Exploring the genetic architecture of Fuchs Endothelial Corneal Dystrophy, a common, age-related, and visually disabling disease

Amanda Nicole Sadan

Thesis submitted for the degree of

Doctor of Philosophy

Supervisors: Associate Professor Alice Davidson and Professor Alison Hardcastle

University College London

2023

Declaration

I, Amanda Nicole Sadan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

_____

Acknowledgments

Firstly, I would like to thank my primary supervisor, Associate Professor Alice Davidson for giving me the opportunity to do this PhD. Words cannot express my gratitude for the continuous support and guidance she has shown me over the years. Thank you for sharing your expertise and dedicating your time in helping me become a better scientist. I couldn't have asked for a better PhD supervisor.

Next, I would like to thank my secondary supervisory, Professor Alison Hardcastle, for her enthusiasm, guidance and always taking the time to make sure I'm doing okay. I would also like to thank Professor Mike Cheetham and Jacqui Van Der Spuy for their advice and contributions during our research update meetings. Thank you also to Naheed Kanuga for her genuine kindness and keeping the lab running.

I would like to give a special thanks to Dr. Niko Pontikos, Dr. Cian Murphy, and Anita Szabó for all their help with the bioinformatics aspect of my PhD. To Dr Marc Ciosi and the other members of Professor Monckton's lab for sharing their expertise and being patient with me whilst I learnt. Many thanks to Steve Tuft, Kiri Muthusamy and Petra Lišková for sharing their clinical knowledge and ongoing support. Without any of you, my PhD would not have been possible.

A sincere thanks to Dr. Christina Zarouchlioti and Dr. Nihar Bhattacharyya, you have both been a great support throughout my PhD and I have learnt so much from the both of you. I would also like extend my thanks to the wider MCN group, it has been a pleasure working with such a lovely and supportive group. Namely, Nathan Hafford Tear for sharing the ups and downs

of the PhD experience with me. To my bench buddy, Olivia Rezek, we have laughed and we have cried together but more importantly, we helped each through this journey. I am so thankful our PhDs brought us together, I have truly found a lifelong friend in you.

I would also like to extend my gratitude to my family, especially my mum and dad, for the unconditional support. You have believed in me when I couldn't and always encouraged me to chase my dreams. I wouldn't be the person I am today without you both. To my Nanny Lucia, thank you for all your faith in me and keeping me in your prayers. To my partner Tom, you have been my lifeline during this PhD. Thank you for all your support, encouragement, forever listening to my problems and caring for me when times were tough.

Finally, I would like to thank my PhD funders, National eye Research Centre and Rosetrees Trust, and to every FECD patient who has generously donated their samples to research. This PhD would not have been possible without them, and I hope that someday this thesis will positively impact them.

Declaration form: referencing doctoral candidate's own published work(s)

in thesis

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- have been uploaded to a preprint server;
- are in submission to a peer-reviewed publication;
- have been published in a peer-reviewed publication, e.g. journal, textbook.

This form should be completed as many times as necessary. For instance, if a student had seven thesis chapters, two of which having material which had been published, they would complete this form twice.

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**
   a. **Where was the work published?**
   Progress in Retinal and Eye Research
   b. **Who published the work?**
   Elsevier
   c. **When was the work published?**
   March 2021
   d. **Was the work subject to academic peer review?** YES
   e. **Have you retained the copyright for the work?** YES
   ☒ *I acknowledge permission of the publisher named under 1b to include in this thesis portions of the publication named as included in 1a.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**
   a. **Where is the work intended to be published?** (e.g. journal name) _____
   b. **List the manuscript's authors in the intended authorship order:**
   _____
   c. **Stage of publication:**
   ☐ Not yet submitted
   ☐ Submitted
   ☐ Undergoing revision after peer review
   ☐ In press

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**

Michael P. Fautsch, Eric D. Wieben, Keith H. Baratz, Nihar Bhattacharyya, Amanda N. Sadan, Nathaniel J. Hafford-Tear, Stephen J. Tuft, and Alice E. Davidson provided literature searches, compilation of figures and tables, writing and editing of the manuscript.

Amanda N. Sadan compiled Figures 6 and 7, and Table 10 used in this thesis.

4. In which chapter(s) of your thesis can this material be found?

Chapter 1 and 3.

5. Candidate's e-signature: A. Sadan
Date: 30/03/2023

6. Supervisor/senior author(s) e-signature: A. Davidson
Date: 30/03/2023

Declaration form: referencing doctoral candidate's own published work(s)

in thesis

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- have been uploaded to a preprint server;
- are in submission to a peer-reviewed publication;
- have been published in a peer-reviewed publication, e.g. journal, textbook.

This form should be completed as many times as necessary. For instance, if a student had seven thesis chapters, two of which having material which had been published, they would complete this form twice.

7. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**
    a. **Where was the work published?**
    b. **Who published the work?**
    c. **When was the work published?**
    d. **Was the work subject to academic peer review?**
    e. **Have you retained the copyright for the work?**

    ☒ *I acknowledge permission of the publisher named under 1b to include in this thesis portions of the publication named as included in 1a.*

    f. **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)
    YES, bioRxiv, https://doi.org/10.1101/2023.03.29.534731

8. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**
    a. **Where is the work intended to be published?** (e.g. journal name)
    American Journal of Human Genetics
    b. **List the manuscript's authors in the intended authorship order:**

    Nihar Bhattacharyya, Nathaniel J Hafford-Tear, Amanda N Sadan, Anita Szabo, Niuzheng Chai, Christina Zarouchlioti, Jana Jedlickova, Szi Kay Leung, Tianyi Liao, Lubica Dudakova, Pavlina Skalicka, Mohit Parekh, Aaron R Jeffries, Michael E Cheetham, Kirithika Muthusamy, Alison J Hardcastle, Nikolas Pontikos, Petra Liskova, Stephen J Tuft, Alice E Davidson

    c. **Stage of publication:**

    ☐ Not yet submitted
    ☐ Submitted
    ☒ Undergoing revision after peer review
    ☐ In press

9. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**

Nihar Bhattacharyya provided conceptualisation, data generation and analysis, manuscript writing and editing, and compiled Figure 44 used in this thesis.

Amanda N Sadan and Nathaniel J Hafford-Tear provided data generation and analysis, and assisted in manuscript editing.

Anita Szabo, Niuzheng Chai and Christina Zarouchlioti assisted in data analysis and manuscript editing.

Jana Jedlickova, Lubica Dudakova provided data generation and assisted in manuscript editing.

Szi Kay Leung assisted in data analysis and manuscript editing.

Tianyi Liao assisted in data analysis.

Pavlina Skalicka and Kirithika Muthusamy assisted in patient recruitment.

Mohit Parekh provided data generation.

Aaron R Jeffries provided data generation and analysis and assisted in manuscript editing.

Michael E Cheetham and Alison J Hardcastle assisted in conceptualisation and manuscript editing.

Nikolas Pontikos provided data analysis and assisted in manuscript editing.

Petra Liskova assisted in patient recruitment, provided manuscript editing and funding acquisition.

Stephen J Tuft provided conceptualisation, assisted in manuscript editing and assisted in patient recruitment.

Alice E Davidson assisted in conceptualization, data generation and analysis, provided manuscript writing and editing, supervision and project administration, funding acquisition.

10. In which chapter(s) of your thesis can this material be found?

chapter 5

11. Candidate's e-signature: A. Sadan
Date: 30/03/2023

12. Supervisor/senior author(s) e-signature: A. Davidson
Date: 30/03/2023

List of publications

Bhattacharyya, N., Hafford-Tear, N. J., **Sadan, A. N.**, Szabo, A., Chai, Niuzheng., Zarouchlioti, C., Jedlickova, J., Leung, S. K., Liao, T., Dudakova, L., Skalicka, P., Parekh, M., Jeffries, A. R., Cheetham, M. E., Muthusamy, K., Hardcastle, A. J., Pontikos, N., Liskova, P., & Davidson, A. E. (2023). Deciphering novel TCF4-driven molecular origins and mechanisms underlying a common triplet repeat expansion-mediated disease. *bioRxiv 2023.03.29.534731*, https://doi.org/10.1101/2023.03.29.534731

Liu, S., **Sadan, A. N.**, Muthusamy, K., Zarouchlioti, C., Jedlickova, J., Pontikos, N., Thaung, C., Hardcastle, A. J., Netukova, M., Skalicka, P., Dudakova, L., Bunce, C., Tuft, S. J., Davidson, A. E., & Liskova, P. (2023). Phenotype and genotype of concurrent keratoconus and Fuchs endothelial corneal dystrophy. *Acta ophthalmologica*, 10.1111/aos.15654. Advance online publication. https://doi.org/10.1111/aos.15654

Dudakova, L., Skalicka, P., Davidson, A. E., **Sadan, A. N.**, Chylova, M., Jahnova, H., Anteneova, N., Tesarova, M., Honzik, T., & Liskova, P. (2021). Should Patients with Kearns-Sayre Syndrome and Corneal Endothelial Failure Be Genotyped for a *TCF4* Trinucleotide Repeat, Commonly Associated with Fuchs Endothelial Corneal Dystrophy?. *Genes*, *12*(12), 1918. https://doi.org/10.3390/genes12121918

Koay, S. Y., **Sadan, A. N.**, Gore, D. M., Goyal, S., & Tuft, S. (2021). Endothelial corneal dystrophy with annular stromal clefts. *Canadian journal of ophthalmology. Journal canadien d'ophtalmologie*, *56*(5), e150–e152. https://doi.org/10.1016/j.jcjo.2021.02.029

Fautsch, M. P., Wieben, E. D., Baratz, K. H., Bhattacharyya, N., **Sadan, A. N.**, Hafford-Tear, N. J., Tuft, S. J., & Davidson, A. E. (2021). TCF4-mediated Fuchs endothelial corneal dystrophy: Insights into a common trinucleotide repeat-associated disease. *Progress in retinal and eye research*, *81*, 100883. https://doi.org/10.1016/j.preteyeres.2020.100883

Davidson, A. E., Hafford-Tear, N. J., Dudakova, L., **Sadan, A. N.**, Pontikos, N., Hardcastle, A. J., Tuft, S. J., & Liskova, P. (2020). CUGC for posterior polymorphous corneal dystrophy (PPCD). *European journal of human genetics : EJHG*, *28*(1), 126–131. https://doi.org/10.1038/s41431-019-0448-8

Hafford-Tear, N. J., Tsai, Y. C., **Sadan, A. N.**, Sanchez-Pintado, B., Zarouchlioti, C., Maher, G. J., Liskova, P., Tuft, S. J., Hardcastle, A. J., Clark, T. A., & Davidson, A. E. (2019). CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy-associated TCF4 triplet repeat. *Genetics in medicine : official journal of the American College of Medical Genetics*, *21*(9), 2092–2102. https://doi.org/10.1038/s41436-019-0453-x

Abstract

Fuchs endothelial corneal dystrophy (FECD) is a common, age-related, genetically heterogenous and visually disabling disease. Expansion (defined as ≥50 copies) of a triplet repeat (termed CTG18.1) in an intron of *TCF4* has been significantly associated with FECD. Other rare genetic causes of FECD have been reported and further genetic heterogeneity is hypothesised.

In this thesis I present the genetic characterization of 990 FECD patients recruited at Moorfields Eye Hospital (London) and General University Hospital (Prague). DNA samples were genotyped for CTG18.1 length using a short tandem repeat assay and triplet-primed-PCR. Genotyping demonstrated approximately 80% of cases had at least one expanded CTG18.1 allele.

FECD cases with one or more expanded alleles were further analysed by an ultra-deep and high-throughput MiSeq-based sequencing approach to determine CTG18.1 allelic structure in the expanded state, and to identify potential sequence variants. Furthermore, data was used to estimate progenitor allele lengths and calculate somatic expansions scores for all expanded alleles. Corresponding samples were also concurrently genotyped using a kompetitive allele-specific PCR assay to consider polymorphisms within DNA repair genes, previously identified to act as *trans*-acting genetic modifiers of other repeat-expansion mediated diseases. Regression analysis identified a significant association between CTG18.1 expansion rates and several polymorphisms in DNA mismatch repair genes.

For expansion negative individuals (<50 repeats; n=141) exome sequencing was performed. Data were interrogated for rare (minor allele frequency ≤0.01) potentially causative variants in FECD-associated genes,

which led to the identification of presumed pathogenic variants in genes including *COL8A2*, *SLC4A11* and *ZEB1*. However, the vast majority of this group remained genetically unsolved and hence a subsequent gene burden case-control analysis was performed and led to the identification of a number of novel and significant associations. In conclusion, this thesis provides new insights into the complex and genetically heterogenous landscape of FECD.

Impact statement

Fuchs endothelial corneal dystrophy (FECD) is a common, progressive and debilitating eye disease affecting the cornea endothelium. In 2012, a CTG triplet repeat expansion (≥50 repeats), situated within an intron of *TCF4* (termed CTG18.1), was associated with FECD and it is now established to be the leading genetic risk factor for the disease. Approximately 75-80% of European FECD patients harbour at least one allele carrying this repeat expansion. However, the genetic causes and/or risk factors for non-CTG18.1 associated disease remain largely elusive. Currently, the only effective treatment for advanced stage disease is surgical intervention which is invasive and relies upon specialist facilities and the availability of healthy donor material, of which there is a global shortage. To develop alternative, effective therapeutic approaches, the underlying genetic cause and molecular mechanisms contributing to FECD pathophysiology need to be comprehensively understood.

In this thesis I have genotyped a large cohort of 990 FECD patients using PCR-based genotyping methodologies to gain insights into FECD genotype-phenotype correlations. Furthermore, I have applied a high-throughput ultra-deep sequencing method to sequence CTG18.1 in expansion-positive FECD samples (n= 630) to quantify somatic expansion rates and explore allelic structure. Moreover, I have genotyped common polymorphisms in DNA repair genes, *MLH1, FAN1, MSH3, PMS1, PMS2, LIG1* and *RRM2B/UBR5*, previously identified to modify somatic instability of the Huntington disease-associated repeat and consequently disease onset and progression. Regression analysis presented examines the potential relationship between DNA repair variants and CTG18.1 instability and provides insights into how the DNA repair pathway may act as *trans*-acting modifiers of FECD phenotypic outcomes. Thus, these data

have advanced our understanding of the role DNA repair pathways may play in controlling CTG18.1 (in)stability and, in future, these finding have the potential to facilitate novel FECD therapeutic approaches.

The large-scale FECD cohort data presented has also contributed to a new clinical indication under the National genomic test directory commissioned by the NHS. It has made a case for CTG18.1 genotyping to be integrated into patient care pathways to enable an efficient, reliable and cost-effective molecular diagnosis for FECD patients. This measure is hoped to advance early disease detection and, in future, enable eligible patients to be identified for emerging CTG18.1-targeted therapies that are currently being developed.

FECD missing heritability is also extensively investigated within the cohort. Genetic aetiology of the genetically refined CTG18.1 expansion-negative subset of the total FECD cohort (n= 141) is explored by exome sequencing. Importantly, this work also represents the first relatively large-scale attempt to genetically characterise FECD cases that do not harbour a CTG18.1 expansion and, through application of a gene-burden approach, provides novel genetic candidates for disease, including *miR-184.*

Table of contents

List of figures

List of tables

Abbreviations

A1 - Minor allele

A2 – Major allele

AIC - Akaike information criterion

AMD - Age-related macular degeneration

ATP - Adenosine triphosphate

BAM - Binary Alignment Map

BCVA - Best corrected visual

BER- Base excision repair

BH-FDR - Benjamini & Hochberg False Discovery Rate

bp – base pairs

BWA-MEM - Burrows-Wheeler Aligner Maximal Exact Match

CADD - Combined Annotation Dependent Depletion

CD - Corneal dystrophies

CEC – Corneal endothelial cell

CED - Corneal endothelial dystrophies

CHED - Congenital hereditary endothelial dystrophy

CI - Confidence interval

CTT - Central corneal thickness

CVS - Comma separated value

DBS - Double-strand breaks

ddH_2O – Double distilled water

DDR - DNA damage response

DGM - Darren’s G. Monckton lab

DM - Descemet’s membrane

DM1 - Myotonic dystrophy type 1

DM2 - Myotonic dystrophy type 2

DMEK - Descemet membrane endothelial keratoplasty

DNA - Deoxyribonucleic acid

ECM – Extracellular matrix

EMT- Epithelial-to-mesenchymal transition

ePAL – Estimated progenitor allele

EtBr – Ethidium Bromide

FAM - Fluorescein

FECD – Fuchs endothelial corneal dystrophy

FRAPOSA - Fast and robutst ancestry prediction by using online singular value decomposition and shrinkage adjustment

FTD/ALS - Frontotemporal dementia and amyotrophic lateral sclerosis

FXTAS - Fragile X syndrome

GATK - Genome Analysis Toolkit

GBR - British population of England and Scotland

GFP- Green Fluorescent Protein

GnomAD - Genome Aggregation Database

GUH - General University Hospital in Prague, Czech Republic

GWAS - Genome-wide association study

HD- Huntington’s disease

HEK293t - Human embryonic kidney 293 cells

HEX - Hexachloro-fluorescein

IC3D - The International Committee for Classification of Corneal Dystrophies

IGV - Integrated Genomics Viewer

KASP - Kompetitive Allele Specific PCR

LB - Luria-Bertani

MAF – Minor allele frequency

MBNL1 - Muscleblind-like 1

MBNL2 - Muscleblind-like 2

MCS - Multiple cloning sites

MEH - Moorfields Eye Hospital, United Kingdom

miRNA – microRNA

MMR- Mismatch repair

mRNA - Messenger RNA

NER - Nucleotide excision repair

NMD – Nonsense mediated decay

NTC – Negative template control

OADP - Online augmentation-decomposition-transformation

OMIM - Online Mendelian Inheritance in ManOR - Odds ratio

PC - Principal component

PCA - Principal component analysis

PCR - Polymerase chain reaction

PPCD - Posterior polymorphous corneal dystrophy

Pre-miRNA - precursor miRNAs

Pri-miRNA - primary miRNAs

PTC - Premature termination codon

R - Spearman's rank correlation coefficient

R1- Forward reads

R2 – Reverse reads

RAN - Repeat-associated non-AUG

RGT - Repeats Genotyping Tool

RNA - Ribonucleic acid

RNA-Seq – RNA sequencing

rSAP - Shrimp Alkaline Phosphatase

SAM - Sequence Alignment Map

SCA8 - Spinocerebellar ataxia type 8

SCAs - Spinocerebellar ataxias

SKAT - Sequence kernel association test

SMRT - Single-molecule real-time

SNPs - Single nucleotide polymorphisms

SNV - Single nucleotide variants

SOC media - Super Optimal Broth

SSB – Single strand breaks

STR - Short tandem repeat

TAE - Tris-Acetate-EDTA buffer

TNR - Trinucleotide repeat

TPM – Transcripts per million

TP-PCR – Triplet primed polymerase chain reaction

UCL - University College London

UCLex - UCLex consortium dataset

UPR - Unfolded protein response

UTR - Untranslated regions

UV - Ultraviolet

VUS - Variant of unknown significance

WGS – Whole genome sequencing

WT - Wild-type

XECD - X-linked endothelial corneal dystrophy

1. Introduction

1.1 Anatomy of the human eye

The human eye is a complex organ within the human body. It is a slightly asymmetric sphere with a diameter of approximately 24mm. **Figure 1** shows a schematic diagram representing the structure of the human eye. There are three distinct layers of the eyeball. The first is the outer fibrous layer which consists of the sclera, cornea and conjunctiva. The second is the middle vascular layer consisting of the choroid, ciliary body and iris. The final layer is the inner layer which includes the retina, which contains the specialised photoreceptor cells, termed cones and rods (Willoughby et al., 2010).

The human eye can alternatively be divided into two segments, the anterior segment and the posterior segment. The posterior segment comprises the back two-thirds of the eye and includes the retina and optic nerve (Ito & Walter, 2014; Willoughby et al., 2010) with the anterior segment including the lens, iris, cornea and ciliary body.

The anterior segment contains two chambers of fluid, the anterior chamber and the posterior chamber. The anterior chamber is located between the cornea and lens and the posterior chamber between the iris and the lens. Both the anterior and posterior chambers are filled with aqueous humour produced by the ciliary body. A third chamber of fluid lies in the posterior segment, called the vitreous chamber. It is much larger and located between the lens and retina. This chamber is filled with a more viscous fluid, the vitreous humour, and has the role of maintaining the shape of the eye (Willoughby et al., 2010).

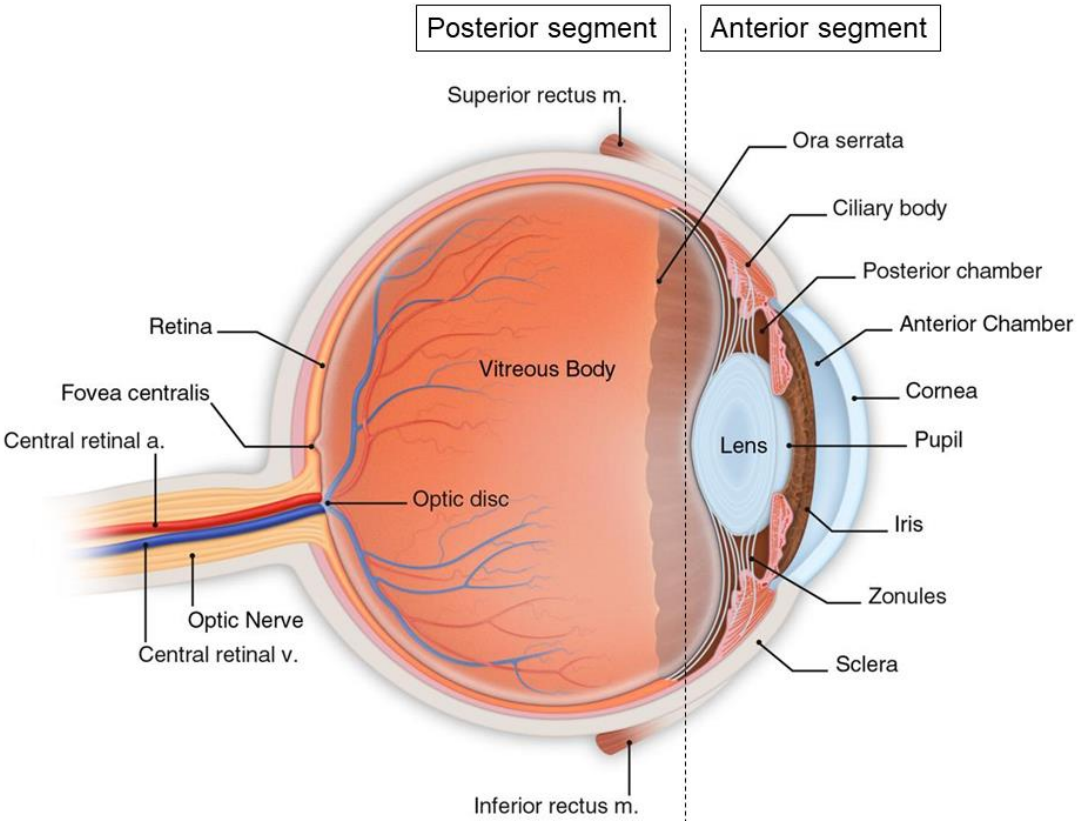


Figure 1 Schematic of the anatomy of the human eye, (adapted from a web resource (American Academy of Ophthalmology, n.d.)(https://www.aao.org/image/anatomy-color-labeled-2).

1.1.1 Cornea

The cornea, located in the outer fibrous layer, is the transparent outermost layer of the eye. There are two main functions of the cornea; the first is to act as a structural barrier, protecting the eye against infections. The cornea also functions to provide two thirds of the eye's total refractive power (DelMonte

& Kim, 2011). The cornea is composed of five distinct layers, seen in the schematic diagram in **Figure 2**.

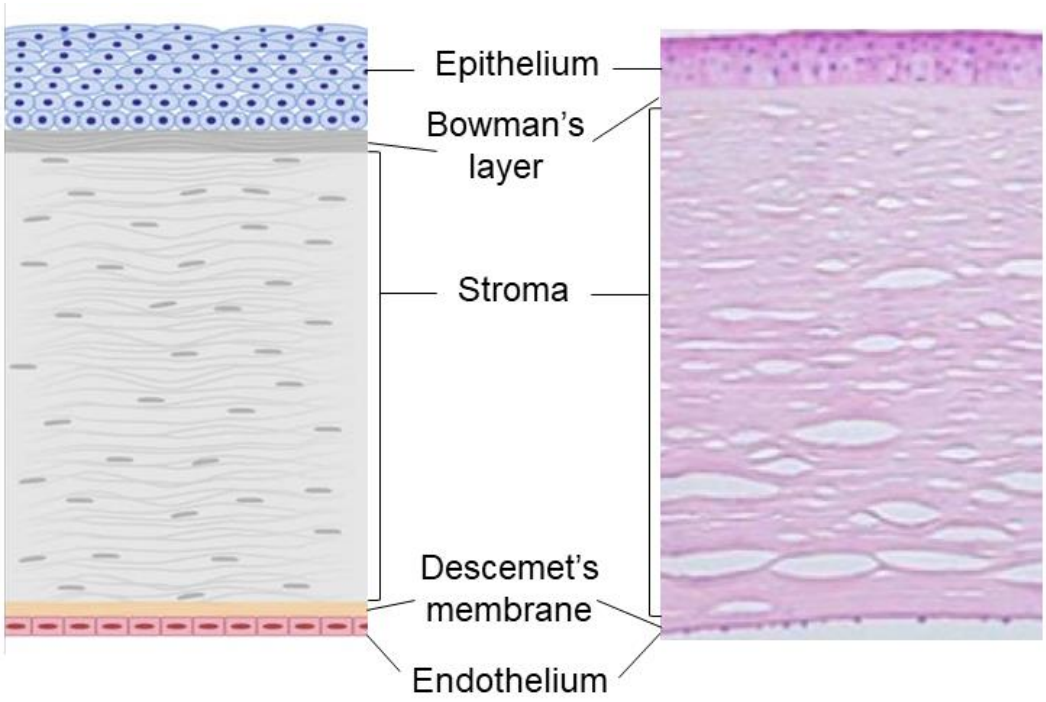


Figure 2 Labelled schematic and histology cross section showing different layers of the human cornea. Adapted from (Feizi, 2018) and web resource (Center, n.d.) (https://www.moyeseye.com/corneal-transplants).

1.1.2 Structure of cornea

The outer most layer of the cornea is the corneal epithelium. It is composed of approximately five to seven layers of epithelial cells which are constantly undergoing involution, apoptosis and desquamation. The cells form a uniformed smooth optical surface and allows for the tear film-cornea interface. The epithelium also holds a protective role to act as a barrier to chemicals, microbes and water (Sridhar, 2018).

Between the epithelial basement membrane and the stroma is the Bowman's layer. The Bowman's layer is an acellular and non-regenerating layer

composed of interwoven collagen fibres to form a strong and dense smooth sheet approximately 8-12 μm thick. The function of this layer is thought to be to provide structural integrity to the cornea, however, this remains unclear (Wilson & Hong, 2000).

The stroma lies below the Bowman's layer which accounts for about 90% of the cornea's total thickness. Keratocytes are the primary cell type present in the stroma and produce collagen, glycosaminoglycans and matrix metalloproteinases, allowing the structure of the stroma to be maintained. The stroma consists mainly of an extracellular matrix (ECM) of predominantly type I and type V collagen fibre networks forming lamellae. The main function of the stroma is to maintain the transparency of the eye and contribute to the refractive index (Eghrari, Riazuddin, & Gottsch, 2015).

The Descemet's membrane (DM) of the corneal endothelium lines the posterior surface of the corneal stroma. The DM is comprised of two distinct layers, the anterior banded zone and the posterior non-banded zone. The anterior banded zone is approximately 30 μm thick and laid down during foetal development. The layer consists of prominently type IV collagen, type VIII collagen α1 and α2, laminin and fibronectin to form an ECM described as wide spaced collagen. This widespread collagen is arranged in an array of broad nodes which reaches full thickness by birth and remains unchanged thereafter. The posterior non-banded zone is produced by endothelial cells and thickens over time to form a broad layer of amorphous ECM (Desronvil et al., 2010; Eghrari et al., 2015; Levy, Moss, Sawada, Dopping-Hepenstal, & McCartney, 1996).

1.1.3 Corneal endothelium

The innermost layer of the cornea is the endothelium. The corneal endothelium is a monolayer of cells lining the posterior surface of the cornea (Willoughby et al., 2010). The cells are polygonal and organised into a honeycomb-like mosaic, seen in **Figure 3.A** (Eghrari et al., 2015). The human cornea begins developing at around five to six weeks of gestation. The central part of the cornea, including the endothelium is derived from neural crest cells while the corneal epithelium from epidermal ectoderm (Zavala, López Jaime, Rodríguez Barrientos, & Valdez-Garcia, 2013). As the monolayer of endothelial cells develop, they begin to flatten which allows the formation of tight junctions between adjacent cells. These cells then arrest in G1 phase of mitosis and thus are unable to regenerate via mitosis. Several aspects contribute to maintain the endothelium in a non-replicative state including cell-cell contact inhibition, the activity of p27kip1, a known G1-phase inhibitor and growth factor TGF-ß preventing entry into the S-phase of the cell cycle (Joyce, 2003).

At birth, endothelium density is at its peak with approximately 6000 cells/mm^2 and gradually decreases during infancy as the eye grows. During adulthood cell density decreases to approximately 2500 cells/mm^2, with an annual reduction of approximately 0.6% (Van den Bogerd, Dhubhghaill, Koppen, Tassignon, & Zakaria, 2018). As a consequence of the overall cell density decreasing, the endothelial cells compensate by increasing in size to maintain the endothelial barrier function (Zavala et al., 2013).

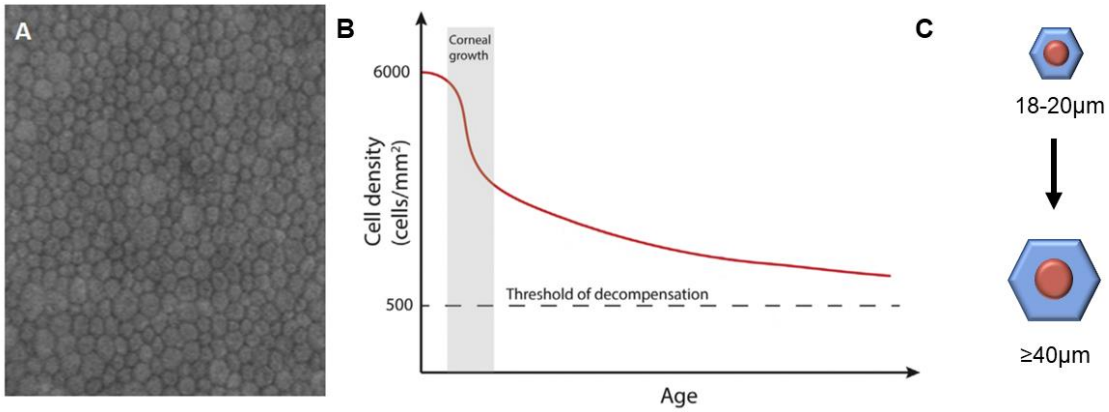


Figure 3 Characteristic of corneal endothelial cells and how they change throughout life. (A) A confocal microscopy image displaying the honeycomb arrangement of the polygonal cells comprising the endothelial cell layer in a healthy individual (Eghrari, Riazuddin and Gottsch, 2015)**. (B)** A schematic diagram showing the progressive loss of endothelial cell density with age and **(C)** enlargement in size of the endothelial cells that occurs as part of the normal aging process in humans (adapted from Van den Bogerd, Dhubhghaill, Koppen, Tassignon, & Zakaria, 2018).

1.1.4 The function of corneal endothelial cells

Hexagonal endothelial cells form tight junctions creating a barrier enabling the endothelial layer to act as a 'leaky pump' via bicarbonate-dependent ATPase pumps. This enables the maintenance of nutrients and optimal hydration levels in the stroma and other more anterior corneal layers (Bonanno, 2012).

The stroma has an imbibition pressure of 60 mmHg, produced by the proteoglycan matrix surrounding the collagen fibres of the matrix, however, stromal hydration must be maintained at 78% water for transparency to be maintained (Geroski, Matsuda, Yee, & Edelhauser, 1985). The endothelial cell layer actively transports ions, using adenosine triphosphate (ATP), to passively move water from the stroma across the DM. Bicarbonate is required in the DM in order to maintain the function of the endothelial pump (Tuft & Coster, 1990).

The density of endothelial cells decreases to approximately 2500 cells/mm^2 in late adulthood, as mentioned earlier (**Figure 3.B**). For adequate function of the pump a minimum number of cells are required and if numbers fall below this threshold, corneal oedema can occur (Eghrari et al., 2015).

1.2 Corneal dystrophies

Corneal dystrophies (CDs) refer to a group of genetically heterogeneous disorders of the cornea. Typically, CDs are bilateral, symmetric, slowly progressive, and not related to environmental or systemic factors. The age of onset, clinical appearance and effect on corneal transparency vary vastly depending on disease. In the majority of cases, the inheritance pattern of CDs is autosomal dominant, but they can also be recessive or X-linked. Causative variants in genes have been identified for several CDs which has aided the understanding of disease pathogenesis. However, in many cases the genetic cause remains unknown. Future discoveries may lead to further revisions of disease classifications but currently CDs are classified into four subcategories based on clinical, pathologic and genetic information by The International Committee for Classification of Corneal Dystrophies (IC3D) (Lin, Chen, & Cui, 2016; Lisch & Weiss, 2019). These categories are as follows: a) epithelial and subepithelial corneal dystrophies, b) epithelial-stromal TGFBI corneal dystrophies, c) stromal corneal dystrophies and d) endothelial corneal dystrophies. **Table 1** shows the current classification of CDs (Lisch & Weiss, 2019) .

Table 1 Current international classification of corneal dystrophies (Lisch and Weiss, 2019).

Disease name	OMIM	Associated genes
Epithelial and subepithelial corneal dystrophies		
Epithelial basement membrane dystrophy (EBMD)	#121820	*TGFBI* (OMIM #601692)
Epithelial recurrent erosion dystrophy (ERED)	#122400	*COL17A1* (OMIM #113811)
Meesmann corneal dystrophy (MECD)	#122100	*KRT3* (OMIM #148043), *KRT12* (OMIM #601687)
Lisch epithelial corneal dystrophy (LECD)	#300778	Unknown; mapped to chromosome Xp22.3
Gelatinous drop-like corneal dystrophy (GDLD)	#204870	*TACSTD2* (OMIM #137290)
Epithelial-stromal TGFBI corneal dystrophies		
Reis-Bücklers corneal dystrophy (RBCD)	#608470	*TGFBI* (OMIM #601692)
Thiel-Behnke corneal dystrophy (TBCD)	#602082	*TGFBI* (OMIM #601692)
Lattice corneal dystrophy, type 1 (LCD 1) and variants	#122200	*TGFBI* (OMIM #601692)
Granular corneal dystrophy, type 1 (GCD 1)	#121900	*TGFBI* (OMIM #601692)
Granular corneal dystrophy, type 2 (GCD 2)	#607541	*TGFBI* (OMIM #601692)
Stromal corneal dystrophies		
Macular corneal dystrophy (MCD)	#217800	*CHST6* (OMIM #605294)
Schnyder corneal dystrophy (SCD)	#121800	*UBIAD1* (OMIM #611632
Congenital stromal corneal dystrophy (CSCD)	#610048	*DCN* (OMIM #125255)
Fleck corneal dystrophy (FCD)	#121850	*PIKFYVE* (OMIM #609414)
Posterior amorphous corneal dystrophy (PACD)/ Cornea plana type 1 (CNA1)	#612868 #121400	Unknown; mapped to chromosome 12q21.33
Cornea plana type 2 (CNA2)	#217300	*KERA* #603288
Brittle cornea syndrome type 1 (BCS type 1)	#229200	*ZNF469* #612078
Brittle cornea syndrome type 2 (BCS type 2)	#614170	*PRDM5* #614161
X-linked megalocornea	#309300	*CHRDL1* #300350
Central cloudy dystrophy of FRANÇOIS (CCDF)	#217600	Unknown
Pre-Descemet corneal dystrophy (PDCD)	N/A	Unknown
Crystalline Pre-Descemet corneal dystrophy (CPDCD)	N/A	Unknown
Endothelial corneal dystrophies		
Fuchs endothelial corneal dystrophy (FECD)	#613267	*TCF4* (OMIM # 602272)
Posterior polymorphous corneal dystrophy 1 (PPCD1)	#122000	*OVOL2* (OMIM #616441)
Posterior polymorphous corneal dystrophy 3 (PPCD3)	#609141	*ZEB1 (OMIM #189909)*
Posterior polymorphous corneal dystrophy 4 (PPCD4)	#618031	*GRHL2 (OMIM #608576)*
Congenital hereditary endothelial dystrophy (CHED)	#217700	*SLC4A11* (OMIM #610206)
X-linked endothelial corneal dystrophy (XECD)	#300779	Unknown; mapped to chromosome Xq252

1.2.1 Corneal endothelial dystrophies (CEDs)

Diseases within this subgroup affect the corneal endothelium and are genetically and phenotypically heterogeneous, include Fuchs endothelial corneal dystrophy (FECD), congenital hereditary endothelial dystrophy (CHED), posterior polymorphous corneal dystrophy (PPCD) and X-linked endothelial corneal dystrophy (XECD) (Lisch & Weiss, 2019).

1.2.2 Fuchs Endothelial Corneal Dystrophy

FECD is the most common CED affecting approximately 4-5% of the population over 40 years of age in the United States and Europe (Fautsch et al., 2021; Zarouchlioti et al., 2018). The disease is a progressive, age-related disorder where symptoms typically manifest during the fifth and sixth decade of life (Friedenwald & Friedenwald, 1925). FECD primarily affects the posterior layers of the cornea and is clinically characterised by the accelerated loss of endothelial cells, progressive thickening of DM resulting in the formation of focal excrescences termed ‘guttae’ (Friedenwald & Friedenwald, 1925; Vedana, Villarreal, & Jun, 2016) (**Figure 4.C**). With advanced disease progression, these features compromise the barrier function of the corneal endothelium leading to corneal swelling, painful epithelial bullae, and progressive corneal clouding resulting in loss of vision (**Figure 4.A**) (Agoldberg, Raza, Walford, Feuer, & Lgoldberg, 2014).

During early stages of the disease symptoms may be treated with topical hypertonic saline to draw excess fluid from the stroma. However, surgical intervention and corneal endothelium transplant is currently the only treatment to restore vision in those with advance stage FECD (Woo, Ang, Htoon, & Tan, 2019).

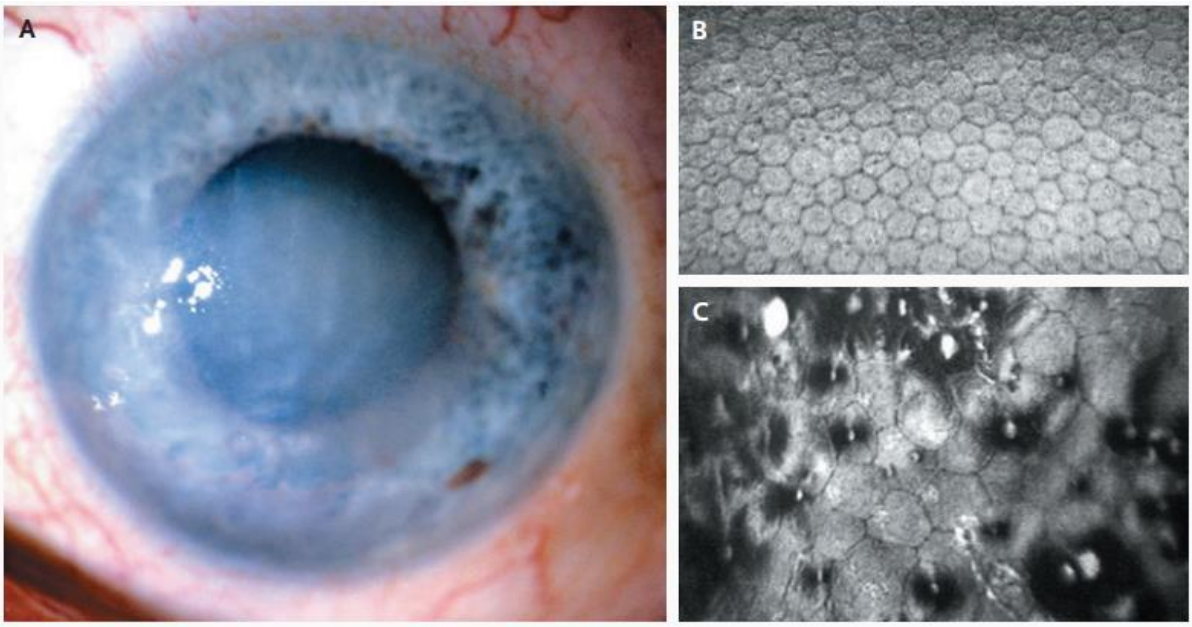


Figure 4 Clinical charateristics of Fuchs Endothelial Dystrophy (FECD) (A) A photograph of a patient with advance stage FECD showing severe corneal oedema. **(B)** A confocal photomicrograph of the endothelium from a healthy individual showing the morphology of a monolayer of densely packed hexagonal shaped cells. **(C)** A confocal photomicrograph of the endothelium of an individual with FECD, at the same magnification, showing larger cells to compensate for the loss of cells with dark patches showing guttae (Baratz et al., 2010).

Originally, FECD was thought to be a genetically complex trait (Doggart, 1957), however, it was proposed to have an autosomal dominant mode of transmission in the early 1970s (Cross, Maumenee, & Cantolino, 1971). This was further supported in the landmark study conducted in 1978 by Krachmer *et al*. confirming that the disease was consistent with an autosomal dominant trait with variable penetrance and expression in 64 families (Krachmer, Purcell, Young, & Bucher, 1978) and a subsequent study including pedigrees comprised of up to four generations (Magovern, Beauchamp, McTigue, Fine, & Baumiller, 1979). Furthermore, the higher prevalence of FECD in females than males was also established, with 46% of probands affected relatives being female and only

19% being male (Afshari, Pittard, Siddiqui, & Klintworth, 2006; Krachmer et al., 1978).

Pedigree-based linkage analysis conducted in the early 2000s identified several genes suggesting a multigenetic phenotype for FECD (Iliff, Riazuddin, & Gottsch, 2012). These include autosomal dominant missense mutations in the *COL8A2* gene, associated with rare early-onset FECD phenotype (Biswas, 2001), (Gottsch et al., 2005). Autosomal dominant missense mutations in *ZEB1* (Mehta et al., 2008; Riazuddin et al., 2010) and heterozygous missense mutations in *SLC4A11* have also been associated with the typical late-onset FECD (Vithana et al., 2008). Mutations in these genes are discussed further in section 4.1.

A genome-wide association study (GWAS) conducted in 2010 revealed an overwhelming significant association within a region spanning the *TCF4* gene located on chromosome 18q21.2. The most highly associated single nucleotide polymorphism (SNP), rs613872, held an odds ratio (OR) of 5.5 for carrying one copy of the risk allele (heterozygous GT) and an OR of 30 for individuals carrying two copies (homozygous GG) (Baratz et al., 2010).

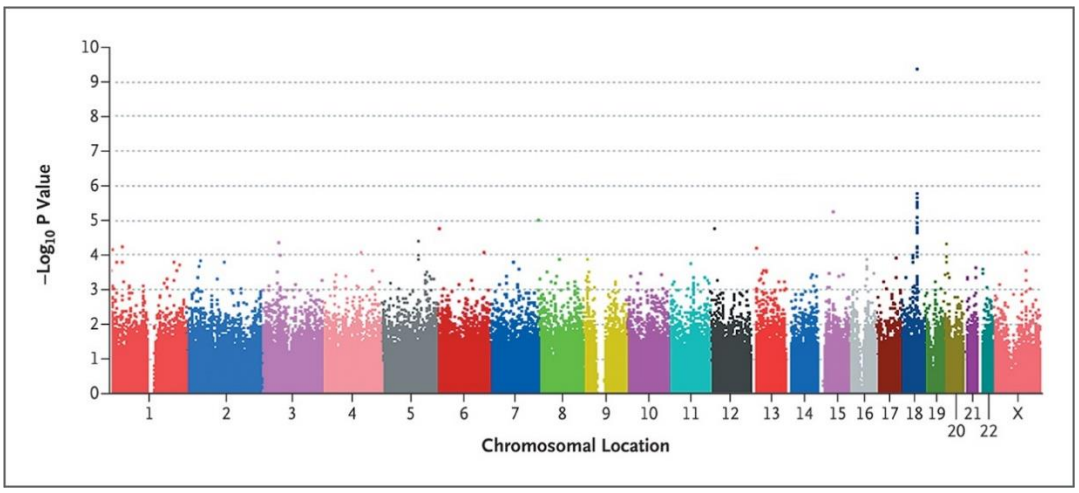


Figure 5 Manhattan plot showing genome-wide significance at a region on chromosome 18, spanning the locus encoding transcription factor 4 (*TCF4*) for association between 338,727 Single-Nucleotide Polymorphisms (SNPs) and Fuchs's endothelial corneal dystrophy (FECD) (Baratz et al., 2010).

TCF4 is a member of the basic-loop-helix (bHLH) family of transcription factors. The bHLH region of TCF4 is highly conserved and serves as an interface for DNA and other proteins, the genomic organisation of TCF4 is shown in **Figure 6**. The bHLH region along with a conserved C domain on the C-terminal end of the bHLH domain enables homo- and heterodimerisation with other transcription factors or transcription modifiers. TCF4 has the potential to dimerise with its own isoforms and with other members of the class II, V, VI bHLH family of transcription factors, enabling transcriptional regulation of a number of genes. Furthermore, depending on the isoform, TCF4 contains several additional protein domains, including the transactivation domains (AD1 and AD2), which enables binding of additional regulatory molecules which can further modify functions of TCF4 isoforms. Due to alternative splicing, the most characterised isoforms are TCF4-A and TCF4-B. TCF4-B isoform contains the full complement of activation domains, nuclear localization and export signals, the TCF4-A isoform lacks AD1 and the nuclear localization signal found in the

N-terminus (Fautsch et al., 2021; Forrest, Hill, Quantock, Martin-Rendon, & Blake, 2014).

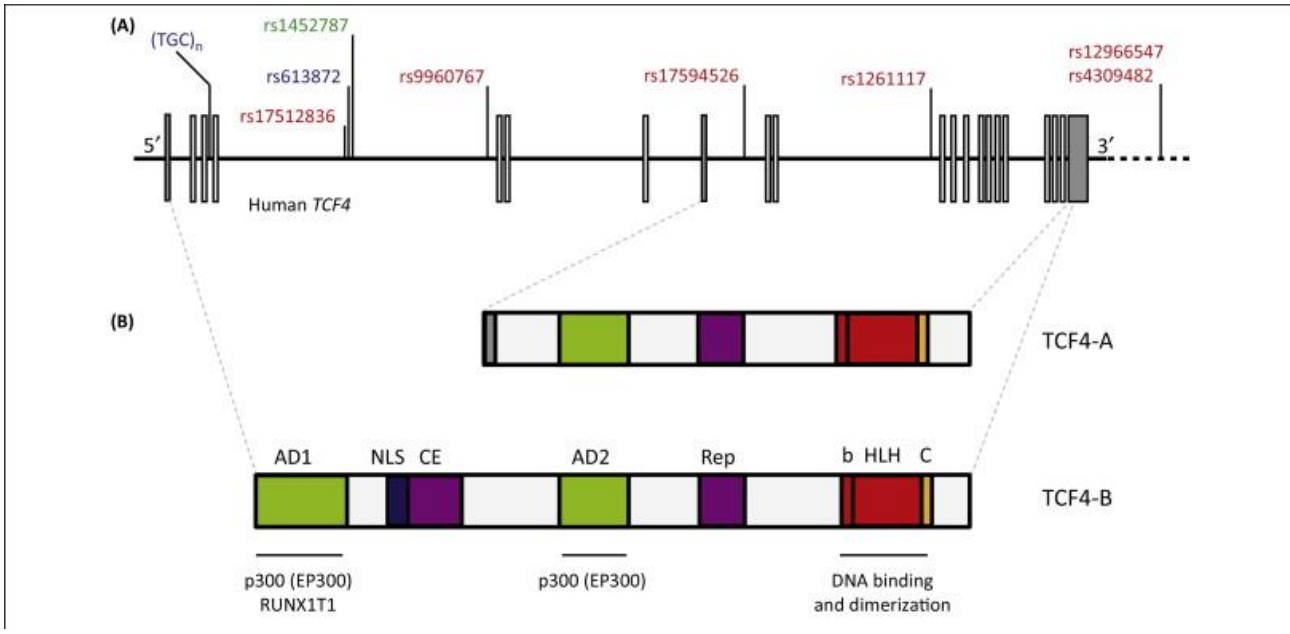


Figure 6 Schematic diagram showing the genomic organisation of the TCF4 gene and the location of common variants, including rs613872 associated with Fuchs Endothelial Corneal Dystrophy. (Forrest et al., 2014)

In 2012, Wieben *et al.* discovered an unstable non-coding CTG triplet repeat, termed CTG18.1, that was in linkage disequilibrium with the SNP rs613872 (Wieben et al., 2012). The CTG18.1 was first identified in 1997, as a candidate genetic marker for bipolar disorder, which revealed the CTG18.1 allele was stable with repeats lengths between 10-37, and very large unstable expansions between up to 2100 repeats. The study concluded that the unstable expansion was not associated with bipolar disorder but revealed the frequency of moderately expanded alleles is approximately 3% in populations of Northern European ancestry (Breschel et al., 1997).

Wieben and colleagues confirmed this finding, identifying 3% of their control group carried an expanded CTG18.1 allele with this range of repeat.

Strikingly they also found that 79% of FECD cases carried at least one allele with an expansion of over 50 repeats (Wieben et al., 2012).

Since Wieben and colleague's original discovery, similar findings have been replicated in many other Caucasian cohorts (Luther et al., 2016; Mootha, Gong, Ku, & Xing, 2014; Skorodumova et al., 2018; Zarouchlioti et al., 2018). While the CTG18.1 expansion is found in FECD in other ethnicities it has been reported to be much less abundant in Asian populations. For example, the occurrence in a Thai population was 39%, 44% in Chinese, 26% in Japanese and 34% in Indian (Nakano et al., 2015; Nanda, Padhy, Samal, Das, & Alone, 2014; Okumura, Puangsricharern, et al., 2019; Xing et al., 2014). This was also found in a study conducted with African Americans with FECD showing only 35% carried an expanded CTG18.1 in comparison to 62.5% Caucasians (Eghrari, Vahedi, Afshari, Riazuddin, & Gottsch, 2017).

In addition to this, ExpansionHunter has been applied to publicly available whole genome sequencing (WGS) data from gnomAD to gain further insights into the frequency of expanded CTG18.1 alleles in the general population (Karczewski et al., 2020). **Table 2** summarises the frequencies of CTG18.1 alleles with ≥50 repeats across different populations. The European (non-Finnish) population had the highest frequency of expanded CTG18.1 alleles with ≥50 repeats in the general population at 7%, with all other populations have a frequency of 4% or less (unpublished data). This data was likely to overestimate the frequency of expanded CTG18.1 alleles as Expansion Hunter was run using the default settings. For a more accurate representation of the expanded CTG18.1 allele in these populations, bespoke validation optimised to the CTG18.1 is needed. Nevertheless, this data supports previous

findings that expanded CTG18.1 alleles are more common in Caucasian populations.

Table 2 Summary of the population frequencies of expanded CTG18.1 alleles with ≥50 repeats in whole genome samples available in gnomAD. Repeat lengths were determined using ExpansionHunter using default settings (unpublished data).

gnomAD population	Number of alleles with CTG18.1 ≥50	Total number of alleles	Frequency of alleles with CTG18.1 ≥50
African/African American	171	6917	0.02
Amish	1	52	0.02
Ashkenazi Jewish	16	394	0.04
East Asian	23	728	0.03
European (Finnish)	88	2105	0.04
European (non-Finnish)	535	7487	0.07
Latino/Admixed American	27	639	0.04
Middle Eastern	2	126	0.02
Other	7	116	0.06
South Asian	25	677	0.04

1.2.3 Genotyping the CTG18.1 repeat

There are currently over 40 disorders caused by expansions of microsatellites, simple sequence repeats, similar to the *TCF4* CTG18.1-expansion positive FECD (Chintalaphani, Pineda, Deveson, & Kumar, 2021). For many repeat associated diseases, directed PCR-based screening, such as a short tandem repeat (STR) assay or triplet-primed polymerase chain reaction (TP-PCR), are a straightforward, sensitive, specific, and inexpensive approach to detect an expanded repeat allele (Mootha et al., 2014; Paulson, 2018; Wieben et al., 2012). These methodologies show FECD patients typically harbour heterozygous alleles with an expansion of 50-200 repeat units in their whole blood-derived genomic deoxyribonucleic acid (DNA). For larger expansions, which cannot be detected by STR or sized by TP-PCR, southern blot hybridisation can be utilised and has shown patients can harbour expansions estimated to be up to several thousand repeat units (Okumura, Puangsricharern, et al., 2019; Soliman, Xing, Radwan, Gong, & Mootha, 2015; Wieben et al., 2012). Although these methods provide a simple and inexpensive method of estimating repeat lengths, they have the inability to accurately determine repeat size and fail to provide sequence level resolution.

Recently new methodologies have been developed enabling sequence level resolution to precisely and accurately quantify somatic repeat length variants, whilst also providing information about genetic variants within and around the repeat (Ciosi et al., 2021). Ciosi et al. applied bulk-PCR sequencing approaches using Illumina MiSeq and PacBio long-read single-molecule real-time (SMRT) sequencing to sequence the *HTT* repeat expansion causing Huntington's disease (HD) (Ciosi et al., 2021). The study genotyped the HTT CAG repeat and quantified somatic mosaicism but both methods came with

their own limitations. The first being that they are PCR-dependent and therefore introduce a higher frequency of sequencing errors due to PCR slippage artefacts that likely reduces alignment efficiency (Ciosi et al., 2021). Additionally, for the MiSeq, there is a limit to the repeat length able to be detected.

