident.

Laboratory analysis of teeth

Upper primary molar teeth from affected members of families F9 and F14 were analysed by three-dimensional µCT and SEM and compared with the relevant control teeth (figure 4, online supplemental figure S3). µCT revealed near normal enamel volume in the affected teeth, but they lacked a hard outer enamel layer and mineral density gradation from higher to lower moving from the outer enamel towards the dental enamel junction (DEJ), by comparison with the control teeth. µCT also revealed a pitted and uneven enamel surface in the probands' teeth, in both primary and permanent dentitions, confirmed by SEM analysis, which showed the presence of pitting, and disruption of the enamel layers appearing as a stack of lamellae, with patches of fused rod-interrod regions hard to distinguish between them (figure 4 and online supplemental figure S3).

Wider clinical phenotype

None of the affected individuals described here were noted to have skin or mucosal abnormalities, corneal problems or any other associated conditions. However, all were recruited in dental clinics as cases of non-syndromic AI and have not been examined by other clinical specialists for subtle skin or corneal presentations.

DISCUSSION

Here, we report 15 pathogenic/likely pathogenic heterozygous *COL17A1* variants as the likely cause of non-syndromic AI in 17 probands, as well as 2 further cases with VUSs that may also be causative. This greatly increases the previous tally of six, within an increasingly clear context that *COL17A1* variants are a frequent and under recognised cause of dominantly inherited AI. The AI phenotype observed is consistent with the limited clinical images, radiographs and other data in the peer-reviewed literature from carriers in JEB families who are heterozygous for *COL17A1* variants.

AI due to heterozygous *COL17A1* variants has been linked to the Witkop classification type 1a pitted hypoplastic AI.[26 36]

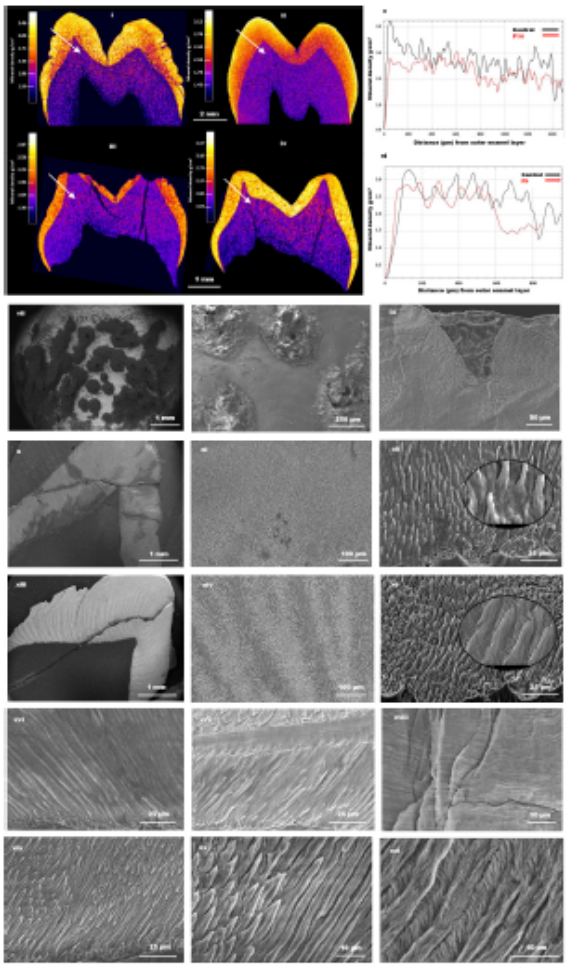


Figure 4 Laboratory analysis of teeth. (i) Micro-computed tomography (μ-CT) imaging of a permanent upper first premolar tooth from the proband of F14 and (iii) a primary upper molar tooth from the affected individual from F9. Panels (ii) and (iv) have images from corresponding control teeth. No significant differences were observed in average enamel mineral density (EMD) between affected and control samples. F14 and its corresponding control had EMD values of 2.58 and 2.60 g/cm^3, respectively, while F9 and its corresponding control were 2.40 and 2.52 g/cm^3, respectively. (v–vi) Line graphs showing the distribution of mineral density from the enamel surface to the dental enamel junction (DEJ), as shown by the arrows in (i–iv). Affected samples (red) lack high mineral density at the surface, as opposed to the control teeth (black). Scanning electron microscopy (SEM) images of the enamel in F14 (vii–viii) and F9 show clear pitting extending towards DEJ.[51] SEM of enamel from F14 (x–xii) shows generally disrupted and poorly formed prismatic microstructure compared with the corresponding control teeth in the images (xiii–xv), respectively. SEM images of the enamel prism in F9 (xvi–xviii) appears as a stack of lamellae, with patches of fused rod-interrod regions hard to distinguish between them. However, enamel from corresponding controls show distinguishable rod interrod regions (xix–xxi).

Witkop described hypoplastic, pitted autosomal dominant type enamel with pits from pinpoint to pinhead size primarily on labial or buccal surfaces of permanent teeth, often arranged in rows and columns, but with comment that some teeth may appear normal in both dentitions. The Witkop classification evolved over time but remained primarily clinically descriptive, with patterns of inheritance included in some instances. Data presented here highlight the advantages of switching to classification where genetic diagnosis has primacy and is correlated to the clinical enamel phenotype that can vary within certain parameters. This is also with recognition that enamel does not have cellular capacity for repair and that the enamel phenotype is altered by post-eruption changes. The enamel present is generally well mineralised but shows disrupted enamel rod morphology, which can be expected to adversely impact enamel functional longevity. Teeth from individuals with JEB due to biallelic *COL17A1* variants were not available for comparative analysis. In summary, in this series affected enamel has hypomaturation characteristics with variable focal hypoplasia (pits and indentations) and in some instances, partial regional hypoplasia of the middle third of the tooth enamel.

The profound adverse impact of JEB on the affected individuals and their families has driven our understanding of how the condition is caused by pathological biallelic variants in *COL17A1*, *LAMA3*, *LAMB3*, *LAMC2*, *ITGA6*, *ITGB4* and *ITGA3*. According to the England and Wales EB database, JEB prevalence is around 1 per million, with most pathogenic variants detected in *LAMB3* (40%–50%), followed by *LAMC2* (15%–20%) and *LAMA3* (10%–15%), with only a small proportion (5%–10%) in *COL17A1* (John McGrath, personal communication). By contrast, the association of AI with heterozygous variants in these genes in families with dominant inheritance and no history of JEB, are less obviously presented in the published literature, which also fails to clarify whether affected individuals are carriers for JEB. While all individuals with JEB have AI (or enamel hypoplasia), there are very few reports of AI in carriers of JEB due to *COL17A1* variants, and it remains unclear what proportion of carriers will manifest enamel or corneal abnormalities.[21 37] Assuming 1 in 10 million people have JEB due to *COL17A1* variants, Hardy-Weinberg equilibrium would predict a carrier frequency of approximately 1 in 1600, not inconsistent with published estimates of the frequency of AI,[38 39] especially given that many such individuals may have been considered to have enamel hypoplasia or enamel opacities rather than inherited AI. This highlights two important related points where a molecular diagnosis can inform clinical decision-making. First, distinguishing between more subtle forms of AI and other enamel development defects. Second, that dental changes offer an opportunity to identify carrier status for JEB in families with no history of this condition.

If it is assumed that all JEB carriers have AI, then one might expect cases of AI due to variants in *LAMB3*, *LAMC2* and *LAMA3* to be more common than those with *COL17A1* variants, given the frequency of the different forms of JEB. Variants in all three genes have been reported in patients with AI in the literature but only in a handful of cases for each,[40 41] while our findings show that variants in *COL17A1* are a relatively common cause of AI. It is therefore evident that further research is needed into the link between dominant AI and recessive JEB due to *COL17A1* variants.

We identified missense, PTC, frameshift and splice site variants in both the endo-domains and the ecto-domains of the protein. Fifteen of the variants described are novel, while two have been previously reported as pathogenic in JEB but not in isolated AI. Patients in this study were not reported to have any associated skin or corneal problems but have not been examined by dermatologists or cornea specialists, meaning that subtle versions of

either condition could potentially have been overlooked. We wanted to understand whether the *COL17A1* variants associated with AI differ from those that cause JEB. By combining a literature search on the NCBI database (https://pubmed.ncbi.nlm.nih.gov/) with data from the HGMD professional database (accessed 30 March 2023),[42] we identified 232 *COL17A1* variants reported to cause JEB (online supplemental table S1). The distribution of mutations and mutation types in JEB and AI are similar (online supplemental figure S1). Variants c.460C>T; p.(Arg154*) and c.1861G>A; p.(Gly621Ser), reported here as causing AI, and variant c.1745–2A>C, c.2407G>T; p.(Gly803*) and c.3327del; p.(Pro1110Argfs*21) reported to cause AI by Prasad and colleagues,[25] have also been identified as pathogenic in JEB,[5 43 44] showing there is overlap in the underlying genetic basis of these conditions.

Only three *COL17A1* variants have been reported in the literature as causing the corneal disease ERED (online supplemental table S3). None of these have been implicated in AI or JEB. The nonsense variant, p.(Arg154*), detected in an AI proband in this study and in JEB, was reported in ClinVar (VCV000931124.3) as causing autosomal dominant ERED. However, no further evidence was provided, and this result remains unpublished at the time of writing, meaning this should be considered unconfirmed at this stage. It is unknown if individuals with ERED due to dominant *COL17A1* variants have AI, but the expectation until demonstrated otherwise is that they will, although many of these patients will not have been thoroughly examined by dentists.

Most variants associated with JEB, AI and ERED are frameshift, splice or PTC. The consequences of these variants have not been determined experimentally, but it seems likely that they will be subject to nonsense mediated decay,[45] meaning no functional protein is produced from those alleles. Many of the individuals with JEB due to *COL17A1* variants are homozygous for such variants, meaning their phenotype is in effect the result of complete knockout of *COL17A1*. It therefore seems likely that many with JEB suffer from near-complete loss of Collagen XVII protein. Since two of the PTCs implicated in JEB were also found in AI, and given the unconfirmed report of one of the same variants in an ERED case, it therefore seems likely that AI and ERED due to heterozygous *COL17A1* pathogenic variants is caused, at least in some cases, by haploinsufficiency. Further evidence for this disease mechanism comes from the work of Yuen and colleagues,[37] who used immunofluorescence staining with antibodies targeted to mouse Col17 to show reduced basement membrane zone and apical–lateral staining in skin from both JEB patients and carriers compared with control skin.

A proportion of the *COL17A1* variants observed in individuals with JEB, AI and ERED are amino acid substitutions. These may also be functional knockouts, but an alternative disease mechanism has been proposed in some of these cases. Missense variants, and most commonly glycine substitutions, have been reported to be associated with milder JEB phenotypes.[17] Substitution of glycine residues within the ectodomain, and particularly within the COL15 collagenous sequence (figure 1), is thought to destabilise the collagenous triple helix, making the protein unstable, with the mutated protein predicted to exert a dominant negative effect on the wild-type protein.[46 47] Interestingly, of the six missense variants reported here in patients with AI, three were glycine substitutions in the COL15 region.

An interesting case describes a patient with JEB who is a compound heterozygote for glycine substitution p.Gly627Val within the COL15 domain and frameshift insertion c.3514ins25 within the COL6 domain.[20] The proband had an abnormal dentition, with complete loss of all teeth by age 14. The proband's daughter, who is a heterozygous carrier of the p.Gly627Val variant, showed no skin abnormalities but had extensive enamel hypoplasia and pitting. The proband's granddaughter, who was also a carrier of the p.Gly627Val variant, manifested dental abnormalities and trauma-induced skin blistering, especially around the knees. The authors concluded that p.Gly627Val has a dominant negative effect on the collagen XVII protein, causing autosomal dominant JEB in the granddaughter.[20 48] As well as providing further evidence of a dominant negative disease mechanism and of overlap between the *COL17A1* variants causing JEB and AI, this case illustrates the importance of a multidisciplinary approach to the clinical care of such patients.

These findings have significant implications for future care of individuals and their families with diagnoses of JEB, AI or ERED due to pathogenic variants in *COL17A1*. Further studies are needed to better understand links between these conditions, but it seems likely that there is overlap between carrier status for JEB and a diagnosis of AI or ERED when they result from heterozygous *COL17A1* pathogenic variants. This may not have been fully appreciated by disparate groups of clinicians treating each condition in isolation. The mucocutaneous lesions of JEB are generally so severe that corneal or dental problems may not have been prioritised in patients and could have been overlooked in their carrier parents or siblings. ERED manifests at around 5 years, but may resolve by the early 20s, meaning many adults with the condition are without symptoms. AI may be dismissed by non-experts as resulting from poor dental hygiene, especially in children with EB, who have considerable difficulty in maintaining oral hygiene for multiple reasons.[49]

To summarise, we identified 17 families with AI due to pathogenic/likely pathogenic heterozygous variants in the *COL17A1* gene, and a further 2 families with variants of unknown significance in *COL17A1* that may also be pathogenic. These findings suggest that the significance of *COL17A1* variants as a cause of AI has not been fully appreciated and this may in fact be a relatively common form of dominantly inherited AI. We detail the spectrum of the enamel phenotype observed and review all the pathogenic *COL17A1* variants known to cause AI. A comparison with those causing the recessive skin disorder JEB suggests they are similar in mutation type and distribution, and there is also direct overlap between the variants implicated in both conditions, and possibly in a third, the dominantly inherited corneal disorder ERED. People with AI or ERED due to heterozygous *COL17A1* variants should be considered carriers for JEB. Furthermore, these results highlight the need for a multidisciplinary approach to the care of families and individuals with JEB, including carriers, and those with dominant AI or ERED due to *COL17A1* variants.

Author affiliations
[1]Leeds Institute of Medical Research, University of Leeds, St. James's University Hospital, Leeds, UK
[2]North East and Yorkshire Genomic Laboratory Hub, Central Lab, St. James's University Hospital, Leeds, UK
[3]School of Dentistry, Clarendon Way, University of Leeds, Leeds, UK
[4]Institute for Fundamental Biomedical Research, B.S.R.C. 'Alexander Fleming', Vari, Attica, Greece
[5]Birmingham Dental Hospital, Mill Pool Way, Edgbaston, Birmingham, UK
[6]LCRN West Midlands Core Team, NIHR Clinical Research Network (CRN), Birmingham Research Park (West Wing), Vincent Drive, Edgbaston, Birmingham, UK
[7]Department of Dentistry of the Child and Adolescent, Universidad de Carabobo, Carabobo, Venezuela
[8]Academic Unit of Oral Health Dentistry and Society, School of Clinical Dentistry, University of Sheffield, Sheffield, UK
[9]Fatima Jinnah Medical University, Punjab Thalassaemia and Other Genetic Disorders Prevention and Research Institute, Lahore, Pakistan

[10]MRC Human Immunology Unit, University of Oxford, Oxford, UK
[11]Paediatric Dentistry, The Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

Twitter Christopher M Watson @ChrisM_Watson, James A Poulter @jamesapoulter and Maria Gabriela Acosta de Camargo @gaviota113

Acknowledgements The authors thank the families involved for their support for this study.

Contributors UH, CMW, CELS, JAP, CI and AM contributed to conception, design, data acquisition and interpretation, drafting and critical review of the manuscript. LL, AH, GN, RB, CJB, AP, JS, RC, MGAdC, HDR, HJ, AA, MM and MA-J contributed to data acquisition, interpretation and critical review of the manuscript. All authors approved the final version of the manuscript. Guarantor: CFI.

Funding This work was supported by Rosetrees Trust Grant PGS19-2/10111, Wellcome Trust grant number WT093113MA and a Leeds Doctoral Scholarship awarded to UH.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by REC 13/YH/0028. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iDs
Ummey Hany http://orcid.org/0000-0002-4486-1625
Christopher M Watson http://orcid.org/0000-0003-2371-1844
James A Poulter http://orcid.org/0000-0003-2048-5693

REFERENCES

1 Franzke CW, Bruckner P, Bruckner-Tuderman L. Collagenous transmembrane proteins: recent insights into biology and pathology. *J Biol Chem* 2005;280:4005–8.
2 Eady RAJ, McGrath JA, McMillan JR. Ultrastructural clues to genetic disorders of skin: the dermal-epidermal junction. *J Invest Dermatol* 1994;103:S13–8.
3 Pulkkinen L, Uitto J. Mutation analysis and molecular genetics of epidermolysis bullosa. *Matrix Biol* 1999;18:29–42.
4 Asaka T, Akiyama M, Domon T, *et al*. Type XVII collagen is a key player in tooth enamel formation. *Am J Pathol* 2009;174:91–100.
5 Kiritsi D, Kern JS, Schumann H, *et al*. Molecular mechanisms of phenotypic variability in junctional epidermolysis Bullosa. *J Med Genet* 2011;48:450–7.
6 Diaz LA, Ratrie H 3rd, Saunders WS, *et al*. Isolation of a human epidermal cDNA corresponding to the 180-kD autoantigen recognized by Bullous pemphigoid and herpes gestationis sera. Immunolocalization of this protein to the hemidesmosome. *J Clin Invest* 1990;86:1088–94.
7 Giudice GJ, Emery DJ, Diaz LA. Cloning and primary structural analysis of the bullous pemphigoid autoantigen Bp180. *J Invest Dermatol* 1992;99:243–50.
8 Franzke C-W, Bruckner-Tuderman L, Blobel CP. Shedding of collagen XVII/Bp180 in skin depends on both ADAM10 and ADAM9. *J Biol Chem* 2009;284:23386–96.
9 Has C, Bauer JW, Bodemer C, *et al*. Consensus reclassification of inherited epidermolysis Bullosa and other disorders with skin fragility. *Br J Dermatol* 2020;183:614–27.
10 Pfendner E, Uitto J, Fine JD. Epidermolysis bullosa carrier frequencies in the US population. *J Invest Dermatol* 2001;116:483–4.
11 Fine J-D. Epidemiology of inherited epidermolysis Bullosa based on incidence and prevalence estimates from the National Epidermolysis Bullosa Registry. *JAMA Dermatol* 2016;152:1231–8.
12 Petrof G, Papanikolaou M, Martinez AE, *et al*. The epidemiology of epidermolysis bullosa in England and Wales: data from the National Epidermolysis Bullosa database. *Br J Dermatol* 2022;186:843–8.
13 Nishie W. Collagen XVII processing and blistering skin diseases. *Acta Derm Venereol* 2020;100:5662.
14 Mellado F, Fuentes I, Palisson F, *et al*. Ophthalmologic approach in epidermolysis bullosa: a cross-sectional study with phenotype-genotype correlations. *Cornea* 2018;37:442–7.
15 Wright JT, Johnson LB, Fine JD. Development defects of enamel in humans with hereditary epidermolysis bullosa. *Arch Oral Biol* 1993;38:945–55.
16 Hintner H, Wolff K. Generalized atrophic benign epidermolysis bullosa. *Arch Dermatol* 1982;118:375–84.
17 Pasmooij AMG, Pas HH, Jansen GHL, *et al*. Localized and generalized forms of blistering in junctional epidermolysis bullosa due to COL17A1 mutations in the Netherlands. *Br J Dermatol* 2007;156:861–70.
18 Gedde-Dahl T. Phenotype-genotype correlations in epidermolysis bullosa. *Birth Defects Orig Artic Ser* 1971;7:107–17.
19 Kirkham J, Robinson C, Strafford SM, *et al*. The chemical composition of tooth enamel in junctional epidermolysis bullosa. *Arch Oral Biol* 2000;45:377–86.
20 McGrath JA, Gatalica B, Li K, *et al*. 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. *Am J Pathol* 1996;148:1787–96.
21 Floeth M, Fiedorowicz J, Schäcke H, *et al*. Novel homozygous and compound heterozygous COL17A1 mutations associated with junctional epidermolysis bullosa. *J Invest Dermatol* 1998;111:528–33.
22 Jonsson F, Byström B, Davidson AE, *et al*. Mutations in collagen, type XVII, alpha 1 (COL17A1) cause epithelial recurrent erosion dystrophy (ERED). *Hum Mutat* 2015;36:463–73.
23 Oliver VF, van Bysterveldt KA, Cadzow M, *et al*. A COL17A1 splice-altering mutation is prevalent in inherited recurrent corneal erosions. *Ophthalmology* 2016;123:709–22.
24 Smith CEL, Poulter JA, Antanaviciute A, *et al*. Amelogenesis imperfecta; genes, proteins, and pathways. *Front Physiol* 2017;8:435.
25 Prasad MK, Geoffroy V, Vicaire S, *et al*. A targeted next-generation sequencing assay for the molecular diagnosis of genetic disorders with Orodental involvement. *J Med Genet* 2016;53:98–110.
26 Bloch-Zupan A, Rey T, Jimenez-Armijo A, *et al*. Amelogenesis imperfecta: next-generation sequencing sheds light on witkop's classification. *Front Physiol* 2023;14:1130175.
27 Prasad MK, Laouina S, El Alloussi M, *et al*. Amelogenesis imperfecta:1 family, 2 phenotypes, and 2 mutated genes. *J Dent Res* 2016;95:1457–63.
28 Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 2009;25:1754–60.
29 DePristo MA, Banks E, Poplin R, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–8.
30 McLaren W, Gil L, Hunt SE, *et al*. The Ensembl variant effect Predictor. *Genome Biol* 2016;17:122.
31 Boyle EA, O'Roak BJ, Martin BK, *et al*. Mipgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* 2014;30:2670–2.
32 Hiatt JB, Pritchard CC, Salipante SJ, *et al*. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res* 2013;23:843–54.
33 Richards S, Aziz N, Bale S, *et al*. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–24.
34 Karczewski KJ, Francioli LC, Tiao G, *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–43.
35 Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, *et al*. Predicting splicing from primary sequence with deep learning. *Cell* 2019;176:535–48.
36 Witkop CJ. Amelogenesis imperfecta, dentinogenesis imperfecta and dentin dysplasia Revisited: problems in classification. *J Oral Pathol* 1988;17:547–53.
37 Yuen WY, Di Zenzo G, Jonkman MF, *et al*. New versatile monoclonal antibodies against type XVII collagen endodomain for diagnosis and Subtyping COL17A1-associated junctional epidermolysis bullosa. *J Eur Acad Dermatol Venereol* 2016;30:1426–7.
38 Altug-Atac AT, Erdem D. Prevalence and distribution of dental anomalies in orthodontic patients. *Am J Orthod Dentofacial Orthop* 2007;131:510–4.
39 Bäckman B, Holm AK. Amelogenesis imperfecta: prevalence and incidence in a northern Swedish County. *Community Dent Oral Epidemiol* 1986;14:43–7.
40 Poulter JA, El-Sayed W, Shore RC, *et al*. Whole-exome sequencing, without prior linkage, identifies a mutation in LAMB3 as a cause of dominant hypoplastic amelogenesis imperfecta. *Eur J Hum Genet* 2014;22:132–5.
41 Gostyńska KB, Yan Yuen W, Pasmooij AMG, *et al*. Carriers with functional null mutations in LAMA3 have localized enamel abnormalities due to haploinsufficiency. *Eur J Hum Genet* 2016;25:94–9.

42 Stenson PD, Mort M, Ball EV, *et al*. The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 2017;136:665–77.
43 Condrat I, He Y, Cosgarea R, *et al*. Junctional epidermolysis bullosa: allelic heterogeneity and mutation stratification for precision medicine. *Front Med (Lausanne)* 2018;5:363.
44 Pasmooij AMG, van Zalen S, Nijenhuis AM, *et al*. A very mild form of non-Herlitz junctional epidermolysis bullosa: BP180 rescue by outsplicing of mutated exon 30 coding for the COL15 domain. *Exp Dermatol* 2004;13:125–8.
45 Chang YF, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* 2007;76:51–74.
46 Väisänen L, Has C, Franzke C, *et al*. Molecular mechanisms of junctional epidermolysis bullosa: col 15 domain mutations decrease the thermal stability of collagen XVII. *J Invest Dermatol* 2005;125:1112–8.
47 Tasanen K, Floeth M, Schumann H, *et al*. Hemizygosity for a glycine substitution in collagen XVII: unfolding and degradation of the ectodomain. *J Invest Dermatol* 2000;115:207–12.
48 Almaani N, Liu L, Dopping-Hepenstal PJC, *et al*. Autosomal dominant junctional epidermolysis bullosa. *Br J Dermatol* 2009;160:1094–7.
49 Marty M, Chiaverini C, Milon C, *et al*. Perception of oral health-related quality of life in children with epidermolysis bullosa: a quantitative and qualitative study. *JDR Clin Trans Res* 2023;8:349–55.
50 Bloch-Zupan A, Rey T, Jimenez-Armijo A, *et al*. Amelogenesis imperfecta: next-generation sequencing sheds light on Witkop's classification. *Front Physiol* 2023;14:1130175.
51 Brookes SJ, Barron MJ, Smith CEL, *et al*. Amelogenesis imperfecta caused by N-terminal enamelin point mutations in mice and men is driven by endoplasmic reticulum stress. *Hum Mol Genet* 2017;26:1863–76.

Heterozygous *COL17A1* variants are a frequent cause of Amelogenesis Imperfecta

Ummey Hany[1], Christopher M. Watson[1,2], Lu Liu[1,3], Claire E.L. Smith[1], Asmaa Harfoush[3], James A. Poulter[1], Georgios Nikolopoulos[4], Richard Balmer[3], Catriona J. Brown[5], Anesha Patel[6], Jenny Simmonds[2], Ruth Charlton[2], María Gabriela Acosta de Camargo[7], Helen D Rodd[8], Hussain Jafri[9], Agne Antanaviciute[10], Michelle Moffat[11], Maisoon Al-Jawad[3], Chris F. Inglehearn[1,12], Alan J. Mighell[3,12]

Affiliations

1: Leeds Institute of Medical Research, University of Leeds, St James's University Hospital, Leeds, UK

2: North East and Yorkshire Genomic Laboratory Hub, Central Lab, St. James's University Hospital, Leeds, UK

3: School of Dentistry, Clarendon Way, University of Leeds, Leeds, UK

4: Institute for Fundamental Biomedical Research, B.S.R.C. 'Alexander Fleming', 16672 Vari, Attica, Greece

5: Birmingham Dental Hospital, Mill Pool Way, Edgbaston, Birmingham, UK

6: LCRN West Midlands Core Team, NIHR Clinical Research Network (CRN), Birmingham Research Park (West Wing), Vincent Drive, Edgbaston, Birmingham, UK

7: Department of Dentistry of the Child and Adolescent, Universidad de Carabobo, Venezuela

8: Academic Unit of Oral Health Dentistry and Society, School of Clinical Dentistry, University of Sheffield, Sheffield, S10 2TA, UK

9: Fatima Jinnah Medical University, Lahore, Pakistan

10: MRC Human Immunology Unit, University of Oxford

11: Paediatric Dentistry, The Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

12: Joint senior authors

Supplementary Files

1. **Table S1:** *COL17A1* variants associated with JEB (Database search) and their distribution in different protein domains.

The *COL17A1* gene encodes a transmembrane protein collagen XVII consisting of 1497 amino acid residues. The amino acid residues in different domains are as defined below [1, 2].

Domains	Exons	Amino Acid residues	Total Amino Acid
Endodomain	exon 2-exon 17	1-466	466
Transmembrane	exon 17	466-489	23
Ectodomain	exon 18-exon 56	490-1498	1008
			1497

By combining a literature search on the NCBI database (https://pubmed.ncbi.nlm.nih.gov/) with access to the HGMD professional database, we identified 232 *COL17A1* variants reported to cause JEB [3]. 172 of these were located in the ectodomain, 3 in the transmembrane domain and 57 in the endodomain.

Variant	Exonic Function	Domain
c.-11-2A>G	Splice	Endodomain
c.2T>A;p.0?	Missense	Endodomain
c.2T>C;p.0?	Missense	Endodomain
c.25C>T;p.(Arg9*)	Premature termination codon	Endodomain
c.51+1G>A	Splice	Endodomain
c.56delT;p.(Val19Alafs*5)	Frameshift small indels	Endodomain
c.82dupA;p.(Thr28Asnfs*15)	Frameshift small indels	Endodomain
c.158del; p.(Glu53fs)	Frameshift small indels	Endodomain
c.202delA;p.(Thr68Leufs*106)	Frameshift small indels	Endodomain
c.209-210insCA	Frameshift small indels	Endodomain
c.213-214dupAC;p.(Arg72Hisfs*103)	Frameshift small indels	Endodomain
c.214C>T:p.(Arg72*)	Premature termination codon	Endodomain
c.366dup;p.(Arg123fs)	Frameshift small indels	Endodomain
c.372_373insA;p.(Glu125Argfs*17)	Frameshift small indels	Endodomain
c.380-1G>A	Splice	Endodomain
c.412C>T;p.(Arg138*)	Premature termination codon	Endodomain
c.418_419delAG;p.(Ser140*)	Frameshift small indels	Endodomain
c.426delT p.(Ile142fs)	Frameshift small indels	Endodomain
c.427C>T;p.(Arg143*)	Premature termination codon	Endodomain
c.433C>T;p.(Arg145*)	Premature termination codon	Endodomain
c.460C>T;p.(Arg154*)	Premature termination codon	Endodomain
c.464-2A>G	Splice	Endodomain
c.505C>T;p.(Arg169*)	Premature termination codon	Endodomain
c.520_521del;p.(Ser174fs)	Frameshift small indels	Endodomain
c.558_567del;p.(Lys187Leufs*7)	Frameshift small indels	Endodomain
c.569-2A>G	Splice	Endodomain

c.621C>G;p.(Tyr207*)	Premature termination codon	Endodomain
c.655dupC;p.(His219Profs*118)	Frameshift small indels	Endodomain
c.675_688dup14;p.(Ser230Cysfs*14)	Frameshift small indels	Endodomain
c.718delA;p.(Thr240Profs*52)	Frameshift small indels	Endodomain
c.772G>A;p.(Gly258Arg)	Missense	Endodomain
c.772G>T;p.(Gly258*)	Premature termination codon	Endodomain
c.779delC;p.(Pro260Glnfs*32)	Frameshift small indels	Endodomain
c.783delC;p.(Asn261Lysfs*31)	Frameshift small indels	Endodomain
c.794C>G;p.(Ser265Cys)	Missense	Endodomain
c.823delC;p.(Ser275fs)	Frameshift small indels	Endodomain
c.884delC;p.(His296fs)	Frameshift small indels	Endodomain
c.980-1G>C	Splice	Endodomain
c.997C>T;p.(Glu333*)	Premature termination codon	Endodomain
c.1139C>T;p.(Ala380Val)	Missense	Endodomain
c.1141+2T>C	Splice	Endodomain
c.1141+5G>A	Splice	Endodomain
c.1179delA;p.(Ala394Leufs*9)	Frameshift small indels	Endodomain
c.1239_1240delGT;p.(Ser414Hisfs*87)	Frameshift small indels	Endodomain
c.1260delC;p.(Thr421Leufs*72)	Frameshift small indels	Endodomain
c.1267+1G>T	Splice	Endodomain
c.1267+2T>C	Splice	Endodomain
c.1268-2A>G	Splice	Endodomain
p.Ile18del389	Large deletion (exon2-15)	Endodomain
c.1268-267_1465+369del834;p.(Asp423Aladel66)	Large Deletion	Endodomain+Transmembrane
c.1284delA;p.(Ser429fs)	Frameshift small indels	Endodomain
c.1285delA;p.(Ser429fs)	Frameshift small indels	Endodomain
c.1336G>A;p.(Gly446Ser)	Missense	Endodomain
c.1365delC;p.(Trp455fs)	Frameshift small indels	Endodomain
c.1372+1G>T	Splice	Endodomain
c.1374C>A;p.(Cys458*)	Premature termination codon	Endodomain
c.1392G>A;p.(Trp464*)	Premature termination codon	Endodomain
c.1395G>A;p.(Trp465*)	Premature termination codon	Endodomain
c.1445T>C;p.(Leu482Pro)	Missense	Transmembrane
c.1465+2T>C	Splice	Transmembrane
c.1480_1482dupAAG;p.(Lys494dup)	In frame small indel	Ectodomain
c.1490_1491delCGinsT;p.(Ala497Valfs*23)	Frameshift small indels	Ectodomain
c.1507delG;p.(Glu503Argfs*17)	Frameshift small indels	Ectodomain
c.1601delA;p.(Asp534Alafs*19)	Frameshift small indels	Ectodomain
c.1601_1602insG;p.(Asp534Glufs*10)	Frameshift small indels	Ectodomain
c.1612delA;p.(Ile538Leufs*15)	Frameshift small indels	Ectodomain
c.1616G>A;p.(Gly539Glu)	Missense	Ectodomain
c.1696C>T;p.(Arg566*)	Premature termination codon	Ectodomain
c.1706delA;p.(Pro569fs)	Frameshift small indels	Ectodomain
c.1745-2A>C	Splice	Ectodomain
c.1745-2A>G	Splice	Ectodomain
c.1750C>T;p.(Arg584*)	Premature termination codon	Ectodomain

c.1772-2A>C	Splice	Ectodomain
c.1817G>A;p.(Gly606Asp)	Missense	Ectodomain
c.1826G>A;p.(Gly609Asp)	Missense	Ectodomain
c.1834G>A;p.(Gly612Arg)	Missense	Ectodomain
c.1834G>C;p.(Gly612Arg)	Missense	Ectodomain
c.1852G>A;p.(Gly618Ser)	Missense	Ectodomain
c.1861G>A;p.(Gly621Ser)	Missense	Ectodomain
c.1877-2A>C	Splice	Ectodomain
c.1880G>T;p.(Gly627Val)	Missense	Ectodomain
c.1880delG;p.(Gly627Alafs*56)	Frameshift small indels	Ectodomain
c.1898G>A;p.(Gly633Asp)	Missense	Ectodomain
c.1992_1995delGGGT;p.(Gly665Profs*17)	Frameshift small indels	Ectodomain
c.2002+2T>G	Splice	Ectodomain
c.2003-1G>C	Splice	Ectodomain
c.2062delC;p.(Arg688Glufs*4)	Frameshift small indels	Ectodomain
c.2062C>T;p.(Arg688*)	Premature termination codon	Ectodomain
c.2227+153_2336-318del	Large Deletion	Ectodomain
c.2228-101_2263+70delins15	Frameshift small indels	Ectodomain
c.2237delG;p.(Gly746Alafs*53)	Frameshift small indels	Ectodomain
c.2240delC;p.(Pro747Glnfs*52)	Frameshift small indels	Ectodomain
c.2251C>T;p.(Gln751*)	Premature termination codon	Ectodomain
c.2282_2283delGG;p.(Gly761Aspfs*40)	Frameshift small indels	Ectodomain
c.2336-1G>T	Splice	Ectodomain
c.2336-2A>G	Splice	Ectodomain
c.2342delG;p.(Thr781fs)	Frameshift small indels	Ectodomain
c.2350C>T;p.(Gln784*)	Premature termination codon	Ectodomain
c.2363-2A>G	Splice	Ectodomain
c.2363dup;p.(Leu789Thrfs*13)	Frameshift small indels	Ectodomain
c.2383C>T;p.(Arg795*42)	Premature termination codon	Ectodomain
c.2407G>T;p.(Gly803*)	Premature termination codon	Ectodomain
c.2434+1G>A	Splice	Ectodomain
c.2441-2A>G	Splice	Ectodomain
c.2441-1G>T	Splice	Ectodomain
c.2468-2A>G	Splice	Ectodomain
c.2468C>A;p.(Pro823Gln)	Missense	Ectodomain
c.2488G>A;p.(Gly830Arg)	Missense	Ectodomain
c.2496dupT;p.(Gly833Argfs*22)	Frameshift small indels	Ectodomain
c.2518del10	Frameshift small indels	Ectodomain
c.2520dupT;p.(Ala841Cysfs*14)	Frameshift small indels	Ectodomain
c.2544delA;p.(His849Ilefs*217)	Frameshift small indels	Ectodomain
c.2551+1G>A	Splice	Ectodomain
c.2561_2565delATTTA;p.(Asn854Thrfs*109)	Frameshift small indels	Ectodomain
c.2563_2564delTT;p.(Leu855Thrfs*109)	Frameshift small indels	Ectodomain
c.2564T>G;p.(Leu855*)	Premature termination codon	Ectodomain
c.2564T>A;p.(Leu855*)	Premature termination codon	Ectodomain
c.2566C>T;p.(Gln856*)	Premature termination codon	Ectodomain