Furthermore, novel amplification-free sequencing methods, utilising a CRISPR-Cas9 system in combination with long-read (SMRT) sequencing has recently allowed us to study the *TCF4* repeat element at a nucleotide level (Hafford-Tear et al., 2019; Wieben, Aleff, et al., 2019). This method has demonstrated the CTG18.1 expansion to be dynamic showing striking levels of repeat length instability and mosaicism in each individual. These studies also highlight that size estimates provided by conventional genotyping assays (e.g., STR and Southern blot) do not provide a robust representation of the dynamic nature of this repeat element in its expanded state (Hafford-Tear et al., 2019; Wieben, Aleff, et al., 2019). This has been demonstrated in myotonic dystrophy type 1 (DM1), where the repeat length is found to be substantially larger in the affected tissue (skeletal muscle) than in blood (Ashizawa, Dubel, & Harati, 1993). Likewise, long-read sequencing and southern blotting has recently demonstrated total ribonucleic acid (RNA) isolated from FECD corneal endothelium showed significantly larger CTG18.1 expansions (>1000 repeats) compared to those characterised in leukocytes from the same individuals (<90 repeats). However in this study they were unable to completely span the CAG repeats in its entirety, and therefore suggest the full extent of the sizes of these expansions are likely to be even longer (Wieben et al., 2021). This finding could indicate why patients with CTG18.1-expanded FECD only develop disease in the corneal endothelium, when TCF4 is ubiquitously expressed.

1.2.4 Phenotype-genotype correlation

For almost all repeat mediated diseases, the expansion in the repeat is thought to have arisen from polymorphisms in the typical repeat length. In general, the larger the repeat length, the more unstable the genomic region becomes and thus resulting in longer repeats. This initial expansion of the repeat is described as a pre-mutation allele, which has been shown to have an increased propensity to further expand upon transmission into the pathogenic range (Paulson, 2018). For example, in fragile X syndrome (FXTAS), a non-coding CGG repeat found in the *FMR1* gene typically has a repeat length between 6-50 units, the pre-mutation length is between 55-200 repeat units where a carrier does not experience phenotypic symptoms but is at risk of the repeat expanding to a full mutation. After transmission, a repeat length of more than 200 repeats results in the FXTAS phenotype in offspring (Mirkin, 2007). The pre-mutation range where the CTG18.1 repeat becomes unstable allowing progression to the full mutation of over 50 repeat units potentially causing FECD, has yet to be determined.

A striking genotype-phenotype correlation has been established between the repeat length and age of symptom onset for many repeat-mediated diseases. This has been established particularly well in CAG/polyglutamine diseases, mainly HD but also bulbospinal neuronopathy, spinocerebellar ataxia type 2 and type 7 (Andrew et al., 1993; Doyu et al., 1992; Figueroa et al., 2017; Johansson et al., 1998). This inverse correlation has also been observed in DM1 ($(CTG)_n$), Frederic ataxia ($(GAA)_n$) and FXTAS ($(CGG)_n$) (Paulson, 2018). However, in more complex repeat mediated diseases such as myotonic dystrophy type 2 (DM2) ($(CCTG)_n$), *C9orf72*-mediated ALS/FTD ($(GGGGCC)_n$), there is very weak evidence of such correlation suggesting unidentified

molecular features may influence the behaviour of these repeats (Paulson, 2018). Whether or not the CTG18.1 repeat in FECD follows this similar genotype-phenotype pattern is yet to be robustly determined (Fautsch et al., 2021).

The severity of repeat mediated diseases often worsens progressively through successive generations, via a process termed anticipation. This phenomenon has been described with other repeat expansion disorders including DM1, HD and FXTAS (Harper, Harley, Reardon, & Shaw, 1992; Ranen et al., 1995; Sutherland et al., 1991). There has been some evidence to suggest the instability of the CTG18.1 expansion increases through parent-child transmissions, however, further work using larger numbers of families need to done before a conclusion can be drawn (Greiner, Terveen, Vislisel, Roos, & Fingert, 2017; Saade, Xing, Gong, Zhou, & Mootha, 2018).

Instability of repeat length within the germline has also been demonstrated to be influenced by parent-of-origin transmission for some specific loci. For example, in some non-coding repeats disorders, such as FXTAS, are almost exclusively expanded from pre-mutation allele length to full mutation through maternal transmission, whereas normal and intermediate length alleles are more likely to expand into pre-mutation during paternal transmission (McMurray, 2010). Paternal transmission of fully expanded disease alleles have been found to result in either no change or contraction of the repeat tract (McMurray, 2010). The expansion is likely to occur in the maternal oocytes, while arrested in meiotic prophase. In female carriers with alleles with a pre-mutation repeat length, expansion of these alleles are already seen to be present in seven-cell pre-implantation embryos (Rifé et al., 2004). As developing oocytes arrest during the first meiotic division, this suggests that the

repeat length alterations occur in quiescent cells before transmission and therefore do not require replication but instead expansion occurs as a result of a repair dependent mechanism (McMurray, 2010). These repair dependent mechanisms include double-strand breaks (DBS), base excision repair (BER) and nucleotide excision repair (NER). DBS occur in quiescent human oocytes during meiosis and occur at the same time expansions arise. During the process of removing chemically damaged bases via the BER process, single-strand breaks (SSB) are generated. It has been shown in mouse models of FXS expansions of the CGC repeat occur when treated with a powerful oxidant, suggesting expansion arise during the removal of oxidised DNA bases. NER has also been implemented in TNR instability in a similar process by which expansions lengthen during the repair process. Repeated rounds of oxidations-repair-expansion leads to the progressive expansion of TNRs in meiotically arrested cells and it can be suggested a similar phenomenon will occur with the *TCF4* CTG18.1 expansion if cells are exposed to oxidative stress (McMurray, 2010).

Males who inherit a full mutation allele from their mother harbour the expanded allele in their somatic cells but do not transmit the expanded allele to their progeny as large repeat tracts in their spermatogonia are shortened around weeks 13 to 17 of foetal development (Malter et al., 1997). Similar patterns of transmission has also been observed in patients with DM1, where the size of the DMPK repeat was significantly greater when transmitted from a female parent to daughter in comparison from when it is transmitted from the male parent to daughter or son (Dean, Tan, & Ao, 2006; Han, Jang, & Park, 2022; McMurray, 2010).

1.2.5 Pathophysiology of CTG18.1 expansion-mediated FECD

The underlying mechanism of how the CTG18.1 expansion results in FECD disease phenotype is not yet completely understood, however, to date, four possible, non-mutually exclusive, mechanisms have been proposed. **Figure 7** illustrates the four proposed mechanisms, dysregulation of TCF4 protein expression, RNA-mediated toxicity, repeat-associated non-ATG (RAN) translation and age and tissue-dependent somatic instability of the repeat element.

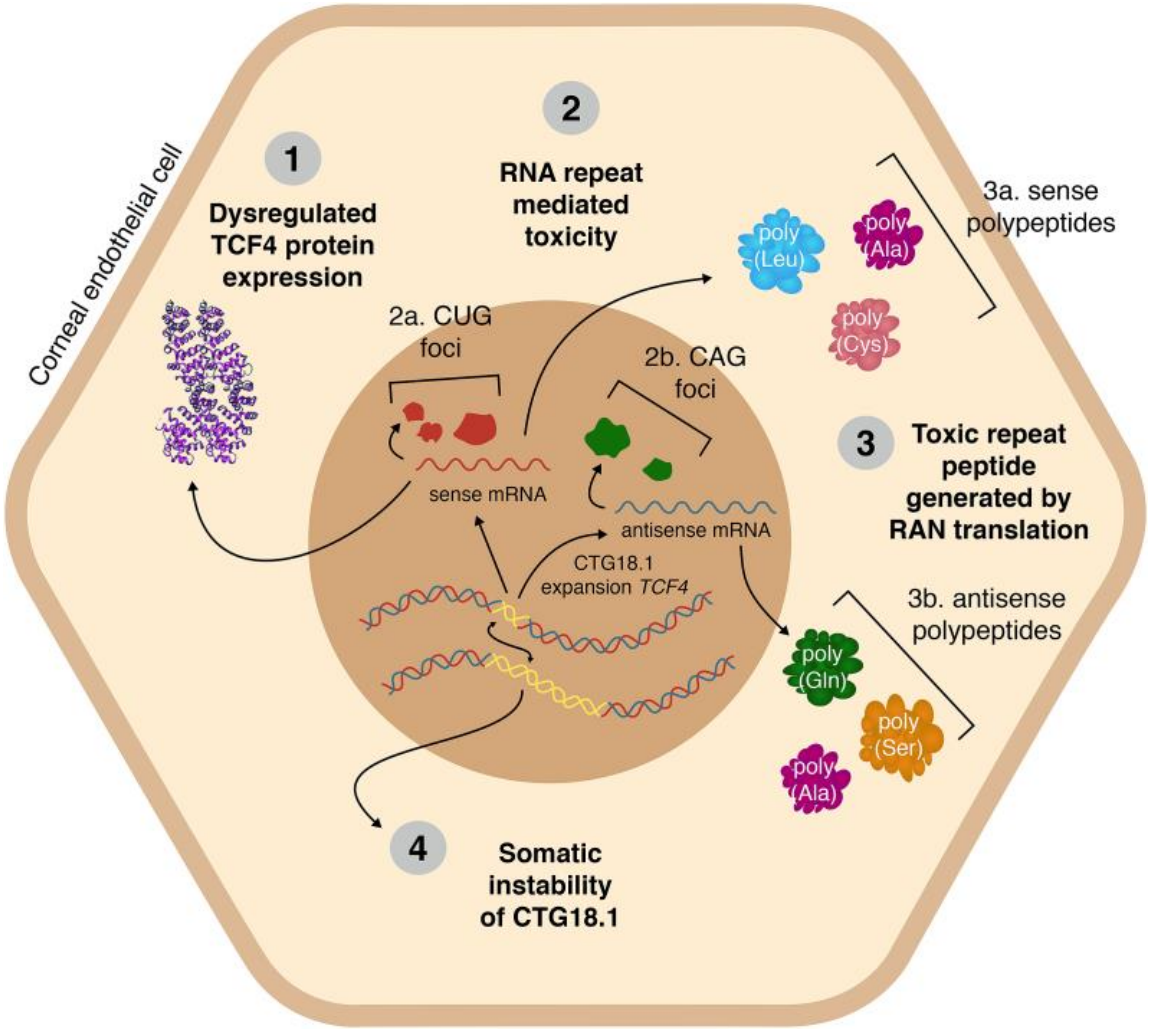


Figure 7 Potential mechanisms of the pathophysiology associated with the CTG18.1 expansion in Fuchs endothelial corneal dystrophy. Four non-mutually exclusive mechanisms have been proposed to drive and/or exacerbate the onset of CTG18.1 expansion-mediated FECD, including **(1)** dysregulated expression of TCF4 transcripts, **(2)** accumulation of toxic (a) sense (CUG)n and (b) antisense-derived (CAG)n repetitive RNA transcripts, **(3)** RAN translation of repetitive RNA transcripts, and **(4)** age and tissue-dependant somatic instability of the repeat element. This figure represents data I illustrated for a collaborative review article (Fautsch et al., 2021).

1.2.6 Dysregulation expression of *TCF4*

Non-coding triplet repeat expansions can influence transcription of the gene it is located within and surrounding genes. For example, Frontotemporal dementia and amyotrophic lateral sclerosis (FTD/ALS) patients harbouring large GGGGCC repeat lengths in *C9orf72* had reduced C9orf72 expression levels, as a result of epigenetic changes such as histone trimethylation, in comparisons to FTD/ALS patients without a *C9orf72* repeat expansion (Belzil et al., 2013). Elucidation of the effect of a given trinucleotide repeat (TNR) expansion on the expression of the gene it is located within is important to fully understand the disease-specific pathophysiology; however, the effect of CTG18.1 expansion on the transcription of *TCF4* is not well characterised.

In an attempt to examine if expression levels of TCF4 are altered in CTG18.1 expansion-mediated FECD, several groups have produced contradicting results. Okumura *et al.* showed that *TCF4* messenger RNA (mRNA) is upregulated in the corneal endothelium of FECD patients, regardless of the presence of an expanded CTG18.1 allele, using a quantitative PCR based approach. They also noted a positive correlation between TCF4 expression level and the length of the TNR repeat (Okumura, Hayashi, Nakano, Yoshii, et al., 2019). Interestingly, Foja *et al*. found the opposite when using a TaqMan probe complementary to exons near to the CTG18.1 expansion region. They demonstrated that FECD patients, with an expanded CTG18.1 allele, showed a reduction in TCF4 expression levels in comparison to healthy controls in corneal endothelial explants (Foja, Luther, Hoffmann, Rupprecht, & Gruenauer-Kloevekorn, 2017). These contradictory data may be attributed to the complex nature of *TCF4* transcription, where over 90 different transcripts

and 64 protein coding isoforms are produced, including multiple isoforms identified within the corneal endothelium (Eghrari et al., 2018). The specificity of PCR primers used to detect *TFC4* isoforms may be ineffective at detecting subtle isoform-specific dysregulation events and potentially reduce any TCF4 TNR expansion specific signals and thus not give an accurate representation of expression levels.

Furthermore, analysis of RNA sequencing (RNA-seq) data, obtained from FECD and control corneal endothelium samples, established intron retention downstream of the expanded CTG18.1 repeat in FECD samples, but not in unaffected controls without an expanded allele (Sznajder et al., 2018). Intron retention has been hypothesised to lead to the introduction of a premature termination codon (PTC) and consequently nonsense-mediated decay (NMD) of specific transcripts, thus isoform-specific downregulation of TCF4 is hypothesised to contribute to the underlying disease mechanism (Sznajder et al., 2018). Conversely, Weiben *at al.* recently demonstrated this mechanism could not explain disease alone as intron retention also occurred in those carrying the CTG18.1 expansion but without phenotypic FECD, suggesting intron retention may be a reliable marker for identifying the presence of a TNR expansion, but is not a reliable marker for FECD status (Sirp et al., 2020; Wieben, Baratz, et al., 2019).

Another speculative model of FECD pathogenesis related to dysregulation of TCF4 is the impact this could have on epithelial-to-mesenchymal transition (EMT) mechanism, in which *TCF4* and *ZEB1* both have important roles (Foja et al., 2017; A. F. Wright & Dhillon, 2010). ZEB1 initiates the EMT process resulting in nuclear translocation of β-catenin and increased β-catenin/TCF4 transcriptional activity (H. T. Wu et al., 2020). In Addition to the

reduced TCF4 levels, Foja *et al*., also identified a reduction in *ZEB1* gene expression in CTG18.1 mediated FECD patients. Dysregulated EMT resulting in aberrant migration of endothelial cells could also be a plausible explanation for the appearance of central cornea guttae as an early hallmark of FECD (Foja et al., 2017; A. F. Wright & Dhillon, 2010). Notably, haploinsufficiency of *ZEB1* and mutations in other EMT regulators, *OVOL2* and *GRHL2*, which result in an altered EMT pathway cause PPCD, another primary CED, further suggesting that EMT dysregulation can underlie corneal endothelial disease (Davidson et al., 2016; Petra Liskova et al., 2018, 2016).

1.2.7 *TCF4* Repeat expansion-mediated RNA toxicity

The intronic location of the *TCF4* $(CTG{\cdot}CAG)_n$ repeat expansion led researchers to hypothesise that RNA toxicity may have a role in the pathogenesis of disease, much like it does for several other repeat expansion diseases such as DM1, DM2, *C9orf72* ALS/FTD and FXTAS (Y. B. Lee et al., 2013; Mankodi, 2001; Tassone, Iwahashi, & Hagerman, 2004). The repetitive elements associated with these diseases are transcribed into toxic gain-of-function RNAs. These RNA transcripts accumulate and form nuclear RNA foci, first described in DM1 (Taneja, McCurrach, Schalling, Housman, & Singer, 1995) and later observed for *C9orf72* ALS/FTD, FXTAS and others (DeJesus-Hernandez et al., 2011; Tassone et al., 2004; Wojciechowska & Krzyzosiak, 2011). Splicing regulators, such as muscleblind-like 1 (MBNL1), are sequestered and co-localised to these RNA foci resulting in downstream disruption of normal mRNA processes (Mankodi, 2001).

It is now well established that *TCF4* sense-derived $(CUG)_n$ transcripts accumulate as nuclear RNA foci within the tissue derived from individuals affected with CTG18.1-mediated FECD (Du et al., 2015; Hu et al., 2018;

Zarouchlioti et al., 2018). Zarouchlioti *et al*. have identified these RNA foci in a tissue-specific manner, only being identified in the cells of tissue derived from the corneal endothelium but absent in patient matched fibroblasts lines, suggesting the tissue-specific nature of FECD (Zarouchlioti et al., 2018). Conversely, Du *et al*. observed RNA foci in two of their patient matched fibroblasts but absent in the third where RNA foci were only present in the corneal endothelial tissue. They concluded that the patient fibroblasts absent for the RNA foci had a notably shorter repeat length suggesting that the length of the CTG18.1 repeat plays an important role in formation of CUG RNA foci in fibroblasts (Du et al., 2015). Similar to DM1, the *TCF4* (CUG)n RNA co-localizes with and sequesters MBNL1 and muscleblind-like 2 (MBNL2), leading to mis-splicing of essential MBNL1-regulated mRNAs (Du et al., 2015; Zarouchlioti et al., 2018). Nuclear RNA foci formed from the accumulation of antisense-derived $(CAG)_n$ transcripts have also been detected in tissue derived from individuals affected with CTG18.1 positive FECD, although much less abundant (Hu et al., 2018).

1.2.8 Repeat-associated non-ATG (RAN) translation

Since 2011, it has been demonstrated that proteins could be synthesised from repetitive RNA transcripts in the absence of an AUG initiation codon, a process described as repeat-associated non-AUG (RAN) translation (Zu et al., 2011). This mechanism was initially described in association with a CAG repeat expansion causative for spinocerebellar ataxia type 8 (SCA8) and DM1(Zu et al., 2011) but has now also been described in several other expansion mediated diseases including DM2, HD and *C9orf72*-mediated ALS/FTD (Ash et al., 2013; Bañez-Coronel et al., 2015; Zu et al., 2017).

Given that the formation of stable sense-derived $(CUG)_n$ and antisense-derived $(CAG)_n$ RNA transcripts from the CTG18.1 expansion within FECD patient, it is very plausible that these repetitive RNA transcripts can could also be translated via RAN translation (Du et al., 2015; Hu et al., 2018; Zarouchlioti et al., 2018). The production of five distinct repeat peptides is possible from the reading frames, poly-Alanine, poly-Cysteine and poly-Leucine from the sense strand and poly-Serine, poly-Alanine and poly-Glutamine from the antisense (**Figure 8**).

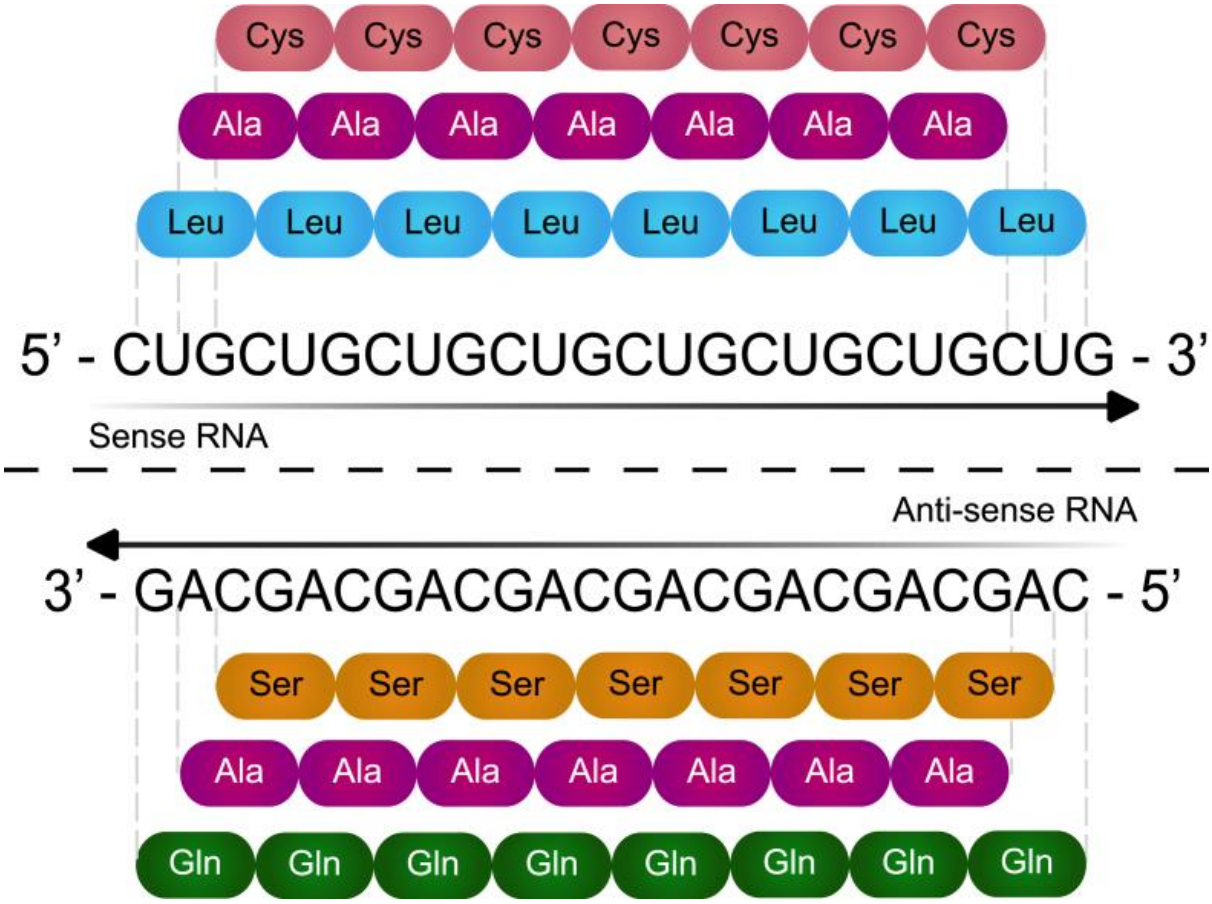


Figure 8 Poly-peptide proteins potentially generated by CTG18.1 repeat-associated non-ATG (RAN) translation. The figure illustrates the potential repeat peptides that may be generated from CTG repeat-associated non-ATG (RAN) translation. The sense strand generates poly-Alanine (Ala), poly-Cysteine (Cys) and poly-Leucine (Leu) and the antisense strand generating poly-Glutamine (Gln), poly-Alanine (Ala) and poly-Serine (Ser). This figure represents data I illustrated for a collaborative review article (Fautsch et al., 2021).

There is evidence to suggest RAN translation potentially occurring in FECD. Most significantly, detection of poly-Cysteine protein, with an antipeptide antibody raised against the putative C-terminal region of poly-Cysteine peptide, within corneal endothelial tissue derived from CTG18.1 expansion-positive FECD patients (Soragni et al., 2018). Although this initial evidence for RAN translation occurring in FECD patient-derived samples is encouraging, further work is needed to confirm the presence of the other species, poly- Alanine, poly-Glutamine, poly-Serine and poly-Leucine; and to continue to explore if these peptides are pathogenic and how they may contribute to the disease. Likewise, the contribution of RNA toxicity and altered levels of the TCF4 itself is yet to be fully elucidated. All three options could be sole contributors to FECD pathogenicity alone although it is likely a combination of two or all the three could explain the pathogenesis of FECD.

1.2.9 Somatic instability of CTG18.1

It is a known phenomenon that disease-associated repeats can expand in length over an individual's lifespan in both an age-dependent and tissue specific manner, termed somatic instability (Morales et al., 2012). The somatic instability of these repeats may contribute to symptom progression of a given disease and has served as a hypothesis to explain the tissue-specificity and phenotypic variability of various repeat-associated diseases including DM1, HD and others (Monckton, Wong, Ashizawa, & Caskey, 1995; Morales et al., 2012; Trang et al., 2015; Wong, Ashizawa, Monckton, Caskey, & Richards, 1995). In HD, mutant CAG repeat sizes vary greatly both within and between somatic tissues of HD patients with the greatest variability occurring in the cortex and striatum, areas of the brain with the most neuropathological involvement (Kennedy et al., 2003; Telenius et al., 1994). Furthermore, the most prominent

mosaicism has been observed in juvenile onset cases of HD suggesting the influence mosaicism has on progression of disease (Kennedy et al., 2003). Long-read amplification-free sequencing methods have recently demonstrated that DNA derived from leukocytes from CTG18.1 expansion-positive FECD patients display high levels of somatic instability, with larger levels of instability found to be positively correlated with increased CTG18.1 length (Hafford-Tear et al., 2019). Furthermore, somatic instability of the CTG18.1 repeat has also now been observed in RNA from the corneal endothelium of three FECD patients (Wieben et al., 2021). Additional research with larger cohort size and other non-ocular tissues are required to further describe somatic instability in CTG18.1 expansions and whether larger somatic expansions within the corneal endothelial tissue drives the progression of FECD.

1.2.10 Diverse molecular mechanisms

1.2.10.1 Bi-allelic CTG18.1 expansions

In rare instances individuals with bi-allelic expansion of TNRs have been reported. The impact of zygosity status on disease phenotype has been controversial to date. Case reports of HD have reported no significant difference in the age of onset or initial symptoms between a patient with a homozygous expanded allele in comparison to patients with a heterozygous expanded allele and in fact reported the patient with the heterozygous expansion to have more severe motor and psychiatric symptoms (Alonso et al., 2002). However, in a larger study, it was reported that disease phenotype in those with bi-allelic expansions progressed more rapidly in comparison to those with a heterozygous expanded allele, although the differences were subtle, age of onset was not earlier (Squitieri et al., 2003). A more severe phenotype caused

by bi-allelic expansion could be attributed to a dosage effect (Squitieri et al., 2003). It is currently unknown whether CTG18.1 biallelic expansions result in earlier disease or a more severe FECD phenotype. Future studies are necessary to comprehensively resolve this uncertainty (Fautsch et al., 2021).

1.2.10.2 Interruptions in CTG18.1 repeat

Although the age of onset has been correlated with repeat length for many repeat mediated disorders, it does not alone explain variability in the age of onset observed for FECD patients. There have been several hypotheses which could explain this variability, the first being interruptions present in the repeat sequence itself. This has been investigated in HD where the CAG repeat in *HTT* is interrupted downstream with a penultimate CAA codon [reference: $(CAG)_n$-CAA-CAG)]. Variation of the CAA resulting in a CAG [i.e. $(CAG)_n$-CAG-CAG)], referred to as loss of interruption, results in carriers presenting with phenotypic symptoms on average 25 years earlier in contrast to those with the reference interruption sequence of the same polyglutamine length, with duplication of the CAA-CAG motif (i.e. $(CAG)_n$-$(CAA\text{-}CAG)_2$) presented in a delayed age of onset (G. E. B. Wright et al., 2019).

For several other repeat expansion diseases, interruptions within the repeat sequence result in a reduced penetrance of the associated phenotype (Matsuura et al., 2006; Stolle et al., 2008). Patterns of interruptions in the CAG repeat have also been observed in spinocerebellar ataxias (SCAs) and distribution of these interruptions in the repeat are believed to have implication on repeat instability and consequently lead to variability in age-of-onset and disease severity (Sobczak & Krzyzosiak, 2004). It has been hypothesised that CTG18.1 interruptions could protect the small proportion of the population,

approximately 4%, who carry an expanded CTG18.1 allele, but do not have FECD. However, sequencing data has not been able to reveal any novel interruptions in the CAG repeat structure of CTG18.1 expanded repeats in clinically unaffected individuals to date (Wieben, Baratz, et al., 2019).

1.2.10.3 Influence of variants in DNA repair genes

There is now sufficient evidence to suggest DNA repair mechanisms have a central role in the pathogenesis of repeat mediated disease. A GWAS performed on a large dataset of HD patients identified variants within genes encoding components of the DNA damage response (DDR), particularly those involved in mismatch repair (MMR) including *MSH3*, *DHFR* and *MLH1*, to be associated with an earlier age of onset (Consortium, 2019; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015). Variants within these genes have been shown to influence the somatic instability of CAG repeats via downstream deficits in the DNA repair mechanism resulting in an accumulation of DNA damage (Massey & Jones, 2018). Alternatively, some variants have been identified to have a protective role in repeat mediated disease, for example, variants identified within *FAN1,* which result in an increase of expression, have significantly been associated with a delayed age of onset and slower disease progression in HD patients (Goold et al., 2019).

1.2.11 Further FECD associated genes

As previously mentioned in **Section 1.2.1**, approximately 79% of FECD in Caucasian populations can be attributed to an expanded trinucleotide repeat in the gene *TCF4,* termed CTG18.1 (Luther et al., 2016; Mootha et al., 2014; Skorodumova et al., 2018; Wieben et al., 2012; Zarouchlioti et al., 2018). The

genetic cause of FECD in patients without a CTG18.1 positive expansion is currently unknown.

Prior to the *TCF4* CTG18.1 repeat expansion being associated with FECD, linkage analysis in large families identified several genes to be causative for FECD. These genes are listed in **Table 3**. Missense mutations in *COL8A2* have been demonstrated to be causative for the rare early-onset FECD phenotype (Biswas, 2001). The common late-onset FECD has also been associated with several genes including *SLC4A11*, *LOXHD1*, *ZEB1* and *AGBL1* (Riazuddin et al., 2012; Riazuddin, Vasanth, Katsanis, & Gottsch, 2013; Riazuddin et al., 2010; Vithana et al., 2008).

Table 3 Genes associated with Fuchs endothelial corneal dystrophy (FECD) (Fautsch et al., 2021).

Associated gene or loci	Protein	OMIM	Genomic coordinates (GRCh38)	Most significantly associated SNP	Reference
Gene harbouring presumed causative variant(s)					
TCF4	Transcription factor 4	613267	18:55,222,184–55,635,956	rs613872	(Baratz et al., 2010; Wieben et al., 2012)
				rs784257	(Afshari et al., 2017)
COL8A2	Collagen Type VIII Alpha 2 Chain	136800	1:36,095,238–36,126,206	NA	(Biswas, 2001)
SLC4A11	Solute carrier family 4 (sodium borate cotransporter), member 11	613268	20:3,227,416–3,241,483	NA	(Vithana et al., 2008)
ZEB1	Zinc finger E box-binding homeobox 1	613270	10:31,318,416–31,529,813	NA	(Mehta et al., 2008)
AGBL1	ATP/GTP-binding protein-like 1	615523	15:86,079,619–87,031,475	NA	(Riazuddin et al., 2013)
LOXHD1	Lipoxygenase homology domain-containing 1	NA	18:46,476,960–46,657,114	NA	(Riazuddin et al., 2012)

The onset of FECD is typically in the fifth or sixth decade of life, however a distinct early-onset FECD phenotype, where disease manifests from as early as the first decade, can also occur (Magovern et al., 1979). Linkage analysis carried out in the early 2000s first identified a missense mutations, p.(Gln455Lys) in *COL8A2*, encoding the α2 chain of type VIII collagen, a major

component of the DM, to segregate within affected family members of three large families with early-onset FECD and PPCD (Biswas, 2001). This has since been replicated, and further *COL8A2* mutations, p.(Leu450Trp) and p.(Gln455Val) have also been identified to be causative for early-onset FECD (Gottsch et al., 2005; P. Liskova, Prescott, Bhattacharya, & Tuft, 2007; Mok, Kim, & Joo, 2009). The *COL8A2* p.(Leu450Trp)-causative early-onset phenotype has also been established to result in definitive characteristics, such as mildly elevated guttae which are associated to an individual corneal endothelial cell (CEC), in comparison to the common late-onset FECD where guttae appear sharply raised and located along the borders between CECs (Gottsch et al., 2005). This mutation also results in a more coarse and distinct distribution of guttae, in contrast to a fine, patchy distribution seen in late-onset FECD (Gottsch et al., 2005). Knock-in mouse models, *COL8A2*$^{L450W/L450W}$ and *COL8A2*$^{Q455K/Q445K}$ have shown to exhibit hallmarks of FECD, including guttae, cell loss and deviations in CEC morphology. The knock-in mouse models also exhibited endoplasmic reticulum stress and activation of the unfolded protein response (UPR) resulting in UPR-associated apoptosis. The studies confirmed the presence of the p.(Leu450Trp) and p.(Gln455Lys) *COL8A2* mutations to cause early-onset FECD in humans (Jun et al., 2012; Meng et al., 2013).

Late-onset FECD was first mapped to a common locus, 18q21.2-q21.32, in 2006 using large pedigrees (Sundin et al., 2006). Although the highly associated *TCF4* SNP rs613872 had been discovered by GWAS in 2010, haplotype analyses suggested the original locus to be independent of this risk factor, suggesting multiple loci in this region may account for the linkage signals on 18q (Riazuddin et al., 2012). A variant in *LOXHD1,* encoding a highly conserved protein consisting of PLAT (polycystin/lipoxygenase/alpha-toxin)

domains and has the role of targeting other proteins to the plasma membrane, was later identified in a large pedigree in 2012. Riazuddin *et al*. then went on to demonstrate an increase of *LOXHD1* expression is observed in the explanted corneal tissue of the FECD patient with the variant in comparison to control tissue, suggesting the variant was causative of the FECD disease phenotype through the mechanism of protein aggregation in the endothelium and DM (Riazuddin et al., 2012). Subsequently, the group went on to reveal a significant enrichment of predicted-pathogenic *LOXHD1* variants in their FECD cohort compared to the control cohort, concluding the observed mutational load of this locus is related to FECD. While there have been other reports of *LOXHD1* variants identified in FECD cohorts, these findings are yet to demonstrate evidence of the pathogenesis of FECD disease caused by *LOXHD1* (Rao et al., 2018; Tang et al., 2016).

In 2006, the rare autosomal recessive endothelial dystrophy, CHED, was identified to be a result of homozygous or compound heterozygous mutations in the gene *SLC4A11*, which encodes a membrane-bound sodium-borate cotransporter (Vithana et al., 2006). Due to the similarities between CHED and FECD, *SLC4A11* was later considered to be a candidate gene for FECD which lead to the finding that heterozygous *SLC4A11* mutations gave rise to the late-onset FECD phenotype (Vithana et al., 2008). It has been suggested that pathology associated with *SLC4A11* is the cause of loss-of-function effect, rather than toxic gain-of-function effect as homozygous *SLC4A11* knock-out mice display corneal oedema, much like with FECD and CHED (Vilas et al., 2013). More recent studies have revealed SLC4A11 protein serves as a cell adhesion molecule, contributing to the adhesion of CECs to the DM, explaining

the increased loss of endothelial cells in both FECD and CHED (Malhotra et al., 2019).

Similarly, like *SLC4A11*, *ZEB1*, also previously known as *TCF8*, had already previously been associated with PPCD and due to the shared common pathologic features, *ZEB1* was explored as a candidate gene for FECD. A possible genotype-phenotype correlation for *ZEB1* mutations was proposed, with loss-of-function, resulting in haploinsufficiency, mutations associated with PPCD and missense mutations with FECD (Mehta et al., 2008). The initial study proposing this idea was not able to provide evidence to support this theory but did not rule out the hypothesis (Mehta et al., 2008). Riazuddin *et al*. later went on to verify this hypothesis, reporting five novel missense mutations in *ZEB1* identified within two cohorts of patients with late-onset FECD (Riazuddin et al., 2010). Nevertheless, *ZEB1* missense mutations have been shown not to significantly impact protein abundance and the functional impact on ZEB1 and their relation to FECD remains to be elucidated (Chung, Frausto, Ann, Jang, & Aldave, 2014).

In 2013, a locus on the chromosomal arm 15q was identified through linkage analysis, utilising a multi-locus model, of a large three generation pedigree. Subsequently, a novel variant, c.3082C>T, resulting in a nonsense mutation, p.(Arg1028Ter), was identified in *AGBL1* (Ensembl transcript ID: ENST00000441037) (Riazuddin et al., 2013). The same variant was later identified in two further unrelated individuals with FECD as well as a missense mutation, c. 2969G>C, p.(Cys990Ser). The group then went on to demonstrate several findings; (1) the p.(Arg1028Ter) nonsense AGBL1 variant protein, which lacked 38 amino acids from the C terminus, was localised predominantly to the nucleus of NIH 3T3 cells, in contrast to wild-type (WT) AGBL1 and

p.(Cys990Ser) missense AGBL1 protein which localised to the cytoplasm only; (2) AGBL1 interacts specifically with TCF4 and (3) both the nonsense and missense *AGBL1* variants significantly reduces binding affinity to TCF4 suggesting the ablation of this interaction may contribute to disease pathogenesis (Riazuddin et al., 2013). However, they did not demonstrate if the variant resulted in NMD of the protein and in addition, the Genome Aggregation Database (gnomAD) genome browser predicts AGBL1 to tolerate loss of function variants, with a 'probability of being loss-of-function intolerant' (pLI) score of 0. Furthermore, it was hypothesised that FECD in the family in which the p.(Arg1028Ter) variant was first identified, might be heterogeneous and multiple causal alleles may be responsible for the disease phenotype. It is also possible this family has a single, unidentified causal allele for the FECD phenotype. Other studies have yet to find the p.(Arg1028Ter) variant or other loss-of-function variants in additional cohorts (Okumura, Hayashi, Nakano, Tashiro, et al., 2019; Skorodumova et al., 2018). Moreover, AGBL1 has been shown not to be expressed in the corneal endothelium which questions whether genetic variants in this gene could have a functional role in the development of the FECD (Frausto, Wang, & Aldave, 2014; Wieben et al., 2018).

In 2017, a GWAS including over 1,400 FECD cases and over 2,500 controls a region on chromosome 18 encompassing *TCF4* was identified to confer the greatest risk for FECD, most significant SNP rs784257, strengthening previous findings. Along with the *TCF4* locus, three further regions were identified to be associated with FECD at genome-wide levels of significance ($P<5 \times 10^{-8}$). These novel loci were situated within intronic region of *KANK4* (rs79742895), an intronic region of *LAMC1* (rs3768617) and an

intergenic region between *LINC00970* and *ATP1B1* (rs1200114) (Afshari et al., 2017).

LAMC1 encodes an ECM laminin glycoprotein and has a key role in cellular adhesions in basement membranes. It has been shown to be highly expressed on the endothelial side of the DM (Afshari et al., 2017). Since the large-scale GWAS was performed by Afshari *et al.,* a rare variant, c. 1468C>T p.(Arg190Trp) (minor allele frequency (MAF) =0.003766), within *LAMC1* has been identified in an FECD patient negative for the CTG18.1 repeat expansion suggesting a possible association between *LAMC1* and FECD (Wieben et al., 2018). *KANK4* codes for a protein which has a role in the regulation of actin stress fibres, however little is known about the cellular function. There is minimal expression of *KANK4* in corneal tissue however, immunostaining has revealed localisation in the endothelial cytoplasm of both controls and FECD samples (Afshari et al., 2017). *LINC00970* was shown to have no expression in the corneal endothelium, whereas *ATP1B1* is highly expressed within the corneal endothelium and encodes for an ATPase Na+/K+ transporting subunit. It is hypothesised to have a role in fluid regulation and ion transport and loss of ATP1B1 may lead to hypertonicity and subsequently cause corneal oedema, a common feature of FECD thus making a viable candidate gene (Afshari et al., 2017).

1.3 Thesis aims and objectives

FECD is a common and debilitating disease however the underlying genetic causes and biological mechanisms behind them are not yet comprehensively understood. The primary objective of the work described in this thesis is to further characterise genotype-phenotype correlations in patients diagnosed with FECD. Firstly, a large cohort of 990 patients were explored with

the aim to further elucidate existing phenotype-genotype correlations, including how the prevalence of the CTG18.1 expansion differs in sexes and populations.

Secondly, I aimed to explore how the structure of CTG18.1 and somatic instability impacted phenotypic outcomes using an ultra-high-throughput sequencing methods. Furthermore, I aimed to investigate how genetic variants in DNA repair genes modified the somatic stability of CTG18.1. Lastly, exome sequence data was interrogated with the aim to explore the missing genetic heterogeneity of FECD patients which do not harbour a CTG18.1 expansion.

2. Methods

2.1 Patient recruitment, clinical phenotyping, and sample collection

A total of 990 unrelated individuals with FECD attending Moorfields Eye Hospital, London, United Kingdom (MEH) or General University Hospital in Prague, Czech Republic (GUH) were recruited to this study. Participants had either undergone corneal transplant surgery for FECD or were showing clinical signs of FECD, such as numerous corneal guttae on slit-lamp biomicroscopy. At this stage of recruitment patients were not excluded if they had previously had cataract surgery or other corneal abnormalities. The study adhered to the tenets of the Declaration of Helsinki and was approved by the Research Ethics Committees of University College London (UCL) (22/EE/0090), Moorfields Eye Hospital, London (13/LO/1084), or the General University Hospital (GUH) Prague (2/19 GACR). Written informed consent was received from all participants included in this study.

2.2 DNA extraction from blood

Blood samples were obtained from the proband, and relatives if available, for genetic analysis. DNA was extracted from peripheral leukocytes by Beverly Scott as part of the UCL Institute of Ophthalmology DNA extraction service using the Qiagen Gentra Puregene Blood Kit in accordance with the manufacturer's protocol.

2.3 Polymerase chain reaction (PCR)

2.3.1 Primer design

All primers were designed using primer3 version 4.1.0 3(primer3.ut.ee) (Koressaar & Remm, 2007; Untergasser et al., 2012). Primers were designed to have a length of 18-24 base pairs (bp) to ensure specific binding and flank the genomic region of interest by at least 100bp. Primer pairs were designed to

have a similar GC content, between 40-60%, and similar annealing temperatures. It was also ensured primer pairs were not complementary to one another or possess self-complementarity to avoid primer dimer formation.

Target sequences were obtained from Ensmbl Genome Browser (https://www.ensembl.org/index.html) and checked for primer specificity and the absence of common polymorphisms within the primer sequence. All oligonucleotides were obtained from Sigma-Aldrich. A full list and details of primers used are listed in **Table S1**.

2.3.2 Standard PCR

PCR was carried out to amplify patients' genomic DNA using an Eppendorf Mastercycler gradient PCR machine. Each PCR reaction was optimised first using a control human genomic DNA (Promega, UK).

PCR amplification was carried out using GoTaq Green Master Mix (Promega) in a 12.5 µL volume reaction. Volumes of reagents are shown in **Table 4**. Thermal cycling parameters included an initial denaturation at 95°C for 2 minutes, followed by 35 cycles of denaturation at 95°C for 30 seconds, annealing at optimal temperature for 30 seconds and extension at 72°C for 30 seconds, followed by a final extension step at 72°C for 5 minutes. Each PCR was initially tested using an annealing temperature of 60°C.

Table 4 Volume and composition of polymerase chain reaction (PCR) master mixes

Reagent	Volume (µL)	Final Concentration
2X GoTaq® Green Master Mix	6.25	1X
Forward Primer (10µM)	0.5	0.4 µM
Reverse Primer (10µM)	0.5	0.4 µM
ddH_2O	4.75	
DNA template (50-100 ng/µL)	0.5	25-50ng
Total	12.5	

2.3.3 Gradient PCR

PCR reactions which did not produce the expected product were optimised using a gradient annealing temperature protocol to determine the optimal annealing temperature for the respective primer pairs. Eight PCR reactions were set up in accordance with **Table 4** and annealing temperature varied between 52.2°C to 68.5°C across the total 8 reactions.

2.3.4 Colony PCR

To screen for the presence or absence of insert DNA in plasmid constructs a colony PCR was performed. Transformants were picked up using a sterile pipette tip and dipped into the PCR reaction in place of a DNA template and a standard PCR was performed following **Section 2.3.2**.

2.3.5 Agarose gel electrophoresis

In a conical flask, the appropriate amount of agarose powder (Bioline, UK) was added in the appropriate volume of 1X Tris-Acetate-EDTA buffer (TAE), prepared by a 1 in 10 dilution of 10X TAE (Severn Biotech Ltd, UK), and dissolved using a microwave to make a 2% (w/v) gel. Once cooled slightly, SafeView (NBS Biologicals) or ethidium bromide (EtBr) (Sigma-Aldrich) was added for nucleic staining at 0.4 µL/mL or 0.5µL/mL of gel, respectively, and poured into a casting plate with well comb inserted. Once the gel was set, it was placed into an electrophoresis tank and submerged in 1X TAE. Samples and DNA ladder were loaded on the gel and resolved at 120 V. The gel was visualised using ChemiDoc Imaging System (BioRad, UK).

2.3.6 PCR Product purification using a vacuum filtration method

To remove unincorporated dNTPs and primers which may interfere with downstream analysis, PCR products were purified using PCR clean-up filter

plates (Merck Milipore, UK). PCR products were made up to 100 µL with double distilled water (ddH2O) and loaded onto the filter plates. The samples were then filtered using a vacuum pump (Merck Milipore, UK) until the sample passed through the size exclusion membrane. A wash step was then carried out by loading 50 µL of ddH2O and filtering again until clear. DNA was then eluted in 20 µL of ddH2O by gently vortexing the sample for 10 minutes before transferring to a 0.2ml PCR tube.

2.3.7 PCR product purification using a gel extraction method

Gel extraction was performed using the QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's protocol. To summarise, the amplified product of interest was excised from the agarose gel under a Ultraviolet (UV) transilluminator with a scalpel blade and placed into a 1.5 ml microcentrifuge tube. Three volumes of Buffer QG (gel:buffer ratio 1:3) were added and the samples were incubated at 50°C, vortexing intermittently, until the gel was completely dissolved. Next, one volume of isopropanol was added to precipitate the DNA. The mixture was then loaded onto a QIAquick spin column to bind the DNA to the column membrane, followed by on-column washing steps using 750 µl of Buffer PE. Finally, the DNA was eluted in 30 µl of Buffer EB.

2.3.8 Sanger sequencing

Sanger sequencing was outsourced to Source BioScience (Cambridge, UK). PCR products were purified as described in **Section 2.3.6** and DNA concentration was measured on NanoDrop2000c (ThermoFisher). PCR samples and primer concentrations were adjusted to provide 5 µL of PCR sample at 10 ng/µL and 5 µL of primer 3.2 pmol/µL per reaction.

2.3.9 Data analysis of Sanger sequencing

Electropherograms produced from Sanger sequencing were aligned to a reference sequence using DNAstar Lasergene SeqMan Pro version 7.1.0.

2.4 *TCF4* CTG18.1 genotyping

2.4.1 Short tandem repeat (STR) assay

A STR assay was performed to genotype the CTG18.1 allele. Genomic DNA was amplified using a 5′ Fluorescein (FAM) conjugated primer (5′-CAGATGAGTTTGGTGTAAGAT-3′) and an unlabelled reverse primer (5′-ACAAGCAGAAAGGGGGCTGCAA-3′). PCR conditions were as followed: an initial denaturation step of 95˚C for 5 minutes, followed by 35 cycles of 95˚C for 30 seconds, 56˚C for 30 seconds and 72˚C for 1 minute and 30 seconds, followed by a final extension step of 72˚C for 5 minutes (Wieben et al., 2012; Zarouchlioti et al., 2018).

2.4.2 Triplet-primed polymerase chain reaction (TP-PCR)

For samples in which only one allele of non-expanded CTG18.1 repeats was detected using STR genotyping, TP-PCR was performed to investigate the presence of potentially a larger expanded allele above the detection limit of the STR-PCR method. TP-PCR utilises three primers, a FAM-labelled P1 forward primer (AATCCAAACCGCCTTCCAAGT) designed upstream of the CTG18.1 repeat, and two reverse primers. The first reverse primer, P3 (TACGCATCCCAGTTTGAGACGCAGCAGCAGCAGCAG), is comprised of 5 units of the CTG repeat and a 5′ tail to serve as an anchor for a second reverse, P2 (TACGCATCCCAGTTTGAGACG), which prevents progressive shortening of the PCR products during subsequent cycles (**Figure 9**). PCR cycling conditions were as followed: an initial denaturation step for 9 minutes at 95˚C,

followed by 10 cycles of 95 ˚C for 30 seconds, 62 ˚C for 30 seconds and 72˚C for 4 minutes. Following, 30 cycles of 95 ˚C for 45 seconds, 62˚C for 45 seconds and 72˚C for 4 minutes were performed and each cycle was extended by 15 seconds each time before a final extension step of 72˚C for 10 minutes (Vasanth et al., 2015).

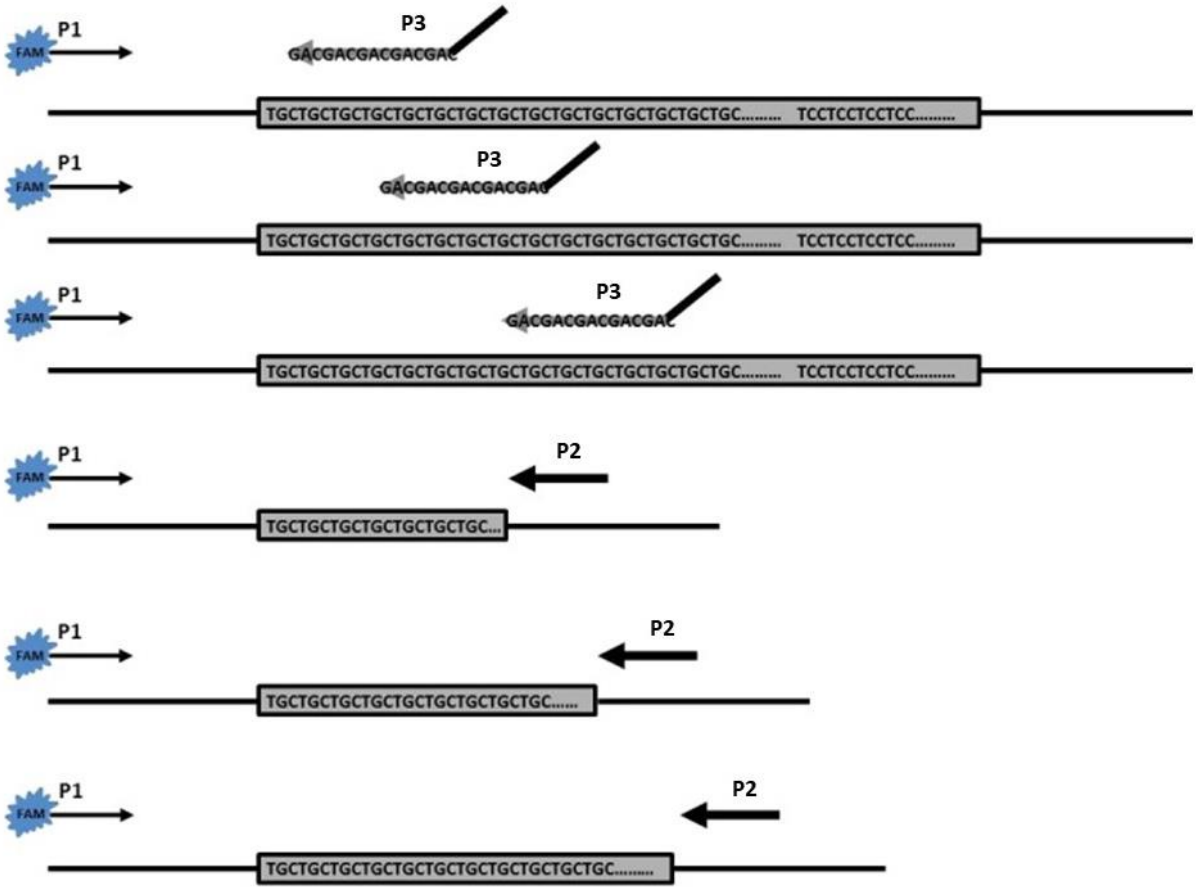


Figure 9 Overview of the triplet primer-polymerase chain reaction (TP-PCR) method to genotype the CTG18.1 locus. The TP-PCR method utilises three primers, P1, P2 and P3. Primer P3 at multiple sites within the CTG repeat in the initial rounds of amplification resulting in a mixture of products. Primer P1 is a locus-specific 5′-6-FAM-tagged primer. Primer P2 amplifies from the end of the mixture of products amplified in the prior cycles. A long extension time is applied to ensure complete extension of the longer products within the PCR product mixture (adapted from Mootha, Gong, Ku, & Xing, 2014).

2.4.3 Post-PCR reaction

STR and TP-PCR PCR products were prepared for capillary electrophoresis by combining with Rox500 Ladder and formamide in a 1:50 ratio allowing a total of ~10µL of this mixture per DNA sample for sequencing.

2.4.4 *TCF4* CTG18.1 genotyping analysis

Following PCR amplification for both the STR and TP-PCR assays, post PCR product separation was performed on the ABI 3730 Electrophoresis 96 capillary DNA analyzer (Applied Biosystems) to determine the number of CTG18.1 repeats present in each amplified allele of the DNA samples analysed. Data analysis was performed using GeneMarker software (SoftGenetics).