c.2585_2586delCC;p.(Pro862Argfs*102)	Frameshift small indels	Ectodomain
c.2635C>T;p.(Arg879*)	Premature termination codon	Ectodomain
c.2666delTT	Frameshift small indels	Ectodomain
c.2690insT;p.(Ser898fs)	Frameshift small indels	Ectodomain
c.2706dup;p.(Phe903Leufs*62)	Frameshift small indels	Ectodomain
c.2716G>A;p.(Gly906Ser)	Missense	Ectodomain
c.2723dupC;p.(Gly909Argfs*56)	Frameshift small indels	Ectodomain
c.2725G>C;p.(Gly909Arg)	Missense	Ectodomain
c.2776delA;p.(Arg926Glufs*140)	Frameshift small indels	Ectodomain
c.2840_2844delTCAAC;p.(Leu947Profs*16)	Frameshift small indels	Ectodomain
c.2861delG;p.(Gly954Alafs*112)	Frameshift small indels	Ectodomain
c.2875delC;p.(Gln959Argfs*107)	Frameshift small indels	Ectodomain
c.2881delA	Frameshift small indels	Ectodomain
c.2897-9G>A	Splice	Ectodomain
c.2897-2A>C	Splice	Ectodomain
c.2944del;p.(Glu982Lysfs84*)	Frameshift small indels	Ectodomain
c.2948-1G>C	Splice	Ectodomain
c.2965delG;p.(Met989fs)	Frameshift small indels	Ectodomain
c.2971G>A;p.(Val991Met)	Missense	Ectodomain
c.2972delT;p.(Val991Glyfs*75)	Frameshift small indels	Ectodomain
c.2975C>A;p.(Ser992*)	Premature termination codon	Ectodomain
c.2993dupC;p.(Gly999Trpfs*22)	Frameshift small indels	Ectodomain
c.3000-3008del;p.(Pro1003_Gly1005del)	In frame small indel	Ectodomain
c.3002-2A>C	Splice	Ectodomain
c.3046C>T;p.(Gln1016*)	Premature termination codon	Ectodomain
c.3053-1G>C	Splice	Ectodomain
c.3067C>T;p.(Gln1023*)	Premature termination codon	Ectodomain
c.3071-6C>A	Splice	Ectodomain
c.3131delC;p.(Pro1044Glnfs*22)	Frameshift small indels	Ectodomain
c.3164delT;p.(Phe1055Serfs*11)	Frameshift small indels	Ectodomain
c.3171_3173delCTC;p.(Tyr1057*)	Frameshift small indels	Ectodomain
c.3175delG;p.(Glu1059Serfs*7)	Frameshift small indels	Ectodomain
c.3193_3208del16;p.(Val1065Leufs*35)	Frameshift small indels	Ectodomain
c.3198C>T;p.(Ser1066Ser	Splice	Ectodomain
c.3205C>T;p.(Arg1069Trp)	Missense	Ectodomain
c.3236delC;p.(Ser1079Cysfs*26)	Frameshift small indels	Ectodomain
c.3269dupT;p.(Leu1091Alafs*5)	Frameshift small indels	Ectodomain
c.3275A>C;(Gln1092Pro)	Missense	Ectodomain
c.3288_3295del8;p.(Arg1097Profs*33)	Frameshift small indels	Ectodomain
c.3292C>T;p.(Gln1098*)	Premature termination codon	Ectodomain
c.3301C>T;p.(Arg1101Cys)	Missense	Ectodomain
c.3327delT;p.(Pro1110Argfs*21)	Frameshift small indels	Ectodomain
c.3408delC;p.(Tyr1137Thrfs*114)	Frameshift small indels	Ectodomain
c.3481dupT;p.(Tyr1161fs*2)	Frameshift small indels	Ectodomain
c.3482_3483del;p.(Tyr1161*)	Frameshift small indels	Ectodomain
c.3487G>T;p.(Glu1163*)	Premature termination codon	Ectodomain

c.3496_3497delTC;p.(Ser1166Leufs*6)	Frameshift small indels	Ectodomain
c.3505C>T;p.(Arg1169*)	Premature termination codon	Ectodomain
c.3509-1G>C	Splice	Ectodomain
c.3513delC;p.(Glu1172fs)	Frameshift small indels	Ectodomain
c.3514ins25	Frameshift small indels	Ectodomain
c.3539dupC;p.(Pro1180Profs*62)	Frameshift small indels	Ectodomain
c.3548delC;p.(Pro1183Argfs*68)	Frameshift small indels	Ectodomain
c.3569dupG;p.(Asn1191Glnfs*51)	Frameshift small indels	Ectodomain
c.3569_3570insT;p.(Asn1191Glnfs*51)	Frameshift small indels	Ectodomain
c.3579G>A;p.(Trp1193*)	Premature termination codon	Ectodomain
c.3600-3601delCT	Frameshift small indels	Ectodomain
c.3613_3616delTTAC;p.(Leu1205Ilefs*45)	Frameshift small indels	Ectodomain
c.3615_3619dupACATA	Frameshift small indels	Ectodomain
c.3619+2T>C	Splice	Ectodomain
c.3673_3674insT;p.(Pro1225fs)	Frameshift small indels	Ectodomain
c.3676C>T;p.(Arg1226*)	Premature termination codon	Ectodomain
c.3686C>T;p.(Pro1229Leu)	Missense	Ectodomain
c.3689dup;p.(V1231Cfs*11)	Frameshift small indels	Ectodomain
c.3730G>A;p.(Asp1244Asn)	Missense	Ectodomain
c.3740G>A;p.(Arg1247Gln)	Missense	Ectodomain
c.3766+1G>A	Splice	Ectodomain
c.3766+1G>C	Splice	Ectodomain
c.3782G>C;p.(Ser1261Thr)	Missense	Ectodomain
c.3786_3789delCATT;p.(Phe1262Leufs*49)	Frameshift small indels	Ectodomain
c.3801insC;p.(Gly1268Argfs*25)	Frameshift small indels	Ectodomain
c.3806delC;p.(Pro1269Leufs*43)	Frameshift small indels	Ectodomain
c.3821_3829delGACCCCCTGinsC;p.(Gly1274Alafs*16)	Frameshift small indels	Ectodomain
c.3827dupC;p.(Gly1277Trpfs*16)	Frameshift small indels	Ectodomain
c.3865dupA;p.(Ser1289Lysfs*4)	Frameshift small indels	Ectodomain
c.3871+1G>A	Splice	Ectodomain
c.3871+1G>C	Splice	Ectodomain
c.3897_3900delATCT;p.(Val1301Glyfs*10)	Frameshift small indels	Ectodomain
c.3899_3900delCT;p.(Ser1300Cysfs*29)	Frameshift small indels	Ectodomain
c.3908G>A;p.(Arg1303Gln)	Missense	Ectodomain
c.3922delA;p.(Ser1308Alafs*4)	Frameshift small indels	Ectodomain
c.4003delTC	Frameshift small indels	Ectodomain
c.4011T>A;p.(Tyr1337*)	Premature termination codon	Ectodomain
c.4041T>G;p.(Tyr1347*)	Premature termination codon	Ectodomain
c.4045dupG;p.(Ala1349Glyfs*18)	Frameshift small indels	Ectodomain
c.4080insGG;p.(Gly1361fs)	Frameshift small indels	Ectodomain
c.4088delT;p.(Leu1363fs)	Frameshift small indels	Ectodomain
c.4100_4101delTT;p.(Phe1367Cysfs*8)	Frameshift small indels	Ectodomain
c.4144del4;p.(Glu1382fs)	Frameshift small indels	Ectodomain
c.4145_4148delAGAG;p.(Glu1382Alafs*40)	Frameshift small indels	Ectodomain
c.4149_4150insG;p.(Met1384fs)	Frameshift small indels	Ectodomain
c.4153C>T;p.(Gln1385*)	Premature termination codon	Ectodomain

c.4156+1G>C	Splice	Ectodomain
c.4156+1G>A	Splice	Ectodomain
c.4159C>T;p.(Gln1387*)	Premature termination codon	Ectodomain
c.4207C>T;p.(Gln1403*)	Premature termination codon	Ectodomain
c.4230delC;p.(Gly1411Alafs*12)	Frameshift small indels	Ectodomain
c.4231G>A;p.(Gly1411Ser)	Missense	Ectodomain
c.4237_4238insC;p.(Ser1413Thrfs*42)	Frameshift small indels	Ectodomain
c.4261+1G>C	Splice	Ectodomain
c.4265_4266insTT;p.(Thr1423*)	Frameshift small indels	Ectodomain
c.4295-1G>C	Splice	Ectodomain
c.4307_4310dupTTCA;p.(Gln1437Hisfs*19)	Frameshift small indels	Ectodomain
c.4319dupC;p.(Gly1441Trpfs*14)	Frameshift small indels	Ectodomain
c.4320delT;p.(Gln1442Lysfs*70)	Frameshift small indels	Ectodomain
c.4321delT;p.(Gln1442fs)	Frameshift small indels	Ectodomain
c.4324C>T;p.(Gln1442*)	Premature termination codon	Ectodomain
c.4335delC;p.(Met1446fs)	Frameshift small indels	Ectodomain
c.4410-4413dupCATT	Frameshift small indels	Ectodomain
c.4424-5insC	Splice	Ectodomain
c.4425_4426insC;p.(Lys1476fs)	Frameshift small indels	Ectodomain
c.4425delT	Frameshift small indels	Ectodomain
c.4460G>A;p.(Arg1487Gln)	Missense	Ectodomain
c.4463-1G>A	Splice	Ectodomain

Table S2: *COL17A1* variants associated with AI (This study and from database search).

Variant	Exonic Function	Domain	References
c.340;p.(Ser114Valfs*60)	Frameshift small indels	Endodomain	Leeds AI group
c.460C>T;p.(Arg154*)	Premature termination codon	Endodomain	Leeds AI group
c.541_550del;p.(Asn181Profs*13)	Frameshift small indels	Endodomain	Leeds AI group
c.1141+1G>A	Splice	Endodomain	Prasad, M. K., et al. (2016)
c.1646G>A;p.(Trp549*)	Premature termination codon	Ectodomain	Prasad, M. K., et al. (2016)
c.1745-2A>C	Splice	Ectodomain	Bloch-Zupan, A., et al. (2023)
c.1861G>A; p.(Gly621Ser)	Missense	Ectodomain	Leeds AI group
c.1873C>T;p.(Arg625*)	Premature termination codon	Ectodomain	Prasad, M. K., et al. (2016)
c.2011G>A;p.(Gly671Ser)	Missense	Ectodomain	Leeds AI group
c.2030G>A;p.(Gly677Asp)	Missense	Ectodomain	Leeds AI group
c.2407G>T;p.(Gly803*)	Premature termination codon	Ectodomain	Prasad, M. K., et al. (2016)
c.2435-1G>A	Splice	Ectodomain	Leeds AI group
c.2912del;p.(Pro971Glnfs*95)	Frameshift small indels	Ectodomain	Leeds AI group
c.2947+2T>C	Splice	Ectodomain	Leeds AI group
c.3277+1G>A	Splice	Ectodomain	Leeds AI group
c.3297C>A;p.(Tyr1099*)	Premature termination codon	Ectodomain	Leeds AI group
c.3327del;p.(Pro1110Argfs*21)	Frameshift small indels	Ectodomain	Bloch-Zupan, A., et al. (2023)
c.3397C>T;p.(Arg1133Cys)	Missense	Ectodomain	Leeds AI group
c.3456del;p.(Pro1154Leufs*97)	Frameshift small indels	Ectodomain	Leeds AI group
c.3462_3463del;p.(Gly1155fs*7)	Frameshift small indels	Ectodomain	Leeds AI group
c.3595G>C;p.(Glu1199Gln)	Missense	Ectodomain	Leeds AI group
c.3605C>T;p.(Ser1202Leu)	Missense	Ectodomain	Leeds AI group
c.4147_4148del;p.(Ser1383Hisfs*71)	Frameshift small indels	Ectodomain	Leeds AI group

Table S3: *COL17A1* variants associated with ERED (From database search).

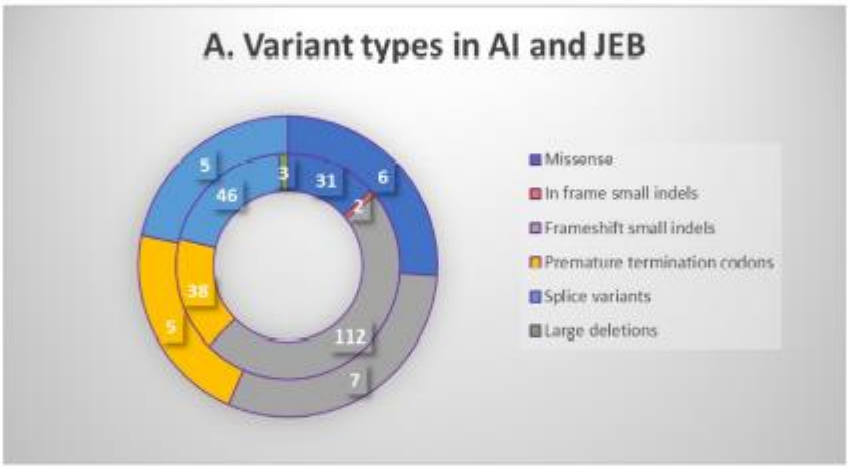
Variant	Exonic Function	Domain
c.2816C>T;(Thr939Ile)	Missense	Ectodomain
c.3156C>T	Splice	Ectodomain
c.3554C>T;(Pro1185Leu)	Missense	Ectodomain

Table S4: Genes included in the smMIP reagent. Reference genome GRCh37/hg19.

Gene Name	Gene Symbol	OMIM	Genomic Coordinates	Cytoband
LAMININ, BETA-3	*LAMB3*	150310	chr1:209,788,218-209,825,820	1q32.2
INTEGRIN, BETA-6	*ITGB6*	147558	chr2:160,958,233-161,056,589	2q24.2
AMELOTIN	*AMTN*	610912	chr4:71,384,298-71,398,459	4q13.3
AMELOBLASTIN	*AMBN*	601259	chr4:71,457,975-71,473,004	4q13.3
ENAMELIN	*ENAM*	606585	chr4:71,494,461-71,512,536	4q13.3
ODONTOGENESIS-ASSOCIATED PHOSPHOPROTEIN	*ODAPH*	614829	chr4:76,481,258-76,491,103	4q21.1
FAMILY WITH SEQUENCE SIMILARITY 83	*FAM83H*	611927	chr8:144,806,103-144,815,914	8q24.3
COLLAGEN, TYPE XVII, ALPHA-1	*COL17A1*	113811	chr10:105,791,046-105,845,638	10q25.1
RECEPTOR EXPRESSED IN LYMPHOID TISSUES	*RELT*	611211	chr11:73,087,405-73,108,519	11q13.4
MATRIX METALLOPROTEINASE 20	*MMP20*	604629	chr11:102,447,566-102,496,063	11q22.2
G PROTEIN-COUPLED RECEPTOR 68	*GPR68*	601404	chr14:91,698,876-91,710,852	14q32.11
SOLUTE CARRIER FAMILY 24 (SODIUM/POTASSIUM/CALCIUM EXCHANGER), MEMBER 4	*SLC24A4*	609840	chr14:92,790,152-92,967,825	14q32.12
WD REPEAT-CONTAINING PROTEIN 72	*WDR72*	613214	chr15:53,805,938-54,051,859	15q21.3
TRANSCRIPTION FACTOR Sp6	*SP6*	608613	chr17:45,922,280-45,928,516	17q21.32
DISTAL-LESS HOMEOBOX 3	*DLX3*	600525	chr17:48,067,369-48,072,588	17q21.33
FAMILY WITH SEQUENCE SIMILARITY 20, MEMBER A	*FAM20A*	611062	chr17:66,531,257-66,597,095	17q24.2
ACID PHOSPHATASE 4	*ACP4*	606362	chr19:51,293,672-51,298,481	19q13.33
KALLIKREIN-RELATED PEPTIDASE 4	*KLK4*	603767	chr19:51,409,608-51,413,994	19q13.41
AMELOGENIN	*AMELX*	300391	chrX:11,311,533-11,318,881	Xp22.2

Figure S1: A comparison of *COL17A1* variants detected in AI and JEB.

A. Collagen XVII variant types reported in AI (outer ring) and JEB (inner ring) patients: Most variants detected in both AI and JEB are those that lead to frameshift and PTC. Fewer variants are reported in AI (n=23) than in JEB (n=232). The differences observed were not significant.



B. Variant distribution in different domains of collagen XVII reported in AI (outer ring) and JEB (inner ring) patients: A total of 23 variants are reported to be associated with AI, 19 of them occurring in the ectodomain and 4 of them in endodomain. A total of 232 variants were reported to be associated with JEB, 172 of them were occurring in ectodomain, 57 in endodomain and 3 of them in transmembrane domain.

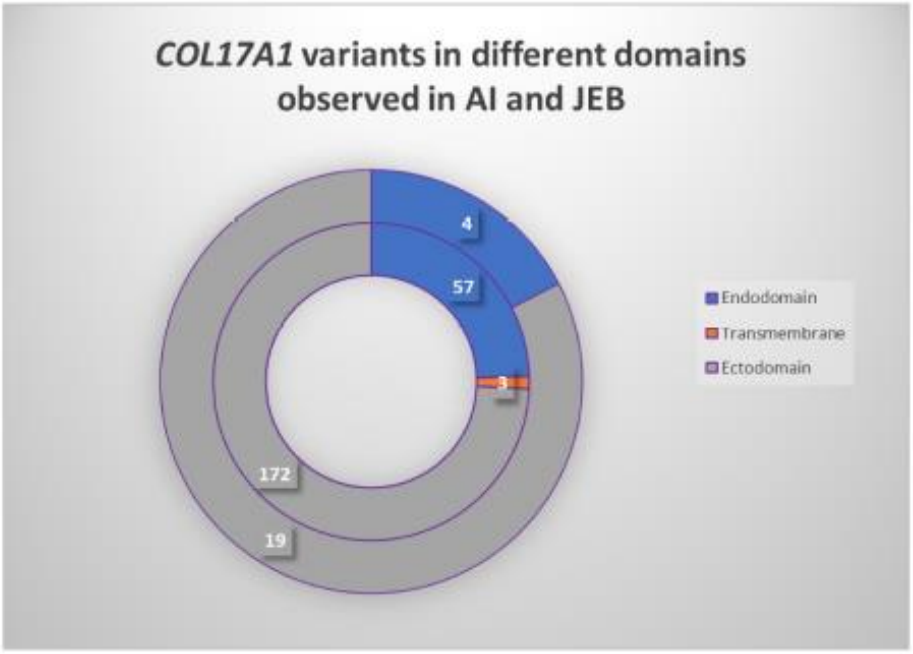


Figure S2: Family pedigree and IGV trace of the variant c.340; p.(Ser114Valfs*60) detected in the family 19, generated by using human reference genome hg19.

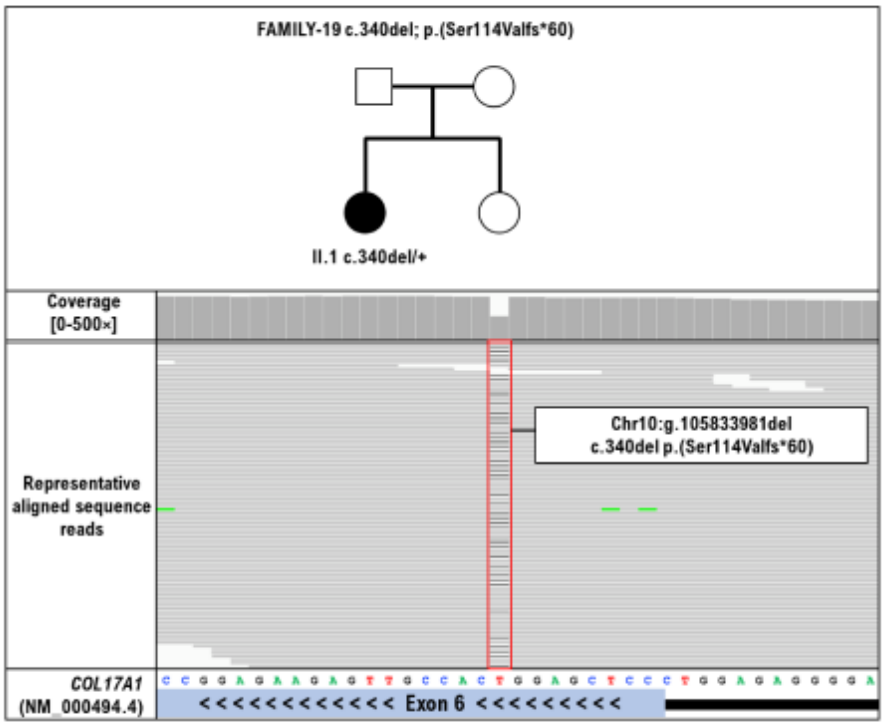
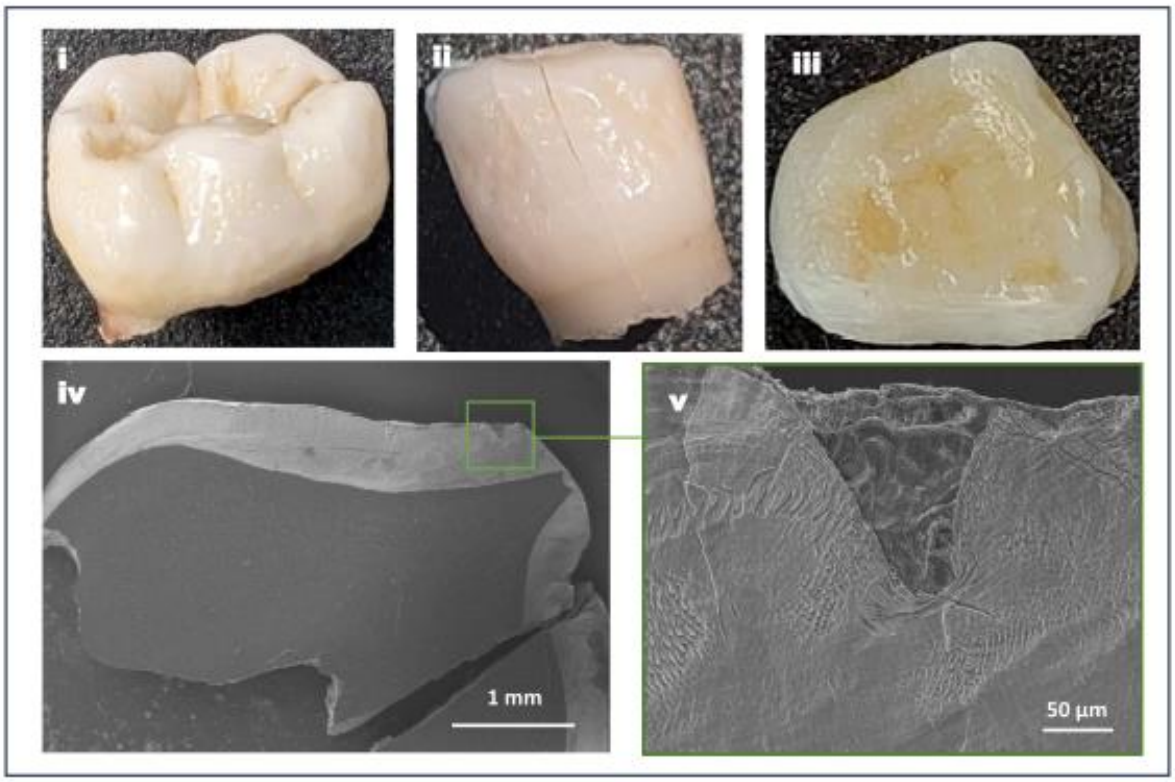


Figure S3: Surface pitting observed in primary teeth from different families.
Whole tooth images from the proband of (i) F14, (ii and iii) F9. SEM cross-section images showing surface pit (iv) F9 (inset), (v) Higher resolution of the pit.



References

1. Giudice, G.J., D.J. Emery, and L.A. Diaz, *Cloning and primary structural analysis of the bullous pemphigoid autoantigen BP180.* Journal of Investigative Dermatology, 1992. **99**(3): p. 243-250.
2. Areida, S.K., et al., *Properties of the Collagen Type XVII Ectodomain: EVIDENCE FOR N- TO C-TERMINAL TRIPLE HELIX FOLDING *.* Journal of Biological Chemistry, 2001. **276**(2): p. 1594-1601.
3. Stenson, P.D., et al., *The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies.* Hum Genet, 2017. **136**(6): p. 665-677.

Supplementary Figure S4: Representative clinical images and dental radiographs for individuals with heterozygous *COL17A1* variants. Some teeth have been restored.

F2: c.460C>T:p.(Arg154*). Mixed primary and secondary dentitions. The enamel is characterised by hypomaturation with patchy variations in colour (white to brown) accompanied by greater opacity than normally expected. The enamel surface is irregular without obvious pitting. The underlying *COL17A1* variant is a known cause of JEB when homozygous.

F3: c.541_550del:p.(Asn181Profs*13). Secondary dentition. The enamel is characterised by hypomaturation with multiple surface pits that in places have contributed to linear defects, particularly in the middle third of the upper incisor teeth where there is regional hypoplasia. By comparison to the middle third, the incisal third of the labial crown surface has minimal, if any hypoplasia.

F4: c.1861G>A:p.(Gly621Ser). Mixed dentition. The enamel is characterised by mild hypomaturation with widespread surface irregularities including pits evident on the surface of the secondary teeth. There is mild hypoplasia of the mid-third of the labial surface of the upper central incisor teeth and minor morphological changes to the crown shapes of the secondary premolar teeth. The underlying *COL17A1* variant is a known cause of JEB when homozygous.

F5: c.2011G>A:p.(Gly671Ser). Secondary dentition. The enamel is characterised by hypomaturation and multiple surface irregularities including focal pits. There are minor morphological changes to the crown shapes of the secondary premolar teeth. Radiographs identify a clear distinction in radiodensity between enamel and dentine, consistent with a hypomaturation phenotype. There are also irregularities in the enamel morphology consistent with the clinical images.

F6: c.2030G>A:p.(Gly677Asp). Mixed dentition. The enamel is characterised by hypomaturation with variable colour changes and multiple surface irregularities. The secondary dentition first molar teeth appear more obviously affected than other teeth. There is mild hypoplasia of the mid-third of the labial surface of the upper central incisor teeth and minor morphological changes to the crown shapes of the secondary premolar teeth.

F7: c.2435-1G>A:p.? Mixed dentition. The enamel is characterised by hypomaturation with multiple surface pits and regional hypoplasia in the middle third of the upper incisor teeth. By comparison to the middle third, the incisal third of the labial crown surface has minimal, if any hypoplasia. There are minor morphological changes to the crown shapes of the secondary dentition premolar teeth that have a consistent pattern with those observed in the upper central incisor teeth.

F9: c.2947+2T>C:p.? Mixed dentition. The enamel is characterised by hypomaturation with patchy variations in colour (white to brown) accompanied by greater opacity than normally expected. widespread mild surface irregularities evident on the surface of the secondary teeth.

F10: c.3277+1G>A:p.? Primary dentition. The enamel is predominantly characterised by hypomaturation with only minimal surface irregularities.

F11: c.3297C>A:p.(Tyr1099*). Secondary dentition. The enamel is characterised by hypomaturation with widespread mild surface irregularities and focal pits that have a variable distribution over the crown surfaces.

F12: c.3397C>T:p.(Arg1133Cys). Secondary dentition. Limited information is available. The attrition to many of the permanent teeth illustrates an enamel thickness within expected limits.

F14: c.3456del:p.(Pro1154Leufs*97). Secondary dentition. The enamel is characterised by hypomaturation with patchy variations in colour (white to yellow/brown) and multiple surface irregularities.

Supplementary Figure S4 *cont*: Representative clinical images and dental radiographs for individuals with heterozygous *COL17A1* variants. Some teeth have been restored.

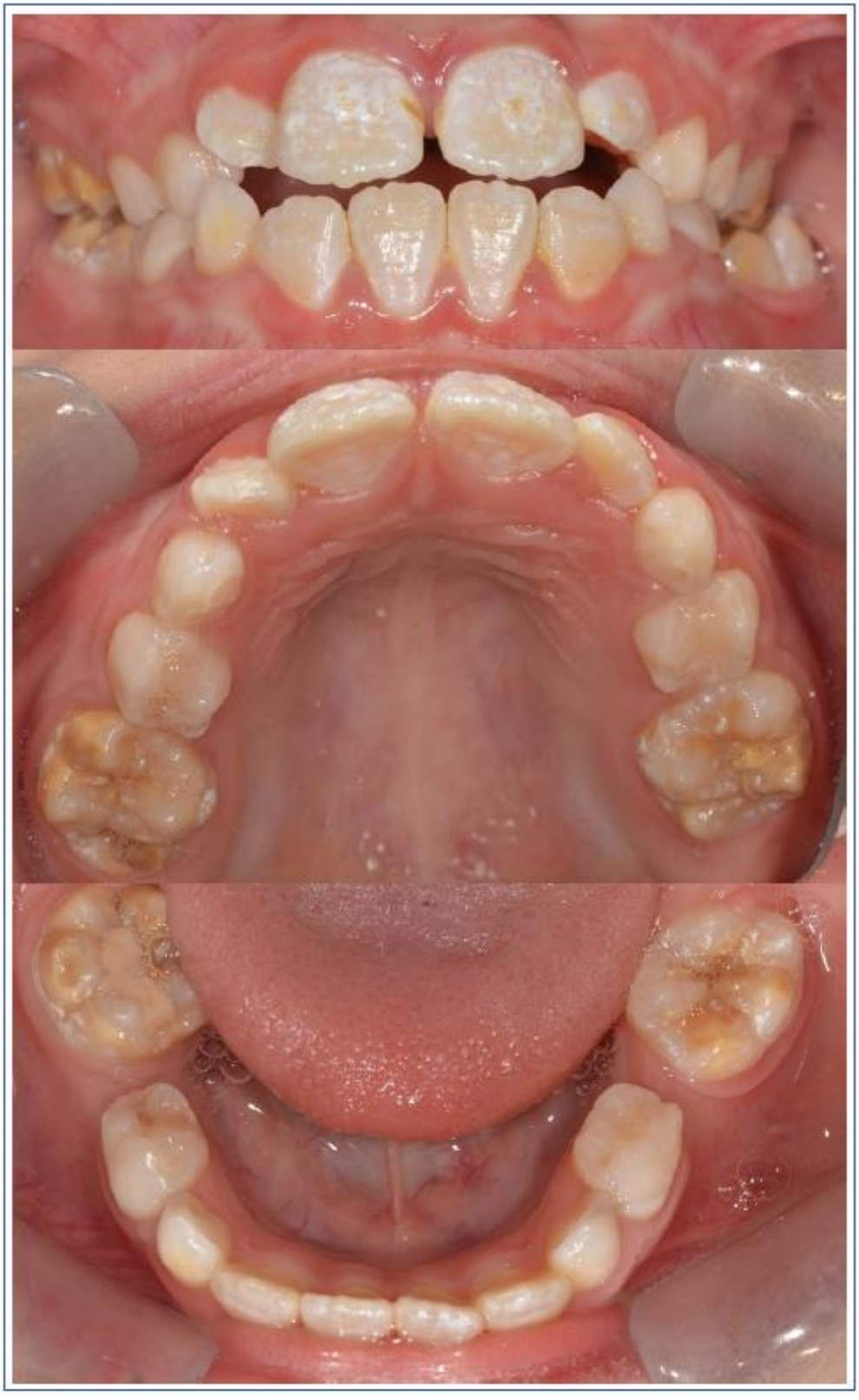
F15: c.3462_3463del:p.(Gly1155Leufs*7). Secondary dentition. The radiograph illustrates enamel thickness being within expected limits and a clear distinction in radiodensity between enamel and dentine, consistent with a hypomaturation phenotype.

F18: c.3605C>T:p.(Ser1202Leu). Secondary dentition. The enamel is characterised by hypomaturation and multiple surface irregularities without obvious focal pits. There are minor morphological changes to the crown shapes of the secondary premolar teeth. A panoramic radiograph of the mixed dentition identifies a clear distinction in radiodensity between enamel and dentine, consistent with a hypomaturation phenotype.

No clinical images or dental radiographs were available for the following families. The clinical phenotype in all families was described as pitted hypoplastic AI.

F8: c.2912del:p.(Pro971Glnfs*95) - Pathogenic
F13: c.3456del:p.(Pro1154Leufs*97) - Pathogenic
F16: c.3595G>C:p.(Glu1199Gln) - Likely Pathogenic
F17: c.3595G>C:p.(Glu1199Gln) - Likely Pathogenic
F19: c.4147_4148del:p.(Ser1383Hisfs*71) - Pathogenic

F2 - c.460C>T:p.(Arg154*) - Pathogenic (Condrat et al 2018)



F2 continued - c.460C>T:p.(Arg154*) - Pathogenic (Condrat et al 2018)

F3 - c.541_550del:p.(Asn181Profs*13) - Pathogenic

F4 – c.1861G>A:p.(Gly621Ser) - Likely Pathogenic

F5 - c.2011G>A:p.(Gly671Ser) - Likely Pathogenic

F5 - c.2011G>A:p.(Gly671Ser) - Likely Pathogenic

F6 - c.2030G>A:p.(Gly677Asp) - Likely Pathogenic

F7 - c.2435-1G>A:p.? - Likely pathogenic

F9 - c.2947+2T>C:p.? - Likely Pathogenic

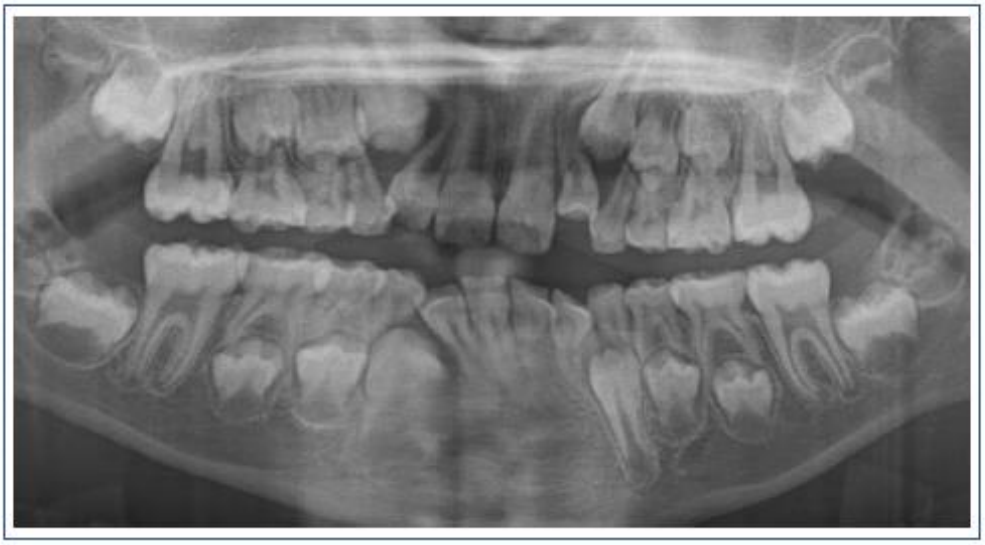
F10 - c.3277+1G>A:p.? - Pathogenic

F11 - c.3297C>A:p.(Tyr1099*) - Pathogenic

F12 - c.3397C>T:p.(Arg1133Cys) - VUS

F14 - c.3456del:p.(Pro1154Leufs*97) - Pathogenic

F15 - c.3462_3463del:p.(Gly1155Leufs*7) - Pathogenic



F18 - c.3605C>T:p.(Ser1202Leu) - VUS

CHAPTER 5 Discussion and Conclusion

5.1 Summary of the project, problems encountered and lessons learned

This study began with the successful development and testing of a genetic screening reagent using smMIPs technology that enabled selective screening of variants in nineteen genes known to be involved in NS AI. The smMIPs method provides a cost-effective approach for multiplex-targeted genomic capture, especially in situations where a group of genes (exons and immediate splice sites) or genomic regions are to be tested in a large cohort. It may not be cost-effective to develop the method targeting a large number of genes when screening a small number of patients, as there is a substantial start-up cost involved in synthesizing smMIPs probes. Initially, eight genes were selected for targeting when developing the smMIPs reagent because there was only limited funding available to cover this start-up cost. These eight genes were chosen because they accounted for a large proportion of the mutation spectrum linked to AI in previous AI cohort screening and were comprised of a relatively low number of exons. A first screen with this limited gene set was successfully accomplished.

The positive impact of this early success with the smMIPs method gave confidence to add more genes to the smMIPs reagent when funding was secured from the Rosetrees Trust. The negative impact of this approach was that adding genes in batches made probe rebalancing more complex. Sequence coverage in the initial smMIP experiment can vary significantly. Adjusting for intraprobe read-depth variability (probe rebalancing) is one part of the method development when usings smMIPs. It is usually done by calculating the average read-depth value per probe from the first trial, then adjusting the volume of each underperforming/overperforming probe, either increasing or decreasing their concentration in the second trial. Probes which give little

or no sequence are replaced. In this way, it required four sequencing experiments to optimise the smMIPs reagent targeting nineteen NS AI genes. Adding genes in multiple batches introduced biases in probe coverage in the subsequent experiments, which complicated the overall probe rebalancing. Here are some recommendations that could have improved the smMIPs method development and could be implemented in future research to achieve better performance.