2.4.5 MiSeq sequencing

2.4.5.1 MiSeq library preparation

Modified CTG18.1 locus-specific MiSeq-compatible PCR primers, previously designed by Mariam Alkhateeb a PhD student in Professor Darren Monkton's lab at the University of Glasgow (Alkhateeb, 2018), were ordered from Eurofins, UK (detailed primer sequenced listed in **Table S3**). The Modified locus-specific primers were composed of P5/P7 MiSeq adapter, Unique barcode indices, sequencing primer binding site, spaces and the CTG18.1 locus-specific inner primers (**Figure 10**). The P5 and P7 MiSeq adapters are composed of oligonucleotides of 29 bp and 24 bp, respectively, that allow the library to be complimentary hybridised onto the flow cell, followed by bridge amplification and cluster generation. Illumina Nextera XT Kit v2 set of 40 5' and 3' barcode indices were incorporated to allow processing of up to 384 samples per MiSeq run. A total of 16 indices in the forward direction and 24 indices in the reverse direction giving 384 unique index combinations for samples to be

multiplexed. Following the index, the sequencing primer binding site was added, complementary to the sequencing primer on the MiSeq platform. Finally, spacers (from 0 bp to 7 bp) between the sequencing primer binding site and the CTG18.1 locus-specific inner primers were included to increase nucleotide diversity. Spacers on the forward primer enable cluster detection and validation on the MiSeq flow cell; without this nucleotide diversity the Illumina machine may experience failure at detecting a signal from the clusters. Reverse primer spacers have been documented to improve sequence quality (Ciosi et al., 2018).

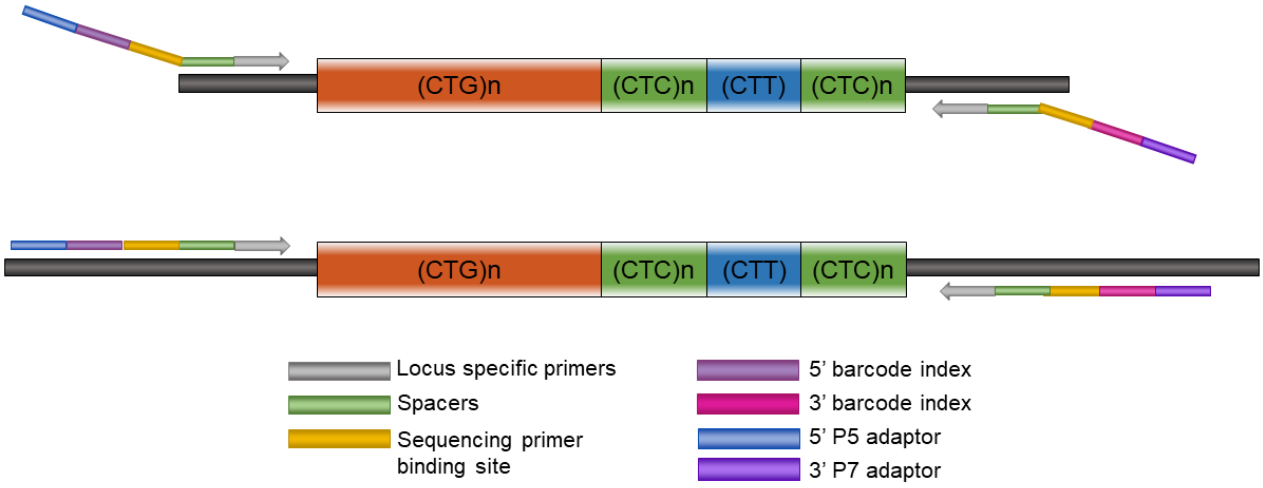


Figure 10 MiSeq primer composition and amplicon structure of CTG18.1 region after PCR. MiSeq primers consisting of adaptor components, P5 or P7 adaptors on the forward and reverse primers respectively, that allow the library to be complimentary hybridised onto the flow cell. 5' and 3' barcode index to identify each sample and allow for multiplexing. The sequencing primer binding site, complementary to the sequencing primer on the MiSeq platform. Spacers to increase nucleotide diversity. Finally, the CTG18.1 locus specific inner primers.

Two MiSeq libraries were set up to enable a total of 768 samples to be analyzed. Each library was prepared in four 96-well plate formats. Within each plate, a total of three controls were included, including a no template negative control and two consistent positive controls used across all plates, one with a mono-allelic expansion and one with bi-allelic non-expansion. Genomic DNA

samples (10 ng) were amplified in 10 µL reactions using 1 µM forward primer (502 to 522) SEF2-C, 1 µM reverse primer (701 to 729) P2+CC and 0.2 µl of (1 U) Taq polymerase (Sigma), 1 X 'Custom PCR master mix + βME' (45 mM Tris-HCL pH 8.8, 11 mM $(NH_4)_2SO_4$, 4.5 Mm $MgCl_2$, 0.113 mg/ml BSA, 0.048% 2-Mercaptoethnol, 4.4 µM EDTA, 1 mM each of dATP, dCTP, dGTP, dTTP) (supplied by Thermo Scientific, ABgene UK) (**Table 5**).

Table 5 Volume and composition for MiSeq library preparation polymerase chain reaction (PCR)

Reagent	1 reaction (10 µL)
Qiagen nuclease free water	2.8 µL
10X 'Custom PCR Master Mix + βME'	1 µL
Taq DNA Polymerase (5 units/µL)	0.2 µL
SEF2-C forward primer (502 to 522) (5 µM)	2 µL
P2+CC reverse primer (701 to 729) (5 µM)	2 µL
Genomic DNA template (5ng/1 µL)	2 µL

PCR parameters were as follows: the lid was heated to 105°C followed by 28 cycles of denaturation at 96°C for 45 seconds; annealing at 56.4°C for 45 seconds and extension at 70°C for 3 minutes. Followed by a final extension at 70°C for 10 minutes.

2.4.5.2 Purification using magnetic AMPure XP beads kit purification

After PCR amplification and negative template controls (NTCs) had been confirmed to be negative, PCR products were purified to remove primer dimers and for size selection purposes. Magnetic AMPure XP beads kit (Beckman Coulter, California, United States) and magnetic stand were used for the PCR product purification. 5 µL of the total PCR products were pooled together, leaving 5 µL of PCR products to be stored for back-up. After samples were

pooled together and mixed well, the pool was distributed into six DNA LoBind 1.5 mL tubes. A 0.6X concentration of AMPure beads were added to the tubes and the solution was mixed by gently flushing up and down, without vortexing. Tubes were placed on the magnetic stand and incubated until the beads had settled and the solution was clear. Subsequently, the supernatant was removed carefully to ensure the beads were not disturbed, whilst keeping the tube on the magnetic stand. Next, 80% freshly prepared ethanol was added to the tubes to completely cover the beads but without disturbing the beads and was removed after 30 seconds. This step was repeated once. After removing the ethanol, the beads were left to air-dry for 5 minutes, being careful not to over dry the beads as this may result in lower recovery of DNA. After this step, the DNA was eluted in 45 µL nuclease free water, by allowing the samples to incubate at room temperature, outside of the magnetic stand, for a total 2 minutes. Following this, tubes were placed back in the magnetic rack until beads have settled and solution was clear. Finally, eluted DNA samples were collected from the tubes without disturbing the bead pellet.

A second AMPure XP clean-up was performed on the purified MiSeq library to concentrate the sequencing library in accordance with the above protocol but using a 0.8X AMPure bead concentration.

2.4.5.3 DNA quantification

DNA concentration was estimated using the Qubit fluorometer and DS DNA HS Assay kit (Double stranded DNA High Sensitivity Assay kit). The kit was used in accordance with the manufacturer's instructions. Qubit quantifies DNA concentration using dsDNA binding dyes, compared to the Nanodrop

which can additionally quantify impurities such as RNA, proteins and other contaminants as it measures absorption of UV at 260 nm.

The prepared library was further quantified using a Bioanalyser (Agilent) to allow for the detection of primer dimers and to estimate the molarity of the prepared library prior to sequencing. This analysis was conducted by Glasgow Polyomics at University of Glasgow (http://www.polyomics.gla.ac.uk/index.html).

2.4.5.4 Sequencing

Once prepared at UCL, MiSeq libraries were sent to Glasgow Polyomics to be sequenced using the MiSeq platform Next Generation Sequencing. All libraries were sequenced up to 400 bp from the forward reads and 200 bp from the reverse reads and 10% PhiX internal control was added to each run to determine the errors in each run. Sequenced PhiX was aligned against the reference sequence to determine the run quality. Glasgow Polyomics returned the output data in fastq file format with each sample named after the corresponding index pairs.

2.4.5.5 MiSeq data analysis

2.4.5.5.1 Preparation of MiSeq data

To summarise, after importing the sequencing files, the reads were demultiplexed using Cutadapt (Version 1.16.8) to remove any index cross-contaminated reads. Following, forward reads were demultiplexed using Cutadapt to remove spacer-related read length variation in the 5’ end of the sequencing reads and so that all reads started at the same position within the sequencing primer binding site. FASTQtrimmer (version 1.1.1) was then used to trim bases at the 3’ end of the reads, depending on spacer length, to remove spacer-related variation in the 3’ sequence and ensure all reads were the same

length across all samples. Lastly, Cudatpat was used to trim the Illumina sequencing adapter at the 3'-end of the reads.

2.4.5.5.2 Repeats genotyping tool (RGT) to genotype the CTG18.1 repeat

After reads were prepared, CTG frequency distributions were processed using Repeats Genotyping Tool (RGT) on forward reads (R1) by Dr. Vilija Lomeikaite and Dr. Marc Ciosi at University of Glasgow. RGT works by extracting the repeat structure from reads, counting the repeat units from the start flank “"GGGCTCTTTCATG" to then end flank “"TTCTAGACCTTCTTTT". Initially, RGT was used to count CTGs only from 5' flank to end of R1. However, this meant that after reviewing the output data, reads which did not contain pure CTG tracts had been discarded. Therefore, a second approach was implemented where an additional parameter was added to enable reads with either PCR and/or sequencing errors to be included in the analysis in which RGT was asked to consider all triplet repeat variation ending at "CTTCTCCTC". To visualise RGT output, R was used to create plots showing total CTG count against abundance of reads, per individual sample analysed. For reverse reads (R2), RGT was used to determine allele structure by using it to count the number of repeat units of “CTC” and “CTG” motifs.

2.4.5.5.3 Alignment of MiSeq reads

Processed reads from **Section 2.4.5.5.1** were divided into unexpanded allele and expanded allele reads by length using Cutadapt by setting the minimum length of the outputted reads to a length that can distinguish unexpanded from expanded allele reads, 200 bp. A synthetic reference sequence set was designed. These templates included the flanking regions upstream and downstream of the CTG and CTC repeat region with varying numbers of repeat numbers including; CTG repeats from 1 to 119 and CTC1

repeats from 1 to 9 to enable accurate allelic structures to be determined. Burrows-Wheeler Aligner Maximal Exact Match (BWA-MEM), within a Galaxy server, was used to align the expanded alleles to the reference sequences using the parameters described by Ciosi *et al* (Ciosi et al., 2021). In brief, the default BWA-MEM parameters were used except for the following three parameters: penalty for a mismatch=1; gap open penalties=2,2; gap extension penalties=2,2. These output files were then converted from Binary Alignment Map (BAM) to Sequence Alignment Map (SAM) files using the BAM-to-SAM tool. To validate the RGT outputs, Tablet (version: 1.21.02.08) was used to visualise the mapped reads against reference sequences.

2.4.5.5.4 Quantifying CTG18.1 somatic expansion levels

From the MiSeq read count distribution, obtained in **Section 2.4.5.5.2**, for each disease-associated expanded allele sequenced the ratio of somatic expansions was quantified using two measures: (1) the proportion of reads larger than the estimated progenitor allele length (ePAL) from ePAL to the end and (2) the proportion of reads larger than 116 from ePAL to the end.

2.5 Kompetitive allele specific PCR (KASP) assay

The Kompetitive Allele Specific PCR (KASP) assay was used to genotype twelve SNPs known to be associated with the somatic expansion of the HTT CAG repeat in blood and/or HD onset. KASP genotyping was outsourced by the LGC group Twickenham, UK (https://www.lgcgroup.com/).

DNA from 631 samples, with CTG18.1 expanded alleles (≥50 repeats), 8 from samples with borderline repeat length alleles (30-49 repeats) and 66 from Age-related macular degeneration (AMD) samples used as controls were sent

at a concentration of 10ng/ul and distributed in 96 well plates with 2 water controls on each plate.

2.5.1 SNPviewer software

KASP genotype data was returned from LGC in the form of comma separated value (CSV) files. SNPviewer, downloaded from the LGC website, was used to view the genotyping clusters plate by plate.

2.5.2 gPLINK

gPLINK was used to analyse genotype SNPs data, calculating minor allele frequency (MAF) and hardy-Weinberg equilibrium test following manufacturer's instructions (Purcell et al., 2007). gPLINK was downloaded from the website (http://zzz.bwh.harvard.edu/plink/index.shtml).

2.6 Predicting ancestry of patients using a genome-wide SNP array

Ancestry of FECD patients were predicted by Anita Szabo, bioinformatician at UCL. Genome-wide SNP data were acquired for all FECD patients recruited in this study. To predict ethnicities of the patients from this data, the tool fast and robutst ancestry prediction by using online singular value decomposition and shrinkage adjustment (FRAPOSA) was utilised using the 1000 Genomes project as a reference panel where the ethnicities of the participants are known. FRAPOSA was ran using the default online augmentation-decomposition-transformation (OADP) approach to predict principal component (PC) scores (Zhang, Dey, & Lee, 2020). After predicting PC scores, the population that samples most likely belong to were predicted using PLINK2.0 (Chang et al., 2015).

2.7 Generation of transcript per million reads mapped (TPM) gene expression levels using RNA-Seq data

Publicly available transcriptomic data derived from three adult and two foetal micro-dissected human corneal endothelial tissues (Chen et al., 2013) were aligned and analysed to determine the relative abundance of genes expressed within the tissue by Dr. Nathaniel Hafford Tear. Quality of FASTQ files was analysed using FastQC and adapter sequences were clipped using trimmomatic software (Bolger, Lohse, & Usadel, 2014). Resulting filtered FASTQ files were aligned to hg38 (release 97) using HISAT2 alignment software (Kim, Paggi, Park, Bennett, & Salzberg, 2019). Gene level counts were generated using featureCounts (Liao, Smyth, & Shi, 2014) and imported to R for DESeq2 analysis. Normalised counts were generated using the DESeq2 normalisation method (Anders & Huber, 2010; Evans, Hardin, & Stoebel, 2018; Love, Huber, & Anders, 2014).

2.8 Exome sequencing

One-hundred-and-forty-one subjects identified to be CTG18.1 expansion-negative (<50 repeats) were selected for WES, on the basis of research funding availability, to investigate the genetic etiology of non-CTG18.1 mediated FECD.

2.8.1 Exome capture and sequencing

Exome sequencing was outsourced to Novogene. Sequencing libraries were generated from blood-derived genomic DNA samples using SureSelect Human All Exome V6 capture kit (Agilent, USA) or SeqCap EZ MedExome Enrichment Kit (Roche). Libraries were sequenced (PE150) on either HiSeq4000 or HiSeq2500 sequencers (Illumina).

2.8.2 Alignment, variant calling, and annotation of exome data read data and variant calling

Alignment of reads and variant calling was carried out by Dr. Nikolas Pontikos and Anita Szabo, Bioinformaticians UCL. Generated data were aligned and annotated in accordance with the previously described pipeline (Pontikos et al., 2017). The short-read sequence data were aligned using novoalign (version 3.02.08), and single nucleotide variants (SNVs) and indels were called according to Genome Analysis Toolkit (GATK) (version 4.1) best practices (joint variant calling followed by variant quality score recalibration) (McKenna et al., 2010). The variants were then annotated using the Variant Effect Predictor (McLaren et al., 2016) and filtered using custom R and python scripts.

The MAF of variants were annotated with their frequencies in GnomAD (Version 2.1.1) (Karczewski et al., 2020), Kaviar genomic variant database (Version 160204) (Glusman, Caballero, Mauldin, Hood, & Roach, 2011) and the internal UCLex consortium dataset (UCLex) (Pontikos et al., 2017).

For each variant, Combined Annotation Dependent Depletion (CADD) scores were generated to predict the deleteriousness of a SNP or insertion/deletions variants. CADD scores are based on diverse genomic features derived from surrounding sequence context, gene model annotations, evolutionary constraint, epigenetic measurements and functional predictions into one metric to predict the pathogenicity of the change (Kircher et al., 2014; Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019). Aligned data were visualised using Integrated Genomics Viewer (IGV) (Broad institute).

2.8.3 UCLex consortium dataset as control dataset

Exome data (n= 5,583) derived from UCLex were used as a control dataset. Data derived from annotated samples which had a diagnosis of ocular disease were removed for the purpose of being used as a control dataset for this study. Sequencing data were aligned and annotated following the same method as described in **Section 2.8.2** (Pontikos et al., 2017)**.**

2.8.3.1 PCA ancestry prediction

Ancestry of control samples from UCLex was predicted using SNPs acquired from exome data carried out by Dr Cian Murphy, Bioinformatician at UCL. From the exome SNP data, principal component analysis was performed using the software FRAPOSA (Zhang, Dey, & Lee, 2020), and samples were plotted on a principal component analysis (PCA) plot based on their close predicted ancestry population. MAF from the 1000 genomes project was used as the reference for SNP analysis acquired from exome data.

2.9 Gene burden testing approach

To discover potential novel genetic causes of CTG18.1 non-expanded FECD samples, a gene burden approach was applied to exome data.

The control exome data were derived from UCLex which has comparable coverage and read-depth as the FECD exome sequencing data cases. All samples in this analysis were of European descent to enable direct comparison of the allele frequency of rare and potentially deleterious variants between case and control subjects. Subsequently, the burden analysis was performed on exome data from 108 non-expanded European FECD cases and 1138 European control samples.

Two approaches were performed for the gene burden analysis: sequence kernel association test (SKAT) (M. C. Wu et al., 2011), a supervised machine learning method that can be used to test for association between rare variants in a region, and a custom-made association test using Fisher-test. Bioinformatic analysis for the SKAT approach was carried out by Dr. Cian Murphy and the custom approach by Anita Szabo.

For both approaches, using the same case-control groups, different thresholds were used to define rare and potentially pathogenic variants on which the gene burden test was performed, **Table 6,** and the outcome of the methods were compared. For the filtering process CADD scores were used to predict deleteriousness and MAF obtained from gnomAD and Kaviar were used to identify rare variants. Internal allele count from UCLex was used to eliminate the potential enrichment of a disease group in the control group.

Table 6 Threshold conditions used to define rare and potentially pathogenic variants in the gene burden analysis.

Threshold condition	CADD	Max MAF (gnomAD exomes, Kaviar)	UCLex AC
1	>20	< 0.01	< 40
2	>20	< 0.001	< 40
3	>20	< 0.0001	< 40
4	>10	< 0.001	< 40

The MAF threshold was applied to the maximum value of the gnomAD exomes and Kaviar allele frequencies. The variants with unknown frequencies were included in the analysis if they passed the remaining filters.

The filtered variants were collapsed by genes and the number of individuals who harboured at least one variant after the filtering (following

dominant inheritance pattern) was calculated in the cases and control groups separately.

In the custom-made approach, Fisher test was used to determine if there was a significant difference in the number of “rare pathogenic” variants between the two groups. The difference was regarded as significant if the Fisher’s test p-value was less than 0.05.

Variants of interest identified by the gene burden analysis were verified by Sanger sequencing, as described in **Section 2.3.8** (primers listed in **Table S1**).

2.10 Luciferase experiment

2.10.1 Modelling miRNA

Secondary structure predictions of miRNAs were generated using RNAfold web service by ViennaRNA, Institute of Theoretical Chemistry (http://rna.tbi.univie.ac.at/) to predict the impact variants may have on miRNA structure the compared to wild type.

2.10.2 Predicting mRNA targets

To identify predicted target mRNA for the MiRNA three databases were used, DIANA Web Server v5.0 (http://diana.imis.athena-innovation.gr/DianaTools/index.php), miRDB (http://www.mirdb.org/) and Target Scan Human (https://www.targetscan.org/vert_80/). The output was compared and mRNA that were expressed in the corneal endothelium (**Section 2.7**) were explored.

2.10.3 Designing miRNA mimics

*mir*Vana™ miRNA mimics were ordered from ThermoFisher. mirVana miRNA mimics are chemically modified double-stranded RNA molecules

designed to mimic endogenous miRNAs. In total, five miRNA mimics were ordered, a WT human species miR-184 mimic and three custom miR-184 mimics containing the variants (+57C>T), (+58G>A) and (+73G>T). The sequence of these miRNA mimics are summarised in **Table 7**. In addition, the *mir*Vana™ miRNA Mimic Negative Control #1 was ordered. This control is a random, undisclosed, sequence miRNA mimic molecule that has been extensively tested in human cell lines and tissues and validated to not produce identifiable effects on known miRNA function.

Table 7 Mature miRNA sequence of mirVana™ miRNA mimics ordered from ThermoFisher

miRNA mimic ID	Mature miRNA Sequence
hsa-miR-184	UGGACGGAGAACUGAUAAGGGU
(+57C>T)-miR-184	UGGAUGGAGAACUGAUAAGGGU
(+58G>A)-miR-184	UGGACAGAGAACUGAUAAGGGU
(+73G>T)-miR-184	UGGACGGAGAACUGAUAAGGUU

2.10.4 Cloning

2.10.4.1 Designing primers to amplify and sub-clone gene-specific 3'UTR regions in the pmirGLO Dual-Luciferase miRNA target expression vector

Primers were designed as described in **Section 2.3.1** and enzyme restriction sites present in the chosen plasmid, SalI and NheI, were added to the 5' of the forward primers and 3' of reverse primers allowing amplified PCR products to be cloned into the plasmid (**Table S2**). Primer efficiency was tested using standard PCR protocol, **Section 2.3.2**.

2.10.4.2 Preparation of Luria-Bertani (LB) broth and LB agar

LB broth was prepared by adding 10 g of LB broth powder (Sigma) to 500 ml of distilled water. To make LB agar, 7.5 g of agar powder (Oxoid) was added.

2.10.4.3 Preparation of ampicillin IPTG/X-gal plates

To prepare Ampicillin IPTG/X-gal plates for blue-white screening. LB agar solution was prepared as in **Section 2.7.1**. When cooled slightly, 500 µl of X-gal (20 mg/ml) and 500 µl of IPTG (100 mM) was added to 500 ml of LB agar. Ampicillin was then added at a final concentration of 100 µg/ml. The solution was poured into sterile agar plates using the Bunsen burner aseptic technique.

2.10.4.4 TA cloning into pGEM®-T easy vector

The pGEM®-T Easy Vector System (Promega) was used for TA cloning. The pGEM®-T Easy Vector is a pre-linearized vector containing multiple cloning sites (MCS) and 3´-T overhangs at the insertion site to provide a compatible overhang for PCR products. The MCS is positioned within the *lacZ* gene. The *lacZ* gene encodes the α-peptide of the enzyme β-galactosidase. β-galactosidase converts lactose into galactose and glucose. Insertion of DNA fragments into the MCS triggers the insertional inactivation of the α-peptide, which enables the differentiation of recombinants and non-recombinants by blue-white screening using IPTG-Xgal plates.

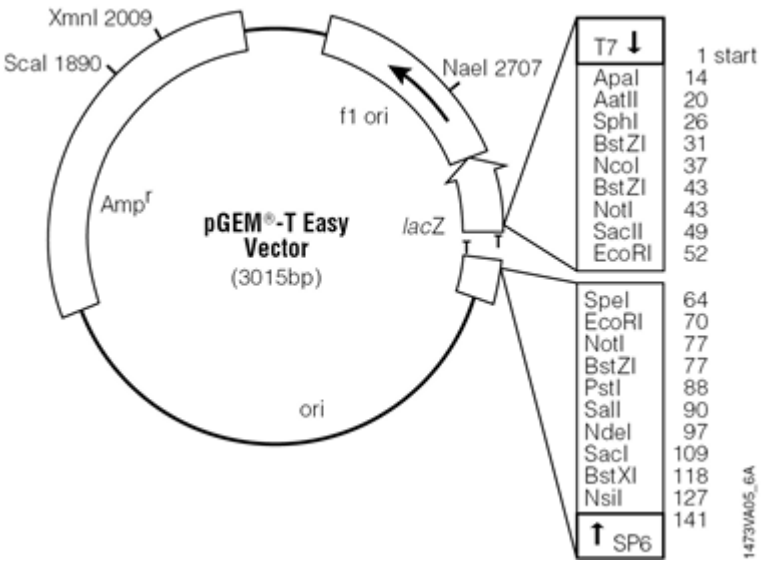


Figure 11 3 pGEM®-T Easy vector map (adapted from Promega)

DNA fragments were amplified using GoTaq® Green Master Mix and purified using filter plates**, Section 2.3.2 and 2.3.6**. PCR products were cloned using pGEM®-T Easy kit (Promega) in accordance with the manufacturer's protocol. In brief, purified PCR products were ligated into the pGEM®-T Easy vector following the ligation reaction set up in **Table 8**. Positive and background control reactions were also set up. After ligation, the reaction was mixed by pipetting gently and incubated overnight at 4°C for the maximum number of transformants.

Table 8 pGEM®-T Easy vector ligation setup

	Volume		
Reagent	**Stand reactions**	**Positive control**	**Background control**
2× rapid ligation buffer	5 µL	5 µL	5 µL
pGEM®-T Easy vector	1 µL	1 µL	1 µL
Purified PCR product	3 µL	--	--
Control insert DNA	--	2 µL	--
T4 DNA ligase (3 Weiss units/µl)	1 µL	1 µL	1 µL
ddH_2O to final volume of	10 µL	10 µL	10 µL

2.10.4.5 Transformation of competent cells (heat shock method)

High Efficiency 5-alpha Competent E. coli cells (NEB) were used for transformation. The competent cells were placed on ice until thawed. Ligation reactions (**Section 2.8.5.3**) were centrifuged briefly and 2 µL of ligation reaction was added to the cells mixing by gently flicking the tube. Cells and ligation reactions were incubated on ice for 20 minutes. After incubation, cells were heat shocked at 42°C for 45-50 seconds before returning to ice for a further 2 minutes. 950 µl of Super Optimal Broth (SOC) medium (NEB) was added to the cells and incubated at 37°C for 1.5 hours with shaking (150 rpm). 100 µl of each transformation was plated onto Ampicillin IPTG/X-gal plates. Plates were then incubated overnight at 37°C. The next day, white colonies were selected, and 50 ml of bacterial cells were grown in bulk in LB broth containing ampicillin shaking overnight at 37°C.

2.10.4.6 Purification of plasmid DNA using ZymoPURE™ Plasmid midiprep Kit

After overnight incubation, 50 ml of bacterial culture was centrifuged at 3,400 xg for 10 minutes to pellet the cells. Plasmid purification using ZymoPURE™ Plasmid Midiprep Kit was performed according to the manufacturer's instructions. In brief, the pellet was resuspended in 8 ml of ZymoPURE™ P1, followed by addition of 8 ml of ZymoPURE™ P2 and mixing by gently inverting the tube before leaving to sit for 2-3 minutes, allowing cells to completely lyse. Following this, 8 ml of ZymoPURE™ P3 added to neutralise the solution and the clear lysate was filtered through the ZymoPURE™ syringe filter into a conical tube to remove the cell debris. Next, 8 ml of ZymoPURE™ Binding Buffer was added and mixed thoroughly by inverting the capped tube. Processing of the lysate was continued using the vacuum protocol. The solution

was added into the ZymoPURE™ III-P column assembly and vacuum was applied until all solution had passed through the column. With the vacuum off, 2 ml of ZymoPURE™ Wash 1 was added to the column and allowed to completely pass through. Next, 2 mL of ZymoPURE™ Wash 2 was added to the column and allowed to completely pass through, repeating this step twice. After this step, the column was placed in a collection tube and centrifuged at 10,000 xg for 1 minute. Finally, the column was placed into a clean 1.5 mL Eppendorf tube and 200 µL of ZymoPURE™ Elution Buffer was added directly to the column matrix allowing to sit for 2 minutes before centrifuging for 1 minute at 10,000 xg to collect the purified plasmid DNA.

2.10.4.7 Restriction enzyme digest

To isolate products from the pGEM vector, DNA was digested using a double restriction enzyme digest. Restriction enzymes, SalI-HF and NheI-HF, were ordered from New England Biolabs. DNA digest was performed following the reaction in **Table 9** and digested for 2 hours at 37°C, followed by a 20 min inactivation period at 80°C. Digested products were run on a 1.5% EtBr gel to check they were fully digested. Products were excised from the gel and purified using Qiagen gel extraction, **section 2.3.7**.

Table 9 Restriction enzyme double digestion set up reaction

Component	50 µl Reaction
DNA	1 µg
10X rCutSmart Buffer	5 µL (1X)
NheI-HF	1.0 µL (20 units)
SalI-HF	1.0 µL (20 units)
Nuclease-free Water	to 50 µL

PmirGlo Vector was dephosphorylated using Shrimp Alkaline Phosphatase (rSAP) to prevent the digested ends re-ligating. After

dephosphorylation, plasmid was purified using filter plates, **Section 2.3.6** and eluted in 25ul.

2.10.4.8 Sub-cloning into pmirGLO Dual-Luciferase miRNA target expression vector

Digested inserts (**Section 2.10.4.7**) were sub-cloned into the pmirGLO Dual-Luciferase miRNA Target Expression Vector (Promega). Vector and insert ratios were determined using the following equation:

$$\frac{Vector\ (ng)\ x\ Insert\ (Kb)}{Vector\ (Kb)}\ x\ Ratio\ of\ insert{:}\ vector = Insert\ (ng)$$

Ligation reactions were performed in accordance with **Table 8**. After ligation, transformation and midi prepping (**Sections 2.10.4.5 and 2.10.4.6)** colony PCRs (**Section 2.3.4**) were performed. The identity of amplified products from colony PCR was confirmed by Sanger sequencing.

2.10.5 HEK293t cell culture

Human embryonic kidney 293 cells (HEK293t) cells were cultured in Dulbecco's Modified Eagle Medium (ThermoFisher), supplemented with 10% Fetal Bovine Serum (ThermoFisher) and Antibiotic-Antimycotic (Gibco), until confluent. Once confluent, cells were washed with PBS to remove traces of medium and lifted using Trypsin. Cells were seeded into 96-well plates at a seeding density of $2.5x10^4$.

2.10.6 Co-transfecting cells with DNA constructs and miRNAs

HEK293t cells were co-transfected using TransIT®-LT1 Transfection Reagent (Mirus) with a final concentration of 50ng – DNA constructs and 1.5 µM of miRNA mimics (synthetic microRNAs) in Opti-MEM™ reduced Serum

Medium (ThermoFisher), following manufacturer's protocol. Cells were incubated at 37°C for 48 hours.

2.10.7 Luciferase assay

To quantify luciferase expression levels, the Dual-Glo® Luciferase Assay (Promega) was used following the Manufacturer's protocol. In brief, Dual-Glo® Reagent was added to wells equal to the volume of culture medium in the well. Cells were incubated for 10 minutes to allow cell lysis to occur, and then firefly luminescence was measured on an Orion L Microplate Luminometer (Titertek Berthol). Following this, Dual-Glo® Stop & Glo® Reagent was added equal to the original culture medium volume to each well, incubated for 10 minutes and *Renilla* luminescence was measured using the same luminometer platform. The ratio of luminescence from the experimental reporter gene (firefly luciferase) to luminescence levels expressed from the control reporter (*Renilla* luciferase) was calculated and normalised against the ratio of control wells.

3. Exploring the epidemiology and genetic architecture of a large British and Czech FECD patient cohort

3.1 Introduction

It is now well established that expansion of an intronic triplet-repeat (defined as ≥50 repeats), termed CTG18.1, is by far the most common genetic risk factor for FECD. Approximately 75-80% of Caucasian FECD patients investigated to date have been identified to carry an expanded allele (Luther et al., 2016; Mootha et al., 2014; Skorodumova et al., 2018; Zarouchlioti et al., 2018) and it has previously demonstrated, that an expanded copy of this CTG repeat confers >76-fold increased risk for developing FECD (Zarouchlioti et al., 2018).

While the incidence of CTG18.1 expansion-positive FECD is still present in other ethnicities it has been found to be much less abundant in non-European populations. For example, the occurrence of expansions in Thai, Chinese, Japanese and Indian FECD cohorts have been reported to be 39%, 44%, 26% and 34% respectively (Nakano et al., 2015; Nanda et al., 2014; Okumura, Puangsricharern, et al., 2019; Xing et al., 2014). Similarly, a study conducted in an African Americans FECD patient cohort reported only 35% of cases harbour an expanded CTG18.1 allele compared to 62.5% of Caucasian patients (Eghrari et al., 2017). **Table 10** provides a summary of CTG18.1 expansion frequency reported amongst FECD patients and control cohorts of varying ethnicities. These studies suggest the prevalence of CTG18.1 expansion-positive FECD varies between different ethnic groups and hence suggests that additional genetic cause(s) are likely more commonly driving disease in non-European populations.

Table 10 Summary of CTG18.1 genotyping studies performed across ethnically diverse Fuchs endothelial corneal dystrophy (FECD) patient and control cohorts. This table represents data I compiled for a collaborative review article (Fautsch et al., 2021).

Ethnicity, as reported in original study	FECD cases with CTG18.1 expansion (%)	Controls with CTG18.1 expansion (%)	Reference
British Caucasian	77.3%†	4.2%†	Zarouchlioti et al. (2018)
Czech Republic	81.1%†	–	Zarouchlioti et al. (2018)
American	79%†	3%†	Wieben et al. (2012); Mootha et al. (2014); Vasanth et al. (2015); Eghari et al. (2017a)
	73%*	7%*	
	62%*	3.6%*	
	63%*		
German	77†;	10.8†	Foja et al. (2017); Okumura et al. (2019a); Luther et al. (2016)
	79%†	11.5%†	
	79%†		
Russian	72%*	5%*	Skorodumova et al. (2018)
Belgian	–	8%*	Del-Favero et al. (2002)
Swedish	–	3%*	Del-Favero et al. (2002)
Croatian	–	6%*	Del-Favero et al. (2002)
Danish	–	3%*	Del-Favero et al. (2002)
Scottish	–	7%*	Del-Favero et al. (2002)
Northern European	–	3%†	Breschel et al. (1997)
Australian	51%*	5%*	Kuot et al. (2017)
Thai	39%*	0%*	Okumura et al. (2019c)
Singaporean Chinese	44%*	1.7%*	Xing et al. (2014)
Japanese	26%†	0%†	Nakano et al. (2015)
Indian	17%†	3%†	Rao et al. (2017)
Inidan (Odisha and West Bengal)	34%†	5%†	Nanda et al. (2014)
African American	35%*	–	Eghari et al. (2017a)
*, >40 repeats used as criteria for expansion; †, >50 repeats used as criteria for expansion; -- not screened.			

Several studies have reported a higher predominance of FECD in females in comparison to males (Jun, 2010; Kitagawa et al., 2002; Ong Tone et al., 2021; Vasanth et al., 2015). In a large study involving 64 families, women were seen to be affected 2.5 times more frequently than men and additionally were reported to be more severely affected (Krachmer et al., 1978). Moreover, Afshari *et al* found a female-male ratio of 3.5:1 when reviewing clinical records of FECD patients who underwent keratoplasty for disease (Afshari et al., 2006). A ratio as high as 3.7:1 towards females has also been noted in a study conducted in Japanese subjects (Kitagawa et al., 2002). Currently, the biological explanation for the higher reported female prevalence of FECD in females has not yet been established. However it has been suggested that sex hormones may play a role in this female predominance. The sex-steroid receptors oestrogen receptor beta, androgen receptor, and progesterone receptor, are all expressed in CECs of both sexes and it is possible that these receptors influence the function of CECs (Suzuki et al., 2001) however the expression profiles of these receptors are yet to be investigated between females and males. Furthermore, recent studies have proposed a potential role of reactive oestrogen quinones in FECD pathogenesis. Certain oestrogens, including oestradiol and estrone, are oxidatively metabolised to form catechol oestrogens, which in turn further oxidised to genotoxic oestrogen quinones. These quinones react with DNA resulting in significant cytotoxicity. These could have a role in FECD pathogenesis as previous studies have shown that acquired DNA damage results in apoptosis of CEC and thus FECD (Miyajima et al., 2019; Ong Tone et al., 2021).

In this results chapter, I explore the epidemiology of a large cohort of FECD patient cohort recruited from MEH and GUH. I present CTG18.1

genotyping data for the full cohort with the aim to explore and further elucidate on corresponding existing phenotype-genotype correlations. Furthermore, I explore the germline transmission of CTG18.1 expansions with families recruited to this study with the aim to gain insights on how the repeat is transmitted through generations.

3.2 Results

3.2.1 Patient recruitment

A total of 990 individuals were recruited to this study, 602 females and 388 males. This represents an updated version of the cohort previously published by Zarouchlioti et al. in 2018 with a total of 450 FECD patients (Zarouchlioti et al., 2018). Participants either had clinical signs of FECD (numerous corneal guttae on slit-lamp biomicroscopy) or had corneal transplantation surgery (either penetrating or endothelial keratoplasty) for FECD. The mean age of the cohort at time of recruitment was 69-years-old. Patients were recruited to the study upon referral to MEH or GUH.

3.2.2 *TCF4* CTG18.1 genotyping

A combination of STR and TP-PCR assays were used to genotype the CTG18.1 alleles in all FECD patients recruited to the study. All samples were genotyped using the STR assay in the first instance. For samples which appeared to be homologous for a smaller allele using the STR genotyping method, TP-PCR was performed to confirm the presence or absence of a larger expanded allele. A highly significant association between expansion of the CTG18.1 trinucleotide repeat (conservatively defined as ≥50 repeats) and FECD was identified (OR = 94.59; 95% confidence interval (CI): 60.50-148.74; $p = 6.52 \times 10^{-78}$) in the European-only portion of the cohort (n = 800; **Table 11**).

CTG18.1 expansion lengths of patients with AMD were genotyped by Zarouchlioti et al. previously and used as an ethnically matched control population for the purpose of this study, **Table 12**. For the AMD cohort, 4.18% (23/550) had one expanded copy (≥50 repeats) of the CTG18.1 allele, in line with reports from other unaffected populations screened for control purposes (Breschel et al., 1997; Wieben et al., 2012), and none were found to have two expanded alleles. In contrast, 77.78% (770/990) of the FECD cohort had one or more expanded copies of the CTG18.1 allele, of which 4.14% (41/990) had bi-allelic expansions. The distribution of CTG18.1 length between FECD cohort and AMD control group can be seen in **Figure 12**.

Table 11 CTG18.1 expansion status in the Fuchs endothelial corneal dystrophy (FECD) Cohort.

	N	NE/NE	E/NE	E/E	≥1 E
Total FECD cohort (mean age = 69)	990	22.22% (220/990)	73.64% (729/990)	4.14% (41/990)	77.78% (770/990)
Females (mean age = 69)	602	26.74% (161/602)	69.60% (419/602)	3.60% (22/602)	73.26% (441/602)
Males (mean age = 67)	388	15.21% (59/388)	79.89% (310/388)	4.90% (19/388)	84.79% (329/388)
Subjects recruited at MEH	584	22.95% (134/584)	72.60% (424/584)	4.45% (26/584)	77.05% (450/584)
European (65.6%)	394	17.77% (70/394)	76.14% (300/394)	6.09% (24/394)	82.23% (324/394)
Non-European (15.8%)	88	39.77% (35/88)	60.23% (53/88)	0.00% (0/88)	60.23% (53/88)
Unknown (18.6%)	102	28.43% (29/102)	69.61% (71/102)	1.96% (2/102)	71.57% (73/102)
Subjects recruited at GUH (Caucasian)	406	21.18% (86/406)	75.12% (305/406)	3.70% (15/406)	78.82% (320/406)
Expanded alleles are defined as ≥50 CTG repeats. Abbreviations are as follows: NE, non-expanded CTG18.1 allele; E, expanded CTG18.1 allele; MEH, Moorfields Eye Hospital; GUH, General University Hospital in Prague.					

Table 12 Summary of CTG18.1 Genotyping Data in the age-related macular degeneration (AMD) as a control cohort.

	N	NE/NE	E/NE	E/E	≥1 E
AMD cohort (mean age = 78)	550	95.82% (527/550)	4.18% (23/550)	0.00% (0/550)	4.18% (23/550)
Females (mean age = 78)	356	96.07% (342/356)	3.93% (14/356)	0.0% (0/356)	3.93% (14/356)
Males (mean age = 78)	194	95.36% (185/194)	4.64% (9/194)	0.0% (0/194)	4.64% (9/194)
Expanded alleles are defined as ≥50 CTG repeats. Abbreviations are as follows: NE, non-expanded CTG18.1 allele; E, expanded CTG18.1 allele; AMD, age-related macular degeneration.					

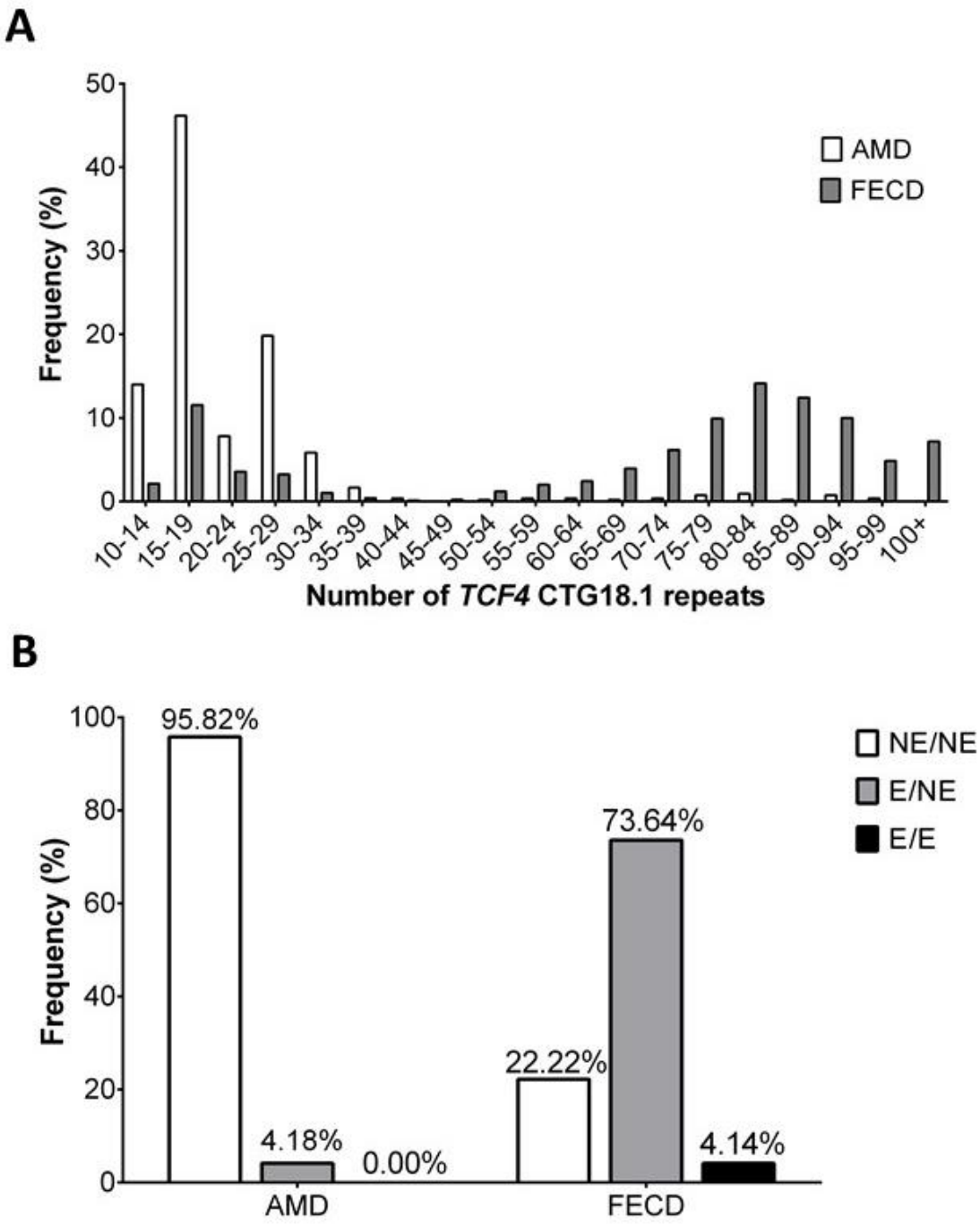


Figure 12 Expansion of CTG18.1 is associated with Fuchs endothelial corneal dystrophy (FECD) in a British and Czech Cohort (A) Frequency histogram comparing relative distribution of CTG18.1 repeat length in FECD and age-related macular degeneration (AMD) cohorts. The longest allele detected, per individual tested, is shown. In total the FECD (grey) and AMD (white) cohorts comprised 900 and 550 individuals, respectively. **(B)** Bar chart illustrating the relative frequency of individuals with both alleles non-expanded (NE/NE), one expanded allele (E/NE), or both alleles expanded (E/E) in both the FECD and AMD cohorts. Expanded alleles are defined as ≥50 CTG repeats. CTG18.1 genotyping data for a subset of the FECD cohort (n=450) and the total AMD cohort has previously been published by Zarouchlioti et al., 2018.

3.2.3 Ethnicity

A UK Biobank Axiom™ Array was used to genotype the entire genome-wide SNPs across the entire FECD cohort, which at the time of this project was conducted comprised of 659 individuals. PhD Student Anita Szabo analysed the output of the array genotyping data to predict the ethnicity of all samples (detailed in methods **Section 2.6**). All FECD samples acquired from the Czech Republic were identified to be of European descent, as expected. FECD samples acquired from MEH had a diverse range of self-reported ethnicities, as listed in **Table 11**. **Figure 13** shows a breakdown of self-reported ethnicities of 571 samples recruited at MEH samples that were also analysed by the genome-wide SNP array. From the reported ethnicities listed in their MEH hospital records, 99 samples did not have a self-reported ethnicity and their ethnicity was listed as 'unknown. A further 36 samples had their ethnicity listed as 'other'.

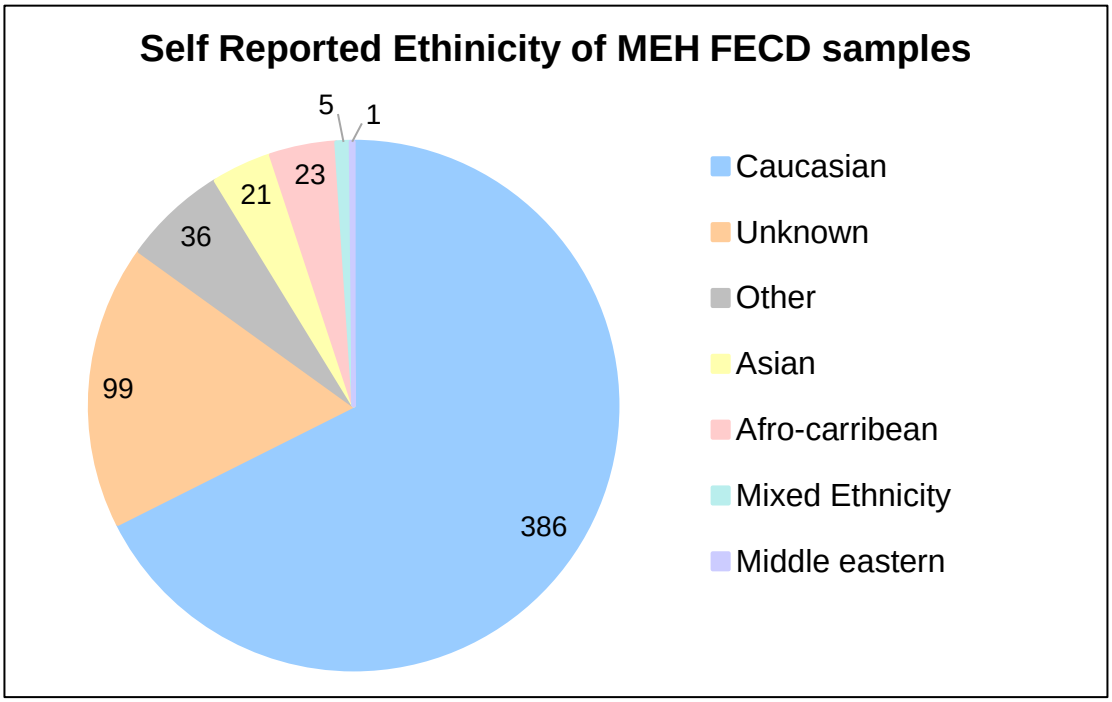


Figure 13 Pie chart showing the breakdown of the self-reported ethnicity of the same 571 Moorfields Eye Hospital samples.

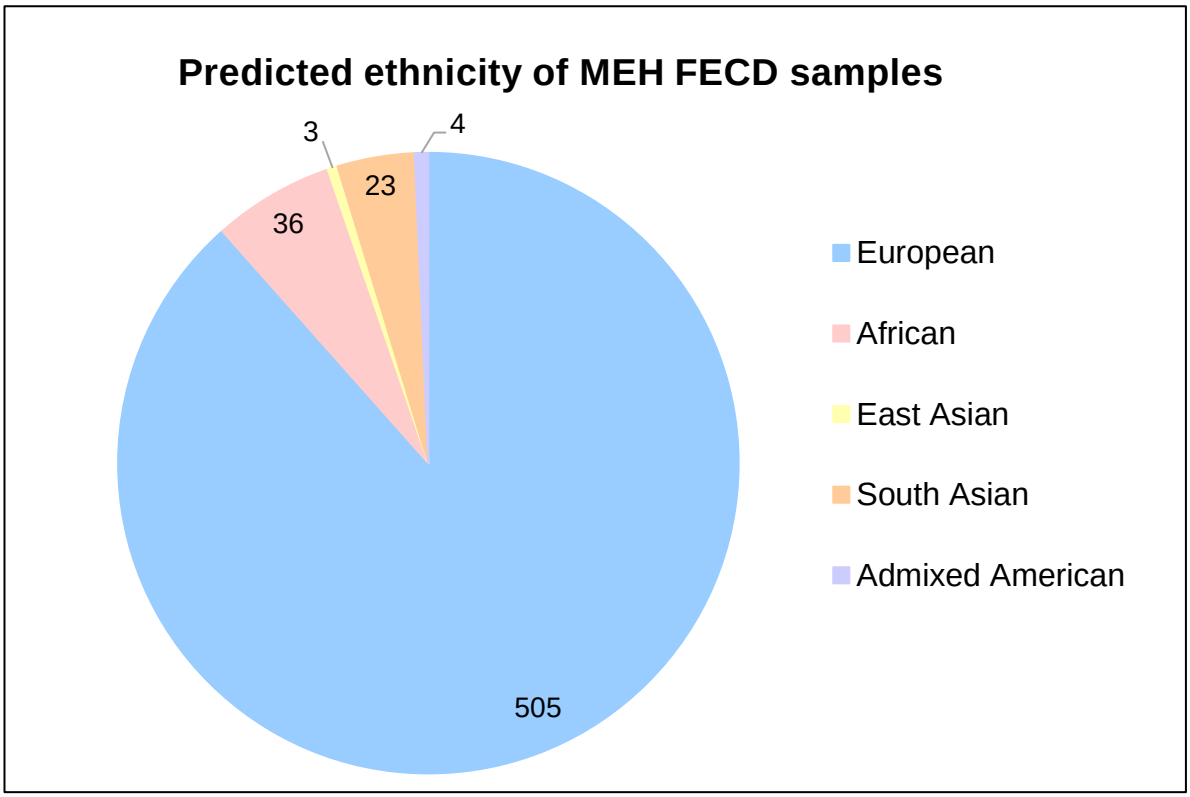


Figure 14 Pie chart showing the breakdown of the genome-wide SNP array predicted ethnicity of the same 571 Moorfields Eye Hospital samples.

Of the 36 samples that had self-reported ethnicities of 'other', predicted ethnicities from the SNP array data suggested of these, 31 samples were of European descent, 2 of African descent, 2 of Admixed American descent and one of southeast Asian descent, **Figure 14**.

Of the 100 samples where there was no self-reported ethnicity and therefore listed as 'unknown' computed ethnicities predicted from the SNP array data suggested these to be of 85 samples to be of European descent, 9 to be of African descent, 2 of southeast Asian descent, 2 of Admixed American descent and one to be of east Asian descent.

Ethnicity predictions generated from the genome wide SNP array data predicted that the vast majority of the outputs matched the self-reported

ethnicity for those of which supplied one. Only a small number of discrepancies were identified. One patient that was self-reported as White British (Caucasian) had a predicted computed ethnicity of Southeast Asian based on the SNP array genotyping data. Another patient self-reported themselves with an ethnicity as Middle Eastern however SNP array genotyping data predicted this sample to be of European descent. Furthermore, five patients had self-reported themselves as having a mixed ethnicity of white/black Caribbean/African however, computed ethnicity predicted three of these patients to be of European descent and two to be of African descent.

There were no discrepancies amongst self-reported and predicted ethnicities, however, those of which were of a dual ethnicity were assigned to a single ethnicity they were predicted to prominently share DNA with. This is a potential drawback of this methodology.

Acquiring the ethnicity of all samples in the cohort is principally important in this study as previous associations between CTG18.1 repeat length and FECD has now been replicated within numerous multi-ethnic cohorts and demonstrate the correlation between CTG18.1 expansion and FECD is typically lower, in other non-Caucasian ethnic groups investigated to date, See **Table 10.** We see this same pattern emerging in our cohort. With the self-reported ethnicity data of the MEH cohort, 82.23% of the Caucasian subjects carried an expanded CTG18.1 repeat (<50 repeats) in comparison to only 60.23% of subjects who were of another ethnicity.

The same pattern emerges with the predicted ethnicity data with European subjects having the highest expansion rates and being the only ethnic group to include cases of bi-allelic expansions, **Table 13**. Although n numbers

are low for non-Europeans within our cohort, they have a notably lower occurrences of *TCF4* repeat expansions.

Table 13 Summary of CTG18.1 genotyping across multi-ethnic sub-groups for Fuchs endothelial corneal dystrophy (FECD) patients in which ethnicity was predicted for.