5.1.1 Improvement in wet lab methods

The quality of the genomic DNA had a significant effect on the method development. The DNA in the Leeds AI cohort had been extracted by different users using different methods, over the last two decades. It was observed that, in the same sequencing run, the same genomic regions were sequenced with differing efficiency in different DNA samples. This could be the result of DNA degradation occurring from laboratory handling (*e.g.* mechanical shearing from the use of a pipette), or repeated rounds of defrosting and refreezing. Using consistently high-quality DNA during methods development may have reduced experimental biases introduced by DNA degradation or contamination, making it possible to obtain more consistent coverage in a larger number of samples in a single sequencing run.

5.1.2 Selecting appropriate data analysis algorithms

The choice of bioinformatics tools and algorithms for read alignment and variant calling also affected initial results. At first, a bioinformatics pipeline was developed using BWA-MIPS (https://github.com/brentp/bwa-mips) to process the smMIPs data. However, the BWA-MIPS algorithm had biases towards specific types of sequences and failed to

retain sequence reads from all nineteen genes due to inappropriate removal of the probe arms. Later, data were processed using MIPVAR (https://sourceforge.net/projects/mipvar/), which retained sequencing reads from all the genomic regions targeted by the reagent. Systematic errors introduced by different algorithms need evaluation at method development before being implemented routinely.

5.1.3 Possible explanations for cases screening negative in smMIPs

Finally, a targeted smMIPs reagent was successfully developed for AI, which achieved a mean capture efficiency of 97% at greater than 20 reads across all the probes for all the nineteen genes. The method was proven to have high throughput diagnostic capability, established by screening 181 genomic DNA samples from previously unsolved AI patients, solving 36% of these cases. After smMIPs screening, 64% (115) of cases remained unsolved. Factors that may have contributed to the failure to find variants in the unsolved cases are detailed below.

Since the smMIPs reagent had a mean capture efficiency of 97%, 3% of target regions were not sufficiently well captured by any probes. Genes like *FAM83H* and *ACP4,* that have high GC content, as well as GC-rich regions in other genes, were underrepresented by the smMIPs screening. Regions with high GC content in the genome are more stable as G and C base pairs are bound by three hydrogen bonds rather than the two in an A and T pairing, which makes GC-rich DNA harder to denature during the sequencing process, leading to uneven coverage. Any variants that may exist in these areas could have been missed.

The current variant filtering criteria excuded synonymous variants. Synonymous variants are changes in the DNA sequence of a gene that does not change the amino acid sequence of the protein. In the past these have typically been considered to have no functional consequences (Kimura, 1977). However, more recent research has revealed that synonymous variants can affect the splicing of mRNA, leading to the inclusion, exclusion or mis-splicing of specific exons in the mature mRNA. Synonymous changes can also affect mRNA secondary structure, potentially influencing the stability of the mRNA and its translation (Vihinen, 2022). Several studies reported synonymous variants were associated with diseases (Kim et al., 2020; Tang et al., 2020).

In addition, the complete genetic spectrum of AI is not known, with further, probably less commonly mutated genes associated with both NS AI and syndromic AI likely to be implicated over time, and these will not be covered by the smMIPs analysis. Also, smMIPs did not cover the non-coding regions (deep intronic, UTRs or promoter variants) of the genes, so any variants in these regions would not be detected. Furthermore, short-read smMIPs data is less sensitive at identifying large structural variants and also has limitations in detecting highly repetitive regions in the genome. Therefore, the likelihood of solving cases may increased by using more comprehensive sequencing and data analysis approaches, as described.

5.2 Main findings

For the patients that tested negative after smMIPs screening, WES analysis was employed in an effort to find novel gene variants that could be responsible for causing

AI. A summary of key findings from the combined smMIPs and WES sequencing efforts are presented below.

5.2.1 COL17A1 variants are a frequent cause of nonsyndromic AI

Following smMIPs and WES, it was noted that *COL17A1* variants were the most common cause of AI in the Leeds AI cohort. Variants in *COL17A1* are well known to cause recessive JEB and dominant corneal dystrophy ERED, as established in the literature (Has et al., 2020; Jonsson et al., 2015). However, the disease database OMIM does not include reference to *COL17A1*-associated AI, while in the literature it is primarily described in association with JEB via a non-specific term 'enamel hypoplasia'. The work detailed in this thesis showed that *COL17A1*-related AI can occur as an isolated disease in families with no previous family history of JEB. This study described in detail for the first time the phenotypes underlying *COL17A1*-associated dominant AI in nineteen families, with implications for likely disease mechanisms and improving patient care. These findings suggest that AI patients should be considered as carriers for JEB, and should be regarded as at risk of developing ERED. Likewise, patients with ERED should also have their teeth examined for AI. For better disease management, a multidisciplinary strategy is proposed to address AI, JEB, and ERED collectively.

5.2.2 Variants in AMBN cause both dominant and recessive AI with contrasting phenotypes

Another intriguing finding that came from this study was that *AMBN* variants can cause both dominant and recessive AI, which are associated with contrasting disease phenotypes. Recessive AI caused by *AMBN* is well documented in the literature, and the molecular mechanism underlying this is likely to be a complete loss of function of

the AMBN protein, though this has not been proved experimentally. Contrastingly, the molecular mechanism behind dominant AI due to *AMBN* variants is unclear. It was not possible to explain why the variant c.209C>G; p.(Ser70*) caused recessive AI when in a homozygous state, but without an obvious phenotype in carriers, in one group of families, while causing dominant AI as a heterozygous variant in a further two families. Interestingly, these two groups also presented with differing phenotypes. In addition, a third phenotype was associated with another group of apparently dominant families heterozygous for the c.76G>A; p.(Ala26Thr) variant. This variant alters an amino acid immediately adjacent to the AMBN secretory signal peptide cleavage site, which may imply a dominant negative mechanism through failure to export the altered protein. No second pathogenic mutation was identified after WES and long-read sequencing of the whole *AMBN* gene in any of the apparently dominant families. However, it remains possible that the second variant, in *AMBN* or another gene, is not detectable by the sequencing and analysis methods used in this study. It is plausible that a non-coding variant in a known AI-related gene or a completely different variant in an as yet unidentified amelogenesis-related gene or genes could contribute to the disease phenotype in these apparently dominant families. Alternatively, there could be a defective regulatory element on the allele in trans with the variant c.209C>G; p.(Ser70*) which modifies the penetrance of this variant, resulting in differing phenotypes in the dominant families (Castel et al., 2018).

5.2.3 Potential novel candidate genes for AI

By further analysing the WES data, a number of potential novel candidate AI genes were also identified. Although the biological significance of these new gene variants for

AI remains to be determined, their co-occurrence in multiple AI families, and segregation with the phenotype where additional family members are available, may indicate a link. In addition, a number of other lines of evidence support their involvement. These vary from one candidate to another and include expression in developing enamel and/or ameloblasts, association with enamel abnormalities in animal models, or sharing functional pathways with known AI-associated genes. The following are examples of genes, newly implicated in this project, that may have a role in AI.

***CFTR* (Cystic fibrosis transmembrane conductance regulator)**

Putative pathogenic heterozygous variants in the *CFTR* gene were detected in probands from five apparently unrelated families with non-syndromic AI. CFTR is a cAMP-regulated chloride (Cl^-) channel protein that functions in maintaining pH homeostasis and is required for mineral deposition in the enamel space. *CFTR* expression is significantly upregulated in the maturation stage of amelogenesis (Arquitt et al., 2002; Bronckers et al., 2010). It is hypothesised that the primary function of CFTR in amelogenesis is to transfer intracellular Cl^- to the enamel matrix (Lacruz et al., 2012). The Cftr-null mice presented with hypomineralized, chalky-white enamel, but the enamel crystals might appear normal under a scanning electron microscope (Arquitt et al., 2002). A hypomineralized enamel phenotype was observed in both a *CFTR*-null and a *CFTR*-heterozygous porcine animal model (Chang et al., 2011). Mutations in *CFTR* cause autosomal recessive cystic fibrosis (CF) in humans, a condition characterized by buildup of sticky mucus in the lungs and digestive system, leading to bacterial infections which were often lethal in teens and twenties before the use of intensive

antibiotics and physiotherapy (Azevedo et al., 2006). CF is often associated with enamel abnormalities, though these have not been reported previously in heterozygous carriers of CF (Ferrazzano et al., 2012).

***TRPM7* (Transient receptor potential cation channel, subfamily M, member 7)**

Probands from three families with non-syndromic AI were found to carry potentially pathogenic heterozygous missense variants in the *TRPM7* gene. TRPM7 is a transmembrane protein that functions to maintain the intracellular Mg^{2+} concentration required for cell proliferation and differentiation (Ryazanova et al., 2010). *TRPM7* expression is upregulated in maturation stage amelogenesis. *Trpm7*$^{-/-}$ is embryonic lethal but a heterozygous mutation in *Trpm7* caused hypomineralized enamel in mice (Nakano et al., 2016; Rostagno et al., 1994).

***FN1* (Fibronectin1)**

Probands from six apparently unrelated families with non-syndromic AI were found to carry potentially pathogenic heterozygous mutations in the *FN1* gene, with two families sharing the same genotype. FN1 is a multifunctional glycoprotein of the ECM, which plays a role in cell adhesion, cell motility, wound healing, and maintenance of cell shape (Owens & Baralle, 1986; Saito et al., 2015). *FN1* is expressed in late maturation stage amelogenesis. FN1 binds to AMBN and then the AMBN-FN molecule adheres to the cell membrane by interacting with the corresponding integrin receptor (Beyeler et al., 2010).

***RELN* (Reelin)**

Probands from two apparently unrelated families with non-syndromic AI carried potentially pathogenic compound heterozygous variants in the *RELN* gene. RELN is a large extracellular matrix glycoprotein that plays an important role in neuronal migration during mammalian brain development (Lambert de Rouvroit & Goffinet, 1998). *RELN* is expressed in human odontoblast cells, and RELN may play a role in the terminal process of recognition-adhesion between nerve endings and odontoblasts (Bleicher et al., 2001; Maurin et al., 2004).

***PCDH11X/Y* (Protocadherin 11)**

Probands from one family were found to a carry a potentially pathogenic heterozygous mutation in *PCDH11X*, and interestingly, a family with apparent Y-linked inheritance was found to carry a heterozygous variant in *PCDH11Y*. *PCDH11X/Y* are a highly homologous gene pair that lie within a region of homology between the X and Y chromosomes. They encode protocadherin X and protocadherin Y, members of the protocadherin subfamily, and both the X and Y transcripts are expressed in near equal proportions according to data found in the GTEx (genotype tissue expression) portal (Lonsdale et al., 2013). These two cell-surface adhesion molecules are thought to play a fundamental role in cell-cell recognition, essential for the segmental development and function of the central nervous system. Disruption of this gene may be associated with developmental dyslexia (Blanco et al., 2000; Yoshida & Sugano, 1999).

***UQCRC1* (Ubiquinol-cytochrome c reductase core protein 1)**

Probands from four apparently unrelated families with non-syndromic AI were found to carry heterozygous and plausibly pathogenic variants in the *UQCRC1* gene, where two families share the same genotype. Human UQCRC1, an oligomeric enzyme, is a nuclear encoded component of the mitochondrial respiratory chain complex. The gene is highly expressed in the mammalian brain, particularly in the substantia nigra, and is located in the p21 region of chromosome 3, just upstream of the *COL7A1* gene which encodes type VII collagen (Hoffman et al., 1993; Shan et al., 2019). Biological functions of UQCRC1 in mammalian cells are unknown.

5.2.3.1 Potential novel CNVs associated with AI

***ODAM* (Odontogenic ameloblast-associated protein)**

ExomeDepth analysis was carried out on the WES data to detect CNVs in the unsolved families. CNV analysis detected large heterozygous deletions in the *ODAM* gene in probands from three unrelated families. Even though the region of the deletion varied among these families, it is interesting to note that they all have in common deletion of last two exons, exon 9 and exon 10. ODAM is an EMP, and belongs to the SCPP family of proteins localized in maturation ameloblasts. ODAM functions in enamel mineralization by regulating the function of MMP20 (Lee et al., 2010; Park et al., 2007).

***PLXNB2* (Plexin B2)**

CNV analysis by ExomeDepth also detected a large homozygous deletion in the *PLXNB2* gene in one of the patients in the Leeds AI cohort. PLXNB2 is a member of a family of transmembrane receptors that participate in axon guidance and cell migration in

response to semaphorins (Perrot et al., 2002). Mutations in *PLXNB2* are not associated with any human disease in the literature to date, though recently a study reported detecting mutations in *PLXNB2* in association with euploid miscarriages (X. Wang et al., 2023). However, other members of the Leeds Amelogenesis group, in collaboration with Prof Agnes Bloch-Zupan, are leading a research project investigating biallelic variants in Plexin B2 (*PLXNB2*) as a possible cause of syndromic amelogenesis imperfecta, hearing loss and intellectual disability. The detection of a novel variant in *PLXNB2* in a further family with the same phenotype provided additional evidence that variants in *PLXNB2* are a cause of syndromic AI (Smith et al 2023, mauscript submitted).

5.3 Future of AI research

Advancing AI research requires a multifaceted approach. First, continued investment in genomics research, with a focus on large-scale sequencing projects, is needed to expand our understanding of genetic variations and their roles associated with AI. Collaborative efforts to collect diverse genetic data from various research groups are also essential to expand the genetic spectrum of AI. At the same time it is vital to stay up to date with and adapt to the rapidly evolving research tools and technologies available in the AI research. The suggestions listed below could be used to advance AI research in the future to enhance disease diagnostics and clinical management of AI.

5.3.1 Improving our genetic and genomic understanding of AI

5.3.1.1 Potential improvements in the analysis of short read sequencing data

In AI research, short read sequencing is used for the identification of genetic variations within the protein-coding regions of the genome. The following strategies could be taken into account for optimizing variant detection by short read sequencing.

Analysing non-coding regions of the genome

Due to the Mendelian nature of AI and our present understanding of gene function, current research focuses primarily on the analysis of variants that affect the protein-coding regions of a series of genes implicated in AI. While mutations in the exons of genes are well established in molecular pathogenesis of genetic diseases, mutations in non-coding regions, UTRs, promoters, silencers, mobile repetitive elements like SINEs (short interspersed element), LINEs (long interpersed elements) or uORFs (upstream open reading frames) can also lead to diseases by disrupting the precise regulation of gene expression, splicing, and other cellular processes (Plaisancié et al., 2018). For a more comprehensive analysis and understanding of the genetic basis of AI, it is essential to analyse both coding and non-coding regions of the genome by WGS. WGS provides a more comprehensive representation of an individual's genome, including all genes, non-coding regions, and regulatory elements. Bioinformatic tools like MELT (mobile element locator tool) and uORF Finder can be incorporated into the WGS data analysis pipeline for identifying and annotating mobile elements and uORFs (Gardner et al., 2017; Scholz et al., 2019).

Using machine learning tools for the prediction of protein function

It is reported that 98% of more than 4 million observed missense variants are classified as VUS. Lack of accurate classification of VUS variants limits diagnosis and clinical management of rare diseases (Cheng et al., 2023). A new tool 'AlphaMissense', a deep machine learning model that builds on the protein structure prediction tool AlphaFold2 (Jumper et al., 2021), can accurately predict the effect of variants without the need for expensive and time consuming experimental methods (Cheng et al., 2023). This tool can be used to gain insights into the potential effects of pathogenic and VUS variants on the disease mechanism of AI.

Custom whole gene sequencing

Custom whole gene sequencing allows analysing a predefined set of genes that are known or suspected to be associated with the disease of interest (Tekin et al., 2016). This approach is more focused and cost effective and can be implemented in AI research to analyse the whole gene, including introns, UTRs and promotor sequences, of a panel of known and candidate genes associated with AI, for a more thorough screen for some patients.

Resolving reference genome coverage and bias

The choice of reference genome can affect the outcome and interpretation of the data analysis. Build hg19 of the human reference genome has been used for the interpretation of genetic variants identified in AI cases in this study. The reference genome assemblies like GRCh37 (genome reference consortium human build 37 or hg19) and GRCh38 (or hg20) have gaps and ambiguities in certain genomic regions,

particularly in regions with complex structural variants, repetitive elements, or regions that are highly polymorphic. These reference genomes were also built based on sequencing data from a small number of individuals and may not accurately represent the genetic variation present in certain populations, leading to reference bias. To address these limitations the T2T (telomere-to-telomere) consortium recently finished the first complete sequence of a haploid human genome, T2T-CHM13, providing a contiguous representation of all the chromosomes including centromeres (Nurk et al., 2022). Furthermore, to address the reference bias issue, a draft pangenome consisting of 47 diploid assemblies from a cohort of genetically diverse individuals is underway (Liao et al., 2023). Using a complete T2T pangenome as a reference genome in AI research will make previously uncharacterised genomic regions accessible, perhaps revealing new genes and genomic features involved in human diseases, and will make it easier to distinguish disease-causing variants from neutral changes that are population specific. As additional genomic data become available, the dataset underlying the pangenome can be expanded and updated to keep the analysis current and relevant.

Investigating digenic/oligogenic inheritance of AI

Although the underlying cause of AI is genetic, there can be other contributing factors or modifiers that influence the severity or manifestation of the disease. For example, proteins encoded by genes at modifier loci can alter the function of the proteins which harbour variants that cause Mendelian disease, without affecting the genes that encode the affected proteins. The bioinformatics platform ORVAL can predict potential digenic/oligogenic interactions between rare variants in multiple genes that may contribute to the disease risk (Renaux et al., 2019). Investigating the role and

interactions of multiple gene variants may identify additional modifier genes that interact with the primary causative gene, contributing to the AI phenotype.

5.3.1.2 Long read sequencing technology

Even after implementing the improvements in short-read sequencing described above, limitations remain in its ability to detect variants causing human diseases. The PCR steps involved in producing short-read sequencing libraries can cause non-uniform coverage, haplotype phasing is rarely possible with sequences assembled from 150 bp fragments, structural variants can prove difficult to detect and assembling sequence in repetitive regions is challenging. These issues are increasingly being addressed by implementing long-read sequencing, in AI research as in other areas of human genetics study.

Implementing long-read sequencing (rather than short-read sequencing) for routine diagnostics may be prohibitively expensive at present. However, with technological advancements the cost per genome is falling and throughput is increasing. A PromethION flow cell can generate ~100 Gb of data in real-time, which is revolutionary. Therefore, it seems likely that, within a few years, long read sequencing technology will be adapted for routine genetic disease diagnostics.

Long-read DNA sequencing

Long-read sequencing of genomic DNA can provide a more comprehensive view of structural variants, and can accurately interpret the length of repeat expansions and phasing of variants (Sedlazeck et al., 2018). Long-read sequencing also allows

sequencing the native DNA without amplification, preserving the structural information in the DNA molecule, the epigenetic markers which can be critical for understanding gene regulation, and chromatin conformation (Gouil & Keniry, 2019).

Long read transcriptome sequencing

Long-read transcriptome sequencing allows sequencing of full-length transcripts, determining the underlying exon combinations directly. Therefore long-read transcriptome sequencing can distinguish variations in usage of different transcript isoforms in different tissues or in normal and disease states, leading to the discovery of rare/novel isoforms and novel gene fusions. Detecting specific transcript isoforms related to AI subtype could lead to the identification of previously missed mutations. Capturing rare transcript isoforms may help in resolving AI cases, when only one heterozygous variant is identified in a recessive gene (Marwaha et al., 2022).

5.3.2 Disease modelling

To understand the function of the proteins identified in this and other AI genetics studies, and the functional consequences of the novel variants implicated in AI, cellular models can be created using the CRISPR/Cas9 gene editing technique in human induced pluripotent stem cells (iPSCs). CRISPR (clustered regularly interspaced short palindromic repeats) refers to specific DNA sequences found in the genomes of bacteria and archaea, and Cas9 is an endonuclease. Cas9 enzyme can use CRISPR sequences as a guide to recognize specific sequences to make precise cuts in the DNA (Gaj et al., 2013). By using this technique, novel mutations associated with AI could be introduced into iPSCs in culture. These modified stem cells could then be differentiated

into ameloblast-like cell types (Miao et al., 2022). By creating mutation-bearing cell models it will be possible to assess how these mutations impact cellular processes and contribute to the development of AI. With the improved understanding of the disease pathophysiology, interventions can be implemented targeting critical cellular and molecular pathways. For example, there are cases of AI resulting from ameloblast endoplasmic reticulum stress (Brookes et al., 2017). These may be treatable using molecular chaperones like 4-PBA, which has been studied for its potential role in facilitating the proper folding and stability of proteins, preventing their misfolding, aggregation, or degradation (Brookes et al., 2014).

5.3.3 Clinical research

The development of treatments for AI has been hampered by the fact that it is in fact not one disease but many. If genetic findings are disseminated to clinicians caring for affected individuals, it may become possible to improve AI treatment strategies significantly. Some mutations cause severe forms of AI, while others result in a milder phenotype. Some mutations affect the quantity of enamel, while others affect its quality. Knowing the exact genetic cause(s) might give new insights into the likely prognosis of the disease for each patient and for their relatives. Based on this information, dentists can make informed choices concerning treatment plans for patients, offering restorative treatments to some while recognising that others may require more extensive interventions, such as full dental coverage or crowns. Also, the existence of groups of patients with defined genetic subtypes of AI will help to facilitate clinical trials and will help patients get recruited to any trials that are running.

5.3.4 Biobank

Biobanks have the potential to advance AI research. The biobank refers to a large collection of well-annotated clinical, molecular and pathological data from patients/volunteers that can be used in medical research for improving understanding of the causes of a wide ranges of disease, for biomarker identification and for designing potential future treatments (Coppola et al., 2019). In AI research, biobanks can provide a centralized resource for the storage and management of clinical data, imaging studies (e.g., dental X-rays), genetic data and biological samples. This resource can help in linking specific genetic variants to the specific clinical presentations of AI for accurate diagnosis and treatment planning.

5.3.5 Shared platform between geneticists and clinicians

AI research will also benefit from having a shared platform for better communication between geneticists and clinicians. This collaboration will foster greater understanding of human genetics in the dental community and will ensure that genetic information is effectively considered and included in the patient's medical evaluation and treatment plan.

5.4 Benefit of genetics testing

As well as directly informing clinical care, obtaining a genetic diagnosis can benefit patients, their families and the clinicians who care for them in a number of ways, as described below.

5.4.1 Psychosocial support

Genetic information empowers patients to make informed lifestyle choices and implement risk mitigation strategies. Genetic testing may provide reassurance to patients who are concerned about their genetic predisposition to certain diseases. A study conducted on the psychological consequences of predictive testing for Huntington's disease showed improvement of the psychological health of persons who received results irrespective of the outcome (Wiggins et al., 1992). Confirmed genetic diagnoses can direct proactive health care, while negative results can reduce worry.

5.4.2 Personalised medicine

Personalised medicine aims to identify specific subtypes of a disease by identification of specific genetic variants that cause the disease, in order to tailor treatment strategies accordingly. As detailed above, at present AI is largely treated as a single disease of varying severity. However, this and other AI research projects have shown that AI can result from mutations in various genes, each of which can affect enamel development differently. To date, mutations in more than 20 different genes have been associated with NS AI. Different genetic subtypes and different specific variants lead to variations in the clinical features of AI. Even within families with the same genetic variant, there can be variability in the phenotypic expression of AI (Aldred & Crawford, 1997). Again the genetic variability in AI interacting with environmental factors, including dietary habits, oral hygiene practices, and exposure to fluoride adds further complexity in the clinical appearance and prognosis of AI (Wright, 2023). All these factors underscore the importance of personalized treatment, which is only made possible by finding out the exact genetic cause underlying AI in each patient. A precise

molecular diagnosis will allow clinicians to develop bespoke treatments targetted at specific sub-types of disease in a personalized medicine approach. Genetics based clinical treatment strategies will help dentists to tailor treatment and support to the unique characteristics and needs of each individual. For example, AI patients carrying muations in *COL17A1* and *LAMB3* genes need to be aware that they are carriers for EB or they may develop ERED in the future, while AI patients with biallelic variants in *FAM20A* will need to have their kidney function assessed.

5.4.3 AI gene-disease catalogue

Creating a comprehensive database of the genetic variants associated with AI will be a valuable resource for finding information about specific genes, variants and their roles associated with AI. This will serve as a centralized repository of knowledge and research findings for students, researchers and healthcare professionals.

5.4.4 Support group network

AI patients will benefit from connecting through social support networks and communities. This can provide emotional support and valuable resources for managing their health conditions (Desine et al., 2021).

5.5 Future of genetics in healthcare

In 2020, the UK government published a ‘Genome UK: the future of healthcare’ strategy led by Genomics England in partnership with the NHS to develop the world's most cutting-edge genomic healthcare system and provide better health results at a lesser cost. One of the key points of this plan is implementing WGS in newborns to

speed up diagnosis and treatment of rare genetic diseases. The program aims to sequence 500,000 whole genomes and to make WGS a part of routine health care. The whole genome sequence can provide a wealth of information about an individual's genetic makeup, and will enable detection of rare diseases at birth that might otherwise go unnoticed until symptoms develop. This will facilitate personalized medical care and enable interventions for improving lifestyle and mitigating risk.

Long term management of genetic data presents substantial ethical challenges as it carries potential consequences for future generations (Eichinger et al., 2021). Determining the ownership and control of access to genetic data is an ongoing debate (Pinxten & Howard, 2014). WGS can reveal information that may result in genetic stigmatization and discrimination, impacting various aspects of a person’s life, including employment, insurance, and education. There are psychological and emotional impacts associated with discovering genetic risks and predispositions. WGS of minors is a complex issue in the context of consent, privacy, and the right to know or not know genetic information. It is important to establish clear ethical guidelines, informed consent procedures, data-sharing frameworks and robust privacy safegurds to ensure that WGS is conducted in a responsible and ethical manner, respecting the freedom and privacy of individuals.

Another key point of the government’s strategy is implementing tailored genomic sequencing of research participants from under-represented ancestry groups. Historically, genomic research has been heavily biased towards individuals of European ancestry (Sirugo et al., 2019). This lack of diversity can result in biased or incomplete

insights into the genetic basis of diseases and traits. Including participants from under-represented groups will help overcome this gap. This will further increase our understanding of genomic diversity, impacting clinical management of genetic disease to improve patient outcomes across all communities.

Recently, using artificial intelligence in the disease diagnosis has gained significant attention. Machine learning algorithms can analyse vast amounts of patient data, including medical records, imaging, and genetic information, to identify patterns and detect diseases at an earlier stage. An artificial inteligence tool, AlphaFold developed by DeepMind, has been designed for predicting the three-dimensional (3D) structure of proteins, deciphering their functions and how they contribute to various biological processes and diseases. Another tool, AlphaMissense, was developed to use the protein structure predictions of AlphaFold for classifying missense variants. These machine learning tools are invaluable in exploring the pathogenicity of unannotated variants without the need of expensive experimental procedures (Cheng et al., 2023).

A promising area for development in future clinical genetics practise is the use of polygenic risk scores (PRS) for disease risk assessment for early diagnosis. PRS quantify an individual's genetic predisposition to polygenic diseases by considering the cumulative effect of multiple genetic variants, potentially extending the benefits of genetic prediction from Mendelian to complex inherited diseases. PRS can help assess an individual's risk for conditions such as coronary artery disease, atrial fibrillation, hypertension, type 2 diabetes, breast cancer, Alzheimer's disease, schizophrenia, bipolar disorder, major depressive disorder and many more (Ala-Korpela & Holmes,

2019; Leonenko et al., 2021). Knowing in advance of an individual's risk of diabetes, hypertension or coronary disease allows them to make choices and gives them the option to change lifestyle practices to reduce the risk of these diseases. However, it may not that easy to take a decision for prophylactic mastectomy for breast cancer prevention or it may be controversial to inform a child if found that they are at high risk of early onset of Alzheimers or schizophrenia (Lewis & Vassos, 2020). There is also a concern that genetic information obtained through PRS could be used for discriminatory purposes, such as in employment or insurance decisions. Legal protections, such as GINA (the genetic information nondiscrimination act) in the United States aim to prevent such discrimination, but the effectiveness of these laws may vary by jurisdiction (Underhill-Blazey & Klehm, 2020).

In conclusion, the field of genetic research has a wide range of opportunities and challenges in the future. On the positive side, genetic research has the potential to transform healthcare and medicine. It presents the possibility of personalized therapies based on an individual's genetic makeup through precision medicine. Considerable progress has already been made in the understanding of the genetic causes of disease and the development of novel treatments thanks to genetic research. It can assist in identifying those who are more susceptible to genetic disease, allowing for early diagnosis and treatment. However, advance in genetic research also comes with ethical, privacy, and security issues, particulary when it comes to the use of genetic data and the possibility of discrimination. For genetic research to continue in an ethical and responsible manner, it is imperative that concerns about privacy, ethical considerations, and equal access to genetic advancements be addressed. As this field

of research continues to develop, finding the right balance between innovation and ethical safeguards will become increasingly important.

5.6 References

Ala-Korpela, M., & Holmes, M. V. (2019). Polygenic risk scores and the prediction of common diseases. *International Journal of Epidemiology*, *49*(1), 1-3. https://doi.org/10.1093/ije/dyz254

Aldred, M. J., & Crawford, P. J. (1997). Molecular biology of hereditary enamel defects. *Ciba Found Symp*, *205*, 200-205; discussion 205-209. https://doi.org/10.1002/9780470515303.ch14

Arquitt, C. K., Boyd, C., & Wright, J. T. (2002). Cystic fibrosis transmembrane regulator gene (CFTR) is associated with abnormal enamel formation. *J Dent Res*, *81*(7), 492-496. https://doi.org/10.1177/154405910208100712

Azevedo, T. D., Feijó, G. C., & Bezerra, A. C. (2006). Presence of developmental defects of enamel in cystic fibrosis patients. *J Dent Child (Chic)*, *73*(3), 159-163.

Beyeler, M., Schild, C., Lutz, R., Chiquet, M., & Trueb, B. (2010). Identification of a fibronectin interaction site in the extracellular matrix protein ameloblastin. *Exp Cell Res*, *316*(7), 1202-1212. https://doi.org/10.1016/j.yexcr.2009.12.019

Blanco, P., Sargent, C. A., Boucher, C. A., Mitchell, M., & Affara, N. A. (2000). Conservation of PCDHX in mammals; expression of human X/Y genes predominantly in brain. *Mamm Genome*, *11*(10), 906-914. https://doi.org/10.1007/s003350010177

Bleicher, F., Couble, M. L., Buchaille, R., Farges, J. C., & Magloire, H. (2001). New genes involved in odontoblast differentiation. *Adv Dent Res*, *15*, 30-33. https://doi.org/10.1177/08959374010150010701

Bronckers, A., Kalogeraki, L., Jorna, H. J., Wilke, M., Bervoets, T. J., Lyaruu, D. M., Zandieh-Doulabi, B., Denbesten, P., & de Jonge, H. (2010). The cystic fibrosis transmembrane conductance regulator (CFTR) is expressed in maturation stage ameloblasts, odontoblasts and bone cells. *Bone*, *46*(4), 1188-1196. https://doi.org/10.1016/j.bone.2009.12.002

Brookes, S. J., Barron, M. J., Boot-Handford, R., Kirkham, J., & Dixon, M. J. (2014). Endoplasmic reticulum stress in amelogenesis imperfecta and phenotypic rescue using 4-phenylbutyrate. *Hum Mol Genet*, *23*(9), 2468-2480. https://doi.org/10.1093/hmg/ddt642

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., & Dixon, M. J. (2017). Amelogenesis imperfecta caused by N-terminal enamelin point mutations in mice and men is driven by endoplasmic reticulum stress. *Hum Mol Genet*, *26*(10), 1863-1876. https://doi.org/10.1093/hmg/ddx090

Castel, S. E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., & Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature Genetics*, *50*(9), 1327-1334. https://doi.org/10.1038/s41588-018-0192-y

Chang, E. H., Lacruz, R. S., Bromage, T. G., Bringas, P., Jr., Welsh, M. J., Zabner, J., & Paine, M. L. (2011). Enamel pathology resulting from loss of function in the cystic fibrosis transmembrane conductance regulator in a porcine animal model. *Cells Tissues Organs*, *194*(2-4), 249-254. https://doi.org/10.1159/000324248

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, *381*(6664), eadg7492. https://doi.org/doi:10.1126/science.adg7492

Coppola, L., Cianflone, A., Grimaldi, A. M., Incoronato, M., Bevilacqua, P., Messina, F., Baselice, S., Soricelli, A., Mirabelli, P., & Salvatore, M. (2019). Biobanking in health care:

evolution and future directions. *J Transl Med*, *17*(1), 172. https://doi.org/10.1186/s12967-019-1922-3

Desine, S., Eskin, L., Bonham, V. L., & Koehly, L. M. (2021). Social support networks of adults with sickle cell disease. *J Genet Couns*, *30*(5), 1418-1427. https://doi.org/10.1002/jgc4.1410

Eichinger, J., Elger, B. S., Koné, I., Filges, I., Shaw, D., Zimmermann, B., & McLennan, S. (2021). The full spectrum of ethical issues in pediatric genome-wide sequencing: a systematic qualitative review. *BMC Pediatrics*, *21*(1), 387. https://doi.org/10.1186/s12887-021-02830-w

Ferrazzano, G. F., Sangianantoni, G., Cantile, T., Amato, I., Orlando, S., & Ingenito, A. (2012). Dental enamel defects in Italian children with cystic fibrosis: an observational study. *Community Dent Health*, *29*(1), 106-109.

Gaj, T., Gersbach, C. A., & Barbas, C. F., 3rd. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol*, *31*(7), 397-405. https://doi.org/10.1016/j.tibtech.2013.04.004

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., Mills, R. E., & Devine, S. E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res*, *27*(11), 1916-1929. https://doi.org/10.1101/gr.218032.116

Gouil, Q., & Keniry, A. (2019). Latest techniques to study DNA methylation. *Essays Biochem*, *63*(6), 639-648. https://doi.org/10.1042/ebc20190027

Has, C., Bauer, J. W., Bodemer, C., Bolling, M. C., Bruckner-Tuderman, L., Diem, A., Fine, J. D., Heagerty, A., Hovnanian, A., Marinkovich, M. P., Martinez, A. E., McGrath, J. A., Moss, C., Murrell, D. F., Palisson, F., Schwieger-Briel, A., Sprecher, E., Tamai, K., Uitto, J., . . . Mellerio, J. E. (2020). Consensus reclassification of inherited epidermolysis bullosa and other disorders with skin fragility. *Br J Dermatol*, *183*(4), 614-627. https://doi.org/10.1111/bjd.18921

Hoffman, G. G., Lee, S., Christiano, A. M., Chung-Honet, L. C., Cheng, W., Katchman, S., Uitto, J., & Greenspan, D. S. (1993). Complete coding sequence, intron/exon organization, and chromosomal location of the gene for the core I protein of human ubiquinol-cytochrome c reductase. *J Biol Chem*, *268*(28), 21113-21119.

Jonsson, F., Byström, B., Davidson, A. E., Backman, L. J., Kellgren, T. G., Tuft, S. J., Koskela, T., Rydén, P., Sandgren, O., Danielson, P., Hardcastle, A. J., & Golovleva, I. (2015). Mutations in collagen, type XVII, alpha 1 (COL17A1) cause epithelial recurrent erosion dystrophy (ERED). *Hum Mutat*, *36*(4), 463-473. https://doi.org/10.1002/humu.22764

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., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

Kim, Y. J., Kang, J., Seymen, F., Koruyucu, M., Zhang, H., Kasimoglu, Y., Bayram, M., Tuna-Ince, E. B., Bayrak, S., Tuloglu, N., Hu, J. C., Simmer, J. P., & Kim, J. W. (2020). Alteration of Exon Definition Causes Amelogenesis Imperfecta. *J Dent Res*, *99*(4), 410-418. https://doi.org/10.1177/0022034520901708

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, *267*(5608), 275-276. https://doi.org/10.1038/267275a0

Lacruz, R. S., Smith, C. E., Bringas Jr, P., Chen, Y.-B., Smith, S. M., Snead, M. L., Kurtz, I., Hacia, J. G., Hubbard, M. J., & Paine, M. L. (2012). Identification of novel candidate genes involved in mineralization of dental enamel by genome-wide transcript profiling. *Journal of Cellular Physiology*, *227*(5), 2264-2275. https://doi.org/https://doi.org/10.1002/jcp.22965

Lambert de Rouvroit, C., & Goffinet, A. M. (1998). A new view of early cortical development. *Biochemical Pharmacology*, *56*(11), 1403-1409. https://doi.org/https://doi.org/10.1016/S0006-2952(98)00209-3

Lee, H. K., Lee, D. S., Ryoo, H. M., Park, J. T., Park, S. J., Bae, H. S., Cho, M. I., & Park, J. C. (2010). The odontogenic ameloblast-associated protein (ODAM) cooperates with RUNX2 and modulates enamel mineralization via regulation of MMP-20. *J Cell Biochem*, *111*(3), 755-767. https://doi.org/10.1002/jcb.22766

Leonenko, G., Baker, E., Stevenson-Hoare, J., Sierksma, A., Fiers, M., Williams, J., de Strooper, B., & Escott-Price, V. (2021). Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nature communications*, *12*(1), 4506. https://doi.org/10.1038/s41467-021-24082-z

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., . . . Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580-585. https://doi.org/10.1038/ng.2653

Marwaha, S., Knowles, J. W., & Ashley, E. A. (2022). A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Medicine*, *14*(1), 23. https://doi.org/10.1186/s13073-022-01026-w

Maurin, J.-C., Couble, M.-L., Didier-Bazes, M., Brisson, C., Magloire, H., & Bleicher, F. (2004). Expression and localization of reelin in human odontoblasts. *Matrix Biology*, *23*(5), 277-285. https://doi.org/https://doi.org/10.1016/j.matbio.2004.06.005

Miao, X., Niibe, K., Fu, Y., Zhang, M., Nattasit, P., Ohori-Morita, Y., Nakamura, T., Jiang, X., & Egusa, H. (2022). Epiprofin Transcriptional Activation Promotes Ameloblast Induction From Mouse Induced Pluripotent Stem Cells via the BMP-Smad Signaling Axis. *Front Bioeng Biotechnol*, *10*, 890882. https://doi.org/10.3389/fbioe.2022.890882

Nakano, Y., Le, M. H., Abduweli, D., Ho, S. P., Ryazanova, L. V., Hu, Z., Ryazanov, A. G., Den Besten, P. K., & Zhang, Y. (2016). A critical role of TRPM7 as an ion channel protein in mediating the mineralization of the craniofacial hard tissues. *Frontiers in physiology*, *7*, 258.