	N	NE/NE	E/NE	E/E	≥1 E
Subjects recruited at MEH	571	22.94% (131/571)	72.68% (415/571)	4.38 (25/571)	77.06% (440/571)
European	505	17.43% (88/505)	77.62% (392/505)	4.95% (25/505)	82.57% (417/505)
African	36	80.50% (29/36)	19.50% (7/36)	0.00% (0/36)	19.50% (7/36)
Eastern Asian	4	75.00% (3/4)	25.00% (1/4)	0.00% (0/4)	25.00% (1/4)
Southeast Asian	23	39.13% (9/23)	60.87% (14/23)	0.00% (0/23)	60.87% (14/23)
Admixed American	3	66.66% (2/3)	33.33% (1/3)	0.00% (0/3)	33.33% (1/3)

Expanded alleles are defined as ≥50 CTG repeats. Abbreviations are as follows: N=number of subjects; NE, non-expanded CTG18.1 allele; E, expanded CTG18.1 allele.

3.2.4 Exploring of sex distribution among FECD patients with and without CTG18.1 expansion

In total 60.8% (602/990) of the total recruited cohort are female, in accordance with previous reports of a female propensity for FECD (Afshari et al., 2006; Krachmer et al., 1978). On this basis we were interested to see how sex distribution varied within genetically refined CTG18.1 expansion-positive and negative subsets of the cohort. Both subgroups still maintained a female bias, however, this effect was more pronounced in the CTG18.1 expansion-

negative group, in which 73.18% (161/220) of cases were female compared with 57.28% (441/770) of cases in the CTG18.1 expansion-positive group (**Figure 15.A**). Overall, a significantly higher proportion of affected females were present within the CTG18.1 negative portion of the cohort in comparison to the CTG18.1 positive cohort (chi-square 18.1724; p= 0.00002)

Next ‘age-at-recruitment’ between these subgroups were compared. The median age at recruitment between the sexes did not significantly differ in either the total cohort (69 years) or the CTG18.1 expansion-positive group (69 years). In addition, although the median age differed greatly, 67 years and 59 years for women and men, respectively, this was not a significant difference (p = 0.0560, Fisher’s exact test; **Figure 15.B**)

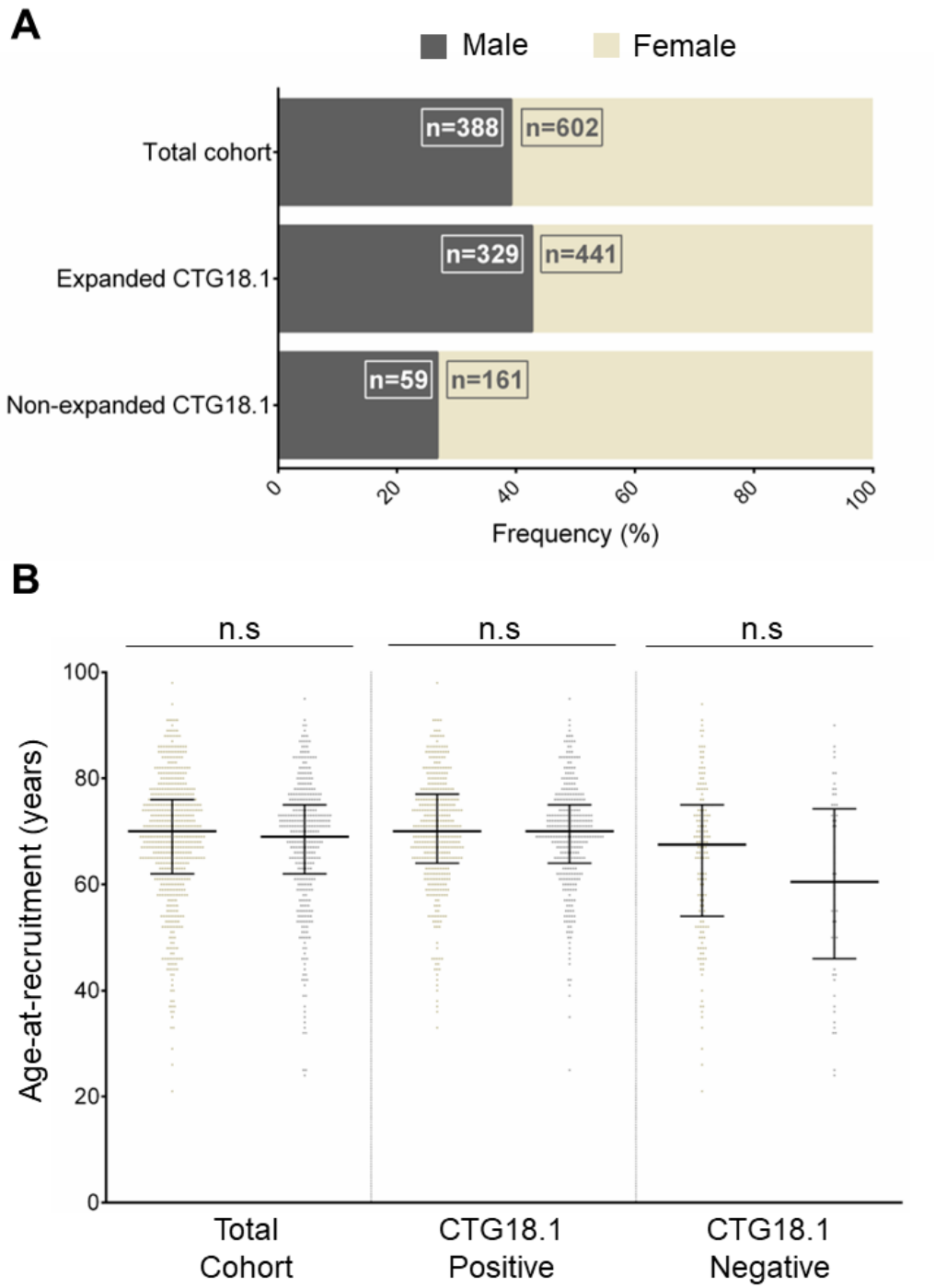


Figure 15 Comparison of sex distribution and age at recruitment among FECD patients with and without CTG18.1 expansions. (A). Stacked bar chart illustrating the sex distribution in total FECD cohort, and within the CTG18.1 expansion-positive and -negative subgroups. A significantly higher ratio of females was identified within the CTG18.1 expansion-negative group compared to both the total cohort and the CTG18.1 expansion-positive group **(B)** Scatterplot illustrating the distribution of age at recruitment for all individuals recruited to this study. Data is shown for the total cohort and the subgroups defined by sex and CTG18.1 expansion status. The median age between the sexes did not significantly differ in any sub-cohort.

3.3.5 Correlation of CTG18.1 length with age at recruitment

Figure 16 illustrates the correlation between the largest CTG18.1 repeat length detected, per individual, and the age at which FECD patients were recruited to this study. A very weak correlation was observed, Spearman's rank correlation coefficient (r) = - 0.083; p = 0.025.

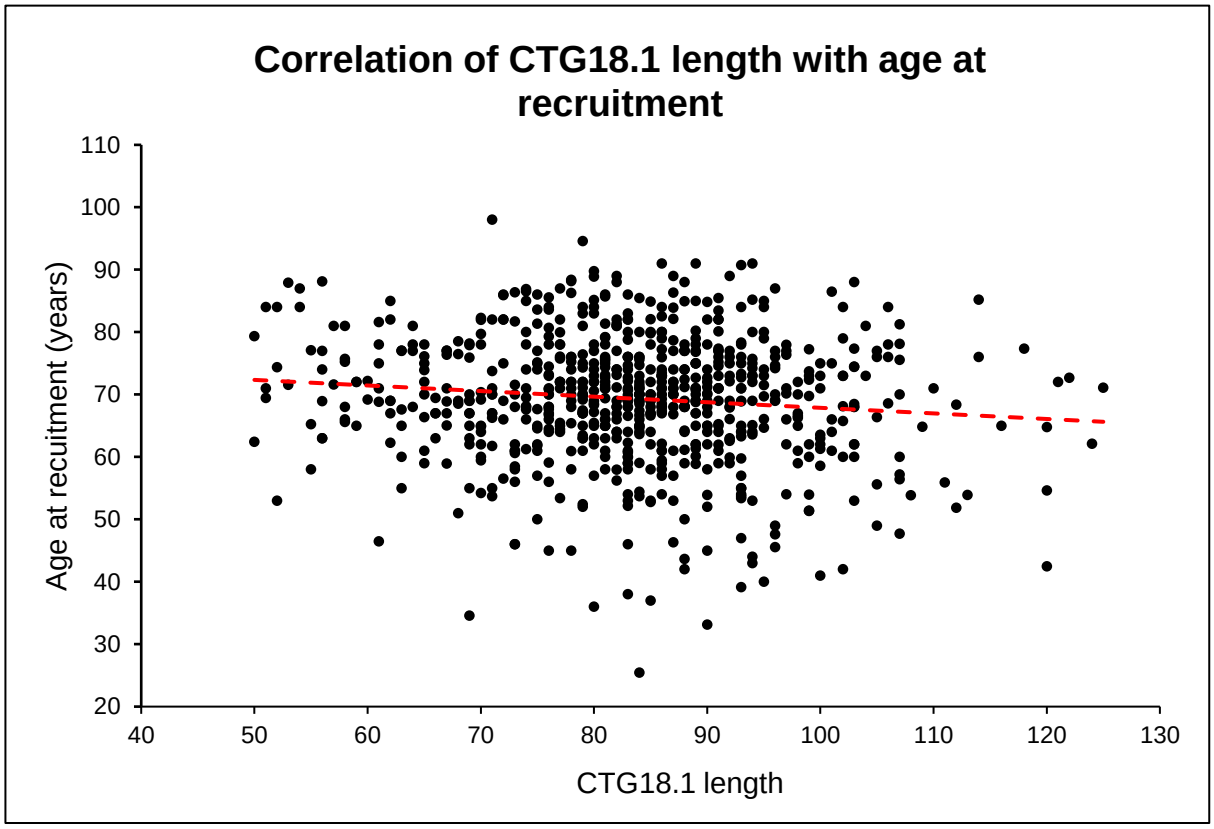


Figure 16 Scatterplot demonstrating the correlation between the repeat number of expanded CTG18.1 allele and age of Fuchs endothelial corneal patient (FECD) at the time of recruitment.

3.2.6 Correlation of CTG18.1 length with age of first surgery

To explore, using an alternative and potentially more relevant proxy for disease severity 'age of first corneal surgery' was next correlated with repeat modal CTG18.1 allele length.

Data of the first corneal graft was collated by clinicians Kirithika Muthusamy and Shane Liu for patients recruited at MEH and Petra Liskova for patients recruited at GUH. To further reduce limitations, patients who had prior cataract surgery were excluded as it is well established that the physical trauma of lens replacement surgery can damage the corneal endothelium and hence influence rates of endothelial decompensation (Walkow, Anders, & Klebe, 2000). A total of 230 FECD patients were included in this analysis, 197 from MEH and 33 from GUH. **Figure 17.A** demonstrated the correlation between age of first corneal transplantation surgery and modal CTG18.1 repeat length amongst FECD patients from both MEH and GUH. A stronger and statistically significant, but still objectively weak, correlation was observed between modal allele length and 'age-of-first-transplant', r =-0.154, p=0.019 (**Figure 17.A**), in comparison to the correlation of CTG18.1 length with 'age-at-recruitment (**Figure 16**). There was a stronger correlation observed amongst patients from GUH in comparison to MEH FECD patients, r = -0.22926, p= 0.19936 and r=-0.152, p=0.034, respectively (**Figure 17.B** and **Figure 17.C**).

Nevertheless, the association between the two variables within the GUH cohort is not considered statistically significant, this is likely due to the small n number as most patients had previously already undergone cataract surgery and therefore eliminated from this association study.

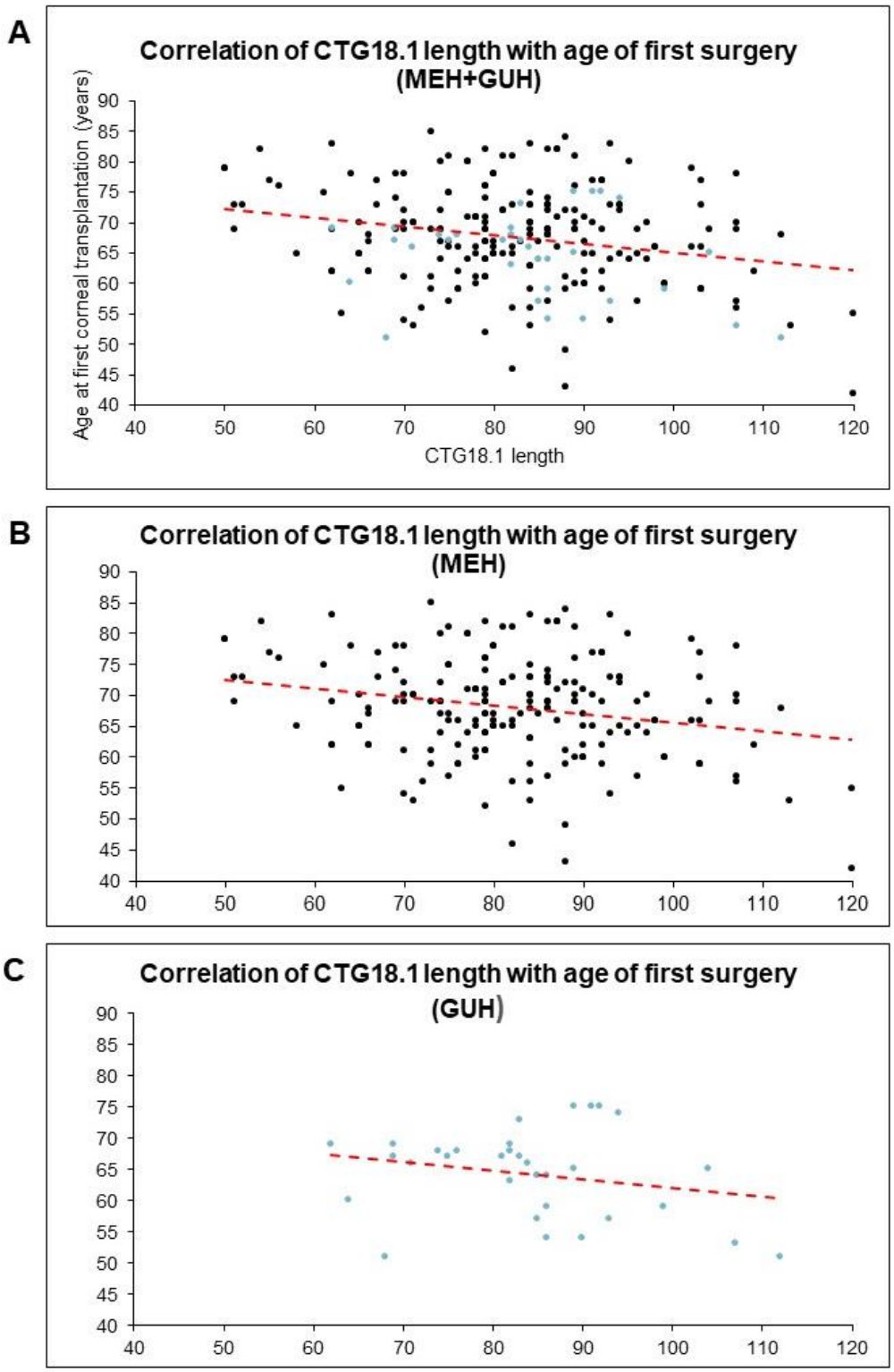


Figure 17 Scatterplots demonstrating the correlation between the repeat number of expanded CTG18.1 allele and 'age-at-first-corneal-transplant surgery' of Fuchs endothelial corneal patients (FECD). (A) Scatterplot showing the correlation between all 230 FECD patients in which 'age-at-first-corneal-transplant surgery' was available. **(B)** Scatterplot showing the correlation between FECD patients recruited at Moorfields Eye Hospital, London (MEH). **(C)** Scatterplot showing the correlation between FECD patients recruited at General University Hospital, Prague (GUH).

3.2.7 Bi-allelic CTG18.1 expansions

In total, 4.14% (41/990/758; **Table 11**) of the FECD cohort was identified to carry bi-allelic CTG18.1 expansions. There was no significant difference in the incidence of male and females carrying homozygous expanded alleles, 3.60% females versus 4.90% males. **Table 14** contains the repeat length of each CTG18.1 allele, age at recruitment and age of surgery if applicable. The mean ‘age-at-recruitment’ for this sub-cohort was 68-years-old, similar to the total FECD cohort. There was no evidence FECD phenotype progressed more rapidly in subjects with bi-allelic versus mono-allelic CTG18.1 expansions. An earlier ‘age-at-recruitment’ was noted for one patient, HOM_2 (43-years-old), however, the endothelium had not yet decompensated to the stage by which surgical intervention was required. Hence this case is an outlier in terms of recruitment to our study, given that the vast majority of case in our total cohort were recruited at the time of undergoing their first corneal transplant.

Table 14 Summary of Fuchs Endothelial Corneal Dystrophy (FECD) patients harbouring bi-allelic CTG18.1 expansions.

Subject ID	Gender	Ethnicity	CTG18.1 genotype	'Age-at-recruitment'	'Age-at-first-corneal-transplant surgery'
HOM_1	M	Unknown	91/102	60	59
HOM_2	F	Unknown	58/94	43	N/A
HOM_3	F	White Czech	52/96	45	N/A
HOM_4	M	White British	85/105	49	49
HOM_5	M	White British	71/113	54	53
HOM_6	F	White British	82/≥82	60	60
HOM_7	F	White Czech	61/≥61	63	N/A
HOM_8	M	White Czech	83/≥83	62	N/A
HOM_9	F	White British	80/120	65	55
HOM_10	M	White Czech	81/≥81	66	66
HOM_11	F	White British	67/94	65	64
HOM_12	M	White British	78/≥78	65	N/A
HOM_13	M	Unknown	81/≥81	66	66
HOM_14	M	White British	76/≥76	67	67
HOM_15	F	White Czech	62/88	61	63
HOM_16	M	White British	73/≥73	68	69
HOM_17	M	White British	74/≥74	68	66
HOM_18	F	White British	68/≥68	69	N/A
HOM_19	M	White British	53/92	69	69
HOM_20	M	White Czech	83/≥83	71	71
HOM_21	F	White Czech	62/97	71	71
HOM_22	F	White Czech	66/125	68	69
HOM_23	F	White British	78/≥78	72	71
HOM_24	M	White British	76/141	73	61
HOM_25	M	White Czech	57/80	74	74
HOM_26	F	White Czech	88/≥88	69	69
HOM_27	F	White Czech	76/≥76	69	75
HOM_28	M	White British	75/≥75	75	N/A
HOM_29	F	White British	86/≥86	76	72
HOM_30	F	White British	87/95	85	79
HOM_31	F	White British	72/114	85	85
HOM_32	F	White British	63/91/121	85	82
HOM_33	M	White British	82/≥82	89	N/A
HOM_34	F	White British	71/≥71	69	69
HOM_35	F	White Czech	67/≥67	71	71
HOM_36	M	White Czech	77/≥77	60	60
HOM_37	M	White British	81/≥81	70	70
HOM_38	F	White Czech	65/≥65	76	77
HOM_39	F	White British	90/≥90	74	74
HOM_40	M	White British	64/≥64	69	69
HOM_41	F	White Czech	66/≥66	73	71

3.2.8 Early-onset CTG18.1 expanded FECD

Typically, CTG18.1-mediated FECD is a late-onset disease affecting those above the age of 40 years; the mean age of individuals at the time of recruitment to this study was 69 years, however this is not reflective of the age-of-onset of disease, given the vast majority of these cases were recruited at the time they underwent their first corneal transplant for FECD. In our cohort we noted four CTG18.1-mediated FECD subjects recruited with an age-of-onset of 39 years or younger, **Table 15**. From this sub-cohort two individuals, EAO_3 and EAO_4, reported a positive family history and family members were recruited, **Figure 18**.

Table 15 Summary of Fuchs endothelial corneal dystrophy (FECD) patients with atypical early-onset phenotype harbouring a CTG18.1 repeat expansion

Subject ID	Gender	Self reported ethnicity	CTG18.1	'Age-at-recruitment'	'Age-at-first-corneal-transplant surgery'
EAO_1	M	Asian Indian	12/69	35	N/A
EAO_2	M	White British	12/93	39	N/A
EAO_3	F	White British	18/90	33	33 years
EAO_4	F	White British	16/84	25	In third decade of life

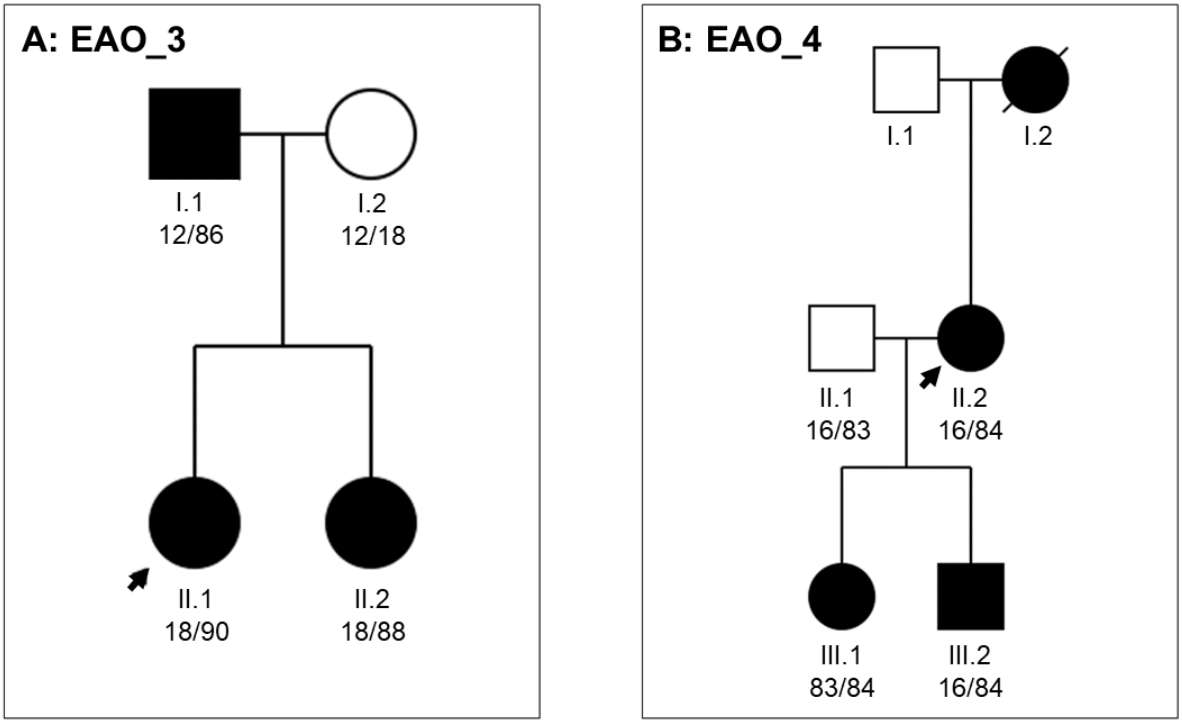


Figure 18 Pedigree's of two families presenting with an atypical early-onset Fuchs endothelial corneal dystrophy (FECD) phenotype and expanded CTG18.1 alleles. (A) EAO_3 33-year-old proband, II.1; **(B)** EAO_4 25-year-old proband II.2.

The proband from the first early-onset family is a 33-year-old female (**Figure 18.A: II.1/Table 15: EAO_3**) who has a CTG18.1 genotype of 18/90, she also presented with symptoms of keratoconus. Her father, a 62-year-old man, with a CTG18.1 genotype of 12/86, also has a FECD phenotype, but his age-of-onset is unknown. Her mother is unaffected and has a bi-allelic non-expanded CTG18.1 alleles with a genotype of 12/18. The proband's sister, 38-years-old, also presents a FECD phenotype, but does not have keratoconus, with a CTG18.1 genotype of 18/88. The CTG18.1 expanded allele in the family has been unstably transmitted from the father to his two children. Within the family the repeat length was of a typical length observed in blood-derived in FECD patient genomic DNA samples, and therefore suggests the length of repeat alone cannot explain the early-onset phenotype observed.

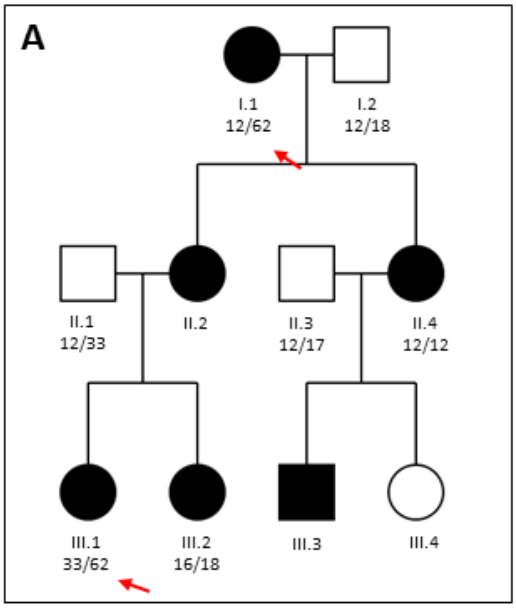
The second early-onset family, **Figure 18.B**, presents three generations with an early-onset phenotype. The proband, now is a 62-year-old female (**Figure 18.B: II.2/Table 15: EAO_4**) with a CTG18.1 genotype of 16/84, underwent her first graft in her third decade of life and since had her left eye regrafted 6 times. She reported her mother, now deceased, also had FECD from an early age. Her son, a 25-year-old, is also affected with an early-onset phenotype and began presenting early features of the disease at the age of 14 years. He carries a CTG18.1 genotype of 16/84 and also had congenital cataracts. Upon recruitment of her daughter, 32-years of age, a clinical eye examination detected early features indicative of FECD at the age of 20-years. Genotyping identified the daughter as carrying bi-allelic CTG18.1 expansions. For segregation reasons, genotyping of the proband's unaffected husband, a 61-year-old male, identified a CTG18.1 genotype of 16/83.

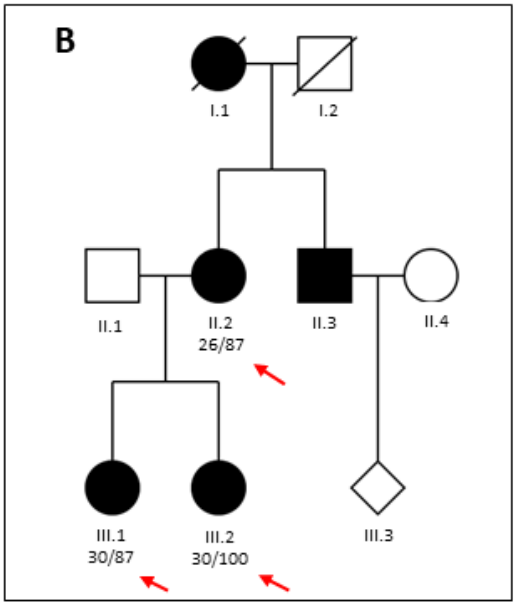
3.2.9 Parent-child transmission of CTG18.1 repeat length

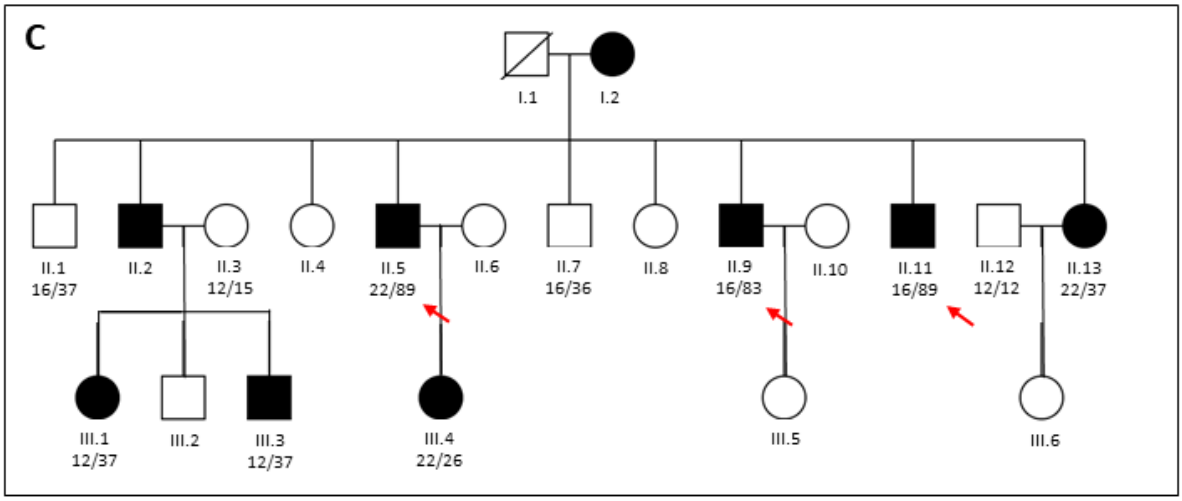
Here I have explored the transmission of CTG18.1 expansions in a total of 10 families, **Figure 19; Table 16.** Due to the difficulty of recruiting large families with FECD and given that approximately 4% of the general population carries an expanded allele, large families previously recruited with PPCD for part of our wider research program (Davidson et al., 2016) were screened for the CTG18.1 expansion and families positive for carrying an expanded CTG18.1 allele are reported here.

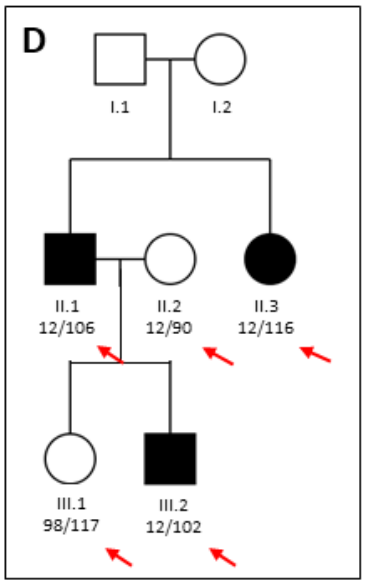
Table 16 Summary of maternal and paternal parent-child transmission of CTG18.1 expansion-positive alleles

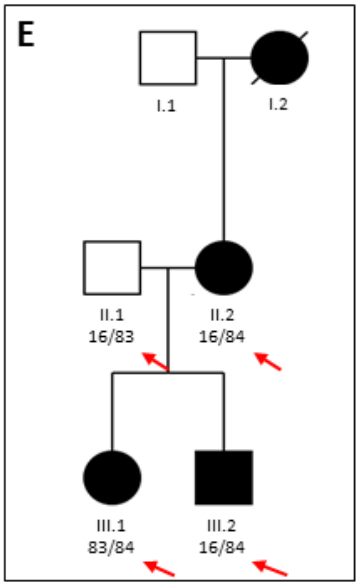
	Repeat contracted	Remained Stable	Repeat expanded
Maternal Transmission	1	1	7
Paternal transmission	2	2	3


A
I.1
12/62
I.2
12/18
II.1
12/33
II.2
II.3
12/17
II.4
12/12
III.1
33/62
III.2
16/18
III.3
III.4


B
I.1
I.2
II.1
II.2
26/87
II.3
II.4
III.1
30/87
III.2
30/100
III.3


C
I.1
I.2
II.1
16/37
II.2
II.3
12/15
II.4
II.5
22/89
II.6
II.7
16/36
II.8
II.9
16/83
II.10
II.11
16/89
II.12
12/12
II.13
22/37
III.1
12/37
III.2
III.3
12/37
III.4
22/26
III.5
III.6


D
I.1
I.2
II.1
12/106
II.2
12/90
II.3
12/116
III.1
98/117
III.2
12/102


E
I.1
I.2
II.1
16/83
II.2
16/84
III.1
83/84
III.2
16/84

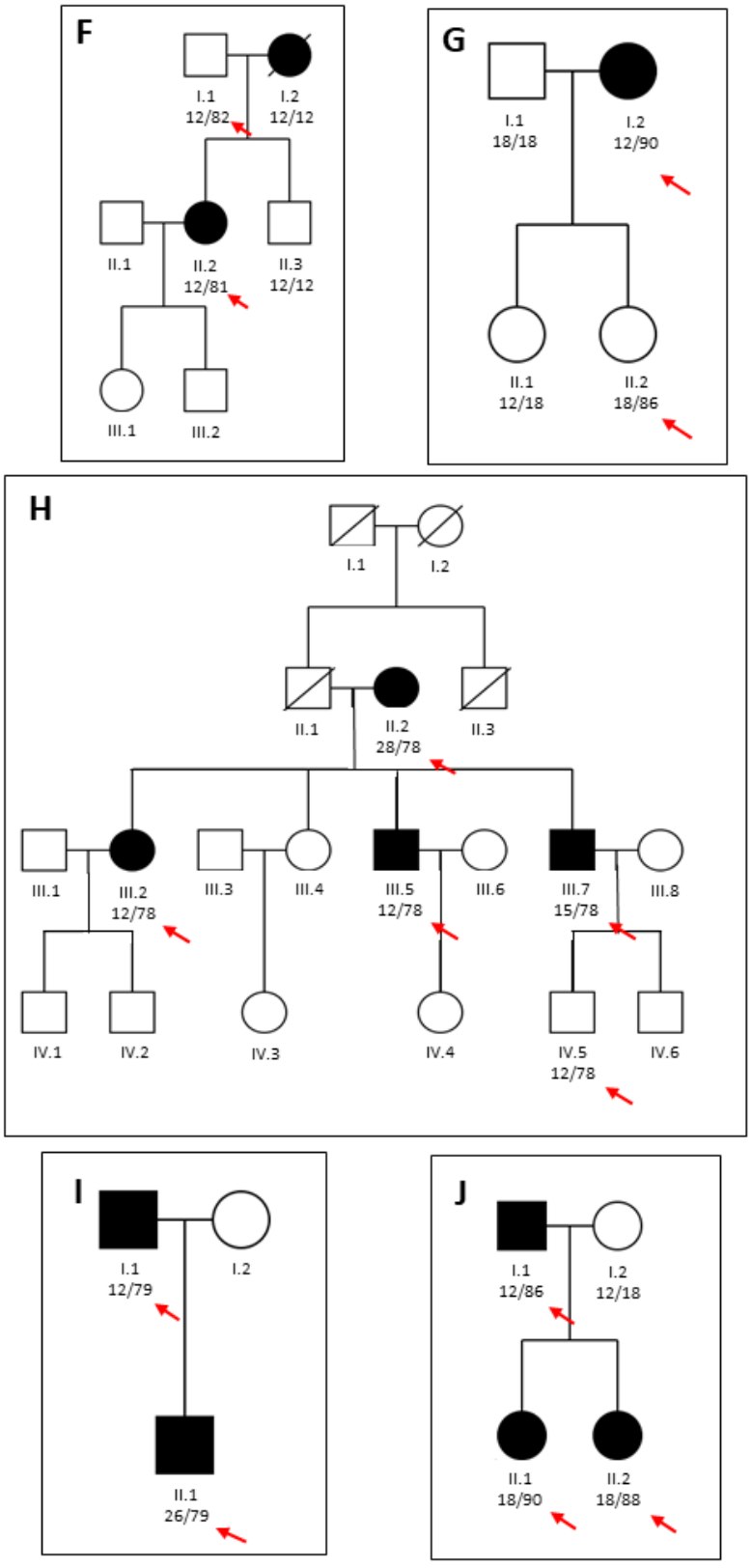


Figure 19 Germline transmission of expanded CTG18.1 repeat alleles within ten families, A-J, affected with either Posterior polymorphous corneal dystrophy (PPCD; families A-C) or Fuchs endothelial corneal dystrophy (FECD; families D-J). Red arrow highlights expanded CTG18.1 alleles within the pedigrees.

An expanded CTG18.1 allele comprising 62 repeat units was observed to be stably transmitted through two generations via the maternal germline in family A (PPCD). Similarly, an expanded CTG18.1 allele was detected to be transmitted across 3 generations via the maternal germline in family B (PPCD). The first transmission observed from individual II.2 to III.1 remained stable with 87 repeats, whereas transmission to individual III.2 the allele was found to expand further to 100 repeat units (**Figure 19.A; 18.B**).

The parental origin of the expanded CTG18.1 allele in Family C (PPCD) cannot be determined as neither parent was recruited to the study. The expanded allele was transmitted to three offspring (II.5, II.9 and II.11), subject II.5 and II.11 both had an expanded alleles of 89 repeat units and subject II.9 had an expanded allele of 83 repeat units. It cannot be confirmed if these alleles were inherited from independent parent or if the allele has expanded or contracted upon transmission from the same parent (**Figure 19.C**).

Transmission of the expanded CTG18.1 alleles present in Family D (FECD; **Figure 19.D**) appeared unstable upon each account of transmission. Both subject II.1 and sibling, subject II.3 were affected by FECD and carried expanded alleles of 106 and 116 repeat units, respectively, and both carried a smaller allele of 12 repeat units. Both parents of these siblings were unable to be recruited to the study to determine the origin of expanded alleles and therefore it is uncertain if the allele has expanded or contracted upon transmission or two separately inherited alleles. Subject II.1 unaffected partner, subject II.2, was recruited to the study and had a genotype of 12/90 along with their two offspring Subjects III.1 and III.2. Subject II.1 had a genotype of 98/117 and Subject II.2 a genotype of 12/102. In both subjects it is not able to

determine which alleles were inherited from which parent but in each account the allele length appeared unstable.

Both maternal and paternal transmission was observed in family E (early-onset FECD; **Figure 19.E**). Maternal transmission of an allele with a repeat length of 84 was observed from individual II.2 to III.1 and III.2. Paternal transmission of an expanded allele with a repeat length of 83 was also observed from II.1 to III.1. The repeat length remained the same on all three accounts of transmission.

The expanded CTG18.1 allele was transmitted through paternal transmission in family F (FECD; **Figure 19.F**). Contraction of the allele was observed from 82 repeat units (II.2) to 81 repeat units (III.2). Maternal transmission of an expanded CTG18.1 allele was observed in family G (FECD**; Figure 19.G**). The repeat length contracted from 90 repeat units (I.2) to 86 repeat units (II.2). Four accounts of maternal transmission of the expanded allele with 78 repeat units was observed in family H, who presented with FECD. Upon transmission the repeat length remained the same on all three accounts (**Figure 19.H**).

Stable, paternal transmission of the expanded CTG18.1 allele, with 79 repeat units, was observed in family I (FECD) from individual I.1 to II.1 (**Figure 19.I**). Two accounts of paternal transmission were observed in family J (early-onset FECD). On both accounts the repeat length appears unstable and expanded further from 86 repeat units (I.1) to 90 repeat units (II.1) and 88 repeat units (II.2).

3.3 Discussion

3.3.1 Patient recruitment

In total, 990 unrelated individuals with FECD were recruited to this study. The recruited patients typically had been referred to MEH or GUH due to reaching an advanced stage of disease and requiring surgical intervention. Thus, the cohort is biased towards cases with an advanced stage of disease. The mean age of this cohort at the time of sample recruitment is 69-year-old. Importantly this should be distinguished from age-of- disease onset. For the majority of the cohort phenotypic data on when the disease began to manifest in an individual is unavailable and difficult to define. There is large interpatient variability in reporting early symptoms of FECD due to a person's lifestyle, visual requirements, personal concerns and the subjective opinion of the clinician, all of which also influence when an individual would be both diagnosed and undergo surgery (Matthaei et al., 2019). Clinical diagnosis is also subjective, to date the most established classification system for clinical staging of FECD was proposed by Krachmer *et al*. in the 1970s (Krachmer et al., 1978). The grading scale is based on the occurrence and distribution of corneal guttae, identified by slit-lamp biomicroscopy, and the existence of corneal edema (Krachmer et al., 1978). However, patients recruited to our study at MEH or GUH were not classified in accordance with this system, as it fails to provide clinically useful information of benefit to patient care pathways and hence is not prioritised within a routine clinical setting.

3.3.2 *TCF4* genotyping

European subjects recruited from MEH and GUH were found to share a similar percentage of patients carrying at least one expanded allele (82.23% and 78.82%, respectively). These findings were also in-line with other relatively

smaller Caucasian cohorts investigated (Mootha et al., 2014; Okumura, Hayashi, Nakano, Tashiro, et al., 2019; Skorodumova et al., 2018). Hence these findings collectively suggest CTG18.1 expansions are the most common genetic risk factor for disease within the Caucasian population. As expected, the non-European sub-cohort (other, **Table 11**) recruited from MEH comprised considerably fewer subjects which carried at least one expanded CTG18.1 alleles (60.23%). Again, similarly to other reported studies it is comparatively less commonly associated with FECD in non-European populations (Eghrari et al., 2017; Nakano et al., 2015; Nanda et al., 2014; Okumura, Puangsricharern, et al., 2019; Xing et al., 2014). This suggests other distinct genetic causes of disease are more prevalent in non-European populations. However, other prominent genetic causes are yet to be identified. In attempts to further investigate this hypothesis, the non-European sub-cohort was further sub-categorised and predicted ethnicity data was generated from genome-wide SNP array analysis. Although n numbers are low for non-Europeans within our cohort, they have notably lower occurrences of *TCF4* repeat expansions, **Table 13**. These findings are indicative of a possibility of a founder effect, separate to the *TCF4* CTG18.1 expansion, occurring in these non-European ethnicities.

From work conducted using CRISPR-guided long-read SMRT sequencing it is now understood that the CTG18.1 expansion behaves dynamically in genomic DNA derived from the blood and display a diverse repeat size range in comparison to non-expanded alleles (Hafford-Tear et al., 2019; Wieben, Aleff, et al., 2019) This study also identified that repeat length instability positively correlates with mosaicism levels; i.e. the larger the repeat the more somatic instability is observed (Hafford-Tear et al., 2019). These findings highlight that traditional methods of genotyping, such a STR, TP-PCR

analysis and southern blotting, only provide crude estimates and mode allele lengths. A limitation to the results presented in this chapter is only that STR and TP-PCR assays were used to genotype, thus we do not have a true reflection of the distribution of allele lengths in this cohort. Another disadvantage to using these methods is the maximum repeat size that can be detected by STR analysis is approximately 120 repeats. Confirmation of larger expansions can be detected using TP-PCR but does not size the largest allele (Warner et al., 1996). A further limitation is that genotyping has only been performed on genomic DNA derived from the blood, and it is still unknown how the CTG18.1 expansion behaves in the post-mitotic cells of corneal endothelium. Tissue-specific dynamics of somatic mosaicism been observed in other repeat mediated disorders including DM1, where larger expanded alleles were detected in skeletal muscle, the main tissue affected by DM1, in comparison to blood of DM1 patients (Anvret et al., 1993; Corrales et al., 2019).

3.3.3 Exploring of sex distribution among FECD patients with and without CTG18.1 expansion

In this study it was identified there was a higher incidence of FECD in females compared to men (60.8% versus 39.2%) complimenting findings in reported literature (Afshari et al., 2006; Krachmer et al., 1978). Although I did find a higher prevalence of FECD in females it was not to the same extent as previous studies have reported a ratio as high as 3.7:1 (Kitagawa et al., 2002; Ong Tone et al., 2021). A plausible explanation to explain these differences is due to the cohort size analysed. The cohort analysed in this study is significantly larger and therefore the predominance in females becomes less apparent compared to smaller cohort sizes. A biological explanation for the higher prevalence of FECD in females is still yet to be established but the role

of sex hormones have been suggested to play a role (Miyajima et al., 2019). Furthermore, I identified a significantly larger proportion of females with a CTG18.1 negative allele in comparison to CTG18.1 positive alleles. This finding demonstrates the sex preference for FECD in females is less prominent in cases with CTG18.1 expansion and thus must be attributed to additional genetic and/or environmental factors.

Additionally, I found no significant difference in the median age-of-recruitment between males and females for either CTG18.1 positive or CTG18.1 negative FECD however, within the CTG18.1 expansion-negative group males have a lower trending age-of-recruitment (**Figure 15**). This finding could be suggestive that a proportion of these cases could share an X-linked early-onset form of the disease supporting a previous report of a severe X-linked endothelial corneal dystrophy (Schmid et al., 2006).

3.3.4 Correlation of CTG18.1 length with age at recruitment

In this study there was weak correlation ($r = -0.083$; $p = 0.025$) observed between the length of repeat and the age-of-recruitment (**Figure 16**), suggesting that increased repeat length may correlate with more severe or earlier onset disease. However, as previously mentioned the age at which an individual was recruited to the study is not a true reflection on stage of disease they were at or the progression of disease. Therefore, the correlation between expansion repeat length and age of diagnosis and/or surgery cannot straightforwardly be made and deep phenotyping of patients recruited in the future is required to acquire the age in which the disease begins to manifest and the disease progression rate. Currently the best metric to measure the severity of the disease is at the point surgical intervention is first required and thus I decided to further explore correlation using this system of measurement. Here I

found a stronger and statistically significant, although still objectively weak, correlation was observed between modal allele length and 'age-of-first-transplant' r=-0.154, p=0.019. However, this carries many limitations as the timing of surgery is heavily influenced depending on the surgeon, availability of donor tissue, surgical waiting times and personal preference such as visual requirements. Notably, the strongest correlation was observed amongst patients from GUH in comparison to MEH FECD patients, r = -0.22926, p= 0.19936 and r=-0.152, p=0.034, respectively (**Figure 17**). The difference in this correlation could be due to the fact that this sub-group of patients were under the care of a single referring clinician, whereas the MEH recruited cases were under the care of numerous clinicians and surgeons. The changeability in surgeons can introduce large interpatient variability due to the subjective opinion of the clinician which influences at what stage of disease an individual's diagnosis is deemed advanced enough to undergo surgery (Matthaei et al., 2019).

In addition, the n number for this study was relatively small, n= 230, due to historical clinical records not being available for all patients recruited in the cohort and excluding patients who had prior cataract surgery, as it is well established that the physical trauma of lens replacement surgery can damage the corneal endothelium and hence influence rates of endothelial decompensation (Walkow, Anders, & Klebe, 2000). Future studies utilising larger and more compressively phenotyped cohorts are anticipated to advance understanding of CTG18.1 phenotype-genotype correlations.

3.3.5 Bi-allelic CTG18.1 expansions

In this study, 41 subjects were identified to carry bi-allelic CTG18.1 expanded alleles. It did not appear apparent that carrying two expanded CTG18.1 alleles resulted in a more severe phenotype. However, phenotypic

data available on the progression of disease in these individuals is scarce and it is not possible to determine if the disease progressed at a quicker rate compared to usual. Deep phenotyping of the disease progression is required to explore any potential correlations with the CTG18.1 zygosity status. A potential explanation for there being no clear additive effect for a person possessing 2 expanded alleles could be that FECD pathogenicity is primarily determined by a single allelic dose of the mutant gene.

3.3.6 Early-onset CTG18.1 expanded FECD

Four probands within the studied cohort were identified to have an age-of-onset of <40 years. However, two of four of these subjects had not yet undergone surgery to restore vision. It could be argued the age-of-onset of these individuals is not unusual and may have been detected earlier due to the patient's lifestyle and visual requirements and by the time surgery is required they may reach an age typical for FECD. However, two of the patients (EAO_1and EAO_2) presented with symptoms in their third decade, very atypical for CTG18.1-mediated FECD, but had not yet required surgical intervention. The repeat length of CTG18.1 expansions in these cases could not explain the severe phenotype in these patients as they fell in the typical range associated with the late-onset disease. These individuals also reported a positive family history and affected family members were able to be recruited.

Proband EAO_3, also presented with dual diagnosis of keratoconus alongside early-onset FECD. It is currently unclear whether co-occurrence of keratoconus and FECD is a random association, or if the two conditions share a common disease pathway (S. Liu et al., 2023). As mentioned, the proband reported a positive family history of FECD. Her father carried an expanded CTG18.1 allele (86 repeats), however his age of disease onset is unknown. Her

sister carried the CTG18.1 allele of 88 repeats and was also reported to have an early onset phenotype. Neither her father nor sister were reported to present with keratoconus. The repeat length of the CTG18.1 allele transmitted within this family appeared to be moderately stable upon transmission, from 86 repeats to 90 and 88 respectively, eliminating the explanation that anticipation was the basis of the severe early-onset phenotype.

Proband EAO_4 presented with a severe early-onset phenotype in which she first presented symptoms in her second decade and had corneal transplant graft in her third decade of life. She reported a positive family history of FECD which was revealed to be complex following further investigation. Her mother was reported to FECD from an early age but was now deceased thus it was not possible to acquire a sample for genotyping. Her son was reported to be affected with an early-onset phenotype and began presenting early features of the disease at the age of 14 years. Her daughter, who reported to not exhibit any FECD symptoms was found to carry bi-allelic expanded CTG18.1 alleles of 83 and 84 and clinical examination detected early features indicative of FECD at the age of 20-years. For segregation purposes, her unaffected husband was also genotyped and found to carry an expanded CTG18.1 allele of 83 repeats. As he did not have a clinical eye examination, it cannot be determined if he is showing early symptomatic features of the disease, being 61-years-old it is possible he may present with the typical late-onset phenotype later in life. Also it may be possible he is part of the small proportion of the general population which carry an expanded allele without a FECD phenotype (Wieben et al., 2012; Zarouchlioti et al., 2018). If this is true, this could be a possible explanation to why the daughter has not presented with a severe early-onset phenotype like her brother and mother, whereby she has also inherited the

causative allele. Furthermore, the son was also reported to have experienced congenital cataracts. Additional work is required to identify the genetic cause of this, however, a digenic inheritance may contribute to the severe disease phenotype in which he and his mother presents with. Again, the length of the CTG18.1 expansion in this family cannot explain the early onset phenotype alone as it is in the expected length range as typical FECD. Further investigation on the possibility of other genetic modifiers, such as polymorphisms in the MMR genes, needs to be investigated to help understand the genotype-phenotype of these atypical cases.

3.3.7 Parent-child transmission of CTG18.1 repeat length

Transmission of an expanded CTG18.1 allele was observed to be inherited via the maternal germline nine times and from the paternal germline seven times via the segregation analysis reported in this study. In most cases of transmission, the repeat was only found to contract or expand from one to four repeat units and could be an artefactual result from the crude estimation from the STR analysis.

On this bias, further expansion of the CTG18.1 allele has only conclusively been observed through maternal inheritance, where an expansion of 13 repeats was noted, which is fitting to the behaviour of the transmission of other non-coding repeat expansions, that expansions of TNR are likely to occur within quiescent cells of the oocyte before transmission (Rifé et al., 2004). Contraction of the CTG18.1 allele was only observed through paternal inheritance of the expanded allele. This again fits with the parent-of-origin effect witnessed in other repeat-mediated disorders, that larger tracts of repeats contract in spermatogonia during development (Malter et al., 1997). However, like the alleles where small units of expansion occurred, allele lengths were only

reduced by one and four units, respectively, and may also be attributed to estimated mode allele length predicted by the STR assay.

3.4 Conclusion

This chapter has further strengthened the findings of previous studies, demonstrating the CTG18.1 expansion is the most common genetic risk factor for FECD, but specifically within the Caucasian population. My findings here support the hypothesis non-Caucasian populations may have a separate genetic risk factor to the CTG18.1 however, FECD was less prevalent in these populations overall. Furthermore, it supports previous findings that FECD is more prevalent amongst females in comparison to males. A further finding of this work included identifying males have a lower trending age-of-recruitment within CTG18.1 non-expansion FECD, suggesting there may be a common and possible X-linked early-onset form of disease within this subset of FECD patients.

4. Exploring CTG18.1 structure, instability, and the potential influence of MMR-associated genetic modifiers.

4.1 Introduction

TCF4 CTG18.1 expansions (≥50 repeats) is the leading genetic risk factor for FECD (**Section 3**). The CTG18.1 locus consists of a triplet-repeat CTG motif of variable length and adjacent CTC repeat immediately 3' of the CTG repeat. Previous research has demonstrated that the CTC repeat length can also be variable, especially on expanded CTG18.1 alleles. It has been demonstrated to consistently comprise one CTT repeat interruption (Alkhateeb, 2018; Hafford-Tear et al., 2019). Sequencing of the Generation Scotland cohort (Alkhateeb, 2018) has demonstrated that $(CTG)_n(CTC1)_n(CTT)(CTC2)_n$ is the most typical allelic structure presented within this population (**Figure 20**). Moreover, this study demonstrated that the CTC1 motif was more variable on expanded alleles with most CTC1 length variation ranging from 4 to 14 copies. On unexpanded alleles, CTC1 comprised 6 copies of the repeat in the majority samples. Interestingly, expanded alleles were identified to be pure with no interruption detected, whereas in the non-expanded alleles a GTG variant at position 10 was commonly detected on alleles with a CTG length of 17. Furthermore, a second variant, a CTC at position 15 was also detected on three non-expanded alleles with a CTG repeat length of 18. In this study the CTC2 repeats were monomorphic and remained stable at 6 repeats on all alleles (Alkhateeb, 2018).