Owens, R. J., & Baralle, F. E. (1986). Mapping the collagen-binding site of human fibronectin by expression in Escherichia coli. *Embo j*, *5*(11), 2825-2830. https://doi.org/10.1002/j.1460-2075.1986.tb04575.x

Park, J. C., Park, J. T., Son, H. H., Kim, H. J., Jeong, M. J., Lee, C. S., Dey, R., & Cho, M. I. (2007). The amyloid protein APin is highly expressed during enamel mineralization and maturation in rat incisors. *Eur J Oral Sci*, *115*(2), 153-160. https://doi.org/10.1111/j.1600-0722.2007.00435.x

Perrot, V., Vazquez-Prado, J., & Gutkind, J. S. (2002). Plexin B regulates Rho through the guanine nucleotide exchange factors leukemia-associated Rho GEF (LARG) and PDZ-RhoGEF. *J Biol Chem*, *277*(45), 43115-43120. https://doi.org/10.1074/jbc.M206005200

Pinxten, W., & Howard, H. C. (2014). Ethical issues raised by whole genome sequencing. *Best Practice & Research Clinical Gastroenterology*, *28*(2), 269-279. https://doi.org/https://doi.org/10.1016/j.bpg.2014.02.004

Plaisancié, J., Tarilonte, M., Ramos, P., Jeanton-Scaramouche, C., Gaston, V., Dollfus, H., Aguilera, D., Kaplan, J., Fares-Taie, L., Blanco-Kelly, F., Villaverde, C., Francannet, C., Goldenberg, A., Arroyo, I., Rozet, J. M., Ayuso, C., Chassaing, N., Calvas, P., & Corton, M. (2018). Implication of non-coding PAX6 mutations in aniridia. *Human Genetics*, *137*(10), 831-846. https://doi.org/10.1007/s00439-018-1940-x

Renaux, A., Papadimitriou, S., Versbraegen, N., Nachtegael, C., Boutry, S., Nowé, A., Smits, G., & Lenaerts, T. (2019). ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Research*, *47*(W1), W93-W98. https://doi.org/10.1093/nar/gkz437

Rostagno, A., Williams, M. J., Baron, M., Campbell, I. D., & Gold, L. I. (1994). Further characterization of the NH2-terminal fibrin-binding site on fibronectin. *J Biol Chem*, *269*(50), 31938-31945.

Ryazanova, L. V., Rondon, L. J., Zierler, S., Hu, Z., Galli, J., Yamaguchi, T. P., Mazur, A., Fleig, A., & Ryazanov, A. G. (2010). TRPM7 is essential for Mg2+ homeostasis in mammals. *Nature communications*, *1*(1), 109.

Saito, K., Fukumoto, E., Yamada, A., Yuasa, K., Yoshizaki, K., Iwamoto, T., Saito, M., Nakamura, T., & Fukumoto, S. (2015). Interaction between fibronectin and β1 integrin is essential for tooth development. *PLOS ONE*, *10*(4), e0121667. https://doi.org/10.1371/journal.pone.0121667

Scholz, A., Eggenhofer, F., Gelhausen, R., Grüning, B., Zarnack, K., Brüne, B., Backofen, R., & Schmid, T. (2019). uORF-Tools—Workflow for the determination of translation-regulatory upstream open reading frames. *PLOS ONE*, *14*(9), e0222459. https://doi.org/10.1371/journal.pone.0222459

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*, *19*(6), 329-346. https://doi.org/10.1038/s41576-018-0003-4

Shan, W., Li, J., Xu, W., Li, H., & Zuo, Z. (2019). Critical role of UQCRC1 in embryo survival, brain ischemic tolerance and normal cognition in mice. *Cell Mol Life Sci*, *76*(7), 1381-1396. https://doi.org/10.1007/s00018-019-03007-6

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell*, *177*(1), 26-31. https://doi.org/10.1016/j.cell.2019.02.048

Tang, M., Alaniz, M. E., Felsky, D., Vardarajan, B., Reyes-Dumeyer, D., Lantigua, R., Medrano, M., Bennett, D. A., de Jager, P. L., Mayeux, R., Santa-Maria, I., & Reitz, C. (2020). Synonymous variants associated with Alzheimer disease in multiplex families. *Neurol Genet*, *6*(4), e450. https://doi.org/10.1212/nxg.0000000000000450

Tekin, D., Yan, D., Bademci, G., Feng, Y., Guo, S., Foster, J., 2nd, Blanton, S., Tekin, M., & Liu, X. (2016). A next-generation sequencing gene panel (MiamiOtoGenes) for comprehensive analysis of deafness genes. *Hear Res*, *333*, 179-184. https://doi.org/10.1016/j.heares.2016.01.018

Underhill-Blazey, M., & Klehm, M. R. (2020). Genetic Discrimination: The Genetic Information Nondiscrimination Act's Impact on Practice and Research. *Clin J Oncol Nurs*, *24*(2), 135-137. https://doi.org/10.1188/20.Cjon.135-137

Vihinen, M. (2022). When a Synonymous Variant Is Nonsynonymous. *Genes (Basel)*, *13*(8). https://doi.org/10.3390/genes13081485

Wiggins, S., Whyte, P., Huggins, M., Adam, S., Theilmann, J., Bloch, M., Sheps, S. B., Schechter, M. T., & Hayden, M. R. (1992). The psychological consequences of predictive testing for Huntington's disease. Canadian Collaborative Study of Predictive Testing. *N Engl J Med*, *327*(20), 1401-1405. https://doi.org/10.1056/nejm199211123272001

Wright, J. T. (2023). Enamel Phenotypes: Genetic and Environmental Determinants. *Genes (Basel)*, *14*(3). https://doi.org/10.3390/genes14030545

Yoshida, K., & Sugano, S. (1999). Identification of a novel protocadherin gene (PCDH11) on the human XY homology region in Xq21.3. *Genomics*, *62*(3), 540-543. https://doi.org/10.1006/geno.1999.6042

APPENDICES

Appendix 1 Running MIPgen to design MIP probes.

Create bed files in the genome browser: To run the MIPGEN command for designing smMIPs probes for a genomic region, genomic coordinates are needed to include as an input file. Using genome browser, a bed file containing all the genomic coordinates was generated as below,

- Tools > Table Browser.
- Click genome in the region
- In the Identifiers section, click Paste list and enter the gene names.
- Then in the Output format: put BED.
- Give the file a name.bed in the Output File section, and then get output.

It will ask a few questions on the next page (so you want to flank the regions etc.) and then press get bed. To include the immediate splice site select exons plus 5-10 bases at each end.

Generate smMIPs probes using the following command

/home/MIPGEN/mipgen -regions_to_scan /home/meduh/AI-mips.bed -project_name /AI-mips -min_capture_size 150 -max_capture_size 150 -bwa_genome_index /home/ref/b37/human_g1k_v37.fasta -snp_file /home/ref/b37/dbSnp151.b37.vcf.gz -tag_sizes 8,0

Generate a UCSC track to visualize online or on IGV

python /home/MIPGEN/tools/generate_ucsc_track.py your_file_name.picked_mips.txt your_file_name.ucsc_track

More detailed instructions can be found in the link https://github.com/shendurelab/MIPGEN.

Appendix 2 Sequences of MIP backbone, Custom read 1, read 2 and index primers.

MIP Backbone	CTTCAGCTTCCCGATCCGACGGTAGTGTNNNNNNNNNNNN
Illumina overhang (P5)	AATGATACGGCGACCACCGAGATCTACAC
Forward PCR Primer	AATGATACGGCGACCACCGAGATCTACACATACGAGATCCGTAA**TCGGGAAGCTGAAG**
Reverse PCR Primer	CAAGCAGAAGACGGCATACGAGATXXXXXXXXACACGCACGA**TCCGACGGTAGTGT**
Illumina Overhang (p7)	CAAGCAGAAGACGGCATACGAGAT
Read 1 sequencing primer	CATACGAGATCCGTAATCGGGAAGCTGAAG
Index sequencing primer	ACACTACCGTCGGATCGTGCGTGT
Read 2 sequencing primer	ACACGCACGATCCGACGGTAGTGT

Appendix 3 MiSeq and NextSeq MIP sequencing protocol using custom primers.

MiSeq

1) MiSeq – custom primers spike-in with existing illumine primers:

   Volume Illumina primers: 680 µL (0.5 µM)
   Spike in custom primers: 3.4 µL of 100 µM (results in 0.5 µM custom + Illumina primers)

NextSeq

2) NextSeq – custom primers spike in with existing illumina primers:

   Volume Illumina primers: 1.8 mL (0.5 µM)
   Spike in custom primers: 9 µL of 100 µM (results in 0.5 µM custom + Illumina primers)

MiSeq protocol

- Label three tubes 1.5 µL tube for each primer (read1, read 2 and index primers), and combine 145 µL HT1 with 5 µL primer (100µM); vortex; spin down.
- Make sure the cartridge is fully thawed.
- Punch holes in the sample well and wells 12, 13, and 14 using a 1000 µL pipette tip
- Load 600 µL of sample DNA (9 pM) to sample well
- Load the full amount of the primer dilution (150 µL) of the read 1 primer to well 12,
- Load the full amount of the primer dilution (150 µL) of the index primer to well 13,
- Load the full amount of the primer dilution (150 µL) of the reverse/r2 primer to well 14
- So in brief: (read 1 = well 12, index = well 13, read 2 = well 14)

NextSeq 500:

- Label three 1.5 ml tubes as 20_read 1, 21_read2 and 22_index.
- Pierce the NextSeq cartridge on the well positions 20 (read1); 21 (read2), 22 (index) each with a clean pipette tip.
- Take 500 µL aliquot from position 20 (read 1) and add it to labelled tube 20_read 1
- Take 500 µL aliquot from position 21 (read2) and add it to labeled tube 21_read 2
- Take 500 µL aliquot from position 22 (Index) and add it to labeled tube 22_index
- Add 9 µL of each custom primer to their corresponding labeled tubes.
- Mix all the tubes, and add the 509 µL back to the respective cartridge positions:
    - Add 509 µL from tube 20_read 1 to position 20 (BP10, Read1)
    - Add 509 µL from tube 21_read 2 to position 21 (BP11, Read2)
    - Add 509 µL from tube 22_index to position 22 (BP12, Index1)

Appendix 4: Setting up the MIPVAR pipeline.

1. Download the MIPVAR pipeline from SourceForge: https://sourceforge.net/projects/mipvar/ Unpack it into the mips-pipe/programs/mipvar directory. The latest release is MIPVAR-0.1.0-package.tar.gz

2. Move the file MIPVAR-0.1.0.jar from the lib directory into the directory above it (so that it's out of the way) and replace it with the jar file that Agne compiled. This resolves the BWA output stream error. The file that Agne compiled is named "MIPVAR.jar". It doesn't matter that the filename is not the same as the name of the file that was removed.

3. Make sure you have BWA (v.0.7.12), BEDTools (v.2.24.0) and samtools (v.0.1.19) installed and available on the $PATH. This can be done using a Conda environment. The conda environment that we're using is named "MipVarPipe". Activate the environment ($source activate MipVarPipe). The versions of BWA and BEDTools are not the latest versions but are releases that were available at the time when the MIPVAR (v.0.1.0) package was uploaded onto SourceForge (14th March 2016). Samtools is needed to index the reference FASTA rather than being needed to run MipVar. To install a specific version of a tool into a Conda environment use *e.g.*

   $conda install bwa==0.7.12
   $conda install bedtools==2.24.0
   $conda install samtools==0.1.19

4. Download GATK v.3.2.2 from Google Cloud (https://console.cloud.google.com/storage/browser/_details/gatk-software/package-archive/gatk/GenomeAnalysisTK-3.2-2-gec30cee.tar.bz2) and unpack it into the mips-pipe/programs/gatk directory. The most recent versions don't seem to work due to a change in the GATK command line argument structure (and there's no option to change the system call in MIPVAR). Version 3.2.2 was detailed in the runConfigExample.txt file in the example dataset that is available to download from SourceForge (Example_input_data).

5. Edit the variables in the runConfig.txt file and save this in mips-pipe/programs/mipvar/MIPVAR-0.1.0-package. It should look like the following screenshot.

   The fasta file may need to be updated to Hg38 at some point (this will require the dbSNP file to also be updated).

   The annotation properties and coverage statistics_properties variable should not be modified.

```
(MipVarPipe) [medcmwa@login2.arc3 MIPVAR-0.1.0-package]$ pwd
/nobackup/medcmwa/mips-pipe/programs/mipvar/MIPVAR-0.1.0-package
(MipVarPipe) [medcmwa@login2.arc3 MIPVAR-0.1.0-package]$ ll
total 184
drwxr-xr-x 2 medcmwa med   4096 Feb 26 12:14 conf
drwxr-xr-x 2 medcmwa med  12288 Feb 26 12:19 lib
-rw-r----- 1 medcmwa med 166727 Feb 26 12:14 MIPVAR-0.1.0.jar
-rw-r--r-- 1 medcmwa med    463 Feb 26 14:58 runConfig.txt
(MipVarPipe) [medcmwa@login2.arc3 MIPVAR-0.1.0-package]$ cat runConfig.txt
reference_file=/nobackup/medcmwa/mips-pipe/resources/references/hg19/ucsc.hg19.fasta
GATK_jar=/nobackup/medcmwa/mips-pipe/programs/gatk/package-archive-gatk-GenomeAnalysisTK-3.2-2-gec30cee/GenomeAnalysisTK.jar
dbsnp_vcf=/nobackup/medcmwa/mips-pipe/resources/references/hg19/dbsnp_138.hg19.vcf
annotationProperties=classpath:/conf/FIONA/${annotation}_AnnotationPipeline.properties
coverage_statistics_properties=classpath:/conf/FIONA/CoverageStatistics.properties
```

Running the MIPVAR pipeline

1. When FASTQ.gz files are generated by the MiSeq they have the following format:

   $zcat /nobackup/meduh/AI-MIPS-01/Sample_AI-163-4486/AI-163-4486_S9_L001_R1_001.fastq.gz | head -n 1

```
(mips) [medcmwa@login2.arc3 Sample_AI-291_5091_original]$ zcat AI291-5091_S4_L001_R1_001.fastq.gz | head
@M00480:123:000000000-JCC5L:1:1101:15384:1945 1:N:0:4
TGGCCTACACTGTCCAGGGCCCACCAGGCCAGCCTGGGCCACAGGGGCCACCCGGCATCAGCAAGGTCTTCTCTGCCTACAGCAACGTGACTGCGGACCTCATGGACTTCTTCCAAAGTAAGGGCATCCAATTAATCGACACTACCGTCGG
+
AAA@AFF11>@1FAF3111AA00BEHHGGG0G0AF00CGCEFFHFAEGGGFHEGGCEEHHHHHHHGHHHHHFGHGHHHFHHHHHHGGEAFHFHGGCE<GHEHHHBGGHHGHFHFHEHHGHHBGGGHHHBFGGDBGFEACCCHHHHGGGGG
```

   In the above example the 4 refers to the index number rather than the index sequence, the latter of which is required for MipVar to run.

   Consequently, the :N:0:**4** needs to be changed to :N:0:**<INDEX SEQUENCE>**. The index sequence is checked by the megapool.txt file, which assigns the sample name to the index. N.B. Reads 1 and 2 and represented differently **1**:N:0:4 and **2**:N:0:4 respectively.

   $zcat AI-239-4864_S1_L001_R2_001.fastq.gz | head

To make the changes first unzip (gunzip fastq.gz) the FASTQ file then adapt the following example sed command:

Forward 1:N:0:1, reverse 2:N:0:1 (last number corresponds to S1, S2 etc in the fastq)
$sed -i s/:N:0:1/:N:0:TGCTAGAG/g AI-239-4864_S1_L001_R2_001.fastq
(where <INDEX> represents the nucleotide sequence of the index.
Then re-gzip the FASTQ file (gzip).

2. Create a megapool.txt file that links the SampleIDs to the Index sequence contained in the header row of each FASTQ file. An example for AI_MIPS_04 is recorded below. Put this in the AI_MIPS_04 directory. (The directory directly above each of the Sample_ folders.) The sample directory column has to match the suffix after the beginning of the "Sample_" as this is what is looked for by the submission script. The file should be named with the format AI_smMIPs_<BATCH_NUMBER>.txt

For the Amelogenesis Imperfecta panel use the prefix "AI"

i.e.: AI_MIPS_04.megapool.txt

```
(MipVarPipe) [medcmwa@login2.arc3 AI_MIPS_04_MipVar]$ cat AI_MIPs_04.megapool.txt
TGCTAGAG        AI-344_5408
TGAGAGCT        AI-376_5718
GTCACTCA        AI-373_5707
ATAAGCGT        AI-291_5091
ACTATCTG        AI-377_4460
ATGGTGAC        AI-356_5579
```

3. Setup a new sampleConfig.txt file for each sample. This is quite fiddly to do. Put the sampleConfig.txt file into each of the Sample_ directories. Prefix the sampleConfig.txt file with the Sample ID. i.e. AI-291_5091.sampleConfig.txt

Change the following variables for each sample: The megapool_file variable only need setting once for each run.