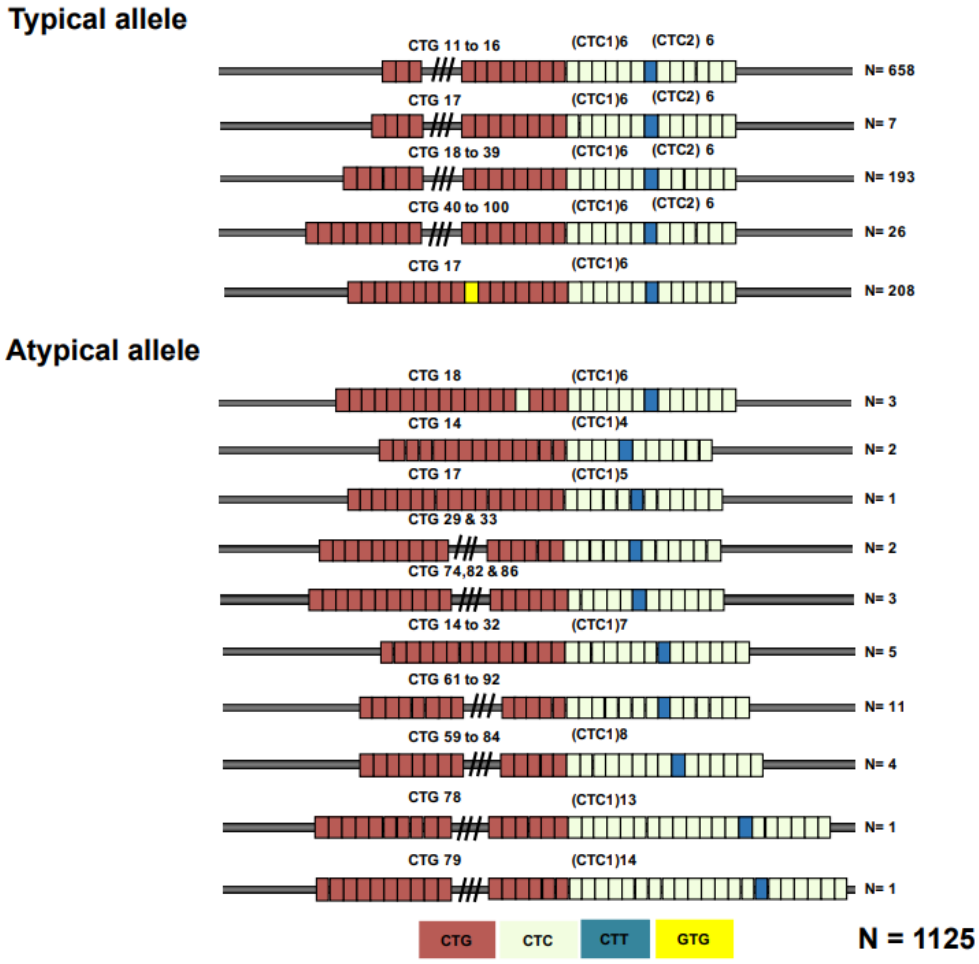


Figure 20 Allelic structure for the CTG18.1 locus for typical (more common) and atypical (less common) alleles revealed by sequencing. A common GTG variant was detected in the CTG repeats region at position 10 on 208 alleles. CTC1 was found to be polymorphic and varies from 4 to 14 CTC repeats, especially on alleles with an expanded CTG repeat tract (≥50 repeats) (Alkhateeb, 2018).

It has been established that disease-associated repeats can somatically expand in both an age-dependent and tissue specific manner (Morales et al., 2012). The somatic instability of these repeats may contribute to the symptom progression of a given disease and has served as a hypothesis to explain the tissue-specificity and phenotypic variability of various repeat-associated diseases including DM1, HD and others (Monckton et al., 1995; Morales et al., 2012; Trang et al., 2015; Wong et al., 1995). In HD, it has been demonstrated that mutant CAG repeat sizes vary greatly both within and between somatic

tissues of HD patients with the greatest variability occurring in the cortex and striatum, areas of the brain with the most neuropathological involvement (Kennedy et al., 2003; Telenius et al., 1994). Furthermore, the most prominent mosaicism has been witnessed in juvenile onset cases of HD suggesting the influence mosaicism has on progression of disease (Kennedy et al., 2003). Long-read amplification free sequencing method has recently demonstrated that DNA derived from leukocytes from CTG18.1 expansion-positive FECD patients display high levels of somatic instability, with larger levels of instability found to be positively correlated with increased CTG18.1 length (Hafford-Tear et al., 2019). Furthermore, somatic instability of the CTG18.1 repeat has also now been observed in RNA from the corneal endothelium of three FECD patients (Wieben et al., 2021) Additional research with larger numbers of samples and other non-ocular tissues are required to further establish this hypothesis and determine whether larger somatic expansions within the corneal endothelial tissue drives the progression of FECD.

In section **1.2.10.3.** I introduced the concept of how genetic variants in DNA repair genes can modify somatic stability of disease associated repeat elements**.** The efficiency of DNA processing can differ as a result of polymorphisms affecting components of the DNA repair pathway. Such *trans*-acting modifiers can modify rates that tandem repeats expand or contract within somatic cells; this phenomenon is known as somatic instability (Massey & Jones, 2018)**.** It has been demonstrated that somatic instability of repeat elements is a tissue specific and age-dependent process with larger repeat tracts being more susceptible to expansions. For both DM1 and HD, repeat lengths and somatic instability rates have been documented to be larger in the affected tissues (skeletal muscles in DM1 and neurons of the striatum and

cortex in HD). Interestingly, high levels of somatic instability have been documented within postmitotic tissue such as brain and muscle indicating that DNA repair mechanisms, instead of replication mechanisms, are pivotal to the process (Gomes-Pereira & Monckton, 2006). Like DM1 and HD, FECD affected tissue is comprised of post-mitotic cells (i.e. corneal endothelial cells) and it therefore can be hypothesized that a comparable scenario may apply for FECD, whereby genetic modifiers influencing the MMR pathway may govern CTG18.1 somatic stability rates.

Moreover, a large GWAS conducted in 2015 using blood-derived DNA from HD patients identified specific loci harbouring genetic variations that alter the age at neurological onset of HD (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015). This finding has since been replicated on a larger scale and loci within DNA repair genes *MLH1*, *FAN1*, *PMS1*, PMS2, *LIG1* and *MSH3* have all been identified to harbour SNPs which modify HD Age-at-Onset. Subsequent studies have also shown these loci are also significant in other repeat-mediated disorders such as SCAs (Bettencourt et al., 2016). Most of these genes encode proteins that play a prominent role within the mismatch repair pathway. It has therefore been hypothesised that DNA repair variants directly affect somatic expansion levels of repeats within individuals, but it is also possible that expanded repeats could exacerbate DNA repair defects (Massey & Jones, 2018).

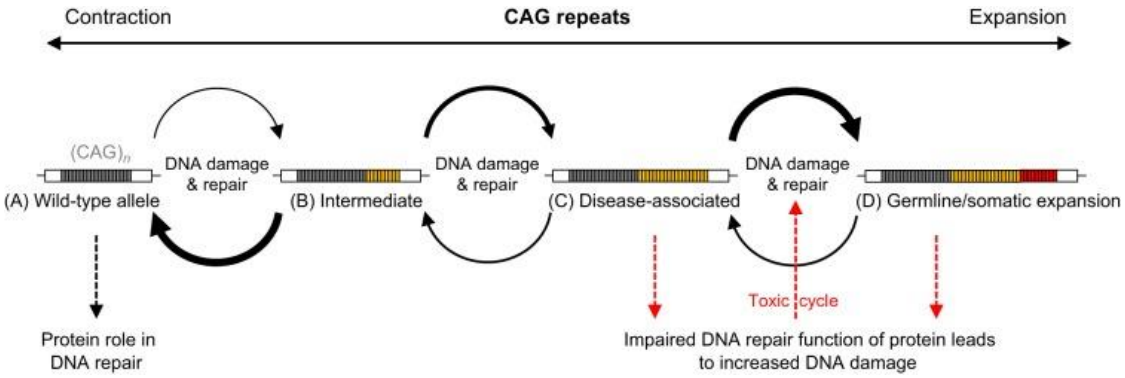


Figure 21 DNA damage and repair can affect CAG repeat length with downstream effects on disease pathogenesis. DNA repeat elements are unstable, and cycles of DNA damage and repair can lead to changes in repeat length over time (Massey & Jones, 2018).

Recently, Ciosi *et al*. sought to investigate how the clinical outcome of HD was impacted by genetic variation, including polymorphisms within the *HTT* CAG/CAA glutamine encoding repeat, somatic instability of the repeat in addition to *trans*-acting variants in DNA repair genes. Ciosi and colleagues employed high-throughput ultra-deep sequencing approach, using MiSeq, on blood-derived DNA from a large number of HD patients to sequence the glutamine-encoding repeat and quantify somatic expansions (Ciosi et al., 2019). Following this, Kompetitive allele-specific PCR (KASP) assay was used to genotype candidate gene polymorphisms within DNA repair genes. Their study revealed the frequency of synonymous CAA repeat interruptions in the *HTT* CAG/CAA repeat and concluded that clinical outcome of HD was better determined by pure CAG length and not total encoded glutamine number. Furthermore, they identified individuals with higher levels of blood-derived somatic CAG repeat instability had worse clinical outcomes, such as an earlier age of onset; and variants within DNA mismatch repair (MMR) genes, *FAN1*, *MSH3* and *MLH1,* were significantly associated with somatic expansion (Ciosi et al., 2019).

In this section, I genotype a panel of common MMR-associated polymorphisms within the cohort, previously determined to influence the stability of other repeat elements. In addition I apply a targeted ultra high-throughput sequencing approach to our CTG18.1 expansion-positive FECD cohort to determine if CTG18.1 somatic instability and or allelic structure impacts on phenotypic outcomes within the cohort. Finally, I combine the results of these studies to assess genotypic associations between variants in MMR genes and how these could influence rates of CTG18.1 somatic instability.

4.2 Results

4.2.1 Genotyping candidate DNA repair genes

4.2.1.1 Selection of candidate DNA repair-associated SNPs

The CTG18.1 expansion is associated with incomplete FECD penetrance and variable expressivity (Mootha et al., 2014). Therefore, if genetic DNA repair modifiers of HD effect FECD pathogenicity through a mechanism common to CAG·CTG repeat expansions, we hypothesise that HD-worsening modifiers would be enriched in a cohort of FECD patients with a CTG18.1 expansion. To investigate this hypothesis, the frequency of HD-worsening modifier alleles was measured to see if they are present at a higher frequency in the FECD CTG18.1 expansion-positive cohort relative to the general population. Twelve candidate DNA repair gene variants, known to be associated with the somatic expansion of the HTT CAG repeat in blood and/or HD onset was selected to be investigated (Ciosi et al., 2019; Consortium, 2019; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015).

These twelve SNP candidates are listed in **Table 17**. The UK10K was selected as a control dataset as all 12 target SNPs were directly genotyped

within the dataset. The UK10K dataset consists of 10,000 individuals, comparing the DNA of 4,000 people whose physical characteristics were well documented to 6,000 people with extreme health problems. The MAFs of SNPs within the British population of England and Scotland (GBR) population was acquired from the 1000 Genomes Project Phase 3 accessible on Ensembl (https://www.ensembl.org/).

Table 17 Summary Genetic modifier haplotypes selected from GWA12345, a genome-wide association study conducted on patients with Huntington's disease (HD) (Consortium, 2019; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015)**.**

Corresponding haplotype based on GWA12345 (Consortium, 2019)	**Gene**	**HD Effect Size (Years/ Minor Allele)**	**Candidate SNP**	**GBR MAF**
15AM2	*FAN1*	1.3	rs35811129	0.302
15AM4	*FAN1*	0.8	rs34017474	0.357
15AM1	*FAN1*	−5.2	rs150393409	0.027
N/A	*MLH3*	N/A	rs175080	0.527
5AM3	*MSH3*	0.6	rs1650742	0.253
5AM1	*MSH3*	−0.8	rs701383	0.247
3AM1	*MLH1*	0.8	rs1799977	0.297
19AM1	*LIG1*	0.9	rs274883	0.176
19AM2	*LIG1*	−0.6	rs3730945	0.407
2AM1	*PMS1*	−0.8	rs3791767	0.165
7AM1	*PMS2*	0.8	rs74302792	0.214
8AM1	*RRM2B, UBR5*	−1.2	rs79136984	0.077
HD, Huntington disease; MAF, minor allele frequency; GBR, British in England and Scotland; SNP, single nucleotide polymorphism.				

4.2.1.2 Genotyping candidate SNPs

Kompetetive allele specific PCR (KASP) assay (https://www.lgcgroup.com/) was chosen to genotype the twelve candidate SNPs listed in **Table 17**. The KASP assay is a homogeneous, fluorescence based genotyping assay that enables accurate bi-allelic discrimination of known SNPs and Indels. The KASP assay has several advantages in comparison to other SNP genotyping platforms such as being cost-effective and having a relatively low error rate in genotyping error in positive control DNA samples (0.7-1.6%) (Semagn, Babu, Hearne, & Olsen, 2014). In addition, the LGC group can conveniently design and optimise the assay and has a reasonable turnaround time of 5-7 weeks.

The LGC group designed oligonucleotides for the KASP assay from the given sequence, **Table 18**. The sequences of polymorphism of interest were indicated within square brackets. Fifteen bp upstream and downstream of the respective SNPs were also included to design the oligonucleotides. KASP service provider used in house control DNA for the validation assay. Optimised KASP assays were already available for 10 of the 12 target SNPs, with the exception of rs150393409 and rs3730945, as they had previously been used in other studies conducted (Bettencourt et al., 2016; Ciosi et al., 2019). For some of the candidate SNPs that had previously has KASP assays optimised for, a proxy SNP, in linkage disequilibrium was chosen due to the presence of being in the UK biobank, the control group used in the previous studies. The proxy SNPs have been listed in **Table 18**.

Table 18 Summary of SNPs investigated using a KASP assay. For each targeted SNP, indicated with a square bracket, 15 bp or flanking up and downstream sequence is shown. KASP assays were designed to either target the candidate SNP or, proxy SNPs in linkage disequilibrium with the target SNP.

Candidate SNP ID	Proxy SNP for KASP assay	Sequence
rs1650742	rs1382539	GTCATTCAGTTGTAA[A/G]GTTTAAATTGNTTC
rs150393409*	N/A	AATTGGCCAAACAGC[A/G]TTCAGTCTGCACTT
rs3730945*	N/A	AGAGGCAGGTGCACA[C/T]AGATGCTTTTCTTT
rs175080	N/A	CTTTCTCTCAAACTA[A/G]GCATCTGTTGTTCT
rs1799977	N/A	GACAATATTYGCTCC[A/G]TCTTTGGAAATGCT
rs274883	N/A	CATGGCCCCCACCCC[A/G]CTCTGGTCTACGCA
rs34017474	N/A	ATAACATGTAAATGC[T/C]TGTTCTACTGATTG
rs35811129	rs3512	TTAAAAGTAAAGGCA[C/G]TTCCAAGAGTAACA
rs79136984	rs3735721	GCTTAGTTGTAAGAA[A/G]AACTATTATTGTAT
rs3791767	rs5742933	GCCTCGCGCTAGCAG[C/G]AAGGTAGTGTGGTG
rs701383	N/A	AAGTCCTGCAGAGCT[G/A]GGAAGTGAGAAAAA
rs74302792	rs852151	CTGATCTCAGAGAGG[C/A]TGAGGAACCAGTGA
*Denotes SNPs in which a new KASP assay was designed and optimised for.		

4.2.1.3 Sample selection criteria

In total 698 DNA samples were aliquoted into 96-well plates and sent to the LGC group for genotyping. These included 632 FECD samples, 394 from MEH and 238 from GUH. Of those, 624 samples had at least one CTG18.1 allele greater than 50 repeats and 8 FECD samples had borderline mono-allelic repeat lengths ranging from 30-50 repeats. In addition, 66 AMD samples were sent for ‘control’ purposes. Forty-three of these had mono-allelic repeat expansions and 23 had borderline repeats on one allele. Two negative controls

were included on each 96-well plate, as recommended by LGC group, to ensure the reliability of the assay.

4.2.1.4 KASP genotyping results

SNP data was received from the LGC group in the format of comma separated value (CSV) files. Data was visualised in the SNPviewer software downloaded from the company website. SNPviewer illustrated the data in the form of a cluster plot to discriminate bi-allelic fluorescent signals for each sample. FAM signal was plotted in the X-axis and HEX signal on the Y-axis. Homozygous allele hybridised with oligonucleotide attached to HEX fluorescence (red), generated HEX signal and the data was plotted closer to the Y-axis. Heterozygous samples generated both FAM and HEX signals and the data was plotted in the centre of the graph (green). No template controls (NTC) were presented with black/pink dots. An example of how the KASP assay results are presented in SNPviewer is shown in **Figure 22** for SNP rs156641 for 94 samples and 2 NTC analysed together on one plate.

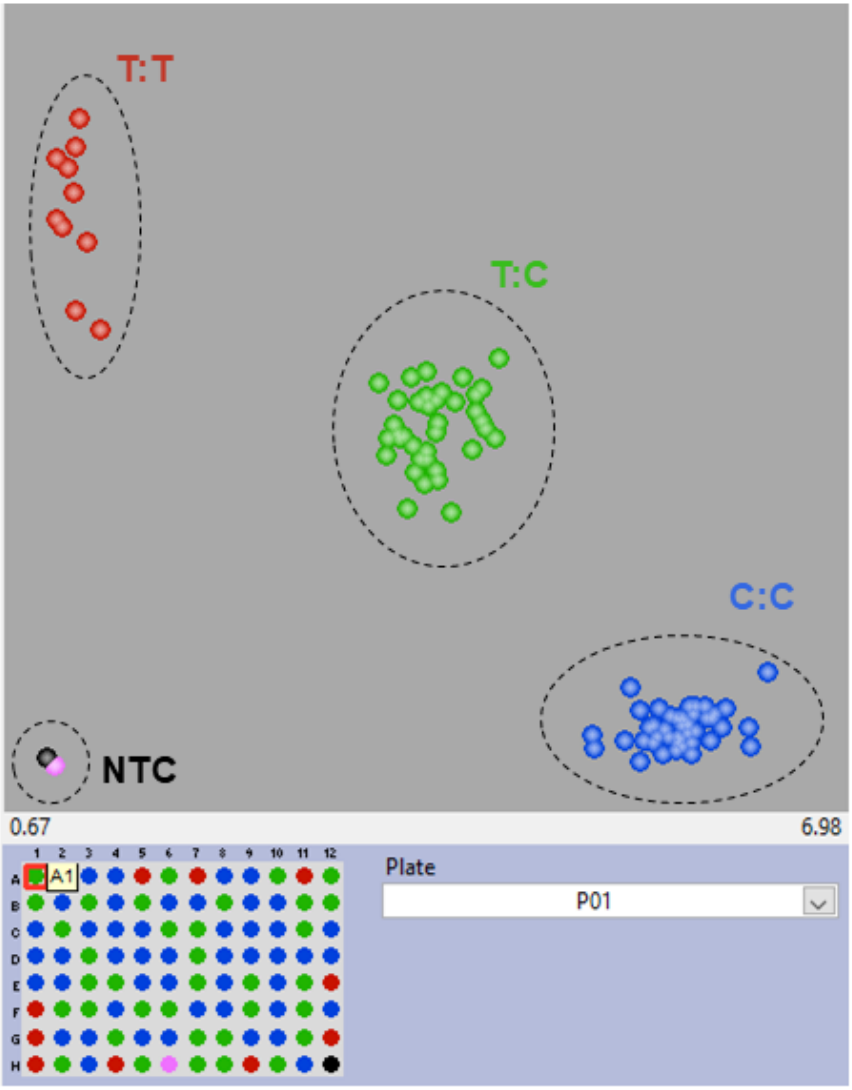


Figure 22 KASP assay results for SNP rs156641 for 94 samples and 2 non-template controls (NTC) visualised in SNPviewer. Individual samples are presented as one data point. FAM signal (C allele) was plotted in the X-axis and HEX signal (T allele) on the Y-axis. Homozygous T:T (red), homozygous C:C (blue) and heterozygous samples for T:C were generated both FAM and HEX signals (green). No template controls (NTC) were presented with black/pink dots.

Genotyping data for SNPs was analysed using PLINK (https://zzz.bwh.harvard.edu/plink/). Minor allele frequency for all samples were computed.

Initially all CTG18.1 expansion-positive samples derived from White British or White Czech samples recruited from MEH and GUH respectively were analysed

and MAF were generated for all 12 targeted SNPs (**Table 19** and **Table 20**). Summary MAF data was compared to population MAF data in both UK10K and the GBR population from 1000 genome project phase 3.

Table 19 Minor allele frequency (MAF) for each genetic modifiers SNP was calculated using blood-derived DNA for white British Fuchs endothelial corneal dystrophy (FECD) patients recruited from Moorfield's Eye hospital (MEH) carrying at least one CTG18.1 expanded allele (≥50 repeats). MAF of FECD samples were compared to the MAF in control dataset, UK10K and a Chi-square test was used to determine if significant differences were present. A significant difference was identified for SNP rs1799977 (3AM1 (*MLH1*)), highlighted in bold. A1: minor allele.

				MEH White British expanded FECD (≤50 repeats)		**UK10K**			
Corresponding haplotype based on GWA12345	**SNP**	**A1**	**A2**	**MAF**	**Number of alleles**	**MAF**	**Number of alleles**	**chi-square statistic value**	***p*-value (<0.05)**
2AM1 (*PMS1*)	rs5742933	C	G	0.2032	566	0.195611	7428	0.1911	0.662009
3AM1 (*MLH1)*	**rs1799977**	**G**	**A**	**0.2624**	**564**	**0.32539**	**7428**	**9.5401**	**0.00201**
5AM1 (*MSH3*)	rs701383	A	G	0.2394	564	0.240576	7428	0.0042	0.948125
5AM3 (*MSH3*)	rs1382539	A	G	0.2429	564	0.270867	7428	2.0844	0.14881
7AM1 (*PSM2)*	rs852151	A	C	0.1802	566	0.159666	7428	1.6426	0.199974
8AM1 (*RRM2B, UBR5*)	rs3735721	G	A	0.07597	566	0.067044	7428	0.6644	0.415003
- *MLH3*	rs175080	A	G	0.484	564	0.462843	7428	0.9471	0.33045
15AM1 (*FAN1)*	rs150393409	A	G	0.0106	566	0.011445	7427	0.0333	0.855233
15AM4 (*FAN1)*	rs34017474	C	T	0.4117	566	0.386241	7428	1.4311	0.231591
15AM2 (*FAN1)*	rs3512	G	C	0.3475	564	0.324044	7428	1.3152	0.25146
19AM1 (*LIG1*)	rs274883	G	A	0.1844	564	0.178514	7428	0.1235	0.72526
19AM2 (*LIG1*)	rs156641	T	C	0.3498	566	0.364028	7428	0.4589	0.498133

Table 20 Minor allele frequency (MAF) for each genetic modifiers SNP was calculated using blood-derived DNA for white Czech Fuchs endothelial corneal dystrophy (FECD) patients recruited from General University Hospital in Prague (GUH) carrying at least one CTG18.1 expanded allele (≥50 repeats). MAF of FECD samples were compared to the MAF in control dataset, UK10K and a Chi-square test was used to determine if a significant difference was present. No significant difference between MAF for FECD patients and control subjects was identified (p = 0.005). A1: minor allele.

				GUH Czech expanded FECD (≤50 repeats)		UK10K			
Corresponding haplotype based on GWA12345	**SNP**	**A1**	**A2**	**MAF**	**Number of alleles**	**MAF**	**Number of alleles**	**chi-square statistic value**	***p*-value (<0.05)**
2AM1 (*PMS1*)	rs5742933	C	G	0.2222	486	0.195611	7428	2.0401	0.153202
3AM1 (*MLH1)*	rs1799977	G	A	0.3277	476	0.32539	7428	0.0112	0.915858
5AM1 (*MSH3*)	rs701383	A	G	0.2573	482	0.240576	7428	0.6877	0.406937
5AM3 (*MSH3*)	rs1382539	A	G	0.2417	484	0.270867	7428	1.9606	0.161444
7AM1 (*PSM2)*	rs852151	A	C	0.1405	484	0.159666	7428	1.252	0.263163
8AM1 (*RRM2B, UBR5*)	rs3735721	G	A	0.0679	486	0.067044	7428	0.0054	0.941638
- *MLH3*	rs175080	A	G	0.4442	484	0.462843	7428	0.6345	0.425727
15AM1 (*FAN1)*	rs150393409	A	G	0.008264	484	0.011445	7427	0.4132	0.520374
15AM4 (*FAN1)*	rs34017474	C	T	0.3621	486	0.386241	7428	1.1193	0.290062
15AM2 (*FAN1)*	rs3512	G	C	0.3222	478	0.324044	7428	0.0072	0.932571
19AM1 (*LIG1*)	rs274883	G	A	0.1626	486	0.178514	7428	0.796	0.372297
19AM2 (*LIG1*)	rs156641	T	C	0.4066	482	0.364028	7428	3.5392	0.059935

For MEH samples only SNP rs1799977 (*MLH1;* A1: minor allele) was identified to be significantly enriched within the FECD expansion-positive cohort compared to subjects included in the only significant comparison between CTG18.1 expanded samples and control subject from UK10K dataset **(Table 19)**. A similar directional trend was identified for SNP rs1382539 (*MSH3*) in this cohort, which is also associated with delayed HD motor onset. As Chi Square results were similar for both white British samples from MEH and Czech samples, it was decided to group all European samples from MEH and GUH to increase n numbers and overall power. Many samples were initially excluded from the analysis within the MEH White British cohort due to having an 'unknown' or 'other' ethnicity. To overcome this limitation, genome wide SNP array data described in **Section 2.6**, was subsequently generated and used to predict ethnicity for the total FECD cohort included in this study. On this basis the analysis was repeated taking into account samples that had previously been assigned 'unknown ethnicity' and later assigned as 'European' increasing the overall cohort size from 283 white British MEH samples and 243 Czech GUH samples to a combined European cohort of 609 FECD samples. Including these additional samples and combining the British and Czech samples allowing the sample size to be maximised for analysis. The analysis was repeated using this combined European cohort and all data from this point forward refers to this dataset. As I am now looking at a European cohort the control data set was changed from the UK10K to gnomAD non-Finnish North-western Europeans to compare the allele frequencies of the candidate SNPs (**Table 21**).

Table 21 Minor allele frequency (MAF) for each genetic modifiers SNP was calculated using blood-derived DNA for all European Fuchs endothelial corneal dystrophy (FECD) patients recruited from Moorfield's Eye Hospital in London (MEH) and General University Hospital in Prague (GUH) carrying at least one expanded allele CTG18.1 expanded allele (≥50 repeats). MAF of FECD samples were compared to the MAF in gnomAD non-Finnish North-western Europeans control dataset and a Chi-square test was used to determine if any significant difference was observed between the case and control groups. No significant differences were observed between MAFs within the FECD patient cohort and gnomAD non-Finnish North-western Europeans control dataset (p = 0.005). A1: minor allele.

				All European expanded FECD (≤50 repeats)		gnomAD non-Finnish North-western Europeans			
Corresponding haplotype based on GWA12345	**SNP**	**A1**	**A2**	**MAF**	**Number of alleles**	**MAF**	**Number of alleles**	**chi-square statistic value**	***p*-value (<0.05)**
2AM1 (*PMS1*)	rs5742933	C	G	0.2126	1218	0.1943	8594	2.2674	0.132123
3AM1 (*MLH1)*	rs1799977	G	A	0.3005	1198	0.3215	50604	2.3668	0.12394
5AM1 (*MSH3*)	rs701383	A	G	0.2434	1212	0.2385	8562	0.1403	0.707958
5AM3 (*MSH3*)	rs1382539	A	G	0.2504	1214	0.2723	8586	2.5887	0.10763
7AM1 (*PSM2)*	rs852151	A	C	0.1513	1216	0.1640	8584	1.2647	0.260767
8AM1 (*RRM2B,UBR5*)	rs3735721	G	A	0.07307	1218	0.06715	8578	0.5911	0.441978
- *MLH3*	rs175080	A	G	0.472	1214	0.4663	50730	0.1549	0.693942
15AM1 (*FAN1)*	rs150393409	A	G	0.009046	1216	0.01035	50810	0.1984	0.656045
15AM4 (*FAN1)*	rs34017474	C	T	0.3957	1218	0.3871	8568	0.3315	0.564772
15AM2 (*FAN1)*	rs3512	G	C	0.3425	1206	0.3297	8572	0.7791	0.377428
19AM1 (*LIG1*)	rs274883	G	A	0.176	1216	0.1756	8560	0.0012	0.97245
19AM2 (*LIG1*)	rs156641	T	C	0.3773	1214	0.3712	8574	0.1653	0.684294

With the European combined MEH and GUH cohort the SNP rs1799977 (*MLH1*) which was previously significant in the all-White British MEH cohort (**Table 19**) was found to be no longer significant (**Table 21**), and the signal is weaker but still trending in the direction which is associated with delayed HD motor onset (Consortium, 2019). Overall, there were no other significant differences observed between MAFs for the 12 candidate SNPs within the FECD patient cohort and the gnomAD non-Finnish North-western Europeans control dataset.

Furthermore, I went on to explore if there were any trends present in an expanded FECD cohort compared to an age-matched cohort (mean age = 78) which carried repeat expansions but did not display clinical symptoms of FECD. The cohort consisted of patients who were diagnosed with AMD who had undergone an eye examination and did not have any records of FECD in their electronic MEH patient notes. However, despite being age-matched, it still remains possible that some of these AMD patients would have gone on to develop signs of FECD post the time each DNA sample had been recruited. As the AMD cohort were all White British ethnicity and recruited at MEH, only White British FECD expansion-positive samples were used for this comparison.

Table 22 Minor allele frequency (MAF) for each genetic modifiers SNP was calculated using blood-derived DNA for white British Fuchs endothelial corneal dystrophy (FECD) patients recruited from Moorfield's Eye hospital (MEH) carrying at least one CTG18.1 expanded allele (≥50 repeats) and compared MAF of an aged-matched cohort which carried either an expanded CTG18.1 allele or an intermediate expanded CTG18.1 allele (30-49 repeats) and did not display clinical features of FECD. The age-matched cohort was obtained from Age-related macular degeneration (AMD) patients recruited from MEH and had undergone extensive clinical eye examinations. A Chi-square test was used to determine if any significant differences were present. No significant difference between MAF for FECD patients and AMD subjects was found (p = 0.005). A1: minor allele**.**

				MEH White British FECD cases (≤50 repeats)		**Age-matched AMD cases with CTG18.1 expansion without FECD diagnosis**			
Corresponding haplotype based on GWA12345	**SNP**	**A1**	**A2**	**MAF**	**Number of alleles**	**MAF**	**Number of alleles**	**chi-square statistic value**	***p*-value (<0.05)**
2AM1 (*PMS1*)	rs5742933	C	G	0.2032	566	0.2045	132	0.0012	0.97201
3AM1 (*MLH1)*	rs1799977	G	A	0.2624	564	0.3409	132	3.2889	0.06975
5AM1 (*MSH3*)	rs701383	A	G	0.2394	564	0.197	132	1.0811	0.29846
5AM3 (*MSH3*)*	rs1382539	A	G	0.2429	564	0.3182	132	3.2536	0.071267
7AM1 (*PSM2)*	rs852151	A	C	0.1802	566	0.197	132	0.2007	0.654141
8AM1 (*RRM2B, UBR5*)	rs3735721	G	A	0.07597	566	0.0625	128	0.2783	0.597844
- *MLH3*	rs175080	A	G	0.484	564	0.4773	132	0.0196	0.888568
15AM1 (*FAN1)*	rs150393409	A	G	0.0106	566	0.02273	132	1.2367	0.266111
15AM4 (*FAN1)*	rs34017474	C	T	0.4117	566	0.3712	132	0.7273	0.393761
15AM2 (*FAN1)*	rs3512	G	C	0.3475	564	0.3258	132	0.2246	0.635533
19AM1 (*LIG1*)	rs274883	G	A	0.1844	564	0.1818	132	0.0047	0.945109
19AM2 (*LIG1*)	rs156641	T	C	0.3498	566	0.3636	132	0.0895	0.764833

As the AMD cohort carry a CTG18.1 expansion-positive allele and do not present with a FECD phenotype, it would be anticipated to observe opposing findings between the FECD cohort and the AMD cohort in the SNP frequency data. It would be expected that the AMD cohort would have an enrichment in SNPs with delaying or protective consequence. Although there were no significant comparisons between the FECD cohort compared to the AMD cohort, two SNPs followed the same trend patterns where the frequency of these SNPs have been associated with delayed HD motor onset. These SNPs were rs1799977 (*MLH1*) and rs1382539 (*MSH3*) (**Table 22**).

4.2.2 Quantifying somatic instability of CTG18.1

4.2.2.1 Optimising Mi-Seq PCR conditions for CTG18.1 locus

PCR for the reaction were previously optimised by Mariam Alkhateeb, a PhD student in Professor Darren Monkton's lab at the University of Glasgow (Alkhateeb, 2018). The CTG18.1 region was amplified using previously published primers, a forward primer, SEF2-C, taken from Fiona Gould (Gould, 2000) and reverse primer, P2+CC, modified from Mootha et al. with an additional CC modification at the 5' end of the primer to increase binding affinity (Mootha et al., 2014).

Initially, to verify the reaction could amplify CTG18.1 of variable lengths, DNA from four FECD patients previously determined to harbour expanded CTG18.1 alleles ranging from 12-79 repeats were selected. PCR products were generated containing the amplified CTG 18.1 region, using 200ng input of four samples were run on a 1.5% Ethidium Bromide (EtBr) gel (see **Section 2.3.5**). Products of the expected size were apparent and no unspecific bands were observed (**Figure 23**).

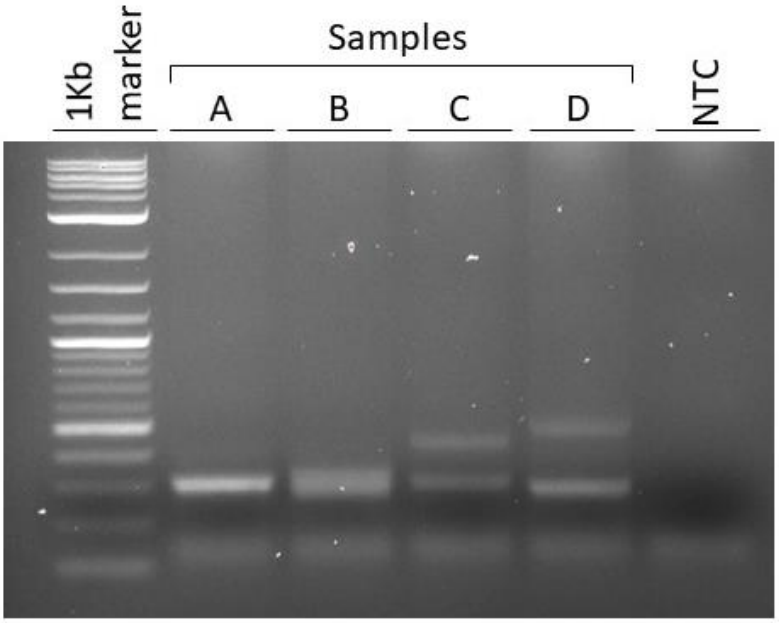


Figure 23 Agarose gel showing PCR products efficiently amplifying the CTG18.1 region using PCR conditions optimised by Alkhateeb, 2018. 200ng DNA from four samples with varying CTG18.1 lengths was amplified, and PCR products were run on a 1.5% ethidium bromide agarose gel. Samples had respective CTG18.1 allele lengths: A – 18/18, B – 12/24, C - 18/61, D – 12/79. All samples produced expected fragment lengths and no non-specific bands were observed. A No template control (NTC) was added to confirm there was no presence of contamination. A 1Kb DNA marker was used as a size reference.

Next, an additional PCR was performed using the same conditions and sequence-specific primers with the addition of MiSeq adaptor components to determine if these modifications affected PCR amplification efficiency **(Table 23,** full primer sequence is listed in **Table S3**). The same PCR conditions were used as previously and two DNA samples were selected for this optimisation assay, one with bi-allelic short CTG18.1 alleles (18/18 repeats) and a second with a monoallelic expansion (12/79 repeats). Four combinations of primer adaptors were tested to verify the PCR still worked, as desired. As previously, PCR products were visualised on a 1.5% EtBr agarose gel (**Figure 24**). All PCR reactions were identified to efficiently amplify the desired CTG18.1 region. However, large primer dimers (approximately 120bp) were produced for all

reactions, likely due to lengths of the MiSeq adaptor components of the primers being used (**Figure 24; Table 23**).

Table 23 Combinations of primers used with MiSeq adaptors attached, including the Nextera XT Index Kit v2 indexes, to verify PCR worked efficiently with these attachments present. Full primer sequence available in Table S3.

Primer combinations	MiSeq Forward primer indexes	MiSeq Reverse primer indexes
1	502	704
2	502	712
3	504	704
4	504	712

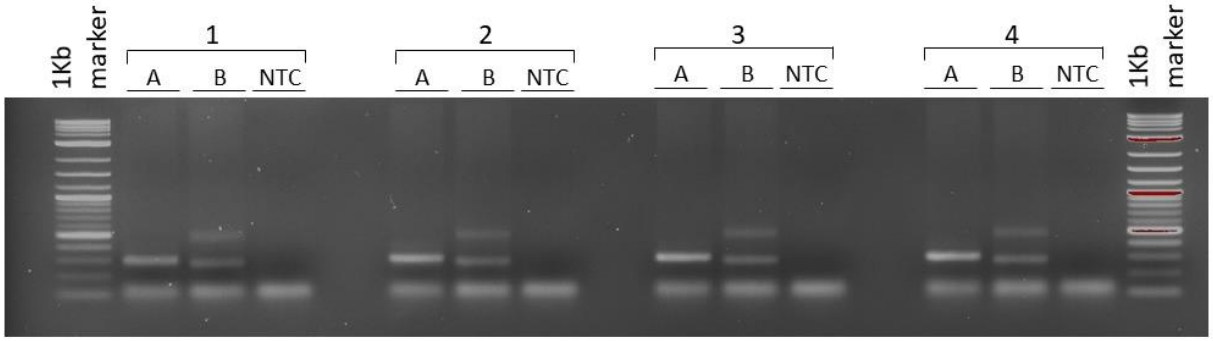


Figure 24 Agarose gel demonstrating PCR products of CTG18.1 region are still effectively amplified using primers with MiSeq adaptor components attached. 200ng DNA from two samples with varying CTG18.1 lengths was amplified, and PCR products were run on a 1.5% ethidium bromide agarose gel. Samples had respective CTG18.1 allele lengths: A – 18/18, B – 12/79. Primer with full MiSeq adaptor components were used with four different combinations of MiSeq barcodes: 1- S502/N704, 2- S502/N712, 3- S504/N704, 4- S504/N712. All combinations efficiently amplified PCR products of correct size and no non-specific bands were observed. A larger primer dimer approximately 120bp was produced for each primer combination. A No template control (NTC) was added to confirm there was no presence of contamination. A 1Kb DNA marker was used as a size reference.

Next, optimal loading DNA concentration was determined to reduce PCR efficiency in amplifying longer alleles compared to shorter ones resulting in relative lower yields of end-products per input molecule for larger alleles (Ciosi et al., 2019). This was achieved using a DNA concentration gradient of samples

(ranging from 10ng to 200g) comprising variable CTG18.1 allele lengths (12/79). PCR products were run on a 1.5% EtBr agarose gel. With increasing DNA input the relative efficiency of amplifying the larger expanded allele was identified to be reduced in comparison to the shorter non-expanded allele (**Figure 25**). On this basis 10ng input DNA was selected as an optimal input amount, given my primary interest in expanded CTG18.1 alleles.

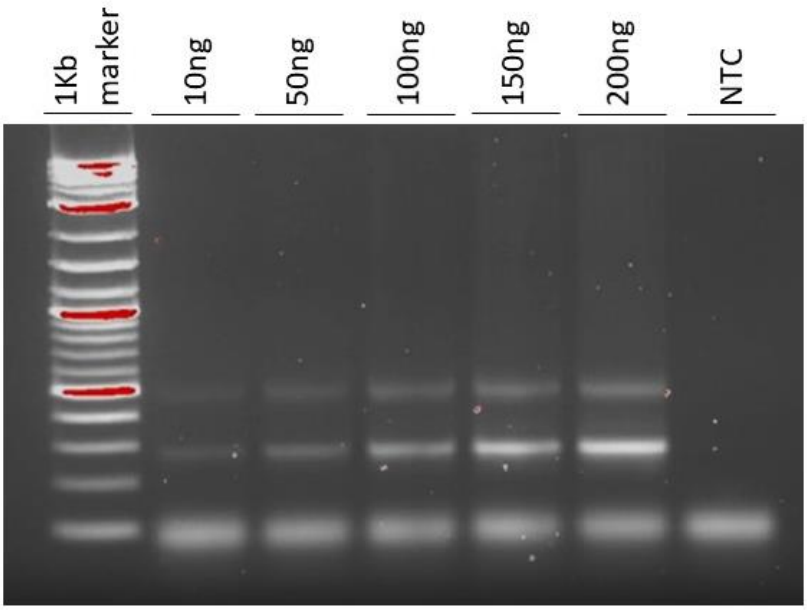


Figure 25 PCR products of input DNA concentration gradient using optimised MiSeq PCR run on a 1.5% ethidium bromide agarose gel. DNA sample with a CTG18.1 genotype of 12/79 with an input DNA concentration ranging from 10ng to 200ng. A relative efficiency of amplifying the larger expanded CTG18.1 allele (top band) is reduced in comparison to the efficiency of amplifying the shorter length CTG18.1 allele (middle band) with increasing DNA input. Lower band observed is the primer dimer is a result of the length of MiSeq primer adaptor components. A No template control (NTC) was added to confirm there was no presence of contamination. A 1Kb DNA marker was used as a size reference.

4.2.2.2 Optimising PCR clean-up for MiSeq sequencing

The MiSeq adaptors components attached onto the CTG18.1 specific locus primers allow up to 384 samples to be multiplexed and sequenced in a single run by incorporating a unique barcode combination into the PCR amplicon. A variable combination of 16 forward primers indices and 24 reverse primers indices produces 384 unique combinations allowing to identify a sample after sequencing. As seen in **Figure 24 and 25**, the MiSeq primers create larger primer dimers due to the length of the adaptor components. Their presence is problematic as the MiSeq sequencing method preferentially amplifies shorter fragments and hence there is a need to remove primer dimers from the libraries prior to sequencing.

To do so, PCR products were purified using magnetic bead clean-up and size selection procedure. To achieve this goal the concentration of beads used in the clean-up procedure first needed to be optimised to remove all traces of primer dimers. A second objective of the clean-up procedure was to concentrate the sequencing library. Again, this required optimization to gain the optimal concentration ratio between the normal and expanded alleles.

An initial optimization protocol was run using a gradient concentration of AMPure XP beads from 1x to 0.4x using a pool of 9 PCR products with an input of 10ng DNA per PCR reaction. Post AMPure purification, cleaned PCR products were run on a 1.5% EtBr agarose gel to visualise if primer dimers had been removed effectively (**Figure 26**). Post-AMPure purification was additionally assessed by capillary electrophoresis on a Bioanalyzer (Agilent) to further check that the fragments had the expected size and that primer dimers were absent.

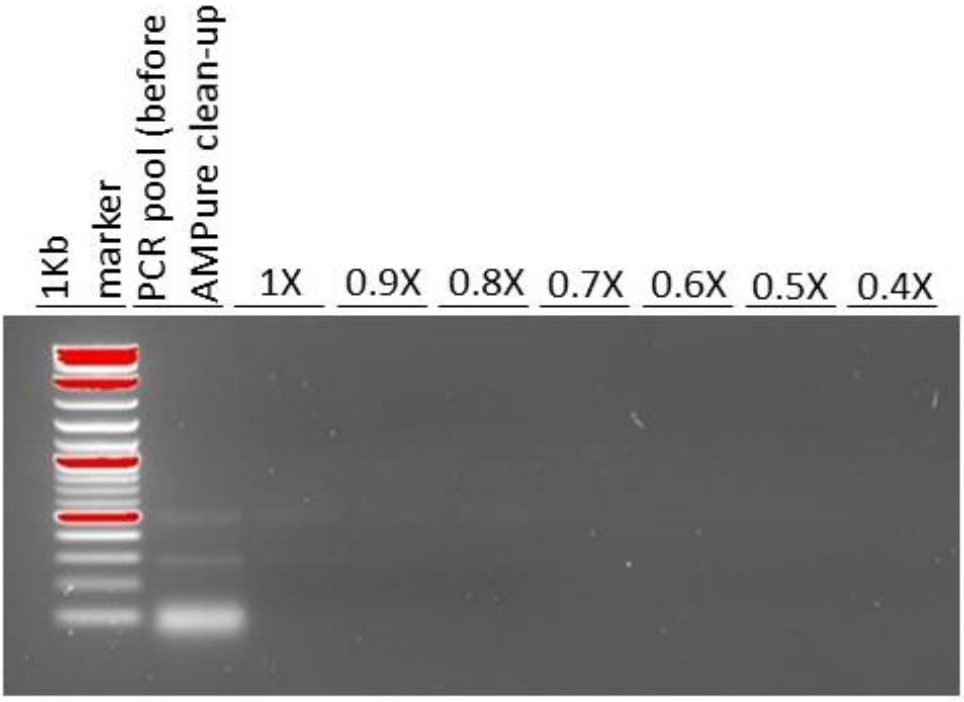


Figure 26 Optimising AMPure XP beads using a concentration from 1x to 0.4x. PCR was carried out products were purified using AMPure XP beads at a concentration from 1x to 0.4x and PCR product run on 1.5% agarose gel. Input DNA was too low to be able to get an accurate representation of how the experiment worked.

As the concentration after the AMPure clean-up was too low to be able to get an accurate representation of what was occurring, this experiment was then repeated on a larger scale using a pool of 48x PCR reactions with an input of 10ng per PCR reaction to increase the post-AMPure product concentration for optimisation purposes. When the protocol is optimised a sequence library with a total of 384 samples will be pooled together, prior to the AMPure bead clean-up and thus post-AMPure concentration was not anticipated to be an issue.

The clean-up was repeated after pooling a larger number of samples and using a concentration gradient range from 0.8X - 0.4X. Purified PCR reactions were run on a 1.5% EtBr agarose gel and the post AMPure cleaned-up PCR products were of a high enough concentration to be visualised on an agarose gel.

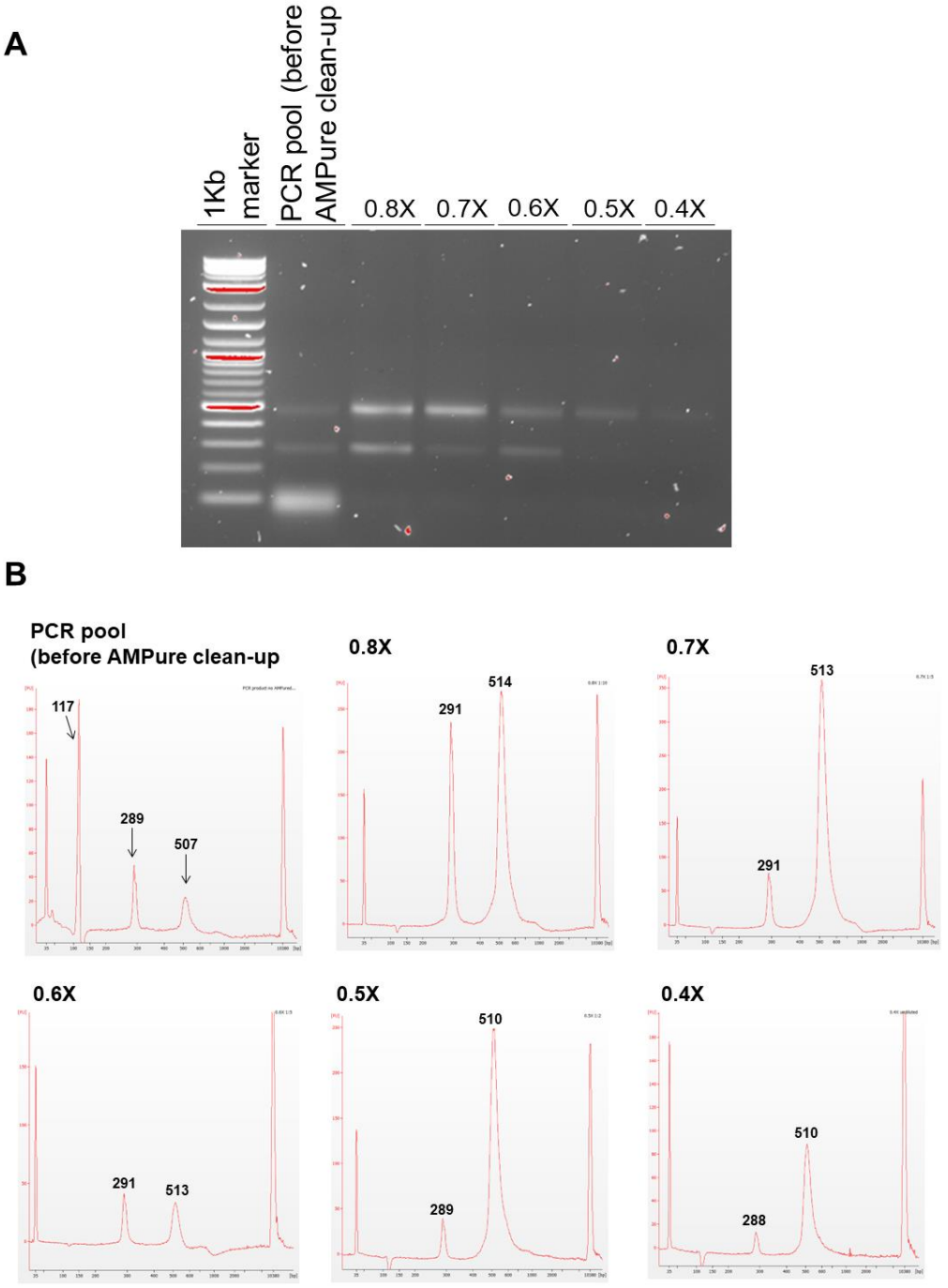


Figure 27 Optimising AMPure XP bead purification method using a variable concentration of input from 0.8x to 0.4x. Post cleaned-up PCR products were assessed by capillary electrophoresis on a Bioanalyzer using a sample with a monoallelic expansion (12/79 repeats). (A) PCR was carried out and products were purified using AMPure XP beads at a concentration from 1x to 0.4x. Purified PCR products were run on 1.5% agarose gel showing primer dimers had been removed with each AMPure bead concentration. **(B)** Samples were assessed by capillary electrophoresis on a Bioanalyzer. Expected product sizes based on were: primer dimers: ~100bp, 12 CTGs allele: ~290bp, 79 CTGs allele: ~ 510bp. The height of the peak indicates the abundance of each product in the sample

In **Figure 27.A**, the majority of primer dimers were effectively removed at all concentration ranges of AMPure beads tested. However, as the concentration of AMPure beads decreased there was an increased loss in the products of interest. For a more in-depth analysis, post AMPure cleaned PCR products for each concentration were assessed by capillary electrophoresis on a Bioanalyzer to further check that the fragments had the expected size and that primer dimers were absent. The Bioanalyzer traces, **Figure 27.B**, showed the PCR product comprised of three size fragments on prior to AMPure clean-up, the first around 100 bp due to primer dimers, one at just below 300 bp representing the shorter length allele and one at around 500 bp representing the expanded allele. Post AMPure clean-up bioanalyzer traces revealed primer dimers had been successfully removed at every concentration. In addition, the bioanalyzer traces revealed how the ratio between the normal and expanded allele changes with the different AMPure bead concentrations. The bioanalyzer trace of the uncleaned PCR product demonstrates the PCR efficiency towards amplifying shorter alleles in relation to larger sized alleles as the trace shows a higher yield of end-products, per input molecule, for the shorter length allele. However, as the AMPure bead concentration is reduced the ratio between end-products are altered. At a concentration of 0.8X the shorter length and expanded alleles are approximately equal but from 0.7X the ratio moves from normal allele and towards the larger expanded allele. This is desirable given the assay is primarily being undertaken to investigate expanded CTG18.1 alleles. The concentration 0.6X was noted to be an anomaly where both alleles appear equal on both the bioanalyzer trace and the agarose gel. From this optimisation run it was decided to go forward with using a concentration of 0.7X for the post PCR clean-up and go ahead with the first full MiSeq PCR of 384 DNA samples.

4.2.2.3 Genotyping the CTG18.1 locus using Illumina MiSeq next generation sequencing

After amplifying the 384 samples with the optimised MiSeq primers and PCR conditions, 5μL of the total 10μL PCR reactions were pooled together and the remaining 5μL was reserved as backup. The total pooled PCR products of ~1920μL were split into two aliquots of 900μL, one aliquot was reserved as back up and one divided into 3 aliquots of 300μL and cleaned up using the AMPure protocol at a concentration of 0.7X.

A diluted aliquot of the cleaned PCR products was then run on bioanalyzer to confirm all primer dimers had been removed and an optimal ratio of shorter length and expanded alleles had been achieved using the previously optimised clean-up protocol.

Figure 28 shows the Bioanalyzer traces for the post PCR reaction pool, prior to being purified and after being cleaned-up with a concentration of 0.7X AMPure beads. The region highlighted in green indicates the primer dimers, in blue the shorter length allele for the 384 samples and in orange expanded alleles. From these traces, it is evident a small concentration of primer dimers remained, despite the clean-up process.

However, the traces provide good indication the AMPure clean-up at 0.7X was very efficient in enriching the post PCR products for the expanded alleles. Nevertheless, as some primer dimer remains, further optimisation and purification was considered to be necessary before proceeding with the MiSeq sequencing.

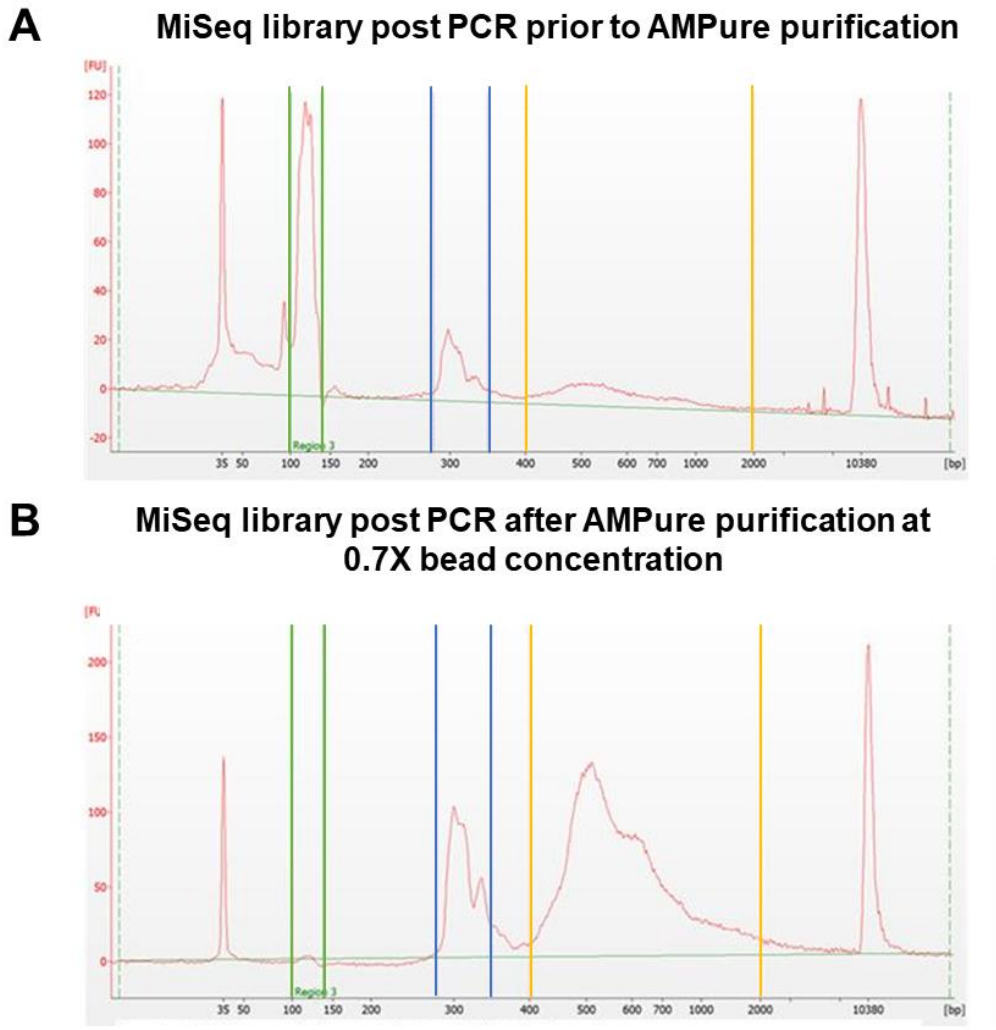


Figure 28 Bioanalyzer assessment on MiSeq Library prior to AMPure bead purification and after purification. (A) 384 Post PCR products samples pooled together forming the MiSeq Library prior to AMPure bead clean up showing primer dimer highlighted in green, normal CTG18.1 alleles in blue and expanded CTG18.1(≥50 repeats) in yellow. **(B)** MiSeq library of 384 samples after being purified using AMPure bead clean-up method at 0.7X concentration. A small proportion of primer dimers, highlighted in green, still remain after purification. Expanded CTG18.1 alleles, highlighted in yellow, have been concentrated.

To further optimise the PCR MiSeq library clean-up, three further clean-up protocols were tested. From the previously purified PCR product, with a concentration of 0.7X AMPure beads, a second 0.7X clean-up was performed and from reserved pooled PCR product aliquot two further clean-ups were performed, one at a concentration of 0.5X and one at a concentration of 0.6X.

After clean-up the corresponding AMPure concentrations were run on the bioanalyzer to assess the most optimal concentration for the clean-up procedure. **Figure 29** presents the bioanalyzer traces for each respective clean-up procedure.

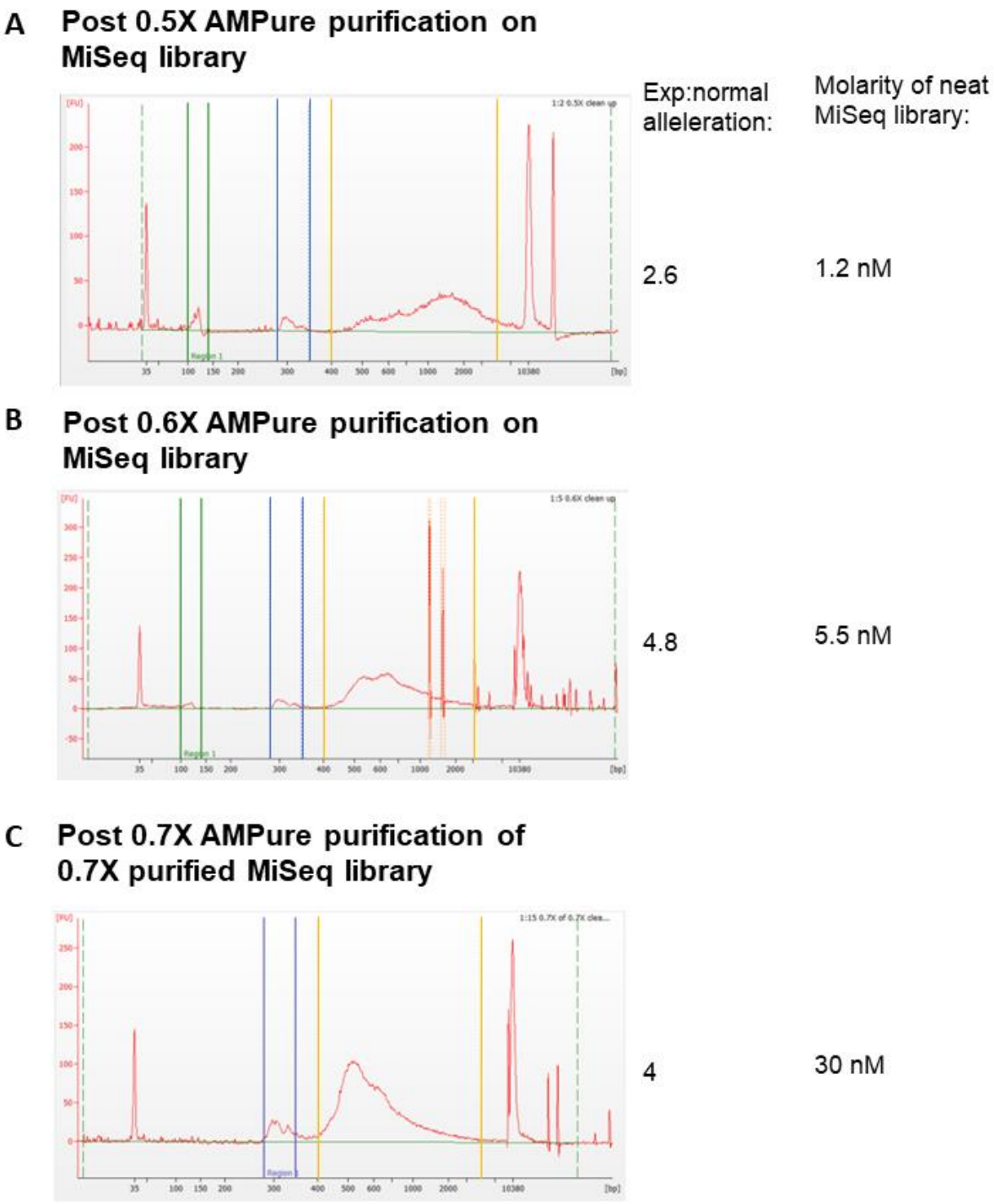


Figure 29 AMPure bead purification at different conditions on 384 pooled PCR products forming MiSeq library. (A) AMPure bead purification at a 0.5X on MiSeq library failed to remove all primer dimers and resulted in too large of a loss of material. **(B)** AMPure bead purification at a 0.6X on MiSeq library did not remove all primer dimers on a single clean-up but produced a good ratio between the normal length and expanded (≥50 repeats) CTG18.1 alleles. **(C)** A second AMPure bead purification at 0.7X on already purified library also at 0.7X concentration resulted in complete primer dimer removal and have a good ratio between the normal length and expanded CTG18.1 alleles. Region highlighted in green: primer dimers, region highlighted in blue: normal CTG18.1 alleles, region highlighted in yellow: expanded CTG18.1 alleles.

A concentration of 0.5X AMPure beads resulted in too large of a loss of material and in addition failed to remove primer dimers completely and hence I concluded it was not possible to proceed with this concentration (**Figure 29.A**). The second clean-up with a concentration of 0.7X on the previously cleaned up library completely eliminated all primer dimers and produced a nice ratio between the shorter length and expanded alleles and was thus considered suitable to be sequenced on the MiSeq platform (**Figure 29.C**). In spite of a small concentration of primer dimers remaining with the clean-up protocol using a AMPure concentration of 0.6X (**Figure 29.B**), this concentration produced a more desirable pattern of enrichment of larger versus shorter alleles and on this basis this library was selected for sequencing on the MiSeq platform. However, due to the small amount of remaining primer dimers a second clean-up using a concentration of 0.8X will be performed following the initial clean-up.

Finally, the selected library was run on the bioanalyzer to confirm all primer dimers were removed and the library was optimal for MiSeq sequencing. **Figure 30** presents the bioanalyzer trace for the library post second clean-up. The bioanalyzer trace confirms primer dimers had been eliminated after the second clean-up and there is a strong enrichment towards the larger expanded alleles.

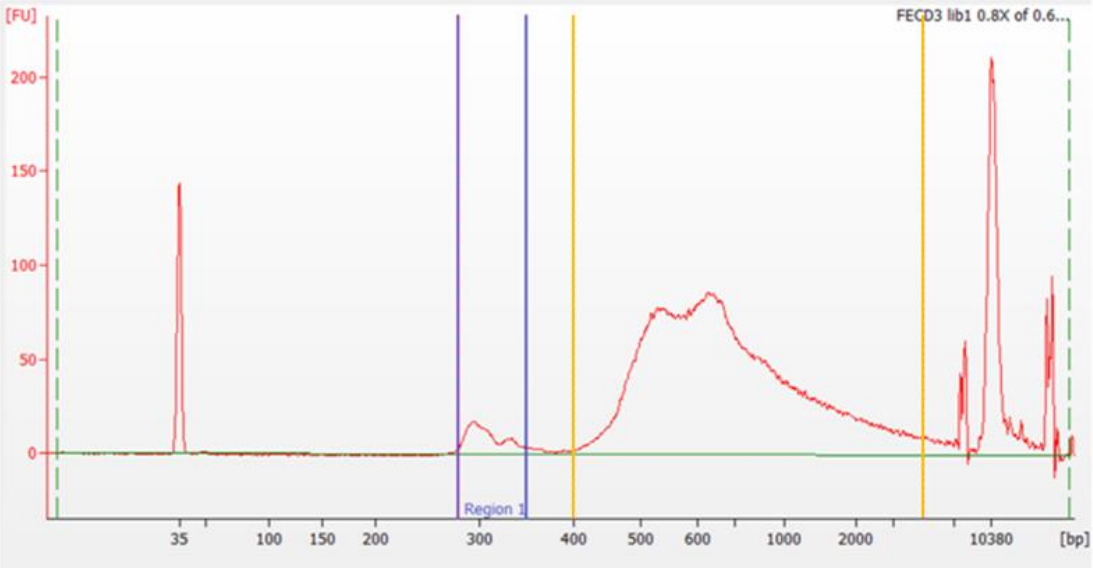


Figure 30 Bioanalyzer analysis of MiSeq library consisting of 384 pooled PCR products after undergoing two AMPure purification clean-ups, the first at 0.6X and second at 0.8X concentration. Bioanalyser trace shows primer dimers have been completely eliminated and there is a strong enrichment towards the CTG18.1 expanded alleles (≥50 repeats), highlighted in yellow, compared to the normal CTG18.1 alleles, highlighted in blue.

4.2.2.4 Genotyping and characterising the CTG18.1 locus using Illumina MiSeq next generation sequencing

FECD samples with monoallelic expansions (≥50 CTG repeats) and biallelic expansions (>50 repeats) previously genotyped using STR (**section 3.2.2**) were selected for further genotyping using Illumina MiSeq next generation sequencing.

MiSeq is a high-throughput method allowing up to 384 samples to be sequenced at one time. Ninety-six well plates were used for ease, each containing 93 FECD samples plus a mono-allelic expansion positive control, bi-allelic non-expansion negative control (<50 repeats) and a NTC. Two MiSeq runs were used to cover the whole cohort of FECD samples recruited to this study (n=630). The same control samples were used on all plates for both runs to confirm any batch effect between the runs. In addition to the FECD samples previously described, FECD samples with intermediate repeat lengths (30-50 repeats) (n=8) and AMD patient samples with either expanded or intermediate alleles but without FECD phenotype (n=66) were included in the sequencing. Sequencing was performed in 5' and 3' direction producing both forward and reverse sequencing reads to allow for genotyping allele structure for each sample.

4.2.2.5 Preparing MiSeq reads for analysis

The sequencing reads obtained were processed on the Galaxy server by the University of Glasgow (https://heighliner.cvr.gla.ac.uk) to determine the quality of the raw reads, demultiplex raw reads, trim MiSeq adapters, alignment, and visualising reads.

As the library was multiplexed, forward reads (R1) needed to be demultiplexed and prepared before genotyping. Reads from individual samples were able to be distinguished and sorted using the sample's unique barcode. Initially reads were demultiplexed using Cutadapt 1.16.8 on the Galaxy server starting with the spacer and primer to remove the oligonucleotide and in addition any cross contamination from other reads at the 5' end of the read. Secondly, spacers and the first 10 bases of the primer at the 5' end of the read were trimmed to remove any spacer-related read length variation and so all reads begin at the same base. Following this FASTQtrimmer (Version 1.1.1) was used to trim spacer length to remove any spacer-related variation from the 3' end of the read resulting in all reads being the same length. Finally, cutadapt 1.16.8 was used to trim the sequencing adapter at the 3' end using an error rate of 0.4.

4.2.2.6 Producing read count distributions for CTG18.1 and calling estimated progenitor allele length (ePAL)

After reads were prepared, CTG frequency distributions were processed using RGT on forward reads (R1). RGT was used to detect the CTG repeat tracts using the settings described in **Section 2.4.5.5.2**. In the first instance, RGT was implemented to count CTGs only from 5' flank to end of R1. However, after reviewing the output, I identified that any reads that did not contain pure CTG tracts were not included resulting in a lot of reads being discarded. To overcome this limitation RGT analysis was repeated adding a function implementing RGT to count all repeat 'sequence structures' up until the first 'CTCCTC' and to remove the downstream CTC repeat to avoid counting issues. Next the CTG frequency distribution was defined in two ways; (1) considering only the reads in which RGT detected a pure CTG tract (approximately 70% of

the reads); (2) counting all triplet repeat variations present in the CTG tract detected by RGT. The latter approach enabled reads with either PCR and/or sequencing errors to be included in the analysis (approximately 30% of reads). Overall the second approach allowed for a higher read count and therefore a more accurate CTG frequency distribution, as exemplified in **Figure 31**.

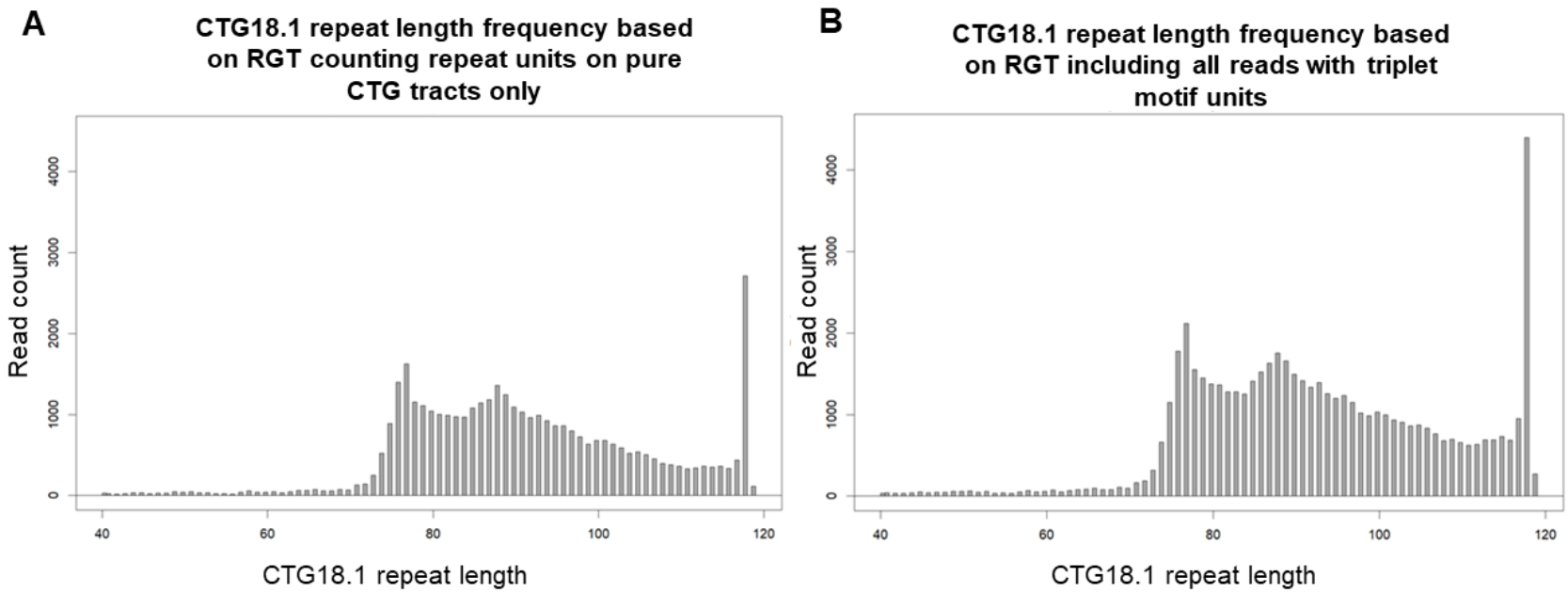


Figure 31 Read count for CTG18.1 repeat length distribution frequency comparing two RGT approaches to count the CTG repeat in samples with the second approach allowing higher reads inclusion. (A) The first RGT approach only counts repeats units on pure CTG tracts, resulting in a high number of reads being discarded. **(B)** The second RGT approach counts pure CTG repeat tracts and additionally counts all repeat units in reads which contain other triple motifs, as a result of PCR and/or sequencing errors, allowing more reads to be included in the analysis.

After acquiring the CTG distribution outputs plots for all samples, using the second RGT approach, estimated progenitor allele length (ePAL), the inherited allele length of the CTG18.1 repeat, for the expanded allele was defined for each sample. Previous research by Ciosi *et al.* applying this methodology to the HD repeat loci has elucidated that the vast majority of reads shorter than the progenitor allele can be attributed to PCR Taq polymerase slippage errors and that the vast majority of reads longer than the progenitor allele represents genuine somatic expansions using single molecule and bulk DNA (Ciosi et al., 2019). On this basis I have applied this model, exemplified in **Figure 32**, to define the ePAL for each sample included in this study.

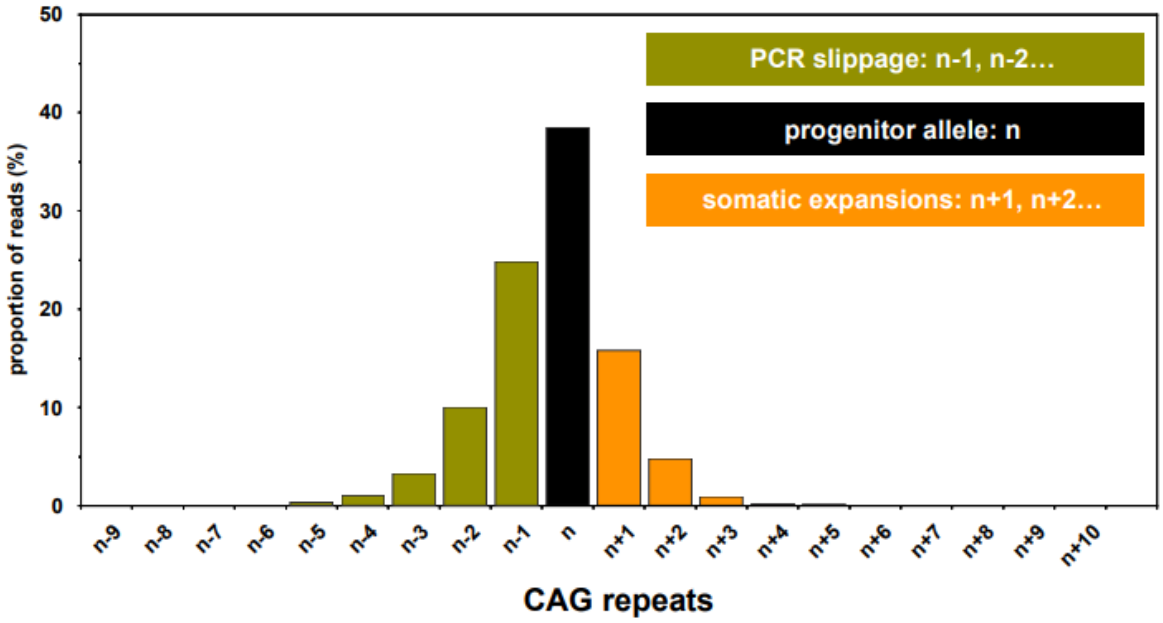


Figure 32 Interpretation of non-progenitor sequence reads in bulk DNA analyses. The data presented support the model in which the vast majority of reads shorter than the progenitor are PCR Taq polymerase slippage errors (n-1, n-2 etc.) and that the vast majority of reads longer than the progenitor allele represent genuine somatic expansions (n+1, n+2 etc.) Figure adapted from (Ciosi et al., 2019).

In total, 615 FECD samples previously defined by the STR assay to have one allele with an expansion over 50 repeats were analysed using the MiSeq assay to determine the ePAL as explained above. The ePAL for each sample was compared to the genotype acquired from the STR assay detailed in **Section 3.2.2**. Supplementary **Table S4** lists ePAL and STR genotype for all expanded alleles included in this study. For the vast majority of samples ePAL and the STR genotype were comparable, with 0-5 repeats differences. **Figure 33** exemplifies STR and MiSeq data distribution outputs for two distinct patient-matched samples. In both examples, the ePAL and STR genotypes were identical.

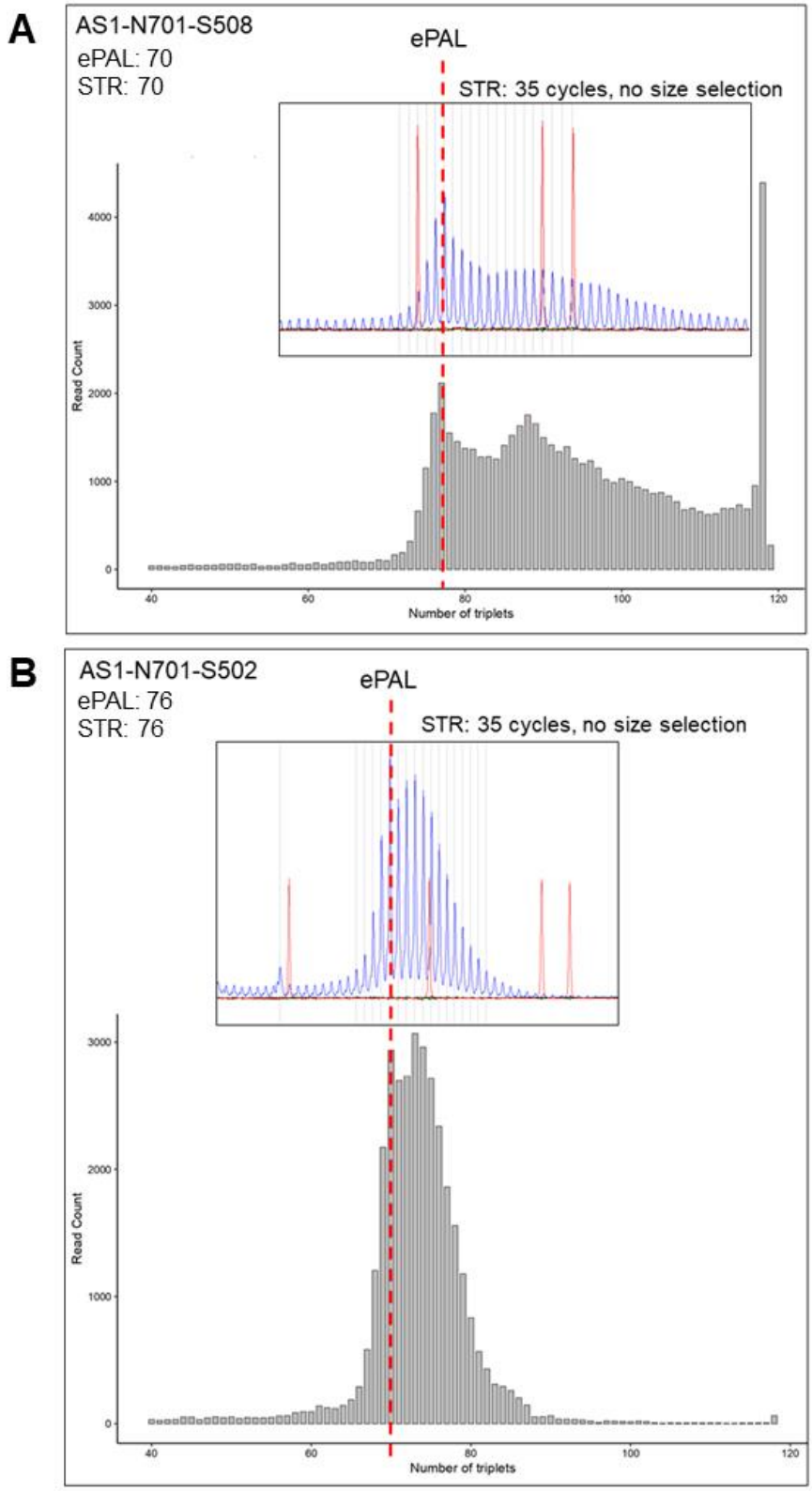


Figure 33 CTG18.1 frequency distribution for expanded CTG18.1 alleles obtained from MiSeq sequencing and STR genotyping for two independent Fuchs endothelial dystrophy samples. (A) MiSeq and STR CTG18.1 distribution plots for a sample with an estimated progenitor allele length (ePAL) of 70. **(B)** MiSeq and STR CTG18.1 distribution plots for a sample with an ePAL of 76. Red dotted line indicates ePAL.

In a few samples in this study, the ePAL varied considerably from the genotype determined by the STR assay. As mentioned earlier, the genotype determined using the STR assay was based on the modal repeat length rather than the ePAL itself. One explanation for the disparity between the two is the bimodal distribution, where two discrete populations can be observed, one centred around the constitutive repeat (i.e., stable repeats) and those further expanded (i.e., unstable repeats) (Larson, Fyfe, Morton, & Monckton, 2015; J. M. Lee, Pinto, Gillis, St. Claire, & Wheeler, 2011). 60.7% of the samples analysed using the MiSeq assay displayed some form of a bimodal distribution. **Figure 34** illustrates a range of six FECD samples with varying degrees of bimodal distribution for the CTG18.1 repeat.

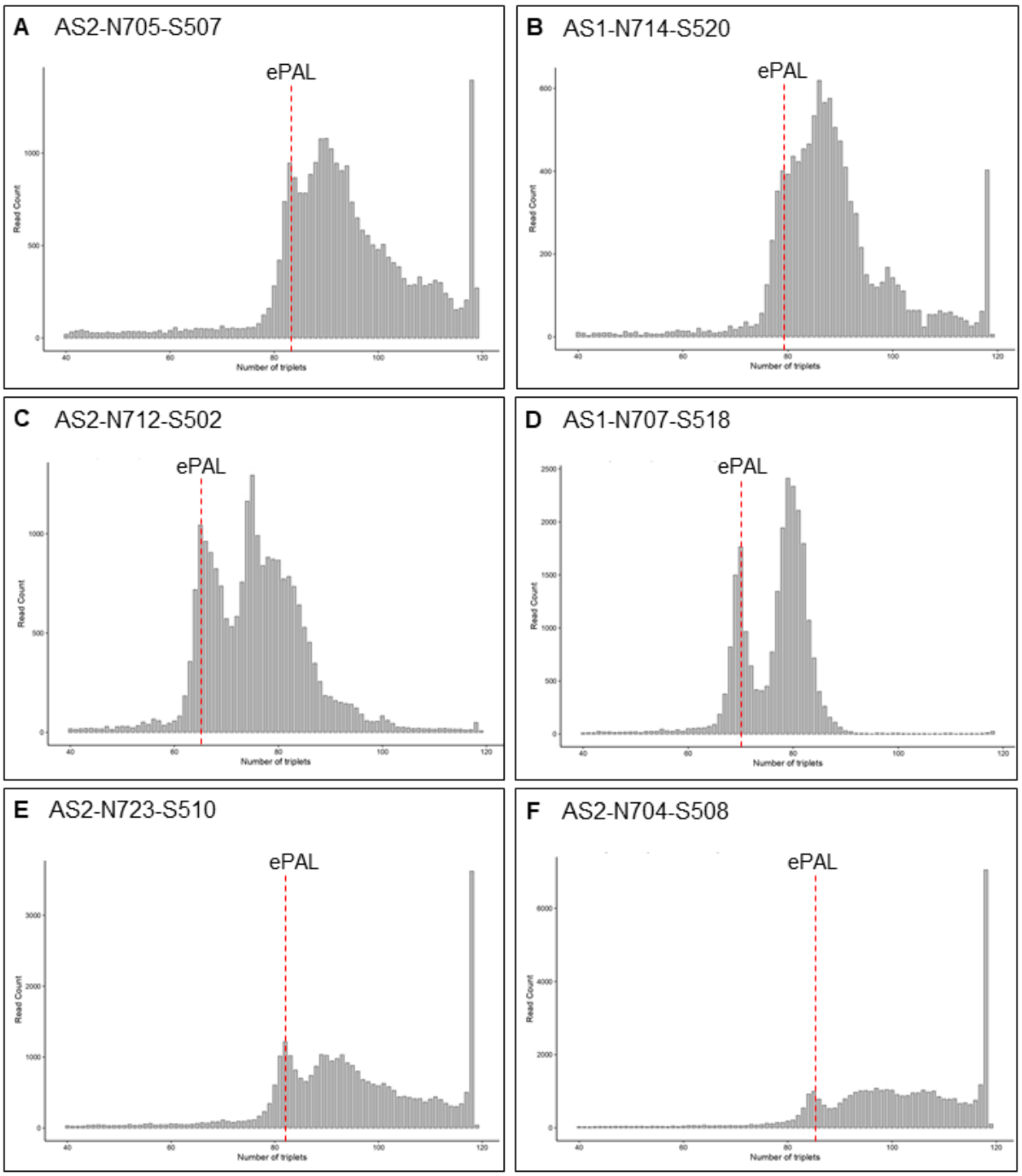


Figure 34 MiSeq CTG18.1 frequency distribution plots for expanded alleles for six independent Fuchs endothelial corneal dystrophy samples showing different degrees of bimodal distribution from the estimated progenitor allele length (ePAL), shown by the red dotted line.

For many samples there was a build-up of reads at 118 repeats. This build-up of reads represents reads which contain ≥118 repeats. Unfortunately, due to the size-detection limitations of the MiSeq assay for reads in this category it is not possible to determine the true size of repeat contained within the original molecule of DNA from which the reads have been generated from. This is an overall limitation of the approach.

For 34/615 samples I was unable to determine ePAL from the generated Miseq data. For seven of these samples this was due to overall low read counts being generated. For another 27 samples ePAL could not be determined because the read distributions largely exceeded the 118 repeats threshold of the Miseq assay, and/or they displayed particularly high levels of somatic instability (**Figure 35**). For example, samples in **Figure 35.A** and **35.B** were able to be genotyped with the STR assay with modal repeat lengths of 88 and 99 respectively. Determining the ePAL for these samples was difficult. As there was such a large number of reads which had 118 repeats or more, this muted the reads which had a lower number of repeats. In both these samples it appears there was a slightly higher number of reads approximately around the repeat length, which was called for the modal STR genotype, however, not enough to be able to confidently determine the ePAL. This pattern of CTG distribution is suggestive of large somatic instability above 118 repeats. Samples in **Figure 35.C** and **35.D** were both unable to call a genotype length from the STR assay but TP-PCR revealed an expanded allele for both samples and thus can be assumed the length of the repeat surpassed the STR threshold. For both samples, the ePAL was unable to be determined also. Here we see a gradual increase in reads towards the 118-repeat length of the MiSeq

assay suggesting the repeat for both these samples are larger than 118 repeats.

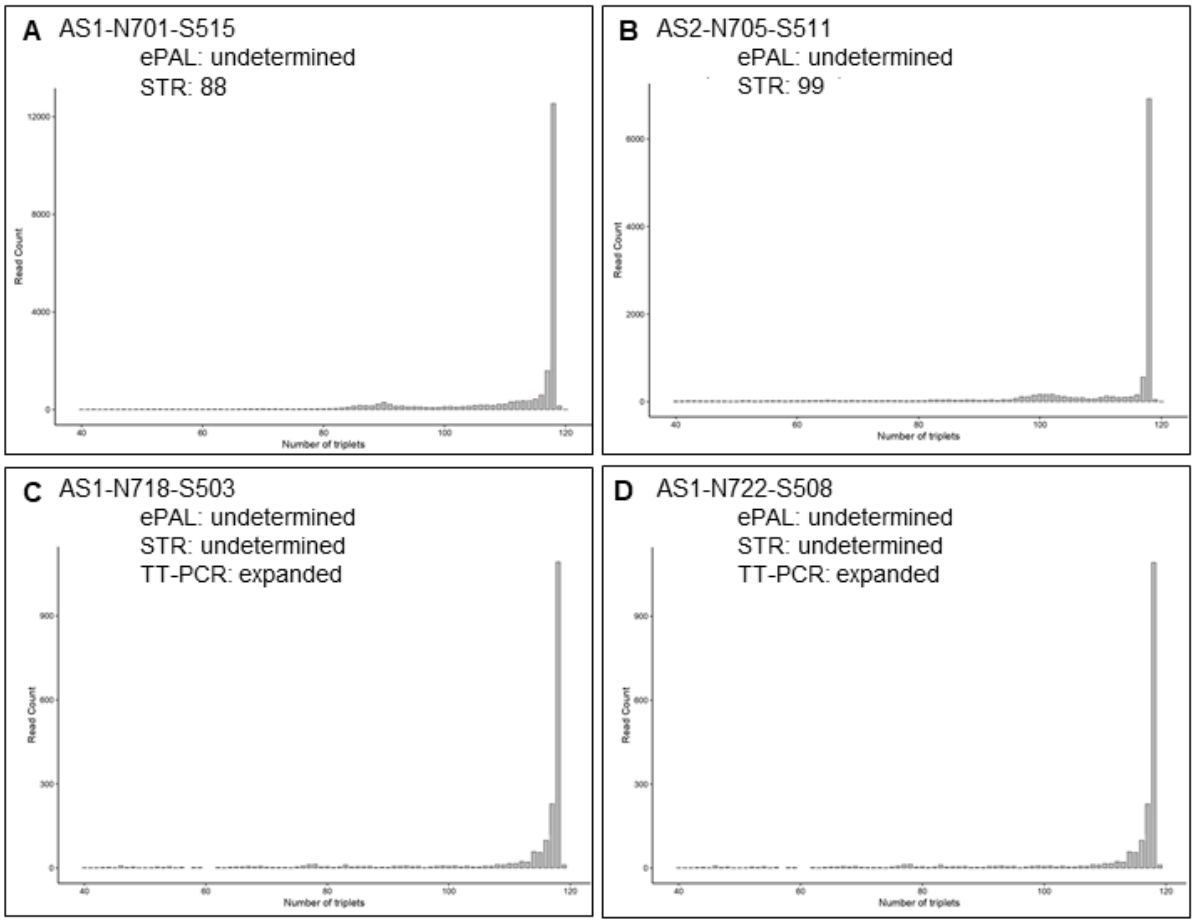


Figure 35 Four samples in which estimated progenitor allele length (ePAL) was not able to be determined from the generated Miseq data determined because the read distributions largely exceeded the 118 repeats threshold of the Miseq assay, and/or they displayed particularly high levels of somatic instability. (A) and **(B)** are examples of samples ePAL was unable to determined because a large number of reads had 118 repeats or more resulting in muting the reads which had a lower number of repeats which may include the ePAL. **(C)** and **(D)** are examples of samples where the assumed length of the ePAL surpassed the STR threshold.

4.2.2.7 Using MiSeq sequencing to quantify somatic instability

In total, 609 FECD samples derived from individuals of European ethnicity had one mono-allelic expanded alleles ≥50 repeats. Within this group it was not possible to determine ePAL for 34 samples. A further 38 samples had

bi-allelic expanded alleles and on this basis ePAL was also not determined for these samples and were excluded for this analysis.

Firstly, there was no significant correlation observed between ePAL and the age the patients were recruited to the study, r = - 0.045, p= 0.293 (**Figure 36**).

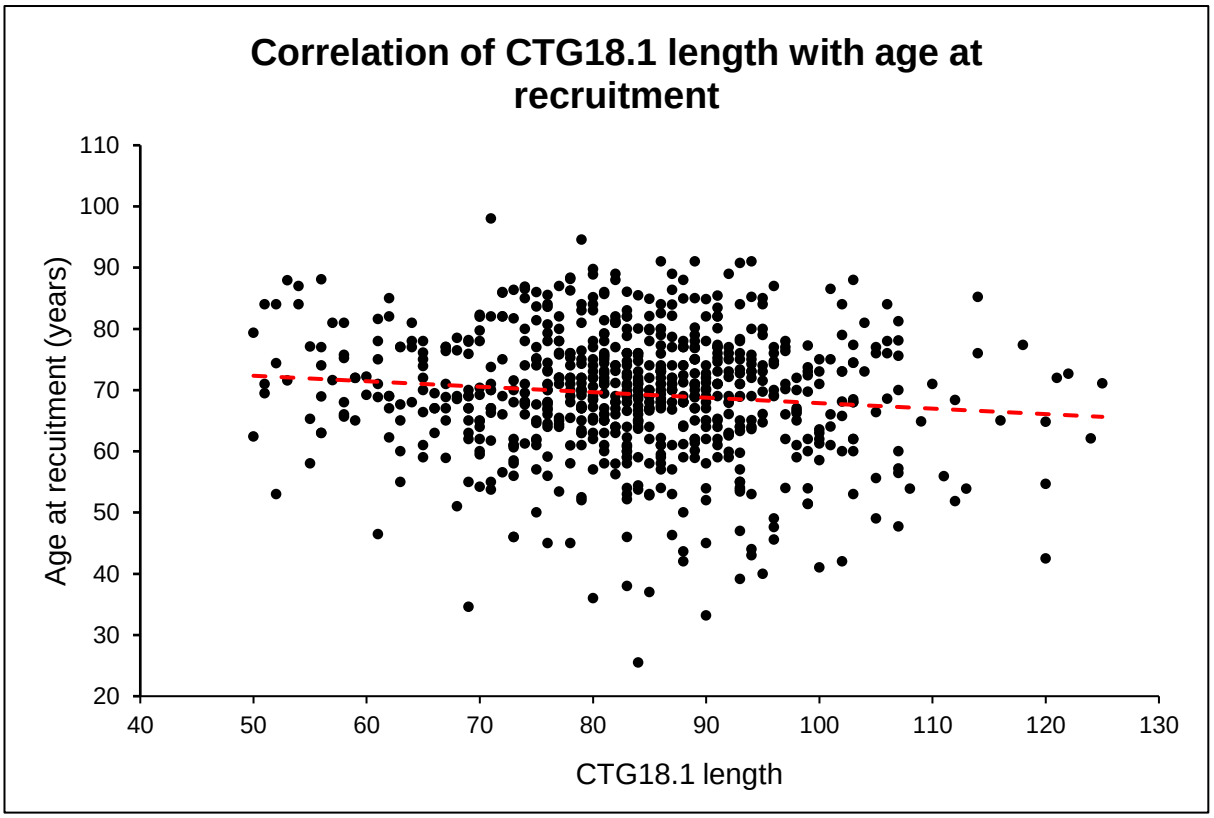


Figure 36 No significant correlation was observed between CTG18.1 estimated progenitor allele length (ePAL) and the age the Fuchs endothelial corneal dystrophy patient was recruited to the study, Spearman's rank correlation coefficient r = - 0.045, p= 0.293.

As explained earlier, using single bulk DNA data showed that the vast majority of reads longer than the progenitor allele represents genuine somatic expansions. (Ciosi et al., 2019) (**Figure 32**). On this basis reads larger than the ePAL were used to measure the degree of somatic expansions occurring in the blood-derived gDNA samples recruited to this study.

Somatic instability levels were calculated for a total of 537 samples. Two alternative measures of instability were considered: (1) the proportion of reads larger than ePAL, from ePAL to the end and (2) the proportion of reads larger than 116 from ePAL to the end.

Firstly, using measure (1), the proportion of reads larger than ePAL from ePAL to the end was used to calculate the ratio of somatic expansions (number of somatic expansion products/number of progenitor allele products) (**Figure 37**).

When looking at the proportion of reads larger than ePAL from ePAL to the end for all 537 samples (**Figure 37.A**), a ceiling effect can be seen for samples with ePAL over 80 repeats created by samples having repeat lengths over MiSeq threshold of 118 repeats. Using this metric of calculating the ratio of somatic expansion worked best for samples with an ePAL ranging from 49 to 79

CTG repeats (**Figure 37.B**).

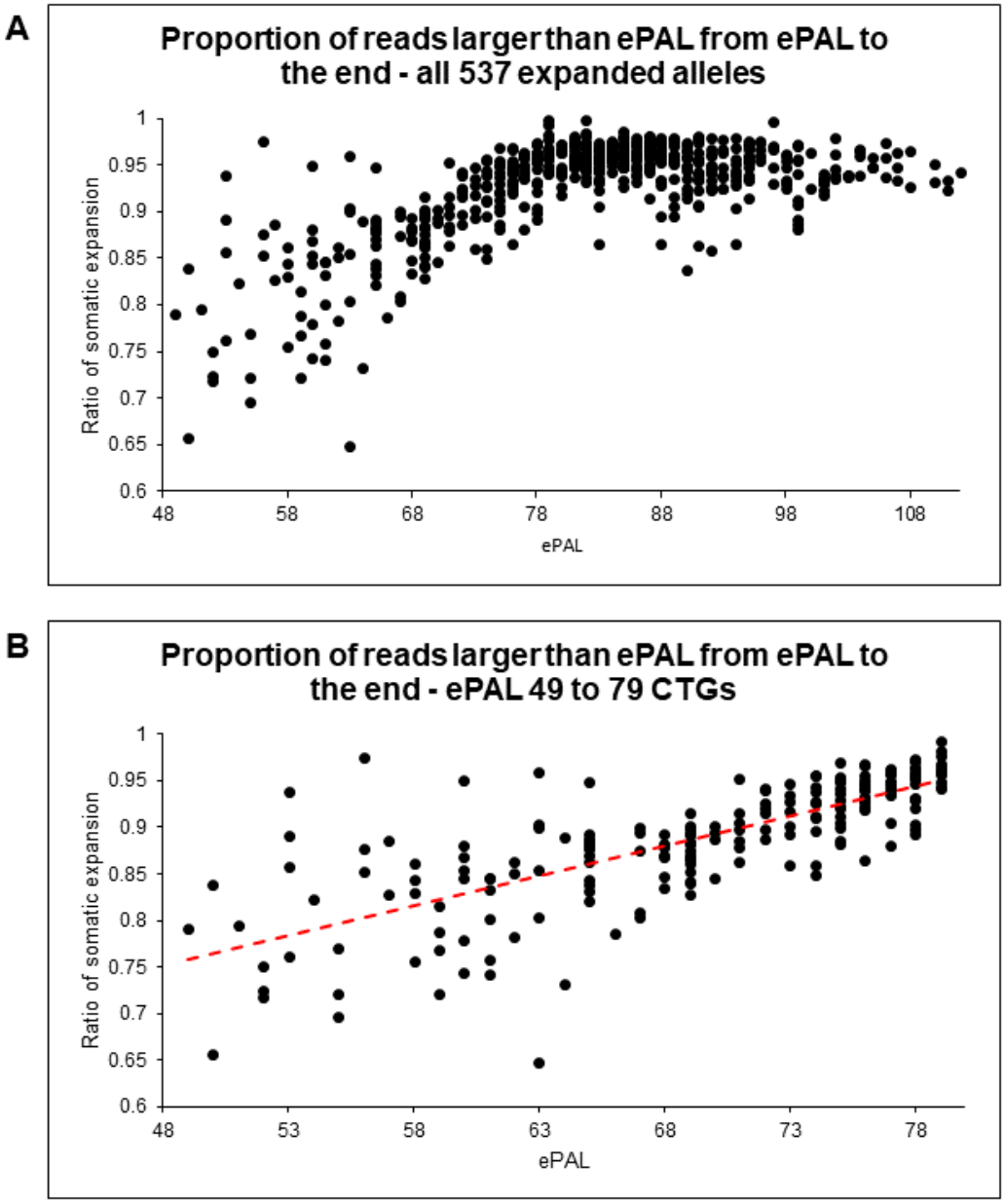


Figure 37 Somatic expansion in CTG18.1 expanded alleles (≥ 50 repeats) in Fuchs endothelial samples measuring the proportion of reads larger than estimated progenitor allele length (ePAL) from ePAL to the end. (A) Ratio of somatic expansion for all 537 expanded alleles in the cohort, showing a ceiling effect occurring in samples with an ePAL over 80 repeat. **(B)** Ratio of somatic expansion in samples with an ePAL between 49 and 79 repeats.

A second metric measure (2) to quantify somatic expansion, was included in this study to account for samples with 80 CTG repeats or more to account for samples with a larger ratio of somatic expansions beyond the

threshold the MiSeq assay can measure. The second metric measures the proportion of reads larger than 116 from ePAL to the end (number of somatic expansion products above 116 repeats/number of progenitor allele products). This metric enables inclusion of the large number of reads at 112 CTG repeats, some samples have which represent CTG lengths above 112 repeats. **Figure 38.A** shows this metric to calculate the ratio of somatic expansion for all 537 samples. Similar to the first metric, calculating the ratio of reads from ePAL to the end, there is a flooring effect for samples with ePALs below 79 CTG repeats but enables resolution of somatic expansion levels for samples with ePALs over 80 to be determined. On this basis method (1) was used for all samples with an ePAL ≤79 repeats and method (2) was used for all samples with an ePAL ≥80 repeats to calculate sample specific somatic expansions scores (**Table S4**).

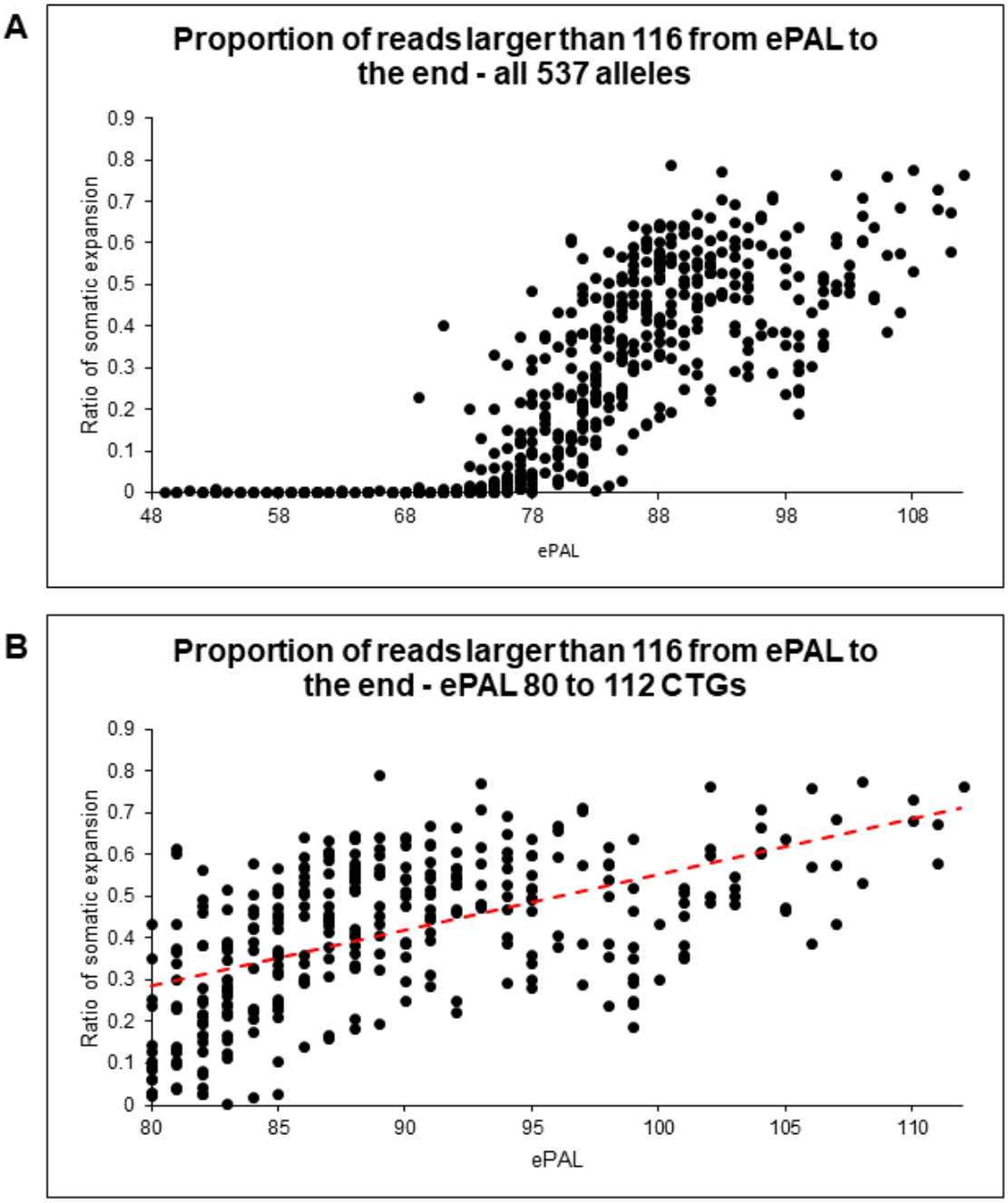


Figure 38 Somatic expansion in CTG18.1 expanded alleles (≥ 50 repeats) in Fuchs endothelial samples measuring the proportion of reads larger than 116 from estimated progenitor allele length (ePAL) to the end (number of somatic expansion products above 116 repeats/number of progenitor allele products). (A) Ratio of somatic expansion for all 537 expanded alleles in the cohort, showing a flooring effect occurring in samples with an ePAL below 79 repeats. **(B)** Ratio of somatic expansion in samples with an ePAL between 80 and 112 repeats.

4.2.3 Exploring genotype-phenotype associations (instability correlated to SNP data)

Targeted SNP data generated from the KASP assay and the somatic instability scores generated from the MiSeq assay, I sought to determine whether the variants within genes analysed correlate with the level of somatic expansion within FECD patients. In the first instance I performed this association study using samples with a self-reported European ancestry ethnicity (n = 459) to avoid any deviations across ancestry which artifactually affect associations. Individual-specific somatic expansion scores were defined as the residual variation in the ratio of somatic expansions corrected for sex, cohort, and an interaction between age at sampling and length of the CTG repeat. Linear regression models of the relationships between allele length and age with somatic expansions in FECD samples for alleles ≤79 and ≥80 was calculated by Dr. Marc Ciosi using The Akaike information criterion (AIC), an estimator of prediction error and thereby relative quality of statistical models for a given set of data (Ciosi et al., 2019). The model which produced the lowest AIC was selected to use in the genotype-phenotype association approach (**Table S5 and S6)**. Finally, the linear regression coefficient (β) was used as a measure to calculate the genotype-phenotype correlations between each SNP and residual variation in the quantity of somatic expansion considered alleles ≤79 and ≥80 independently, before combining the analysis to produce the final association results. p-value adjusted for multiple testing using the Benjamini & Hochberg (BH) False Discovery Rate (FDR) correction for multiple testing. The association between SNP data and somatic expansion scores are presented in **Table 24.** The results reveal a significant association between one *MSH3* SNP (rs701383) and two *FAN1* SNPs (rs34017474 and rs3512) with somatic

expansion scores, indicating that the minor allele at these residues are associated with a higher quantity of somatic expansions in blood-derived DNA within the studied FECD cohort.

Table 24 The genetic association data for 12 SNPs, selected from the GWA12345, a genome-wide association study conducted on patients with Huntington's disease (HD) (Consortium, 2019; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015), and somatic instability scores calculated using blood-derived DNA for white European Fuchs endothelial corneal dystrophy (FECD) patients recruited from Moorfield's Eye hospital (MEH) and General University Hospital in Prague (GUH) carrying one CTG18.1 expanded allele (≥50 repeats). Significant directional affect was identified for SNPs rs701383 (*MSH3),* rs34017474 (*FAN1*) and rs3512 (*FAN1*), highlighted in bold.

Chr	Gene	SNP	A1	A2	MAF	Total number of alleles*	*β*	Unadjusted p-value	BH-FDR corrected p-value
2	*PMS1*	rs5742933	C	G	0.213	459	-0.141	0.086	0.259
3	*MLH1*	rs1799977	G	A	0.289	454	-0.099	0.182	0.394
5	***MSH3***	**rs701383**	**A**	**G**	**0.252**	**456**	**0.202**	**0.01**	**0.039**
5	*MSH3*	rs1382539	A	G	0.242	458	-0.101	0.197	0.394
7	*PSM2*	rs852151	A	C	0.156	458	-0.072	0.426	0.731
8	*RRM2B, UBR5*	rs3735721	G	A	0.07	459	-0.005	0.969	0.989
14	*MLH3*	rs175080	A	G	0.469	457	-0.012	0.863	0.989
15	*FAN1*	rs150393409	A	G	0.009	458	-0.182	0.611	0.916
15	***FAN1***	**rs34017474**	**C**	**T**	**0.396**	**459**	**0.176**	**0.01**	**0.039**
15	***FAN1***	**rs3512**	**G**	**C**	**0.339**	**456**	**0.198**	**0.004**	**0.039**
19	*LIG1*	rs156641	T	C	0.358	457	0.014	0.837	0.989
19	*LIG1*	rs274883	G	A	0.168	458	-0.001	0.989	0.989

Chr: chromosome. SNP: Single nucleotide polymorphism. A1: minor allele. MAF: Minor allele frequency β: regression coefficient. BH-FDR corrected p-value adjusted for multiple testing using the Benjamini-Hochberg false discovery rate correction.