Mapping_folder
Forward_file
Reverse_file

```
(mips) [medcmwa@login2.arc3 Sample_AI-291_5091]$ cat AI-291_5091.sampleConfig.txt
sample_name=AI_MIPS_04
mapping_folder=/nobackup/medcmwa/AI_MIPS_04_MipVar/Sample_AI-291_5091
machine=MISEQ
forward_file=/nobackup/medcmwa/AI_MIPS_04_MipVar/Sample_AI-291_5091/AI291-5091_S4_L001_R1_001.fastq.gz
reverse_file=/nobackup/medcmwa/AI_MIPS_04_MipVar/Sample_AI-291_5091/AI291-5091_S4_L001_R2_001.fastq.gz
barcode_file=null
forward_random_tag=0
reverse_random_tag=8
megapool_file=/nobackup/medcmwa/AI_MIPS_04_MipVar/megapool.AI_MIPS_04.txt
remove_duplicates=TRUE
mer_file=/nobackup/medcmwa/mips-pipe/resources/interval_lists/2021-02-11_All_probes.picked_mips.txt
trimming=TRUE
annotation=RESEARCH
multiple_hcdiffs=FALSE
gene_panels=
downsample_value=5000
count_probes=TRUE
"run_tools=FASTQ_GENERATOR,BAM_FILE_GENERATOR,VARIANT_CALLING"
```

4. Load the submission script using the “megapool” file to select which samples need to be run:

$source activate MipVarPipe
$qsub -t 1-6 submission.sh
$qstat

This will write out a MIPS_out_<date> direcotry into each of the Sample_dirs It will also attempt to run Annovar on the produced VCF.

```
(mips) [medcmwa@login2.arc3 AI_MIPS_04_MipVar]$ cat submission.sh
#!/bin/bash

# Use current environment variables and modules
#$ -V

# Set current working directory
#$ -cwd

# Request 2 hour of runtime
#$ -l h_rt=02:00:00

# Request 11 GB of memory per core
#$ -l h_vmem=11G

# Request 3 cores
#$ -pe smp 3

# Email at the beginning and end of the job
#$ -m be

export TMPDIR=/nobackup/medcmwa/mips_temp/
export TMP_DIR=$TMPDIR

RunFolder=$(pwd)
Sample_Tag_ID=$(awk "NR==$SGE_TASK_ID" megapool.AI_MIPS_04.txt)
SampleID=$(echo $Sample_Tag_ID | cut -f 2 -d " ")
echo $SampleID

java -Dsnappy.disable=true -Dlogging.config=/nobackup/medcmwa/mips-pipe/programs/mipvar/MIPVAR-0.1.0-package/conf/logback.xml -cp "/nobackup/medcmwa/mips-pipe/programs/mipvar/MIPVAR-0.1.0-package/lib/*" org.umcn.gen.mip.pipeline.RunMIPPipeline -sampleConfig ${RunFolder}/Sample_${SampleID}/${SampleID}.sampleConfig.txt -environment EMPTY -overridePipelineConfig /nobackup/medcmwa/mips-pipe/programs/mipvar/MIPVAR-0.1.0-package/runConfig.txt

perl /nobackup/medcmwa/mips-pipe/programs/annovar/convert2annovar.pl -format vcf4 --includeinfo --withzyg Sample_${SampleID}/MIP_out_*/*.variants.vcf > Sample_${SampleID}/${SampleID}.avinput

perl /nobackup/medcmwa/mips-pipe/programs/annovar/table_annovar.pl Sample_${SampleID}/${SampleID}.avinput /nobackup/medcmwa/mips-pipe/programs/annovar/humandb/ -buildver hg19 -out Sample_${SampleID}/${SampleID} -remove -protocol refGene,cytoBand,exac03,avsnp147,dbnsfp30a -operation gx,r,f,f,f -nastring . -csvout -polish -xref /nobackup/medcmwa/mips-pipe/programs/annovar/humandb/gene_xref.txt --otherinfo
```

Additional notes

The file 2021-02-11_All_probes.picked_mips.txt was created from the following MIPGEN files:
180220.picked_mips.txt
AI-mips8.picked_mips.txt
COL17A1.picked_mips.txt
new-genesv1.picked_mips.txt

To create a useable reference sequence:

The reference FASTA file was first indexed using BWA v.0.7.12 (the same version that is installed in the conda environment:

$bwa index ucsc.hg19.fasta

The FASTA file was then indexed using samtools
$samtools faidx ucsc.hg19.fasta

GATK requires a Dictionary file. Picard v.1.102.0 is included with MipVar but this is too hard to install via conda. Therefore used the version of Picard (v.2.21.3) that was installed in the original MIPs environment to generate the Dictionary.

$picard CreateSequenceDictionary R=ucsc.hg19.fasta

Trimming parameter

Decided to turn off the “Trimming” i.e. trimming=FALSE

This has the following effect in the IGV. Upper track is the no-trim track. Lower track shows trimmed reads (which reduces coverage because bits of the read are removed – see grey horizontal line coming out of the back of the reads). With the trimming turned off there appears to be a number of additional variant calls that arise at the ends of the reads.

Appendix 5: Exome library prep using Twist comprehensive exome kit.

The following protocol is a brief description, it will need consulting the detailed protocol 'library preparation EF 2.0' when preparing library for sequencing.

Precapture library prep

Genomic DNA libraries were prepared for exome sequencing using Twist library preparation kit EF 2.0 (104207: 96 reaction) following the manufacturer's guidelines. An enzymatic fragmentation was carried out for DNA fragmentation and Twist universal adaptor system (101308) was used for multiplexing. A summary of the protocol is given below.

50 ng high-quality genomic DNA (gDNA) was used to make libraries using enzymatic fragmentation protocol. 18 minutes of fragmentation time was used to target a size of 200bp and subsequently end-repaired to generate dA-tailed gDNA fragments. Twist universal adaptors were ligated to the dA-tailed gDNA fragments followed by a paramagnetic beads clean-up using DNA purification beads. Next, Twist UDI (unique dual index) was added to dA-tailed gDNA fragments by a PCR reaction using 6 PCR cycles. The Indexed gDNA libraries were purified by paramagnetic beads and the size of the libraries was checked by tapestation using DNA BR 5000 kit (Figure 1).

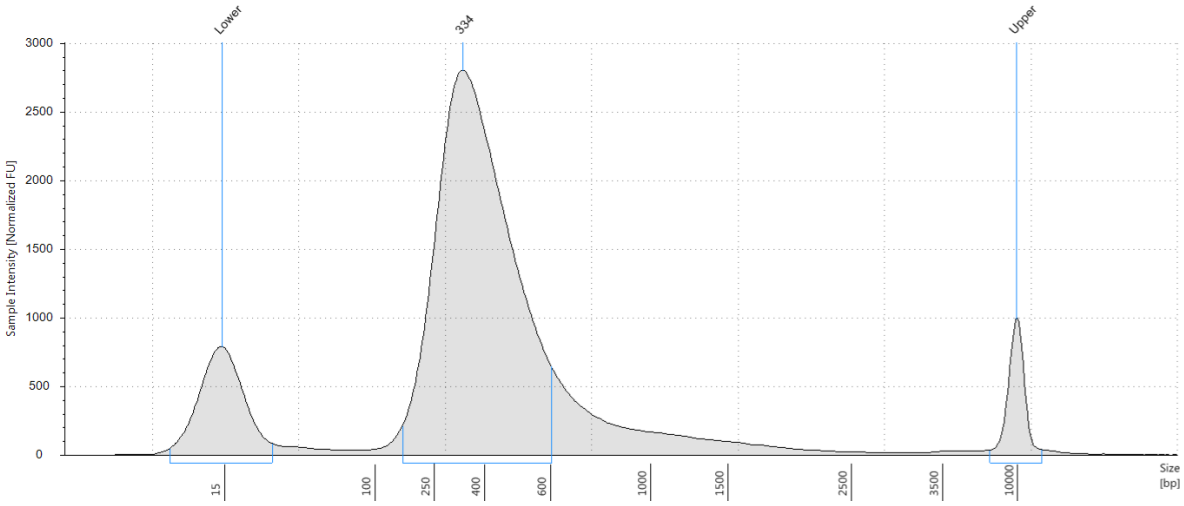


Figure 1: A tapestation trace of an indexed gDNA library. 50 ng high quality gDNA was used for the library prep followed by an 18 minute fragmentation at 37°. The final library was PCR amplified using 6 cycles.

Hybridization

The DNA libraries were quantified by Qubit BR reagent. A hybridization capture reaction was made by pooling 187.5 ng of indexed gDNA library from each of 8 samples. In this way, 12 hybridization capture reactions were made each pool containing 1500 ng DNA.

Number of indexed samples per pool	Amount of each indexed library per pool	Total mass per pool
8	187.5 ng	1500 ng

After pooling, all the capture reaction tubes (12 tubes) were dried using a vacuum concentrator using no heat. The hybridization reaction mix was made using the comprehensive exome panel according to the manufacturer's protocol and added to the dried capture reactions. The hybridization reaction was carried out at 70° for 16 hours on a thermocycler. Hybridized targets were enriched by binding to Dyna beads streptavidin. These beads are biotinylated oligonucleotide probes (baits) that are complementary to the genetic sequence of interest. The biotinylated probes immobilize the oligonucleotide onto the bead surface and can be separated by

magnetic precipitations. The isolated and purified target regions were PCR amplified using 8 cycles to target a panel size of 10-50Mb. The post-capture libraries were purified with DNA purification beads at a 1:1 ratio according to the manufacturer's protocol and visualized on a tapestation using DNA 1000 BR kit (Figure 2). Sequencing was done using P3 300 cycle kit on the NextSeq2000 instrument following manufacturer's protocol.

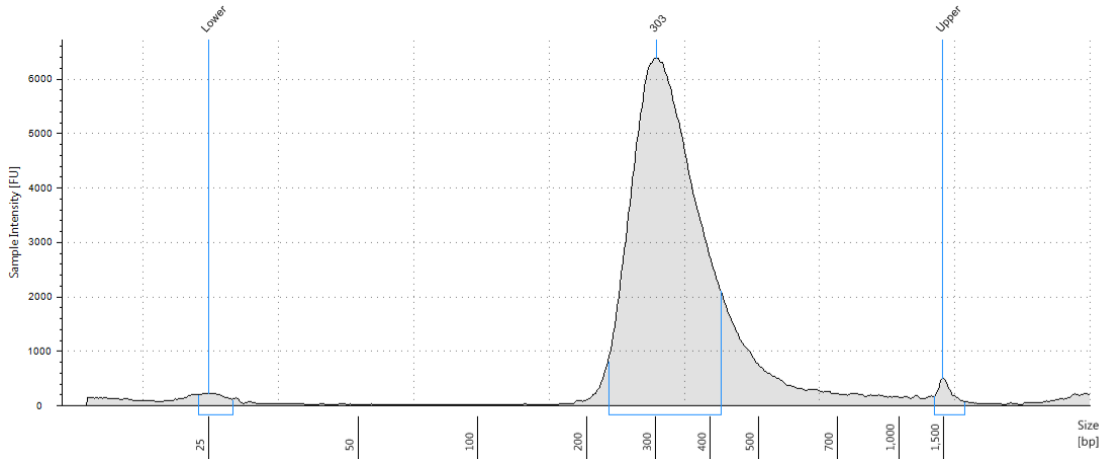


Figure 2: Tapestation trace of a post capture gDNA library

Appendix 6: Exome data analysis pipeline.

1. Trim the adaptors and do quality control:

#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

2. 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 - > **/path/sample**_bwa.bam

3. Next sort the alignment:

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

4. remove original bam to save space:

rm -i /**path/sample**_bwa.bam

5. 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

6. Delete pre-deduplicated bam to save space:

rm -i /**path/sample**_bwa.sort.bam

7. 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

8. 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

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

rm -i **sample**_bwa.sort.dedup.bam

gzip **sample**_bwa.sort.dedup.intervals

10. 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

11. print reads

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

12. Delete old bam (the non-recal file)

rm -i **sample**_bwa.sort.dedup.indelrealn.bam

13. Generate g.vcf file for each sample 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 /data/bssy/results/**sample**_bwa.sort.dedup.indelrealn.recal.bam -o /data/bssy/results/**sample**.g.vcf

14. Convert the raw.g.vcf to a raw.vcf: (Optional: 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/bssy/results/**sample**.g.vcf -o /data/bssy/results/**sample**.combined.raw.vcf -nda --showFullBamList -nt 8

15. Merge g.vcf files to a combined .vcf:

java -Xmx4g -jar /home/marc1_b/bs16gn/GenomeAnalysisTK-3.5.jar -T GenotypeGVCFs \
-R /nobackup/bgycels/human_g1k_v37.fasta \
-V /nobackup/bgycels/george/4880.g.vcf \
-V /nobackup/bgycels/george/5079.g.vcf \
-o /data/bs16gn/unsolved/combinedAI164_4481_4483.noBED.vcf

echo "g.vcfs merged split variants for filtering" && \

16. Split Variants SNP

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T SelectVariants -R /home/ref/b37/human_g1k_v37.fasta -selectType SNP --variant /data/bs16gn/path/to/file.vcf -o /data/bs16gn/path/to/file-snps.vcf

17. Split Variants INDEL

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T SelectVariants -R /home/ref/b37/human_g1k_v37.fasta --variant /data/bs16gn/path/to/file.vcf -selectType INDEL -selectType MNP -o /data/bs16gn/path/to/file-indels.vcf

18. Hard filtering SNP

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

19. Hard filtering INDEL

java -Xmx4g -jar /home/GATK/GenomeAnalysisTK-3.5-0/GenomeAnalysisTK.jar -T VariantFiltration -R /home/ref/b37/human_g1k_v37.fasta -V /data/bs16gn/path/to/file-indels.vcf --filterExpression "QD < 20 || FS > 200.0 || ReadPosRankSum < -20.0" --filterName "indel_hard_filter" -o /data/bs16gn/path/to/file.fltd-indels.vcf

20. CombineFilteredVariants

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

21. Filter variants in gnomad >/= 1%

perl /home/vcfhacks-v0.2.0/filterVcfOnVcf.pl -f /home/ref/ExAC/gnomad.exomes.r2.0.1.sites.vcf.gz -w -y 0.0001 -b -i /data/bs16gn/path/to/file.fltd-combined.vcf -o /data/bs16gn/path/to/file.fltd-combined.gnomAD.vcf

echo "filtering finished, now doing VEP annotation" && \

22. Annotate variants with VEP

perl /home/variant_effect_predictor/variant_effect_predictor.pl --offline --vcf --everything --dir_cache /home/variant_effect_predictor/cache_b37 --dir_plugins /home/variant_effect_predictor/cache_b37/Plugins --plugin SpliceConsensus --fasta /home/variant_effect_predictor/cache_b37/homo_sapiens/92_GRCh37/Homo_sapiens.GRCh37.75.dna.primary_assembly.fa.gz -i /data/bs16gn/path/to/file.fltd-combined.gnomAD.vcf -o /data/bs16gn/path/to/file.fltd-combined_gnomad.vep.vcf -fork 6

23. 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

24. Rank on CADD Score

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

-d gives unsorted list

#upload the .tsv file to CADD to filter variants that were not filtered automatically

25. 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

26. 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

27. 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

echo "finished AR filtering, now starting AD filtering"

28. Select dominant variants: get functional 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

29. Rank on CADD Score

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

-d gives unsorted list

#upload the .tsv file to CADD to filter variants that were not filtered automatically

30. 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

31. 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

echo "finished with AD filtering, check output"

Appendix 7: Data processing commands for long read sequencing.

Include both pass and fail reads to easily define the “total” number of sequenced reads

$mkdir analysis. $cd analysis
$zcat ../fastq_sup/*.fastq.gz > all_reads.fastq
$gzip all_reads.fastq
$source activate nanopore_tools
$porechop -i AI-312A-5319.fastq.gz -o AI-312A-5319.allRds.porechop.fastq.gz > porechop.log

Filter by read length (+/- 500bp surrounding amplicon)

$gunzip -c AI-312A-5319.porechop.fastq.gz | NanoFilt -l 8020 --maxlength 9020 --logfile NanoFilt.length.log | gzip > AI-312A-5319.porechop.8020_9020.fastq.gz

$NanoStat --fastq PKD1_CB-8352.allRds.porechop.7003_8003.fastq.gz

```
0(ONT_tools) [medcmwa@login1.arc4 analysis]$ NanoStat --fastq PKD1_CB-8352.allRds.porechop.7003_8003.fastq.gz
General summary:
Mean read length:                7,375.4
Mean read quality:                   9.8
Median read length:              7,388.0
Median read quality:                 9.8
Number of reads:                29,561.0
Read length N50:                 7,390.0
Total bases:                218,024,401.0
Number, percentage and megabases of reads above quality cutoffs
>Q5:    28763 (97.3%) 212.1Mb
>Q7:    25333 (85.7%) 187.1Mb
>Q10:   13910 (47.1%) 103.2Mb
>Q12:   6141 (20.8%) 45.7Mb
>Q15:   139 (0.5%) 1.0Mb
Top 5 highest mean basecall quality scores and their read lengths
1:      16.7 (7466)
2:      16.7 (7497)
3:      16.6 (7519)
4:      16.4 (7480)
5:      16.3 (7572)
Top 5 longest reads and their mean basecall quality score
1:      8001 (4.8)
2:      8000 (6.5)
3:      7999 (7.0)
4:      7998 (5.4)
5:      7998 (4.5)
```

Typically use Q10 as the cutoff; stick with this for this analysis

$gunzip -c AI-312A-5319.porechop.8020_9020.fastq.gz | NanoFilt -q 10 --logfile NanoFilt.quality.log | gzip > AI-312A-5319.porechop. 8020_9020.Q10.fastq.gz

$minimap2 -ax map-ont -t 4 /nobackup/meduh/hg38/Homo_sapiens_assembly38.fasta AI-312A-5319.porechop.8020_9020.fastq.gz 2>> minimap2.log >> AI-312A-5319.porechop.8020_9020.Q10.sam

$samtools view -bh AI-312A-5319.porechop.8020_9020.Q10.sam > AI-312A-5319.porechop.8020_9020.Q10.bam

$samtools sort -o AI-312A-5319.porechop.8020_9020.Q10.sort.bam AI-312A-5319.porechop.8020_9020.Q10.bam
$samtools index AI-312A-5319.porechop.8020_9020.Q10.sort.bam

Remove supplementary and secondary alignments from the BAM so that they don't affect the mapped read count. -F 2304 Exclude (not primary alignment, supplementary alignment)

$samtools view -F 2304 -bh AI-312A-5319.porechop.8020_9020.Q10.sort.bam > AI-312A-5319.porechop.8020_9020.Q10.sort.ExSuppSec.bam

$samtools index AI-312A-5319.porechop.8020_9020.Q10.sort.ExSuppSec.bam