4.2.4 Using MiSeq to genotype downstream polymorphic CTC repeat and define allele structure

RGT was implemented to genotype the flanking CTC repeat downstream of the CTG18.1 repeat using the MiSeq reverse reads (R2). Previous research has demonstrated that the CTC repeat length was variable, especially on expanded CTG18.1 alleles, but has been consistently found to contain one CTT repeat interruption (see **Figure 20** in **Section 4.1**) (Alkhateeb, 2018; Hafford-Tear et al., 2019).

The typical allele for the CTG18.1 locus is $(CTG)_n(CTC1)_n(CTT)(CTC2)_n$. **Figure 39.A** shows visualised mapped reads using the visualisation tool Tablet (Milne et al., n.d.) for reads generated from an unexpanded allele with the allele structure $(CTG)_{14}(CTC)_6(CTT)(CTC)_6$. **Figure 39.B** shows a schematic model of the allele structure for this allele.

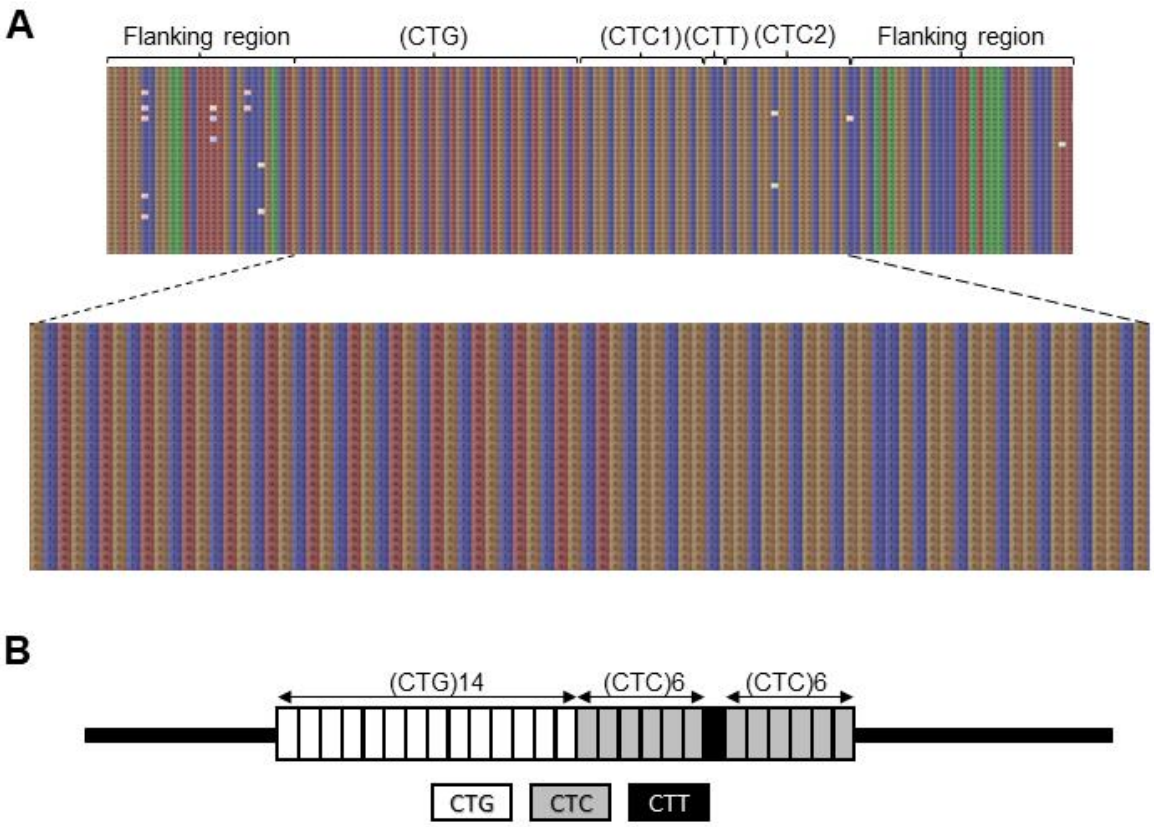


Figure 39(A) Mapping of MiSeq sequencing reads for CTG18.1 locus using Tablet to show structure of CTG18.1 locus and flanking region. CTG18.1 consists of CTG repeats followed by CTC1 and CTC2 repeats interrupted by one CTT. (B) Schematic diagram illustrating CTG18.1 locus.

MiSeq results revealed CTC repeats at the CTG18.1 locus were highly polymorphic. Allele structures were determined for all samples for which an ePAL could be called were determined (**Table S4**). In total, this included 563 samples and 1,126 alleles. **Table 25** summarises the allele structures for the non-expanded alleles and **Table 26** the allele structures for expanded alleles in this cohort. Schematics of the allelic structures are presented in **Figures 40 and 41** for non-expanded alleles and expanded alleles, respectively.

Table 25 Allele structures identified on CTG18.1 non-expanded alleles amplified from FECD patient derived gDNA samples. Structures have been labelled with a unique identifier A- O. Occurrence (n) corresponds with the total number of alleles observed for each structure. Occurrence (%) corresponds to the percent of non-expanded alleles with each respective structure analysed within the total FECD patient cohort (563 alleles).

Structure identifier	Non-expanded allelic Structure	Occurrence (n)	Occurrence (%)
A	$[CTG]_{11\text{-}16}[CTC]_6[CTT]_1[CTC]_6$	332	58.97%
B	$[CTG]_9GTG[CTG]_7[CTC]_6[CTT]_1[CTC]_6$	92	16.34%
C	$[CTG]_{18\text{-}29}[CTC]_6[CTT]_1[CTC]_6$	91	16.16%
D	$[CTG]_{30\text{-}49}[CTC]_6[CTT]_1[CTC]_6$	23	4.09%
E	$[CTG]_{17}[CTC]_6[CTT]_1[CTC]_6$	9	1.60%
F	$[CTG]_{13\text{-}14}[CTC]_7[CTT]_1[CTC]_6$	4	0.71%
G	$[CTG]_{32\text{-}44}[CTC]_5[CTT]_1[CTC]_6$	3	0.53%
H	$[CTG]_{38\text{-}39}[CTC]_7[CTT]_1[CTC]_6$	2	0.36%
I	$[CTG]_{41}[CTC]_4[CTT]_1[CTC]_6$	1	0.18%
J	$[CTG]_7[CTC]_6[CTT]_1[CTC]_6$	1	0.18%
K	$[CTG]_9[CTC]_6[CTT]_1[CTC]_6$	1	0.18%
L	$[CTG]_{10}[CTC]_6[CTT]_1[CTC]_6$	1	0.18%
M	$[CTG]_9GTG[CTG]_8[CTC]_6[CTT]_1[CTC]_6$	1	0.18%
N	$[CTG]_9GTG[CTG]_8[CTC]_5[CTT]_1[CTC]_6$	1	0.18%
O	$[CTG]_{14}[CTC]_1[CTG]_3[CTC]_6[CTT]_1[CTC]_6$	1	0.18%

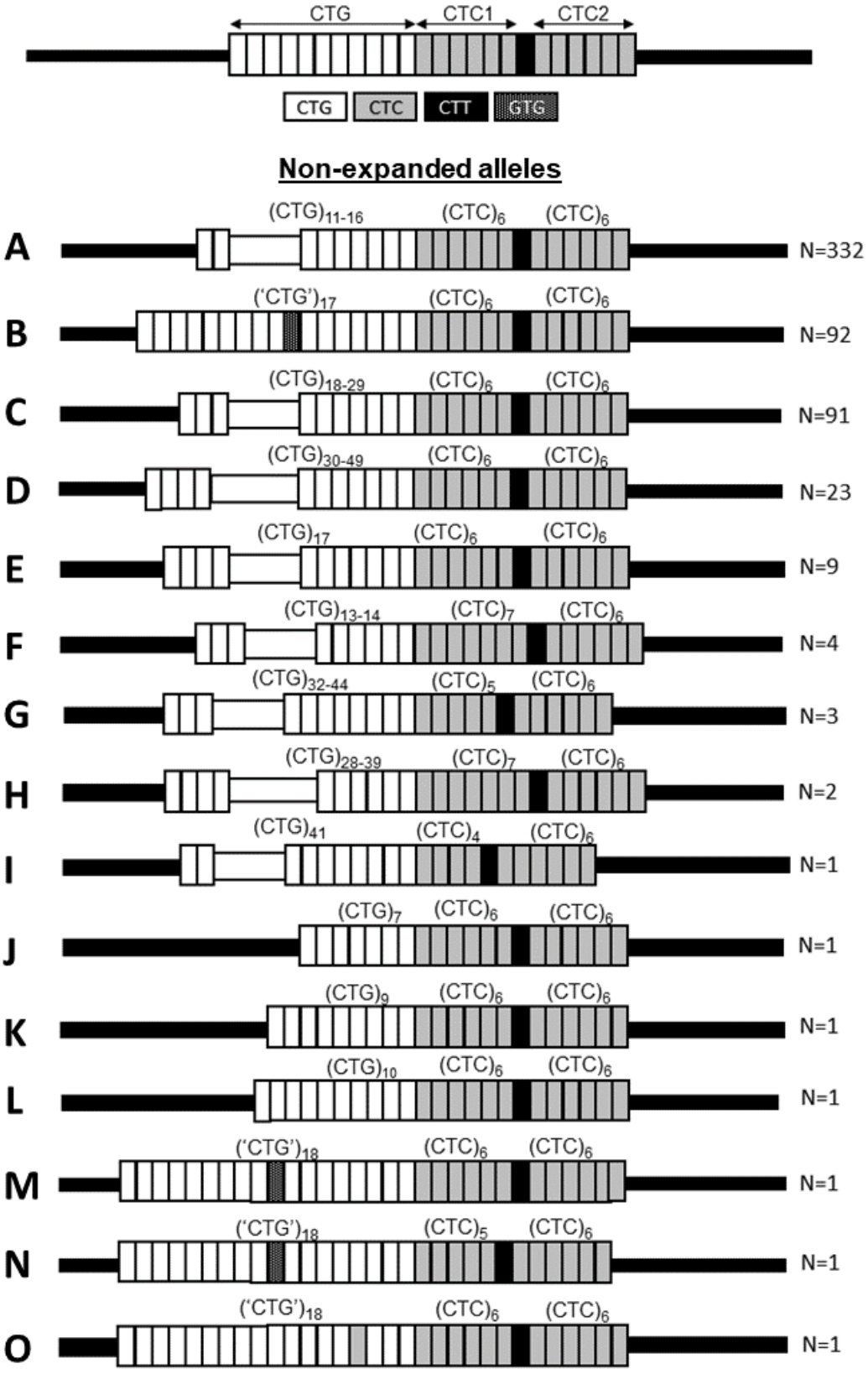


Figure 40 Schematic representation of allele structures identified on CTG18.1 non-expanded alleles within a FECD patient cohort Numbers (N=) on the right corresponds to the total number of alleles observed of each structure.

Table 26 Allele structures identified on expanded alleles amplified from FECD-patient-derived gDNA samples harbouring mono-allelic CTG18.1 expansions. Structures have been labelled with a unique identifier A- R. Occurrence (n) corresponds with the total number of alleles observed for each structure. Occurrence (%) corresponds with the percent of expanded alleles with each respective structure with the total cohort (total expanded alleles analysed =563).

Structure identifier	Expanded allelic Structure	Occurrence (n)	Occurrence (%)
A	$[CTG]_{49\text{-}112}[CTC]_6[CTT]_1[CTC]_6$	280	49.73%
B	$[CTG]_{58\text{-}111}[CTC]_5[CTT]_1[CTC]_6$	108	19.18%
C	$[CTG]_{55\text{-}107}[CTC]_7[CTT]_1[CTC]_6$	107	19.01%
D	$[CTG]_{59\text{-}110}[CTC]_4[CTT]_1[CTC]_6$	28	4.97%
E	$[CTG]_{60\text{-}86}[CTC]_8[CTT]_1[CTC]_6$	15	2.66%
F	$[CTG]_{80\text{-}111}[CTC]_3[CTT]_1[CTC]_6$	9	1.60%
G	$[CTG]_{70\text{-}95}[CTC]_6[CTT]_1[CTC]_8$	3	0.53%
H	$[CTG]_{59\text{-}93}[CTC]_9[CTT]_1[CTC]_6$	2	0.36%
I	$[CTG]_{102\text{-}103}[CTC]_6[CTT]_1[CTC]_5$	2	0.36%
J	$[CTG]_{97}[CTC]_1[CTT]_1[CTC]_6$	1	0.18%
K	$[CTG]_{93}[CTC]_2[CTT]_1[CTC]_6$	1	0.18%
L	$[CTG]_{87}[CTC]_5[CTT]_1[CTC]_7$	1	0.18%
M	$[CTG]_{73}[CTC]_6[CTT]_1[CTC]_7$	1	0.18%
N	$[CTG]_{97}[CTC]_7[CTT]_1[CTC]_5$	1	0.18%
O	$[CTG]_{65}[CTC]_1CGCTC[CTC]_4[CTT]_1[CTC]_6$	1	0.18%
P	$[CTG]_{80}[CCG]_1[CTC]_5[CTT]_1[CTC]_6$	1	0.18%
Q	$[CTG]_{11}[CAG][CTG]_{95}[CTC]_4[CTT]_1[CTC]_6$	1	0.18%
R	$[CCG][CTG]_{77}[CTC]_6[CTT]_1[CTC]_6$	1	0.18%

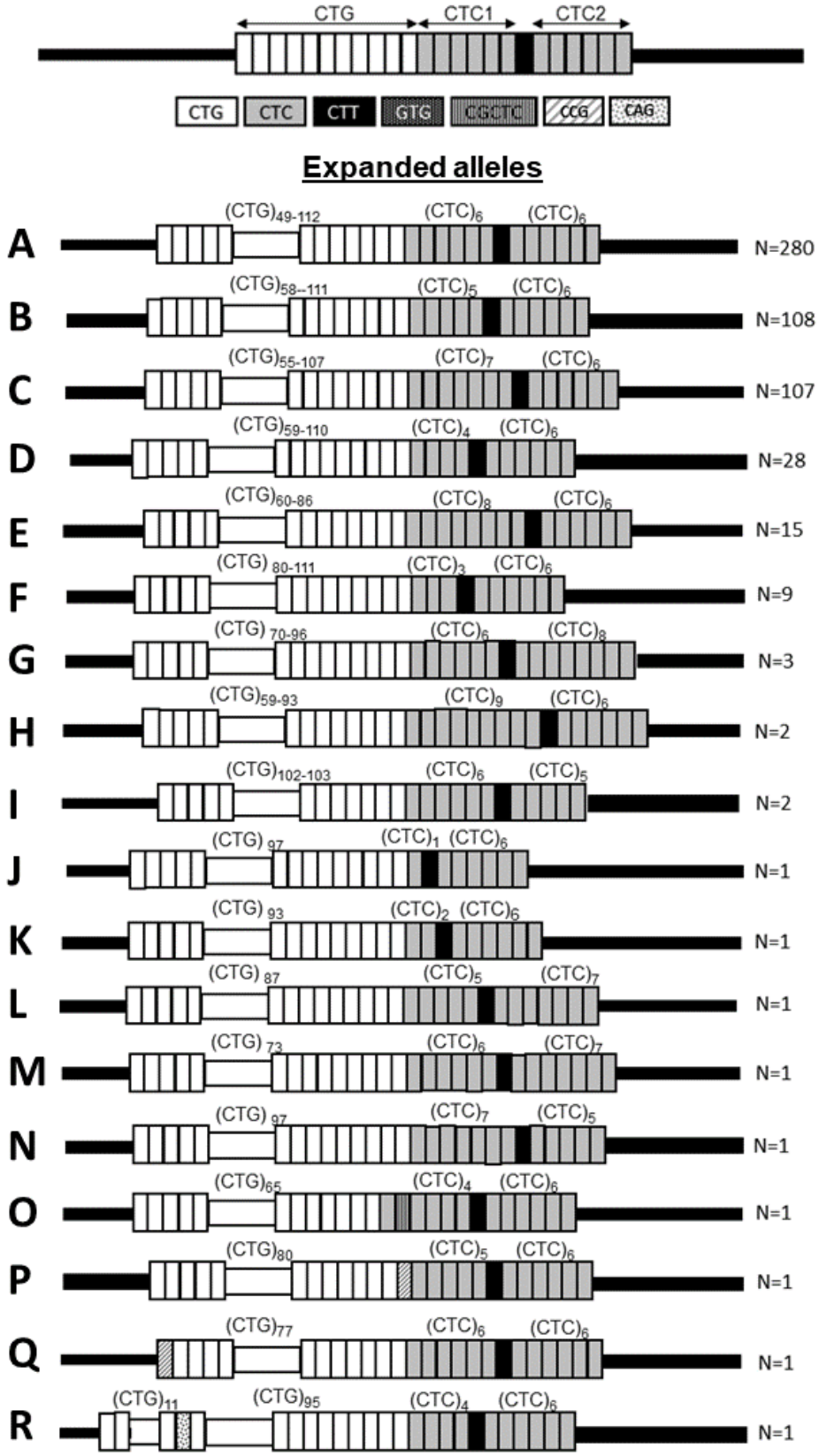


Figure 41 Schematic representation of allele structures identified on expanded CTG18.1 alleles within a FECD patient cohort. Numbers (N=) on the right corresponds to the total number of alleles observed of each structure.

4.2.4.1 Non-expanded CTG18.1 allele structures with a FECD patient cohort

For the non-expanded alleles, the most commonly occurring allele structure was structure A, **Table 25**, **Figure 40**, with almost 59% of samples with mono-allelic expansions harbouring this structure on their unexpanded alleles. This allele followed the typical allele structure with the CTG repeat length ranging from 11 to 16 repeats followed by the downstream $(CTC)_6(CTT)_1(CTC)_6$ structure. Following this, the second most common allele structure was structure B, seen in 16.3% of samples. This allele structure has a single 'GTG' variant within the CTG tract at position 10 and has also been described previously as $('CTG')_{17}(CTC)_6(CTT)_1(CTC)_6$ (Alkhateeb, 2018). Interestingly this allelic structure was absent from all expanded alleles (**Table 26, Figure 41**). In contrast, structure E, a consistent tract of 17 CTG repeats, followed by a downstream $(CTC)_6(CTT)_1(CTC)_6$ structure was only seen in 9 (1.60%) samples in this cohort (**Table 25, Figure 40**). Variation's structure B was also seen in two non-expanded alleles, structures M and N. Both of these structures differed from Structure B by having an additional CTG repeat. Structure N differed further by having a downstream CTC repeat of $(CTC)_5(CTT)_1(CTC)_6$. In addition, a similar allele structure, structure O, was seen in one sample in our cohort. This structure had a single 'CTC' motif interruption in the CTG tract at position 15. This allele structure was also seen three times in the DMGV and General Scotland cohorts (Alkhateeb, 2018).

Both longer and shorter variations in CTG repeat length from the most commonly observed $(CTG)_{11\text{-}16}(CTC)_6(CTT)_1(CTC)_6$ structure was seen in the cohort. Structure C, a pure CTG tract of 18-29 repeats was seen in 91 (16.16%)

samples, and less commonly 30-49 repeats in 23 (1.6%) samples. Furthermore, CTG repeat lengths of 7, 9 and 10 repeats were each seen in one sample.

The downstream $(CTC)_6(CTT)_1(CTC)_6$ motif appeared to be largely stable on unexpanded alleles with only 11/563 alleles deviating from this structure. Five samples had a longer $(CTC)_1$ repeat, $(CTC)_7(CTT)_1(CTC)_6$ structure, four of which had CTG repeat lengths of 13-14 (structure F) and two of which had CTG repeat lengths of 38 and 39 (structure H). Four samples had shorter (CTC1) repeats. Three of these with shorter (CTC1) repeats had a structure of $(CTC)_5(CTT)_1(CTC)_6$ with CTG repeat lengths ranging from 32-44 (structure G) and one sample had a structure of $(CTC)_4(CTT)_1(CTC)_6$ with a 41 CTG repeat length (structure I). (CTC2) remained stable at 6 repeats on all the unexpanded alleles investigated in this cohort.

4.2.4.2 Expanded CTG18.1 allele structures within an FECD patient cohort

Almost half of the samples with mono-allelic expansions harboured structure A on their expanded alleles. This allele structure had a pure CTG tract ranging from 49 to 112 repeats followed by a downstream $(CTC)_6(CTT)_1(CTC)_6$.

In contrast to the unexpanded alleles, there was more variation in the downstream (CTC1)(CTT)(CTC2) structure, with over half of the samples deviating from the typical $(CTC)_6(CTT)_1(CTC)_6$ (**Table 26, Figure 41**). Notably, on the non-expanded alleles there was only variation in the (CTC1) repeat, ranging from 4 to 7 CTC repeats. (CTC2) remained stable at 6 CTC repeats on all the unexpanded alleles. On the expanded alleles, (CTC1) varied from 1 to 9 CTC repeats and (CTC2) varied from 5 to 7 CTC repeats, suggesting expansions in the CTG tract resulted in the downstream (CTC1)(CTT)(CTC2) tract becoming more unstable.

Furthermore, four samples had interruptions within the CTG18.1 locus. One allele harboured Structure O, 65 CTG repeats followed by a single CTC motif, a CGCTC motif and then $(CTC)_4(CTT)_1(CTC)_6$. It appeared that the CGCTC motif interrupted the (CTC1) in the downstream (CTC1)(CTT)(CTC2) structure. Structure P was seen on one allele and had a CCG motif interruption downstream of 80 CTG repeats and before the $(CTC)_5(CTT)_1(CTC)_6$.

One sample had a CAG motif interruption within the CTG tract, sample Q. This allele had 11 CTG repeats followed by the CAG interruption before continuing the CTG tract for 95 repeats and a downstream $(CTC)_4(CTT)_1(CTC)_6$. At this time, it is unknown what effect, if any, this interruption had on the FECD phenotype in the given proband. The patient harbouring this allele was a 76-year-old female from the Czech Republic and did not present with any atypical phenotypic findings.

A further sample had an expanded allele with a CCG motif interruption at the beginning of a 77 repeat CTG tract followed by a downstream $(CTC)_6(CTT)_1(CTC)_6$, structure R. As this motif doesn't actually interrupt the CTG tract, it would be thought to have little effect on the FECD phenotype and instead shorten the repeat length from 78 to 77 repeats. This patient was a 67-year old white British female but did in fact have an atypical FECD phenotype. The patient displayed asymmetrical features atypical of FECD, with relatively thick corneas and scattered guttata on the posterior corneal surface of her left eye.

4.3 Discussion

4.3.1 Genotyping genetic modifiers SNPs in FECD

Variants within DNA repair, particularly those involved in MMR have been associated with an earlier age of HD onset (Consortium, 2019; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015) by influencing somatic instability rates of CAG repeats via downstream deficits in the DNA repair mechanism resulting in an accumulation of DNA damage (Massey & Jones, 2018). Incomplete penetrance CTG18.1 expansion mediated FECD has previously been described (Mootha et al., 2014). In this study I sought out to investigate if genetic DNA repair gene modifiers associated with HD also effect FECD pathogenicity through a mechanism common to CAG·CTG repeat expansions. I hypothesised an enrichment of HD-worsening modifiers may be present in the CTG18.1 expansion-positive FECD cohort. To test this hypothesis I selected 12 SNPs within genes, *FAN1*, *MLH3*, *MSH3*, *MLH1*, *LIG1*, *PMS1*, *PMS2* and *RRM2B, UBR5*, which have been extensively investigated in HD but also been associated with other repeat-mediated diseases including DM1 and SCAs (Bettencourt et al., 2016; Ciosi et al., 2019; Flower et al., 2019; Schmidt & Pearson, 2016).

In the first instance, White British and Czech FECD were analysed independently, and no significant difference was discovered with the exception of rs1799977 (*MLH1*) in the white British FECD cohort (**Table 19**). rs1799977 (*MLH1*) has been associated with a later residual HD onset of 0.8 years, compared to the majority of HD subjects (Consortium, 2019). From these data we can therefore hypothesise that the FECD patients which carry the functional modifier allele in these SNPs have a slightly delayed onset of disease, but this is yet to be investigated.

In attempts to maximise the power of this cohort, all patients with a European ancestry were pooled together and analysed as they displayed similar Chi Square values for the SNPs. This included all self-report European samples and samples in which the ethnicity was predicted to be European by a genome wide SNP array. Unfortunately, this resulted in the rs1799977 (*MLH1*) which was previously significant in the all-White British MEH cohort being no longer significant, but showed a weaker signal, still trending in the direction which is associated with delayed HD motor onset. Given the trending direction of effect size, it is possible that with a higher n number significance in this SNP and possibly others too may be detectable in future.

SNP analysis was also performed between white British FECD samples and an age-matched cohort in which carried repeat expansions but did not display clinical symptoms of FECD (**Table 22**). The cohort consisted of patients who were diagnosed with AMD that had undergone an eye examination and did not have any records of FECD in their electronic MEH patient notes. As the AMD samples carry the CTG18.1 expanded allele but do not present with FECD symptoms it would be expected these samples would carry SNPs which delay or are protective of developing disease. Although there were no significant comparisons between the FECD cohort compared to the AMD cohort, two SNPs, rs1799977 (*MLH1*) and rs1382539 (*MSH3*), followed the hypothesised trend patterns in the AMD cohort where the frequency of these SNPs have been associated with delayed HD motor onset (Ciosi et al., 2019; Consortium, 2019). These findings support the hypothesis that these two SNPs have some sort of protective function in delaying the onset of disease since it may be the case that the AMD samples could remain completely asymptomatic or go on to develop FECD symptoms later in life.

4.3.2 Using MiSeq to quantify somatic instability

Illumina MiSeq next generation sequencing was employed on the FECD samples found to have one or two CTG18.1 expanded (≥50 repeats) alleles using the STR and TP-PCR assays from **Section 3.2.2**. MiSeq sequencing is advantageous over the STR method, which simply sizes amplified fragments, as it also generates sequence level resolution, enabling the detection of the presence and frequency of interruptions within the repeat, confirmation of allele structure, and the ultra-deep sequencing approach enables quantification of somatic instability of the repeat in blood. Furthermore, it is a high-throughput technique that enables relatively large-scale and cost-effective sequencing data to be generated. This method enables sequence level resolution to explore the hypothesis that the presence and/or absence of variant repeats, seen in HD and DM1 may also occur in FECD.

One major limitation to the MiSeq assay is the repeat length threshold it is able to sequence efficiently. Theoretically MiSeq sequencing can sequence up to ~ 182 CTG repeats at the CTG18.1 locus assuming sequencing from the forward and reverse primer is 35 bp and 13 bp, respectively from the either side of the allele $(CTG)_n(CTC)_6(CTT)(CTC)_6$. The standard method recommended by Illumina is to sequence 300 bp from the forward strand and 300 bp reverse which would only enable up to 88 CTG repeats to be sequenced. Although MiSeq can theoretically sequence up to 600 bp in one direction, experience from Darren's G. Monckton lab (DGM) has revealed that, when sequencing 600 bp in one direction, after ~ 400 bp the sequencing quality drops notably (unpublished data Dr. Sarah Cumming, Dr. Marc Ciosi and Dr. Asma Alshammari from DGM lab). This experience suggests that sequencing 600 bp in one direction is not a good approach. Nonetheless, sequencing 400 bp

forward and 200 bp from reverse had previously yielded good data for both DM1 and HD. On this basis, sequencing the CTG18.1 locus with 400 bp reads, could potentially sequence approximately 121 CTG repeats, with the assumption that the allele structure is $(CTG)_n(CTC1)_6(CTT)(CTC2)_6$. 400bp forward reads captured the CTG sequence and 200bp reverse read captured the CTC sequence for each sample. Data presented in this chapter demonstrates that I was only able to accurately sequence up to 112 CTG repeats without reducing the quality of the data produced, with 118 repeats being the absolute limit. I was unable to determine ePAL for 34 samples which had mono-allelic expansions determined previously by either the STR or TP-PCR assays. This was either due to the ePAL exceeding the MiSeq threshold or large levels of somatic instability causing a build-up or repeats at '118' repeats which muted the reads at lower CTG values. Furthermore, this limitation to sizing also meant I was unable to capture the complete levels of mosaicism for those with larger repeats, in which No-AMP SMRT sequencing has been previously demonstrated to achieve (Hafford-Tear et al., 2019), however this approach is much lower throughput and much more expensive and could therefore not feasibly be applied to the total cohort.

For the majority of samples, the MiSeq assay and STR assay produced the same or similar genotypes (**Table S4**). Differences between the ePAL called and the STR genotype can be explained by the genotyping designated from the STR assay were defined as the modal repeat length rather than the progenitor allele. The ePAL is the allele size transmitted by the affected parent to the affected offspring and while the ePAL and the modal length measure can be the same in many cases, the modal length has the ability to fluctuate due to somatic instability. Calling the ePAL has previously been documented to

significantly improve genotype-phenotype analyses over the traditional modal length measure, as it greatly reduces the confounding effects of somatic instability (Cumming et al., 2019; Morales et al., 2012). In future, ePAL data here will be correlated with phenotypic outcome measures to determine if this also is the case for FECD.

Bimodal repeats in the CTG distribution were seen in over 60% the FECD samples analysed in this study, where two discrete populations can be observed, one centred around the progenitor allele (i.e., stable repeats) and those further expanded (i.e., unstable repeats). The distribution can appear as one mode having gained in repeat length and the other mode having decreased in size; this is a result in somatic instability occurring (J. M. Lee et al., 2011). The extent at which samples appeared bimodal varied greatly between samples. This can be expected as the bimodal distribution takes place over time as somatic instability greatens with some patients being at different stages of disease. The variation could be dependent on the size of the allele inherited and also the age of the patient. Bimodal distributions have been witnessed in HD where different cell types display distinct distribution characteristics. Bimodal CAG repeat length distributions were more apparent in liver consisting of CAG repeats centred around the constitutive repeat, and are stable repeats, and those which are further expanded and are unstable repeats. In contrast, in striatum, the bimodal distribution of repeats is much less obvious with the population of unstable repeats being greater and therefore more widely distributed (Ciosi et al., 2021; J. M. Lee et al., 2019). It would be expected that the distribution of CTG18.1 would also vary in cell types for FECD, with the corneal endothelium being the affected tissue, expecting to see higher levels of somatic instability resulting in a more widespread distribution and a less

apparent bimodal distribution that we can see here in peripheral blood samples. Furthermore, the bimodal distribution I observed could also be explained since blood-derived DNA was used in this study, which consists of a number of different cell types.

As expected, I observed large levels of somatic expansion occurring in the blood-derived gDNA samples analysed consistent with our previous observations (Hafford-Tear et al., 2019). As mentioned earlier a limit to the MiSeq assay is the length of repeats it is able to sequence. We know from previous studies that higher levels of repeat instability are associated with increased CTG18.1 allele length (Hafford-Tear et al., 2019; Wieben et al., 2021). The same trend can be witnessed in this study despite the sequencing threshold, and you can see a build-up of reads at '118' repeats. This build-up of reads represents reads which contain repeats of 118 or longer. In samples with longer ePALs the abundance of reads at 118 are greater than samples with short ePALs, strengthening the finding that there are greater levels of somatic instability with a longer CTG18.1 allele length (**Figure 35**).

As the MiSeq assay is not able to give a full representation of the extent of somatic instability for longer alleles, I attempted to measure somatic expansions using two methods to produce the most informative analysis with the data available. The two methods to measure levels of somatic expansions were (1) The proportion of reads larger than ePAL from ePAL to the end and (2) the proportion of reads larger than 116 from ePAL to the end. The second method allows estimating the extent of reads longer than the MiSeq is able to sequence and therefore gives a better indication of the levels of somatic instability for these samples it was not possible to capture. Both of these measures positively correlated with ePAL but measure (1) effectively captured

variation for alleles which had an ePAL below 79 repeats in length, where there were no or limited levels of somatic instability that surpassed the MiSeq threshold of 118 repeat. Measure (2) was able to effectively capture variation for alleles with an ePAL longer than 80 repeats due to the proportion of reads at 118 indicating longer lengths of instability.

4.3.3 Genotype-phenotype association between somatic expansion scores and genetic modifier SNPs

ePAL alone is unable to explain the variable expressivity of FECD phenotype. **Figure 36** shows no significant correlation was observed CTG18.1 repeat length and the age at recruitment, indicating that there are additional factors involved which modify phenotypic outcomes of FECD. Variants in DNA repair genes, specifically MMR genes have been established to modify residual variation in HD outcomes not accounted for by measured CAG length (Ciosi et al., 2019; Consortium, 2019; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015). As a starting to point to investigate if *trans*-acting modifiers, such as those previously associated with HD, have an implication on FECD phenotype, I investigated the association between 12 genetic SNP modifiers and somatic expansion scores with a large FECD patient cohort (n=459). I identified a significant directional effect for SNPs rs701383 (*MSH3),* rs34017474 (*FAN1*) and rs3512 (*FAN1*), **Table 24**.

The minor allele at rs3512 (*FAN1*) has previously been associated with higher levels of somatic expansions in the HTT repeat in blood-derived DNA from HD patients and has further been correlated with a later onset of HD (Ciosi et al., 2019; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015). Likewise, this variant has also been associated in the same direction for SCAs (Bettencourt et al., 2016). I have now identified this same directional

association between blood-derived CTG18.1 somatic expansion rates and the rs3512 (*FAN1*) variant. This finding supports the hypothesis that this variant has a shared mechanism in governing somatic instability consistent across repeat expansions and not limited to CAG repeats alone (Bettencourt et al., 2016). The FAN1 nuclease plays a role in repair of DNA interstrand cross-links (MacKay et al., 2010).

The *MSH3* SNP rs701383 has previously been defined as the top genome-wide significant SNP in a HD GWAS to identify disease-modifying factors, and was found to have the same directional association as identified in this study, with the minor allele at this SNP being associated with higher levels of somatic expansions (Consortium, 2019). Furthermore, in HD this SNP has been associated with hastened HD onset by an average of 0.8 years (Consortium, 2019). Polymorphisms within *MSH3* have also been implicated with increased levels of somatic instability of the expanded CTG repeat in the blood DNA of DM1 patients (Morales et al., 2016). *MSH3* encodes the DNA mismatch repair protein MutSβ, along with *MSH2,* and has an involvement in recognising and repairing slippage mistakes in microsatellite sequences and in addition is involved in double-strand break repair via homologous recombination (Tseng-Rogenski et al., 2020).

I did not detect associations between *LIG1*, *PMS1*, *PMS2*, *MLH1*, *MLH3, RRM2B* and *UBR5* polymorphisms and CTG18.1 somatic expansion rates. This could possibly be due to a limitation of the relatively small sample size analysed in this study. These findings suggest that *trans*-acting factors, such a genetic variation in DNA repair genes, do have a role in governing somatic instability and future work is needed to establish how this affects FECD phenotypic outcomes.

4.3.4 Using MiSeq to define CTG18.1 allelic structure

Using both the forward and reverse reads produced by the MiSeq sequencing, the complete allele structure of the CTG18.1 locus could be determined for samples where ePAL was able to be determined. I did this for both the non-expanded normal alleles and the expanded alleles of samples which carried mono-allelic expansions of ≥50 repeats. In the non-expanded alleles, structure B (**Table 25**, **Figure 40**) was the second most commonly seen structure in 92/563 alleles. This structure harboured a single 'GTG' variant within the CTG tract at position 10, interestingly, there was no variation of this structure seen in the expanded alleles. This suggested this 'GTG' variant could have a protective function preventing the CTG tract from expanding further. In previous studies it has been shown that variation of the downstream CTC repeat, consisting of a basic allele structure of $(CTC1)_n(CTT)_1(CTC2)_n$, has been associated with expanded CTG tracts (Alkhateeb, 2018). My findings in this study strongly support this with the majority of non-expanded alleles, 98%, maintaining a stable $(CTC)_6(CTT)_1(CTC)_6$ structure. In the expanded alleles, only 49.9% of alleles had this $(CTC)_6(CTT)_1(CTC)_6$ structure, with the rest of the alleles having some form of variation. Most of the variation occurred in the (CTC1) and again expanded alleles having a larger degree of variation, from the non-expanded alleles ranging from 4 to 7 CTC repeats and the expanded alleles ranging from 1 to 9 CTC repeats. Again, this is supportive that the expansions in the CTG tract influence variation in the downstream CTC repeat. Currently it is unknown how the CTG repeat influences the downstream CTC repeat and what effect it has on the FECD phenotype.

For other repeat-mediated disorders sequence interruptions have been noticed and have an impact on the dynamics of the pathophysiology of the

disease. For example, 3-5% of DM1 patients have been found to have interruption of other motifs, including CCG, CTC or GGC within the *DMPK* CTG repeat. Furthermore, these interruptions have been found to have implications on the clinical phenotype, where patients with variant repeats may exhibit delayed onset, unusually mild symptoms, or atypical patterns of symptoms. This is thought to be a result of increased stability of the repeat in the germline (Cumming et al., 2018; Musova et al., 2009; Santoro et al., 2013). In this study I sought out to find variants within the CTG repeat in FECD, and if any, whether they had any impact on the FECD phenotype. Surprisingly, I only discovered two samples, one of which had a CAG motif interruption at position 12 of a CTG tract of 95 repeat, structure Q, **Figure 41,** and a second sample that had CCG motif interruption at the very beginning of the CTG repeat tract, structure R**, Figure 41**.

The patient which harboured a CAG motif interruption at position 12 in the CTG18 repeat tract was a Czech female diagnosed at 76-years of age. Clinical information gathered suggested there was nothing unusual about her phenotype but she had yet to undergo the need for corneal transplantation surgery. As this is an isolated case, it is not possible to draw any conclusions as to whether this interruption has an impact on the phenotype or progression of disease severity. However, as this patient is yet to reach a stage of disease requiring surgical intervention it could be indicative this variant may slow disease progression.

For the sample which had a CCG motif at the start of the CTG repeat, it would be expected that this variant would not have an impact on the disease phenotype as it does not actually interrupt the repeat tract and instead shorten the repeat length from 78 to 77 repeats. However, this patient did present with

an atypical FECD phenotype displaying asymmetrical features, with relatively thick corneas and scattered guttata on the posterior corneal surface of her left eye. Again, as this is an isolated case, it is not possible to draw any conclusions as to whether this motif is the cause of this patient's atypical phenotype features.

Thirty-eight samples included in the MiSeq study harboured biallelic expansions, identified through the STR and TP-PCR assays. Initially I had planned to analyse these samples with the aim to further investigate how bi-allelic expansion may affect FECD phenotype however due to the timing I was unable to do so. To characterise the alleles in these samples, they would first have to be phased using the downstream CTC repeat. This could be done if the alleles carried independent CTC genotypes.

In addition, I had also sequenced AMD samples using the MiSeq assay which harboured either a CTG expansion (≥50 repeats) or an intermediate expansion (30-50 repeats). I planned to investigate these samples with the aim to explore why they did not present with a FECD phenotype when carrying a CTG expansion. As previously explained, these samples harboured a late-onset disease, typical of when FECD would begin to present and had undergone extensive eye examinations where no symptoms of FECD were present. There is a possibility they may develop symptoms later on in life, but this would still be much later than the typical age of onset for FECD. FECD characteristics such as guttae can be present for many years before noticeable symptoms (Goar, 1933) and these AMD patients did not present with any symptoms recorded in their electronic notes. I did manage to briefly explore the CTG repeat sequences of these samples and did not find any interruptions within the repeat tract that could explain why they did not develop a FECD phenotype. With

more time, I would have liked to explore these samples in greater detail, including the levels of somatic instability they harbour and if they had variations present in the downstream CTC repeat.

4.4 Conclusion

This study has used an ultra-high throughput MiSeq method to demonstrate the CTG18.1 repeat is a dynamic unit when expanded. Here I have provided new evidence that the CTG18.1 is somatically unstable and instability of the repeat increases with larger ePAL lengths. Additionally, in this section I have explored the frequency and influence of genetic variants within DNA repair genes amongst FECD patients and found a significant enrichment of the minor allele SNP rs1799977 (MLH1) in the white British FECD, previously associated with a delayed onset of disease in HD. Furthermore, two polymorphisms, rs1799977 (*MLH1*) and rs1382539 (*MSH3*), were found to be enriched within the cohort without FECD symptoms where the frequency of these SNPs has been associated with delayed HD motor onset. This finding is suggestive the potential of protective function delaying the onset of disease.

Furthermore, association studies combining the somatic instability data and SNP frequencies within DNA repair has provided an insight into the effect trans-acting modifiers had on the CTG18.1 repeat. Minor alleles at the *MSH3* SNP rs701383 and *FAN1* SNPs rs34017474 and rs3512 were all significantly associated with increased levels of somatic instability. These findings supports the hypothesis that there are underlying modifiers which play a role in variable FECD expressivity which should be further investigated.

5. Exploring the genetic architecture of non-expanded CTG18.1 FECD using exome sequencing

5.1 Introduction

In this chapter, exome sequencing has been employed to explore the genetic aetiology of FECD in a large genetically refined CTG18.1 expansion-negative FECD cohort. Importantly, this work represents the first relatively large-scale attempt to genetically characterise FECD cases that do not harbour a CTG18.1 expansion. As described earlier, **Section 1.2.11**, several genes have previously been associated with FECD through linkage analysis, indicating FECD has high genetic heterogeneity, however, the frequency in which these genes are causative for FECD is unknown given these studies have been conducted within small groups of patients or isolated families. In recent years exome sequencing has become routinely used in research for identifying rare variants which may contribute to the pathogenesis of Mendelian and complex, multifactorial disorders, replacing traditional linkage and candidate gene re-sequencing approaches previously used for novel Mendelian disease gene discovery (Bamshad et al., 2011; Cirulli & Goldstein, 2010). It is especially useful for conditions with high genetic heterogeneity, where Sanger sequencing of all potential causative genes would be time consuming and expensive (Rabbani, Tekin, & Mahdieh, 2014).

Given that FECD can approximately affect up to 5% of the population, and 80% of FECD can be attributed to the *TCF4* CTG18.1 expansion, the genetic cause for FECD in this cohort must be seen at a frequency of 1% or less in the population. Assuming FECD is highly heterogeneous, there are likely several causative variants within this cohort and thus variants above 1% (<0.01) in control datasets were excluded for the purpose of this study. Furthermore,

given that the vast majority of CTG18.1-expansion negative patients recruited to this study are simplex cases it is not currently possible for us to perform meaningful linkage analysis at present. With the lack of large multiplex families, locus heterogeneity, and the incomplete penetrance which comes with FECD using exome sequencing alone to efficiently prioritise the vast number of variants generated by exome sequencing can be challenging. However, alternative approaches to using exome sequencing data by the means of gene-based burden testing, in which the aggregate frequency of “qualifying variants” is compared between case and control subjects for each gene can provide a powerful method in identifying novel candidate genes (M. H. H. Guo et al., 2016; M. H. Guo, Plummer, Chan, Hirschhorn, & Lippincott, 2018).

In this chapter I use exome sequencing with the aims of exploring the genetic heterogeneity and identifying novel genetic candidates in the first relatively large-scale CTG18.1 expansion-negative FECD cohort.

5.2 Results

A total of 220 FECD cases were identified in Chapter 3 to not carry an expanded CTG18.1 allele (defined as ≥50 repeats; **Sections 3.2.2**). Here a total of 141 of these samples (total number available for analysis time this investigation was performed) were selected for exome sequencing to explore potential alternative genetic causes of disease. Exome capture, library preparation and sequencing were outsourced to the commercial sequencing provider Novogene and the raw sequencing data generated were aligned and annotated courtesy of Dr Nikolas Pontikos and Anita Szabo, bioinformaticians at UCL, as described in the methods, **Section 2.8.2**.

The MAF of variants were annotated in accordance with gnomAD, composed of both exome and genome sequences, (Version 2.1.1; n=141,456) (Karczewski et al., 2020), Kaviar genomic variant database (Glusman et al., 2011) and the internal UCLex database (Pontikos et al., 2017). Importantly, UCLex (n= 5,583) exome sequencing data was aligned and annotated using the same informatic pipeline. Thus, comparison of FECD case data against this internal dataset enabled identification of alignment and annotation artefacts.

5.2.1 Rare variants identified in genes previously associated with FECD

Initially, rare variants (MAF = <1%) in previously identified FECD-associated genes (**Table 3**) using frequency data available in gnomAD and UCLex, were identified across all 141 exome datasets generated from the molecularly unsolved FECD cases. Variants were then prioritised based on their predicted functional impact, pathogenicity and frequency in control datasets. Aligned sequencing reads (BAM files) were visualised in IGV to confirm variants were not artefacts. For each variant of interest identified, primers were designed to amplify the region containing the variant and validation was performed by Sanger sequencing. No variants were excluded through Sanger sequencing.

In total, 64/141 probands were identified to harbour 44 rare (MAF ≤1%) variants in FECD-associated genes assessed (**Table 27**). After independent validation by Sanger sequencing, variants were classified as pathogenic, potentially pathogenic, variant of unknown significance (VUS) or likely benign, based on population frequency data available, reports in additional FECD probands, in addition to in silico prediction scores. Furthermore, the expression profiles of these given genes within the corneal endothelium was determined using publicly available RNA-Seq data (**Section 2.7**) (Chen et al., 2013) to contextualise the potential for different categories of variants to induce disease

specifically within the corneal endothelium. Notably, *COL8A2*, *SLC4A11*, and to a lesser extent *ZEB1* are all expressed within healthy adult corneal endothelium, whereas *AGBL1* and *LOXHD1* are not (see TPM values listed in **Table 27**).

Table 27 Summary of rare, potentially deleterious variants identified in FECD-associated genes from a total of 141 FECD cases analysed by exome sequencing. In total 141 FECD cases were analysed and found to harbour a total of 44 variant based on the filtering criteria applied; MAF < 0.01 in publicly available gnomAD genomes, exomes and Kaviar, CADD score > 10.

Subject ID	Functional change	Genomic coordinates (Hg19)	Change	In silico predictions			GnomAD Frequency (Total)	UCLex Frequency (AC/AN)	UCLex Frequency without FECD cases	Reported as FECD-associated	Variant interpretation
				CADD	DANN	Reveal					
COL8A2 (ENST00000397799.2) **247.16TPM**											
BR1; BR64	MS	1-36563919-G-T	c.1363C>A, p.(Gln455Lys)	12.52	0.883	0.551	0 (0/0)	0 (0/0)	0 (0/0)	Yes	Pathogenic
CZ49	MS	1-36563558-G-A	c.1724C>T, p.(Pro575Leu)	22.5	0.999	0.719	0.001403 (381/271566)	0.0006464 (7/10830)	0.000554119 (6/10828)	No	Likely Benign
CZ14	MS	1-36563981-C-T	c.1301G>A, p.(Arg434His)	22.7	0.908	0.282	0.001141 (222/194554)	0.0015637 (16/10232)	0.001466276 (15/10230)	Yes	Potential pathogenic
SLC4A11 (ENST00000380059) **2938.99TPM**											
BR60	MS	20-3215432-C-A	c.326G>T, p.(Arg109Leu)	31	0.998	0.609	0.0001096 (31/282798)	0.0002660 (3/11280)	0.000177336 (2/11278)	No	VUS
BR63	MS	20-3211845-C-T	c.1121G>A, p.(Arg374Gln)	19.24	0.997	0.102	0.0006471 (183/282792)	0.0000888 (1/11266)	0 (0/11264)	No	Likely Benign
BR30	MS	20-3211846-G-A	c.1120C>T, p.(Arg374Trp)	15.66	0.937	0.189	0.001160 (328/282796)	0.0030179 (34/11266)	0.002929688 (33/11264)	No	Likely Benign
BR38;BR63+	MS	20-3214851-T-C	c.530A>G, p.(Asn177Ser)	14.73	0.938	0.259	0.004128 (1167/282680)	0.0010667 (12/11250)	0.000800285 (9/11246)	No	Likely Benign
BR34	SS	20-3218242-G-C	c.173-8C>G	10.98	0.618	0	0.0003076 (87/282850)	0.0003604 (4/11098)	0.000270368 (3/11096)	No	Likely Benign
ZEB1 *(*ENST00000361642) **4.11TPM**											
BR47	MS	10-31803541-C-T	c.698C>T, p.(Thr233Met)	28.9	0.999	0.199	0.0001916 (48/250576)	0.0000928 (1/10780)	0 (0/10778)	No	Likely Benign
CZ35	MS	10-31809258-T-C	c.998T>C, p.(Ile333Thr)	21.4	0.975	0.16	0.00004385 (11/250872)	0.0000888 (1/11264)	0 (0/11262)	No	Likely Benign

BR73	MS	10-31815913-G-C	c.3099G>C, p.(Glu1033Asp)	16.98	0.165	0.15	0.00002035 (5/245750)	0.0000892 (1/11212)	0 (0/11210)	No	Likely Benign
BR35;CZ29	MS	10-31810514-A-G	c.2254A>G, p.(Thr752Ala)	18.81	0.981	0.024	0.001374 (388/282378)	0.0016167 (18/11134)	0.001437556 (16/11130)	No	Likely Benign
BR21;BR34	MS	10-31810823-C-A	c.2563C>A, p.(Gln855Lys)	19.55	0.886	0.125	0.001807 (510/282184)	0.0008978 (10/11138)	0.00071852 (8/11134)	Yes	Likely Benign
BR19;CZ19	MS	10-31810782-A-C	c.2522A>C, p.(Gln841Pro)	26.6	0.996	0.326	0.007636 (2153/281936)	.0054926 (62/11288)	0.005317263 (60/11284)	Yes	Likely Benign
BR6;BR32; BR63;BR38+	MS	10-31809921-A-G	c.1661A>G, p.(Lys554Arg)	21	0.995	0.053	0.00512 (1446/282254)	0.0015940 (18/11292)	0.001152074 (13/11284)	Yes	Likely Benign
BR11	IFD	10-31750006-AGAT-A	c.105_107del, p.(Asp35del)	20.4	0	0	0.002589 (731/282304)	0.0012610 (14/11102)	0.001171171 (13/11100)	No	Likely Benign
CZ26	SS	10-31809047-T-G	c.794-7T>G	17.51	0.838	0	0(0/0)	0.0000918 (1/10894)	0 (0/10892)	No	VUS
***AGBL1* (ENST00000635782) 0.00TPM**											
BR17	MS	15-86791003-A-T	c.490A>T, p.(Ile164Phe)	21.9	0.981	0.333	0.000004020 (1/248740)	0.0000948 (1/10550)	0 (0/10548)	No	VUS
CZ42	MS	15-86800157-C-T	c.671C>T, p.(Thr224Met)	14.05	0.971	0.081	0.00009271 (26/280436)	0.0004683 (5/10676)	0.000374742 (4/10674)	No	Likely Benign
BR11	MS	15-86822926-A-G	c.1994A>G, p.(Tyr665Cys)	25.1	0.998	0.285	0.005605 (1572/280458)	0.0040209 (43/10694)	0.003928171 (42/10692)	No	Likely Benign
CZ49;CZ51	MS	15-86800154-C-T	c.668C>T, p.(Pro223Leu)	26.6	0.999	0.312	0.008278 (2321/280380)	0.0107698 (115/10678)	0.010586472 (113/10674)	No	VUS

BR5	MS	15-87099481-A-G	c.2884A>G, p.(Lys962Glu)	10.56	0.723	0.007	0.003450 (966/280000)	0.0008585 (9/10484)	0.000763213 (8/10482)	No	Likely Benign
BR79	NS	15-87217666-C-T	c.3082C>T, p.(Arg1028Ter)	34	0.846	0	0.001753 (486/277246)	0.0034240 (34/9930)	0.003323932 (33/9928)	Yes	Potentially pathogenic
***LOXHD1* (ENSP00000300591.6/ENST00000536736.5 *) 0.01TPM**											
CZ44	MS	18-44057153-C-A	c.3338G>T, p.(Cys1113Phe)	14.06	0.893	0.005	0.00001277 (2/156564)	0.0001915 (2/10442)	9.57854E-05 (1/10440)	No	Likely Benign
CZ43	MS	18-44121750-A-C	c.569T>G, p.(Leu190Arg)	15.06	0.895	0.084	0.000006310 (1/158476)	0.0000929 (1/10764)	0 (0/10762)	No	Likely Benign
CZ39	MS	18-44190795-T-G	c.703A>C, p.(Lys235Gln)*	25.7	0.889	0.457	0.00003184 (1/31408)	0.0000955 (1/10472)	0 (0/10470)	No	VUS
BR13	MS	18-44121778-G-A	c.541C>T, p.(Leu181Phe)	24.2	0.916	0.3	0.0002001 (38/189872)	0.0006489 (7/10788)	0.000556277 (6/10786)	Yes	VUS
BR65	MS	18-44102126-G-A	c.1690C>T, p.(Arg564Cys)	18.74	0.892	0.069	0.001380 (262/189836)	0.0017489 (19/10864)	0.001657153 (18/10862)	No	Likely Benign
BR34	MS	18-44113256-C-T	c.911G>A, p.(Arg304Gln)	28.1	0.993	0.422	0.0002414 (46/190544)	0.0004623 (5/10816)	0.000369891 (4/10814)	No	Likely Benign
BR41;BR48	MS	18-44171980-G-A	c.1570C>T, p.(Arg524Cys)*	32	0.992	0.661	0.002651 (509/192004)	0.0018416 (20/10860)	0.001658069 (18/10856)	No	VUS
BR43	MS	18-44085877-G-T	c.2469C>A, p.(Asn823Lys)	21.2	0.953	0.05	0.002793 (528/189046)	0.0028366 (30/10576)	0.002742576 (29/10574)	No	VUS
BR11	MS	18-44159694-C-T	c.1708G>A, p.(Asp570Asn)*	22.9	0.994	0.209	0.001198 (228/190254)	0.0004625 (5/10810)	0.000370096 (4/10808)	No	VUS

BR69	MS	18-44221971-C-T	c.274G>A, p.(Val92Ile)*	15.59	0.938	0.175	0.001952 (371/190110)	0.0002847 (3/10536)	0.000189861 (2/10534)	No	Likely Benign
BR11	MS	18-44152069-T-C	c.2027A>G, p.(Asp676Gly)*	14.4	0.926	0.022	0.002338 (445/190330)	0.0006478 (7/10806)	0.00055535 (6/10804)	No	Likely Benign
BR38	MS	18-44159660-A-G	c.1742T>C, p.(Val581Ala)*	17.6	0.952	0.048	0.002804 (533/190088)	0.0006475 (7/10810)	0.000555144 (6/10808)	No	Likely Benign
BR21	MS	18-44114362-G-A	c.815C>T, p.(Thr272Met)	15.54	0.739	0	0.006408 (1220/190388)	0.0022026 (24/10896)	0.002111254 (23/10894)	No	Likely Benign
BR38;BR70	MS	18-44149569-C-A	c.2080G>T, p.(Asp694Tyr)*	25.5	0.935	0.218	0.003622 (688/189944)	0.0007371 (8/10854)	0.000552995 (6/10850)	No	Likely Benign
CZ4	SS	18-44113118-C-T	c.1042+7G>A	10.97	0.692	0	0.00003167 (6/189474)	0.0001868 (2/10706)	9.3423E-05 (1/10704)	No	Likely Benign
BR83	SG	18-44109190-G-A	c.1147C>T, p.(Arg383Ter)	39	0.998	0	0.0006522 (124/190118)	0.0010123 (11/10866)	0.000920471 (10/10864)	No	Potentially pathogenic
TCF4* (ENST00000566286*/ENST00000544241**/ENST00000354452)**											
BR65	MS	18-53255710-C-A	c.57G>T, p.(Arg19Ser)*	18.3	0.9264	0	0 (0/0)	0.0001489 (1/6716)	0 (0/6714)	No	VUS
BR65	NS	18-53255709-T-A	c.58A>T, p.(Lys20Ter)*	16.45	0.8388	0	0 (0/0)	0.0001489 (1/6716)	0 (0/6714)	No	Potentially pathogenic
BR49	MS	18-53070920-A-G	c.26T>C, p.(Ile9Thr)**	17.29	0.8959	0.04	0.000006524 (1/153274)	0.0000993 (1/10070)	0 (0/10068)	No	VUS
BR63	SS	18-53255701-C-T	c.66G>A, c.(Glu22=)*	18.39	0.9549	0	0.0007313 (118/161362)	0.0005963 (4/6708)	0.000447361 (3/6706)	No	Potentially pathogenic

CZ33	MS	18-52928743-G-A	c.944C>T, p.(Ala315Val)***	24.6	0.9987	0.422	0.0006238 (176/282150)	0.0004459 (5/11214)	0.000356761 (4/11212)	No	Likely Benign

MS: missense, SS: splice site variant; NS: nonsense variant, TPM: transcripts per million, VUS: variant of unknow significance, CADD: Combined Annotation Dependent Depletion, FECD: Fuchs endothelial corneal dystrophy, MAF: minor allele frequency, gnomAD: The Genome Aggregation Database, AC/AN: allele count/allele number, +: homozygous.

5.2.1.1 *COL8A2*

A pathogenic *COL8A2* missense variant, c.1363C>A, p.(Gln455Lys), previously associated with the early-onset FECD (MIM# 136800), was identified in two unrelated British individuals (BR1 and BR64; **Table 27**). Both cases presented with an early-onset phenotype, in keeping with previous reports of *COL8A2* mutation associated disease (Biswas, 2001; Gottsch et al., 2005) and had family history of vision loss.

Proband BR64 (IV.3 figure X), is a 51-year-old British Caucasian female diagnosed with FECD at the age of 20 years and underwent corneal transplantation surgery at the age of 30 years. As previously mentioned, a positive family history was reported for this proband including her father (III.3) and paternal first cousin (IV.1), **Figure 42.B**. Her paternal great-grandmother (I.1) was also reported to have lost her vision at a young age, but no further details regarding the reason for this vision loss was available. Segregation analysis, performed using genomic DNA from available relatives revealed the proband's unaffected mother did not carry the variant, **Figure 42.A**.

Unfortunately, genomic DNA samples from affected relatives of the proband were not available for segregation analysis.

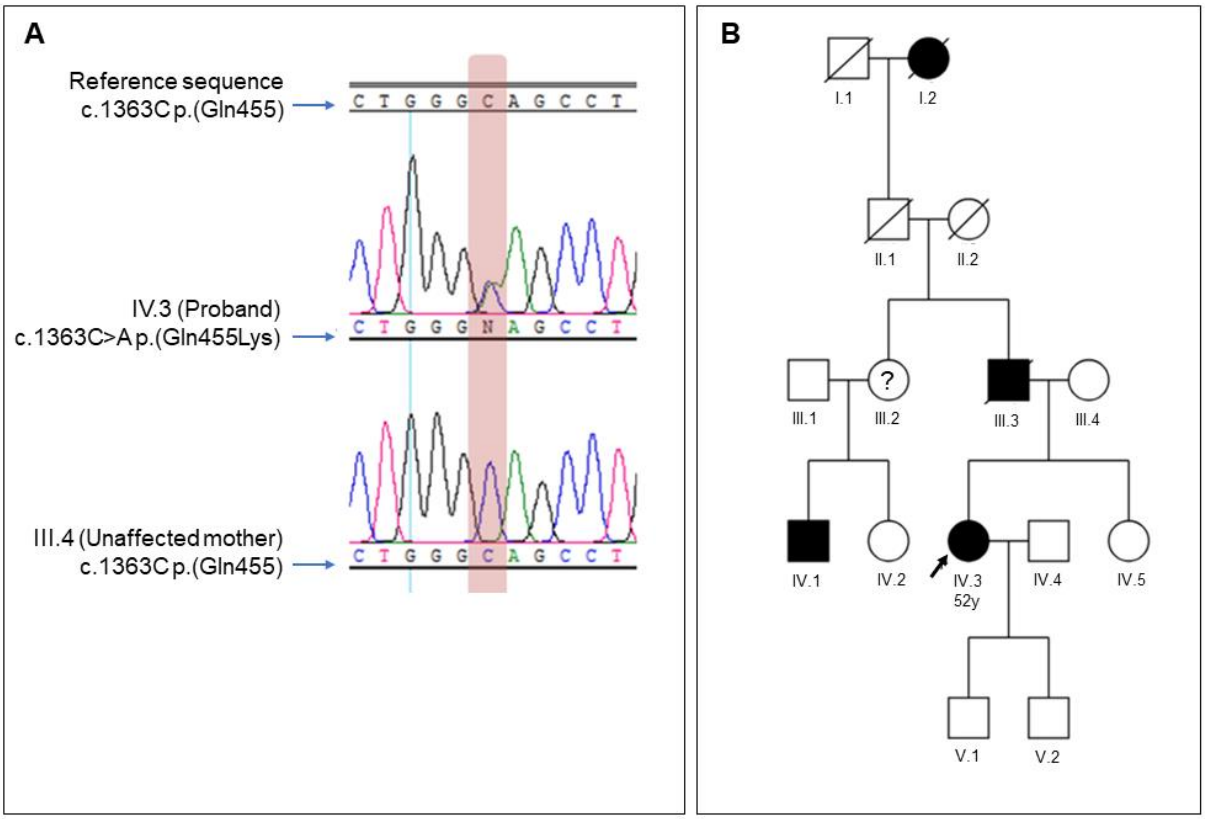


Figure 42 Identification and segregation analysis of COL8A2 c.1363C>A p.(Gln455Lys) variant identified Proband BR64. (**A**) Sanger Sequencing chromatogram confirming the presence of variant c.1363C>A, p.(Gln455Lys) in the proband. The proband's unaffected mother was identified to be wild-type for the variant. (**B**) Family Pedigree for proband BR64 (IV.3).

Proband BR1 (II.1 **Figure 43**), is a 28-year-old Polish Caucasian male diagnosed with an unspecified primary corneal endothelial dystrophy at the age of 25 and underwent corneal transplantation a year later. A positive family history for this patient was also reported as his mother was diagnosed with FECD at the age of 29 (**Figure 43.B**). Sanger sequencing confirmed the presence of the variant in the proband's genomic DNA (**Figure 43.A**). Genomic DNA of family members was unavailable for segregation analysis.

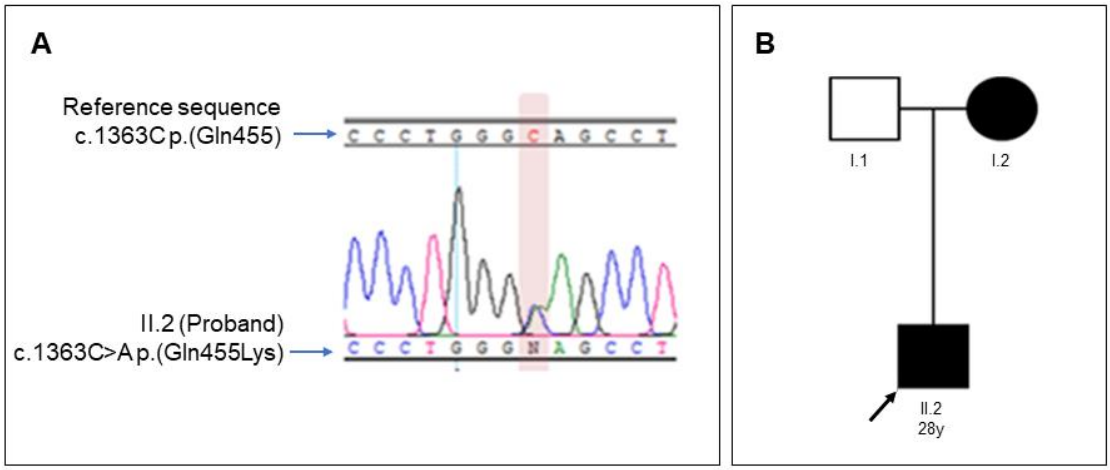


Figure 43 Identification of *COL8A2* c.1363C>A, p.(Gln455Lys) variant in Proband BR1. **(A)** Sanger Sequencing chromatogram confirming *COL8A2* variant c.1363C>A, p.(Gln455Lys) in the Proband. **(B)** Family Pedigree for proband (II.1).

On the basis of these findings subsequent additional targeted screening for the *COL8A2* p.(Gln455Lys) mutation was carried out on all CTG18.1 expansion-negative FECD patients, which had not undergone exome sequencing analysis, and presented as an early-onset (<40-years-old) FECD phenotype (n=12). A further individual, a 21-year-old female was also identified to carry this mutation. The patient had reported a family history of FECD, reporting that her father is also affected. Unfortunately, further clinical information regarding family history and any familial samples were unable to be acquired to conduct segregation analysis.

A further two missense variants were identified in two unrelated individuals; Proband CZ14 was found to have the variant c.1301G>A, p.(Arg434His) and CZ49 the variant c.1724C>T, p.(Pro575Leu). Notably the p.(Arg434His) variant has previously been reported in an individual with late-onset FECD (Gottsch et al., 2005). The affected proband reported here (CZ14) was also found to display a late-onset disease suggesting that this change may

still be disease-associated given the associated in silico predation scores and low frequency in the control population (gnomAD total frequency: 0.001141) but not result in an early-onset phenotype like other previously described *COL8A2* mutations. In addition, CZ14's two daughters had been examined in detail and found to have no symptoms of FECD. Furthermore, the eldest daughter, 47 years of age at examination, was found to be wild type for this variant, meanwhile, the youngest daughter, 43 years of age at examination did carry the p.(Arg434His) change. Despite carrying the variant, she displayed no guttate and had a normal endothelial count.

The p.(Pro575Leu) variant, was identified in a Czech proband (CZ49) who was noted to have a slightly earlier manifestation than expected as guttae were noted at 44 years of age. This variant has not yet previously been reported in FECD in current literature. No family history was available for this proband.

5.2.1.2 *SLC4A11*

A total of 13 rare (MAF ≤1%) *SLC4A11* variants were identified in this cohort. Of these, four were missense changes which had a CADD >10 (c.326G>T, p.(Arg109Leu), c.1121G>A, p.(Arg374Gln), c.1120C>T, (p.Arg374Trp), and c.530A>G, p.(Asn177Ser)) and one further variant was located near an exon boundary (c.173-8C>G) and hence categorised as a splice-site change (**Table 27**). The remaining eight were either missense variants with a CADD score >10 or synonymous variants and thus were predicted to be benign (**Table S7**). To the best of my knowledge, none of the rare *SLC4A11* variants identified by this study have previously been reported FECD-associated. However, only one variant, p.(Arg109Leu) had a notably high in silico disease prediction score, CADD score 31, and found to alter a highly

conserved residue located in functional domains of the encoded solute carrier and have thus been assigned as potentially pathogenic.

5.2.1.3 *ZEB1*

In total, 15 rare (MAF≤1%) *ZEB1* variants were identified including seven missense variants (CADD score >10) (c.698C>T, p.(Thr233Met), c.998T>C, p.(Ile333Thr), c.3099G>C, p.(Glu1033Asp), c.2254A>G, p.(Thr752Ala), c.2563C>A, p.(Gln855Lys), c.2522A>C, p.(Gln841Pro), c.1661A>G, p.(Lys554Arg)), one in-frame deletion (c.105_107del, p.(Asp35del)), one splice site variant (c.794-7T>G) (**Table 27**). Five additional synonymous changes were also identified (**Table S7**). Four of the variants have previously been reported as FECD-associated including; p.(Asn78Thr), p.(Lys554Arg), p.(Gln841Pro) and p.(Gln855Lys) (Minear et al., 2013; Riazuddin et al., 2010). Notably, p.(Asn78Thr) and p.(Lys554Arg) (Riazuddin et al., 2010), in addition to two rare synonymous changes p.(Ser202=) and p.(Ala420=) (**Table S7**) are observed in the same four unrelated cases of African American ancestry and were found to be in close linkage disequilibrium (D' 1.0, R2 >0.9), suggesting that they occur on the same ancestral haplotype. All four of these variants had a MAF of above 5% in the GnomAD African/African American population.

5.2.1.4 *AGBL1*

Thirteen rare variants (MAF ≤1%) were identified in *AGBL1*, including one nonsense change and five missense variants with a CADD score >10 (c.3082C>T, p.(Arg1028Ter), c.490A>T, p.(Ile164Phe), c.671C>T, p.(Thr224Met), c.1994A>G, p.(Tyr665Cys), c.668C>T, p.(Pro223Leu), c.2884A>G, p.(Lys962Glu)) (**Table 27 and S7**). The same nonsense variant identified in this study, p.(Arg1028Ter), has previously been reported once to

segregate with FECD under a multi-locus model (Riazuddin et al., 2013). All other rare missense and synonymous variants identified here have not previously been reported as FECD-associated. However, several are seen in more than one unrelated individual; c.3149A>G, p.(Asn1050Ser), c.484G>A, p.(Val162Met) and c.668C>T, p.(Pro223Leu). Nonetheless, given that *AGBL1* is not expressed within adult corneal endothelial cells (0.00TPM) it is difficult to hypothesise how either a premature termination codon or the missense changes identified could induce a functional effect within the affected corneal endothelial cells. Consequently, all rare *AGBL1* variants identified by this study have been assigned as VUS (**Table 27**).

5.2.1.5 *LOXHD1*

Twenty-five rare variants (MAF ≤1%) including 14 missense variants, one splice-site change, one stop-gain and 9 synonymous variants were identified in *LOXHD1* which is recognised to be a highly polymorphic gene (**Table 27 and S7**). Of these variants only one missense variant, c.541C>T, p.(Leu181Phe), has previously been reported to be associated with FECD (Riazuddin et al., 2012). However, we have assigned all identified changes to be VUS given that *LOXHD1*, like *AGBL1*, is not expressed within healthy adult corneal endothelial cells (TPM 0.1).

5.2.1.6 *TCF4*

To date, expansion of the CTG18.1 repeat is the only *TCF4*-specific mutation known to be causal of FECD. Intriguingly, in this study rare (MAF ≤ 0.01) *TCF4* coding variants were identified within the following two patients from the expansion negative FECD cohort.

Proband BR65 is an African female diagnosed with FECD at the age of 57-years old and had no family history of FECD. Two, *in cis,* variants within the coding region of *TCF4* were identified in this individual. The first variant was a missense heterozygous variant, c.57G>T, p.(Arg19Ser), and the second, a nonsense variant, c.58A>T, p.(Lys20Ter). Both variants were novel and not present in any of the control datasets.

Additionally, an exonic splicing variant, c.66G>A, p.(Glu22=) was identified for individual BR63, a black African male diagnosed at 49-years old with no family history of FECD. Although this variant was present in GnomAD at a very low frequency of 7.31^{x10-4}.

When visualising these variants using IGV all three variants were observed to cluster within a single *TCF4* exon (**Figure 44**). Sanger sequencing confirmed the variants were present in the genomic DNA of each proband.

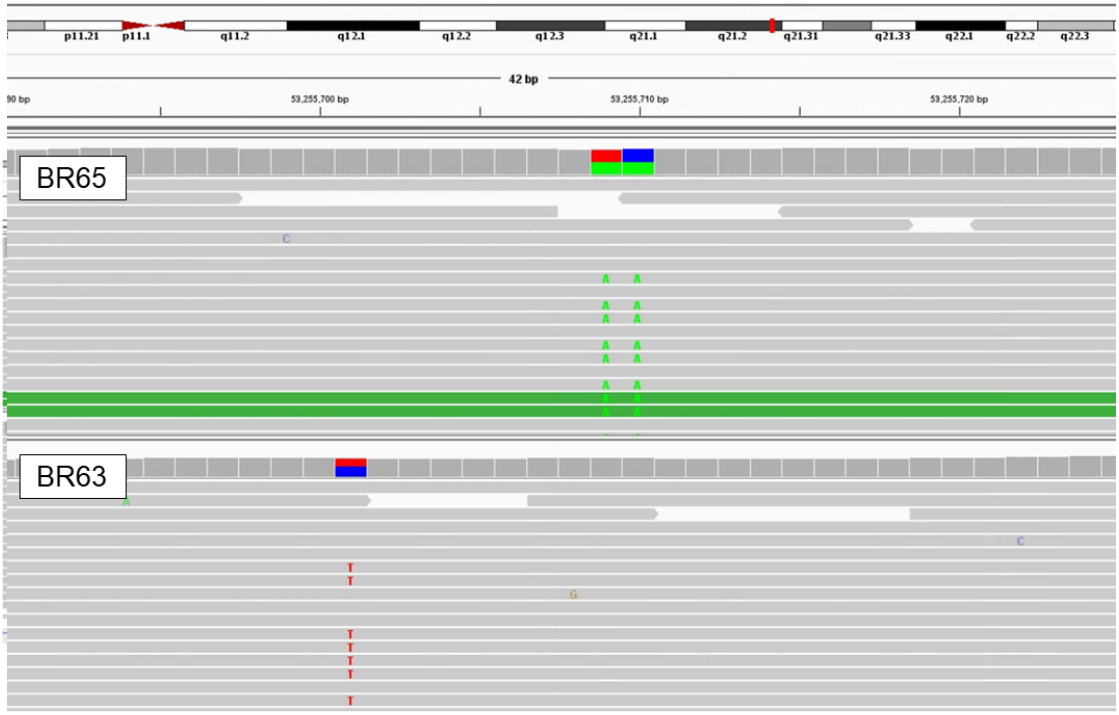


Figure 44 Visualisation of identified TCF4 coding variants (ENST00000566286) c.57G>T, p.(Arg19Ser), c.58A>T, p.(Lys20Ter) (BR65) and c.66G>A, p.(Glu22=) (BR63) by exome sequencing. Reads were visualised in Integrated Genomics Viewer.

The *TCF4* gene displays a vast array of alternatively spliced exons and multiple within 5' untranslated regions (UTR), with more than 90 different transcripts reported in Ensembl to date (Fautsch et al., 2021; Sepp, Kannike, Eesmaa, Urb, & Timmusk, 2011; Zerbino et al., 2018) (**Figure 44.A**). The variants identified here are only predicted to affect six transcripts and are only present as coding variants in one transcript (ENST00000566286.5), in which they can cause potential haploinsufficiency. For all other transcripts, the variants are located within 5'UTR and could therefore also potentially exert regulatory effects on their respective expression within the corneal endothelium (**Figure 45.B and 45.C**).

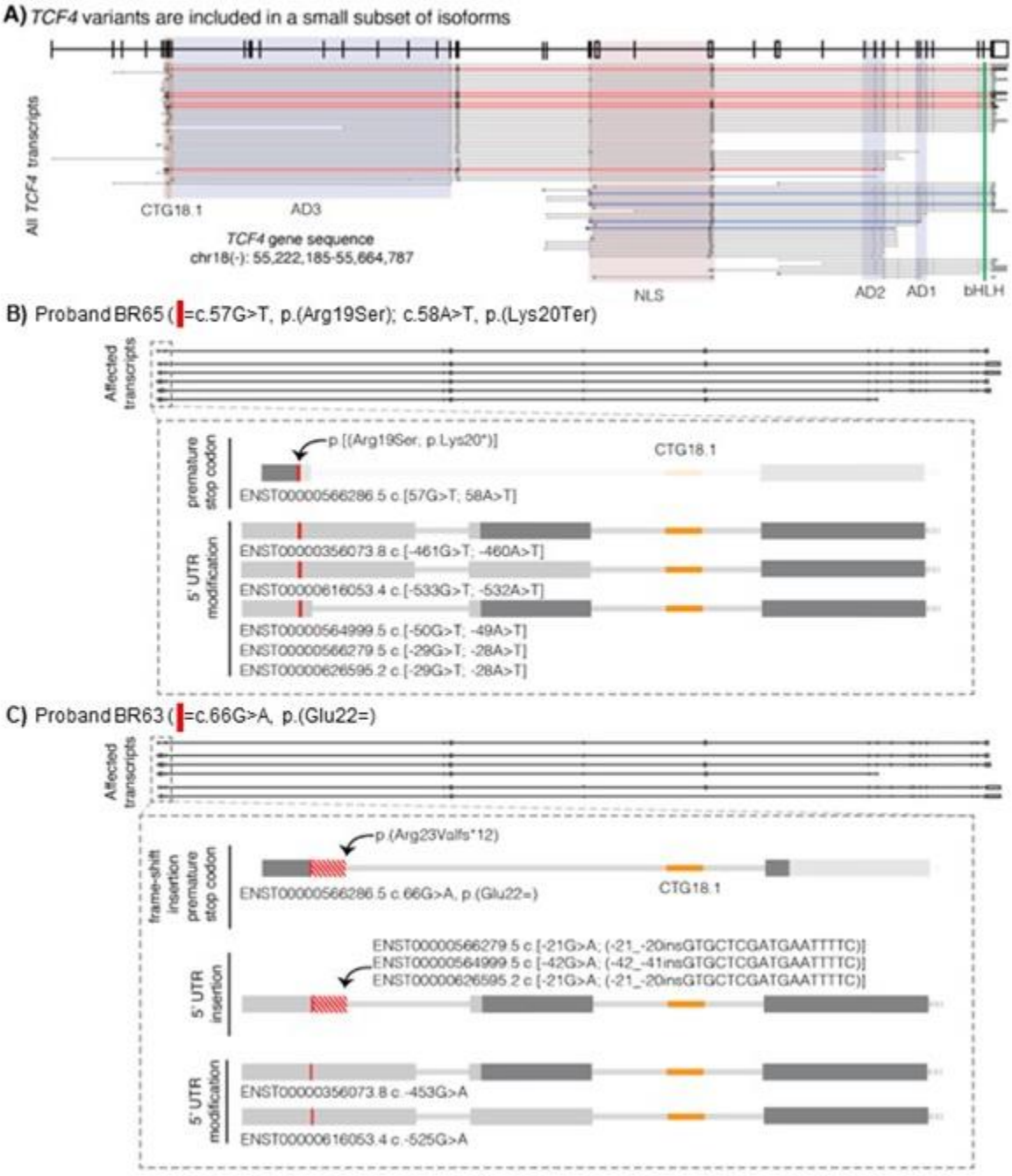


Figure 45 Schematic of *TCF4* and rare variants identified in two CTG 18.1 expansion-negative FECD cases. **(A)** All 93 annotated *TCF4* transcripts depicted, and only 10 include exons containing rare variants identified in *TCF4* expansion-negative FECD cases. The six transcripts that encompass variants present within Proband BR65 and BR63 are highlighted in red (ENST00000616053.4, ENST00000356073.8, ENST00000564999.5, ENST00000566279.5, ENST00000566286.5, ENST00000626595.2). Green bar denotes TCF4 bHLH region, blue regions highlight TCF4 activation domains, red box shows the bipartite TCF4 NLS signal and orange shows the genomic region containing the CTG18.1 repeat. **(B)** Proband BR65 harbours two heterozygous, *in cis*, rare variants, c.57G>T, p.(Arg19Ser) and c.58A>T, p.(Lys20Ter), affecting consecutive nucleotides included within 6 distinct

transcripts. For protein coding transcript ENST00000566286.5, the variants are predicted to introduce a missense variant immediately followed by a premature stop codon after 20 amino acids. For the remaining 5 transcripts, the rare variants are located within 5' untranslated regulatory regions (UTR). **(C)** Proband BR63 harbours a single rare heterozygous variant, c.66G>A, p.(Glu22=), encompassed within the same 6 transcripts as proband A. For protein coding transcript ENST00000566286.5, the variant introduces a synonymous variant altering the last nucleotide of exon 1. This is predicted to weaken the native splice donor site resulting in activation of a cryptic downstream splice donor site and subsequently a frameshift insertion, followed by a premature termination codon (PTC). The variant is located within 5'UTR regions for 5 additional transcripts. For 3/5 of these (ENST00000566279.5, ENST00000564999.5, ENST00000626595.2) the variant is similarly predicted to affect splicing and introduce a 17bp insertion into the 5'UTR region. Figure adapted from (Bhattacharyya et al., 2023).

Notably, the p.(Lys20Ter) is a nonsense substitution therefore produces a functionally null allele. The synonymous change, p.(Glu22=) alters the last nucleotide of exon 1 within the transcript, and thus it is likely to alter splicing. To investigate this, I performed in silico analysis, using the tools SpliceAI and SpliceRover, to predict the impact of this variant (Jaganathan et al., 2019; Zuallaert et al., 2018). Splice AI predicts the variant introduces loss of the splice donor site for exon 1 (SpliceAI Δ score 0.78). SpliceRover also predicts that c.66G>A weakens the native splice donor site for exon1 ENST00000566286.5 (from 0.320 to 0.004) and that this could result in the activation of a cryptic splice donor downstream (from 0.098 to 0.233) of the wildtype donor site, which would introduce a short frameshift insertion, followed by a PTC, c.66_67insGTGCTCGATGAATTTTC, p.(Arg23Valfs*12).

5.2.2 Variants identified in GWAS associated genes

More recently in 2017, common polymorphisms located within an intronic region of *LAMC1*, an intergenic region between *LINC00970/ATP1B1* and an intronic region of *KANK4* have all been significantly associated with FECD by

GWAS (Afshari et al., 2017). Given this, it was intriguing to investigate if any rare, potentially functional and disease-associated variants occurred within these given genes. In total, 30 rare variants (MAF ≤1%), including 13 missense, 16 synonymous and one substitution, were identified within the 141 CTG18.1 expansion-negative exomes analysed. Only synonymous VUS were identified in *ATP1B1* which encodes an ATPase Na+/K+ transporting subunit that is abundantly expressed within healthy CECs (TPM 77). Unfortunately, the long non-coding RNA *LINC00970* was not captured by the exome sequencing approached applied. However, this transcript is not expressed within healthy corneal endothelial cells and hence it seems unlikely that coding variants in the transcript could directly induce disease.

Table 28 Summary of rare, potentially deleterious variants identified in GWAS-hit genes from a total of 141 FECD cases analysed by exome sequencing. In total 141 FECD cases were found to harbour a total of 11 variants based on the filtering criteria applied; MAF < 0.01 in publicly available gnomAD genomes, exomes and Kaviar, CADD score > 10.

Subject ID	Functional change	Genomic co-ordinates (Hg19)	Change	In silico predictions			GnomAD Frequency (Total)	UCLex Frequency (AC/AN)	UCLex Frequency without FECD cases	Reported as FECD-associated	Variant interpretation
				CADD	DANN	Reveal					
KANK4 0.77TPM											
BR12	MS	1-62739014-C-T	c.1762G>A, p.(Ala588Thr)	23.3	0.998	0.077	0.00001204 (3/249220)	0.0001784 (2/11212)	8.92061E-05 (1/11210)	No	VUS
BR57	SS	1-62733957-AC-A	c.2231+1del	34	0	0	0.003907 (1101/281802)	0.0045382 (51/11238)	0.004449982 (50/11236)	No	VUS
***LAMC1* (ENST00000258341) 13.71TPM**											
BR14	MS	1-183084684-G-A	c.1240G>A, p.(Gly414Ser)	28.8	0.999	0.872	0.000003977 (1/251470)	0.0000886 (1/11288)	0 (0/11286)	No	VUS
CZ50; CZ51	MS	1-183106945-A-G	c.4456A>G, p.(Met1486Val)	22.2	0.909	0.326	0.00004937 (12/243048)	0.0001774 (2/11272)	0 (0/11268)	No	VUS
BR65	MS	1-183096522-C-T	c.3106C>T, p.(Arg1036Trp)	28.4	0.999	0.217	0.00002476 (7/282756)	0.0000891 (1/11222)	0 (0/11220)	No	VUS
BR36	MS	1-182992946-G-C	c.95G>C, p.(Cys32Ser)	18.67	0.972	0.088	0.0001603 (35/218332)	0.0007939 (6/7558)	0.000661726 (5/7556)	No	VUS
CZ32	MS	1-183077444-C-G	c.757C>G, p.(Leu253Val)	24.7	0.999	0.755	0.00008487 (24/282794)	0.0000983 (1/10170)	0 (0/10168)	No	VUS
BR31	MS	1-183094585-G-A	c.2701G>A, p.(Val901Met)	23.8	0.999	0.296	0.0002086 (59/282792)	0.0000886 (1/11284)	0 (0/11284)	No	VUS

BR40	MS	1-183105709-G-A	c.4303G>A, p.(Ala1435Thr)	14.34	0.862	0.148	0.0009831 (255/259372)	0.0010004 (11/10996)	0.000909587 (10/10994)	No	VUS
BR72	MS	1-183086559-C-T	c.1669C>T, p.(Arg557Trp)	22.5	0.973	0.145	0.001780 (503/282506)	0.0003540 (4/11298)	0.000265581 (3/11296)	No	VUS
CZ5;BR73	MS	1-183102632-G-A	c.3796G>A, p.(Glu1266Lys)	22.3	0.951	0.047	0.003280 (927/282636)	0.0043096 (48/11138)	0.004131489 (46/11134)	No	VUS

MS: missense, SS: splice site variant, TPM: transcripts per million, VUS: variant of unknown significance, CADD: Combined Annotation Dependent Depletion, FECD: Fuchs endothelial corneal dystrophy, MAF: minor allele frequency, gnomAD: The Genome Aggregation Database, AC/AN: allele count/allele number.

KANK4 encodes a protein of unknown function that is minimally expressed within healthy corneal endothelium (TPM 0.77). Four rare *KANK4* variants, c.1762G>A, p.(Ala588Thr), c.1550G>A, p.(Arg517Lys), c.797A>G, p.(Asp266Gly) and c.1229C>T, p.(Thr410Met), were identified in four unrelated FECD cases, including three missense variants and one substitution that is predicted to abolish a splice donor site and could hence potentially represent a loss-of-function allele **(Table 28**). Notably however, *KANK4* has a pLI constraint metric of 0 (gnomAD v2.1.1) indicating that it is highly tolerant to haploinsufficiency and hence the c.2231+1del variants are likely functionally benign.

LAMC1 encodes laminin gamma 1 an extracellular matrix glycoprotein abundantly expressed by healthy corneal endothelial cells (TPM 13.71TPM). In total, nine rare LAMC1 missense variants were identified including six unrelated probands with CADD scores ≥ 20; c.1240G>A, p.(Gly414Ser), c.2191G>C, p.(Glu731Gln), c.3106C>T, p.(Arg1036Trp), c.757C>G, p.(Leu253Val), c.2701G>A, p.(Val901Met) and c.1669C>T, p.(Arg557Trp) (**Table 28**). Wieben *et al.* have previously reported the occurrence of a single rare heterozygous missense variant (c.1468C>T p.(Arg490Trp)) in a CTG18.1 expansion negative FECD patient (Wieben et al., 2018). Due to lack of available segregation data in the relevant families we were only able to assign the identified missense variation with CADD scores ≥ 25 as VUS. However, given that laminin gamma 1 is known to play an important functional role within, the basement membrane secreted by corneal endothelial cells (termed Descemet's membrane), in addition to the fact that a common intronic variant (rs3768617 6.9×10^{-16}) has been significantly associated with FECD, we hypothesise that these changes may be disease-associated. Future functional approaches to determine how the

amino acid substitutions may alter the functional role of the protein with the cornea, in addition to segregation analysis of these variants within the families of the affected individuals should help to further elucidate if the identified changes reported here are pathogenic.

5.2.3 Candidate gene identification and segregation familial FECD samples

Proband BR24 is a 35-year-old Italian male presenting an early-onset phenotype. Initially a novel *COL8A1,* c.619A>T, p.(Ile207Phe) with a CADD score of 8.716 (Ensembl transcript ID: ENST00000261037.7) variant was noted in this proband due to the similarities of the early-onset phenotype caused by *COL8A2* mutations and the shared functional roles of extracellular matrix encoding genes expressed by corneal endothelial cells. This was considered to be a strong candidate gene and first degree family members were recruited for segregation analysis. However, the variant was found not to segregate in affected relatives (**Figure 46**) and thus was eliminated as a potential disease-causing variant in this family.

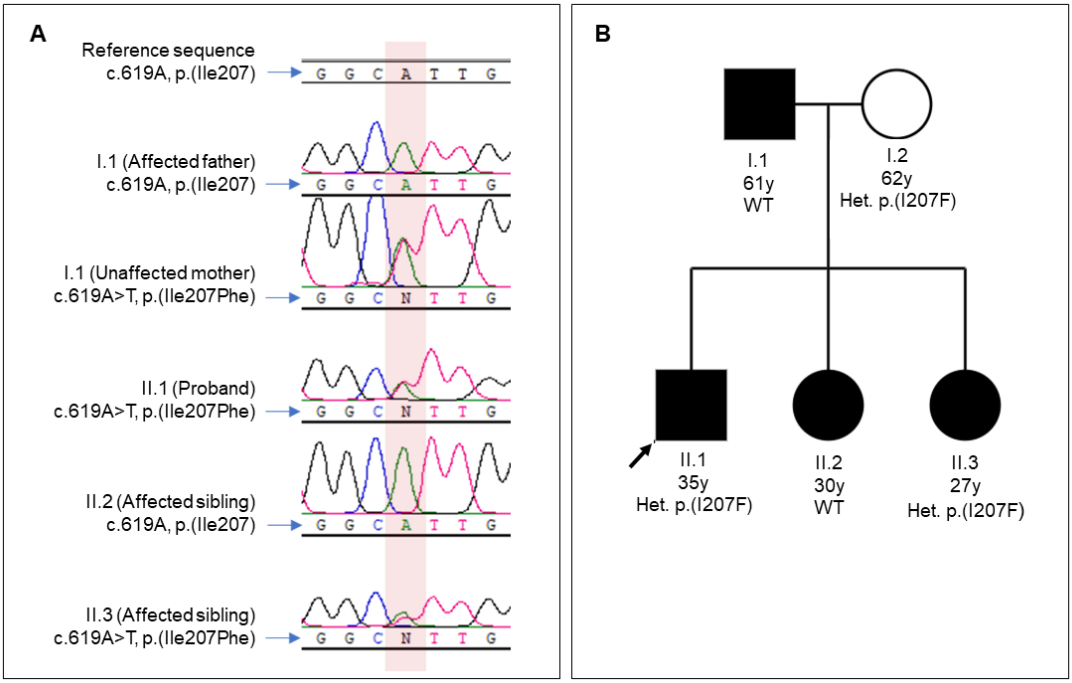


Figure 46 Pedigree of family, presenting an early-onset FECD phenotype identified to harbour a rare *COL8A1* variant. The identified variant, c.619A>T, p.(Ile207Phe) in proband (II.1) was demonstrated not to segregate with disease in this family.

Following this, as genomic DNA of affected siblings, II.2 and II.3 **Figure 46**, was available, exome sequencing was additionally performed, and shared variants were interrogated in all genes to search for novel genetic causes of CTG18.1 expansion-negative FECD. Variants were considered if they have a MAF frequency ≤0.01, significant functional change and a CADD score suggestive of being pathogenic (>10). Genes in which candidate variants fell within were then considered based on expression levels within the corneal endothelium and function, if known.

A unique missense variant, c.658C>T p.(Arg220Cys) within the gene *LYPD3* (Ensembl transcript ID: ENST00000244333) with a CADD score of 34 was identified to be shared between all three affected family members. Further segregation analysis demonstrated that the variant was inherited from the

affected father. *LYPD3* was identified to be relatively highly expressed in the endothelium of the cornea with a TPM value of 83.89.

5.2.4 Gene burden analysis

5.2.4.1 Gene burden analysis approach

As the majority of the non-expanded samples were not identified by exome sequencing to have FECD causative variants in currently identified disease-associated genes, a gene burden approach was applied to the cohort. This aimed to identify novel candidate genes by comparing the number of individuals carrying rare, deleterious variants in genes between case and control subject groups. Utilising exome data generated from the non-expanded FECD samples (i.e. cases) and comparing these to the internal UCLex dataset (Pontikos et al., 2017) (i.e. controls) we were able to apply statistical methods to identify if an enrichment of variants were present in any genes sequenced by the exome approach.

The reason samples from UCLex database were used as controls in this burden test, and not other large exome-sequencing databases, is because both UCLex and the FECD non-expanded exomes had been processed using the same bioinformatic pipeline, eliminating potential technical alignment artefacts or annotation artefacts. Furthermore, publicly available databases typically only release variant-level data, by using UCLex samples allows access to individual level genotype data (i.e. phased data) and thus allows us to also apply more sophisticated statistical approaches such as SKAT (M. H. Guo et al., 2018).

One limitation to this approach is all samples should ideally be of the same ancestry to be able to directly compare the frequency of rare, and potentially deleterious, variants in case subjects compared to control subjects.

In order to work around this limitation, all case and control samples ancestry was predicted using SNP data acquired derived from exome data carried out by Dr Cian Murphy (**Section 2.8.3.1**). From the exome-derived SNP data, samples were plotted on a PCA plot and their predicted ethnicity was calculated based on their proximity to samples with predefined ancestry, **Figure 47**.

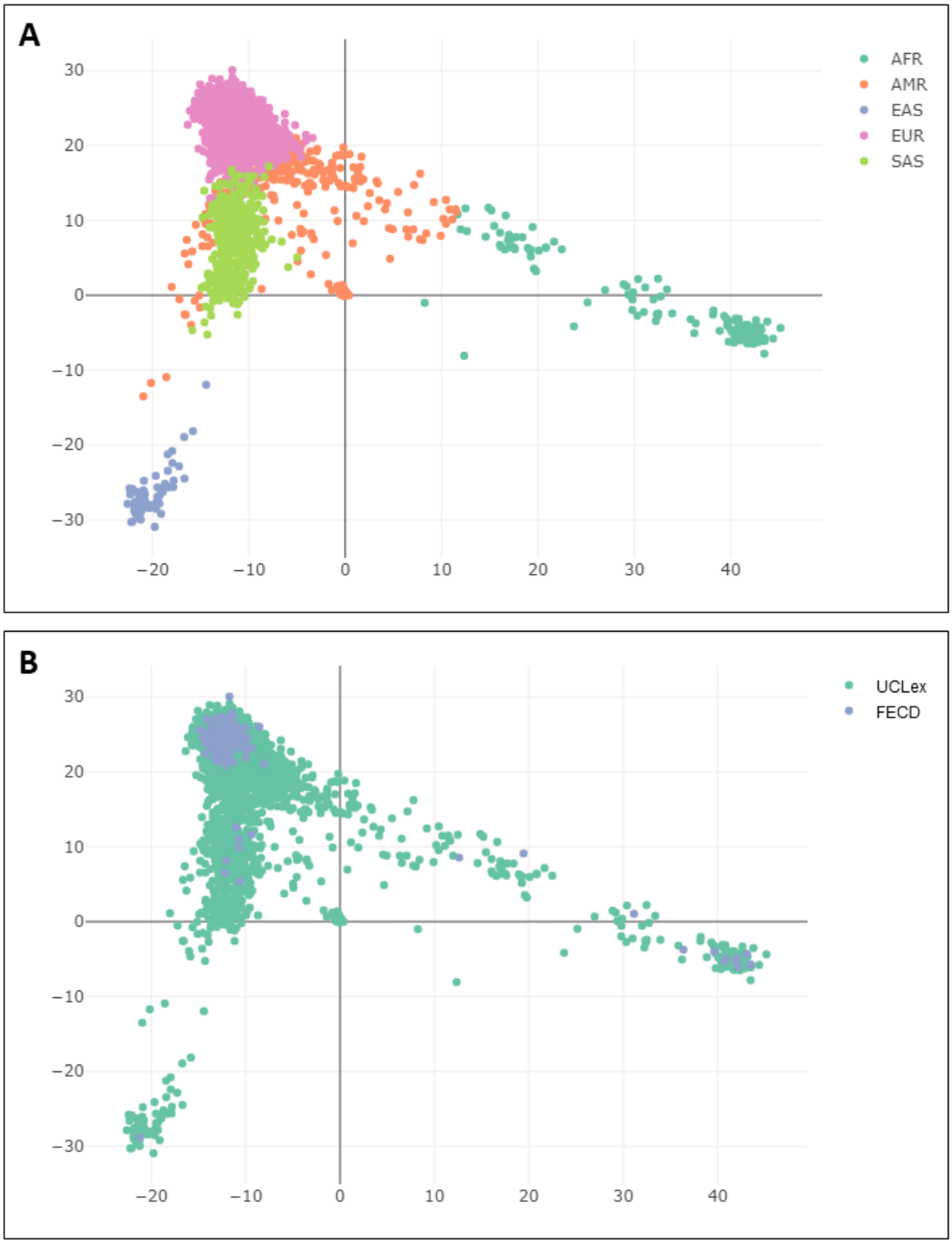


Figure 47 Principal Component Analysis (PCA) was performed using SNP data acquired from exome sequencing data to predict sample ethnicity (A): PCA plot showing ancestry populations from exome sequencing data for the UCLex control database and FECD non-expanded cohort, built using 2 first principal components. Pink: European (EUR), Orange: Ad Mixed American (AMR), green: African (AFR), Asian (SAS), Light green, South Asian (SAS),

blue: East Asian (EAS). **(B)** PCA plot showing FECD samples in blue and UCLex samples in green.

After assigning predicted ancestry to all UCLex and FECD samples, non-European samples were filtered, leaving at total of 108 FECD (case) samples and 1,138 UCLex (control) samples to be included in the gene-burden analysis approach.

Two approaches were used for the gene burden analysis: (1) a SKAT (M. C. Wu et al., 2011), which is a supervised machine learning method that can be used to test for association between rare variants in a region, and (2) a custom-made association test encompassing a Fisher-test. For both approaches, the gene burden test was run four times with four different filtering thresholds applied to the variants. The following conditions are described in **Table 29**.

Table 29 Filtering conditions applied to sequence kernel association test (SKAT) and custom gene burden analysis.

Condition	CADD	Max (gnomAD exomes MAF, Kaviar MAF)	UCLex AC
1	>20	< 0.01	< 40
2	>20	< 0.001	< 40
3	>20	< 0.0001	< 40
4	>10	< 0.001	< 40

The highest MAF for either gnomAD exomes or Kaviar allele frequencies was used as the filtering threshold for variants. Variants with unknown gnomAD exome MAFs and unknown Kaviar MAF information were included in the analysis if they passed the remaining filters.

The filtered variants were grouped by genes and the number of individuals who harboured at least one variant after the filtering (following a dominant inheritance pattern model) were added together in the cases and control groups separately. Only those genes where this sum was greater in the cases group compared to the control group proportionately, were kept.

A Fisher test was used to determine whether there was a significant difference between the number of “rare pathogenic” filtered variants present between the case (FECD) and control (UCLex) groups. The difference was regarded as significant if the Fisher’s test p-value was less than 0.05.

Significant candidate genes were plotted for each condition in **Figures 48-51**. Genes were present in different colours depending on the following classifications. Genes significant in both with a TPM < 20; significant in both with a TPM >= 20; significant in both with an unknown TPM; significant in only the SKAT approach; and significant in only the custom approach. The top 50 candidate genes which appear significant in both the SKAT and custom approach, for each condition, are listed in **Tables S8-S11.**

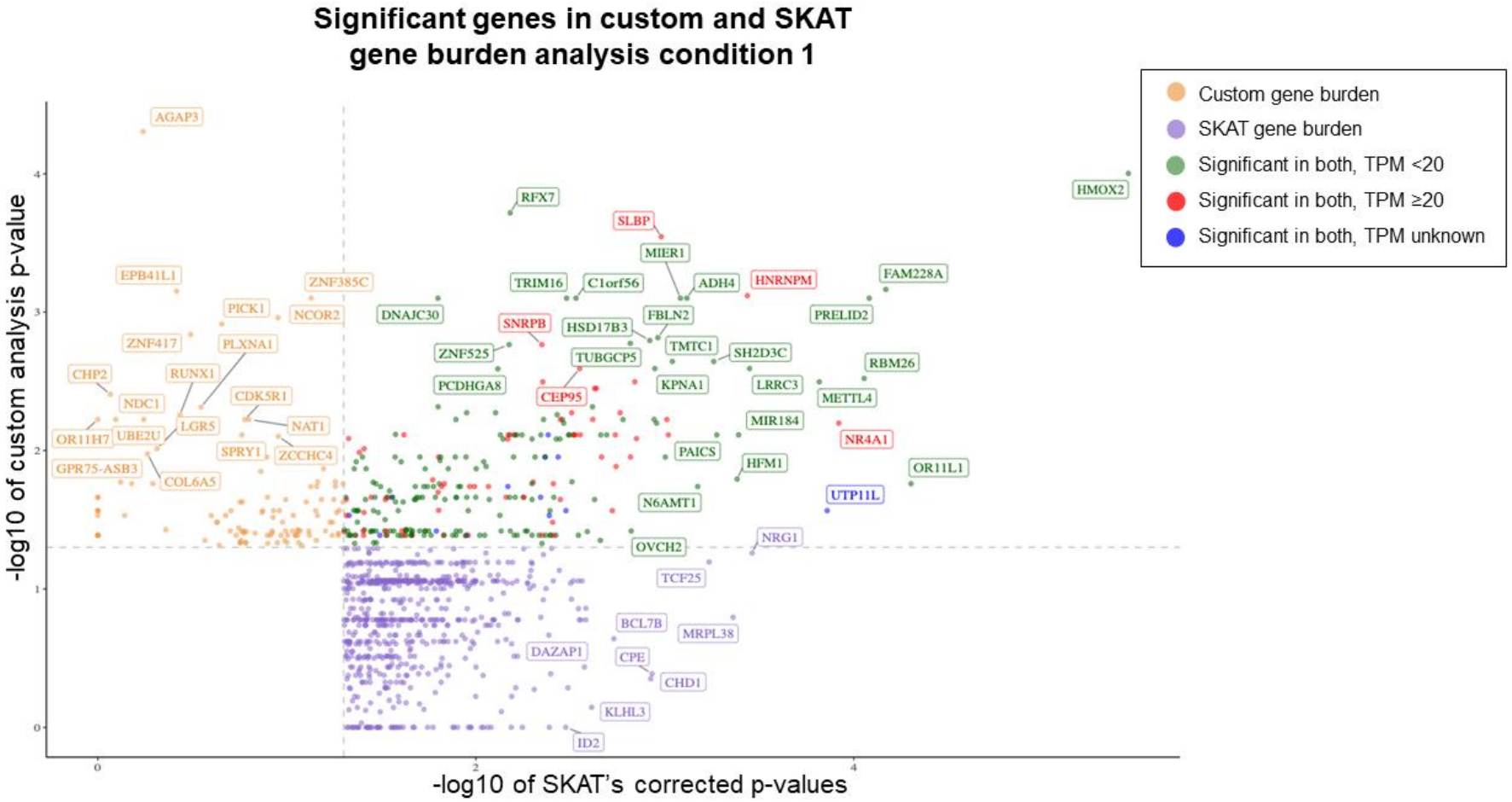


Figure 48 A summary of genes significantly enriched for rare and potentially deleterious variants within a cohort of European FECD cases compared to European controls derived from UCLex exome consortium dataset. Two complementary exome-wide gene burden approaches were applied, including a custom approach (y-axis) and a SKAT gene burden analysis (x-axis). For condition 1, CADD score >20, MAF < 0.01 (gnomAD exomes MAF, Kaviar MAF) were applied. Candidate genes identified to be significantly enriched for rare variants within the FECD case group by both approaches are highlighted in the top right quadrant of the plot.

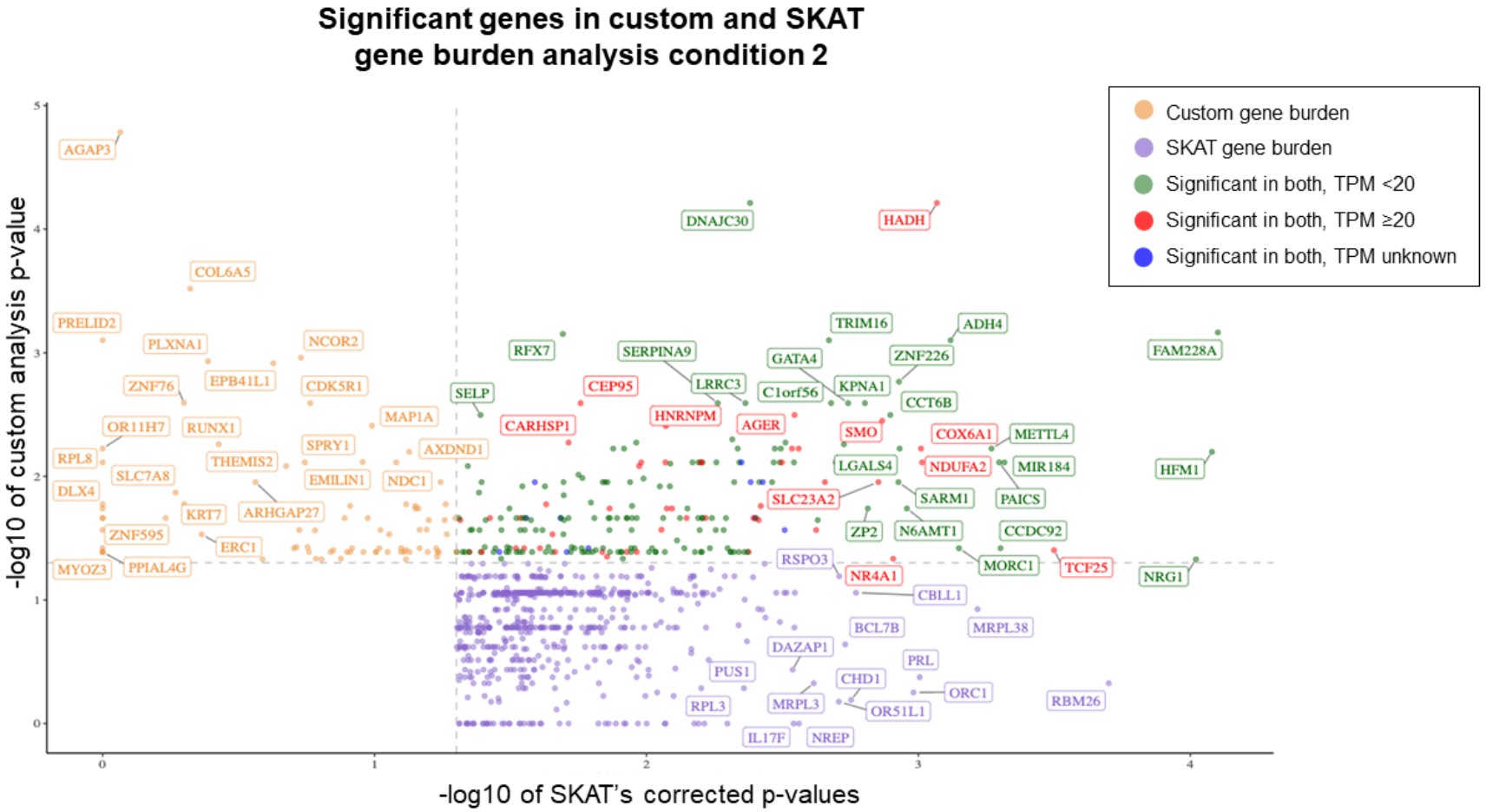


Figure 49 A summary of genes significantly enriched for rare and potentially deleterious variants within a cohort of European FECD cases compared to European controls derived from UCLex exome consortium dataset. Two complementary exome-wide gene burden approaches were applied, including a custom approach (y-axis) and a SKAT gene burden analysis (x-axis). For condition 2, CADD score >20, MAF < 0.001 (gnomAD exomes MAF, Kaviar MAF) were applied. Candidate genes identified to be significantly enriched for rare variants within the FECD case group by both approaches are highlighted in the top right quadrant of the plot.

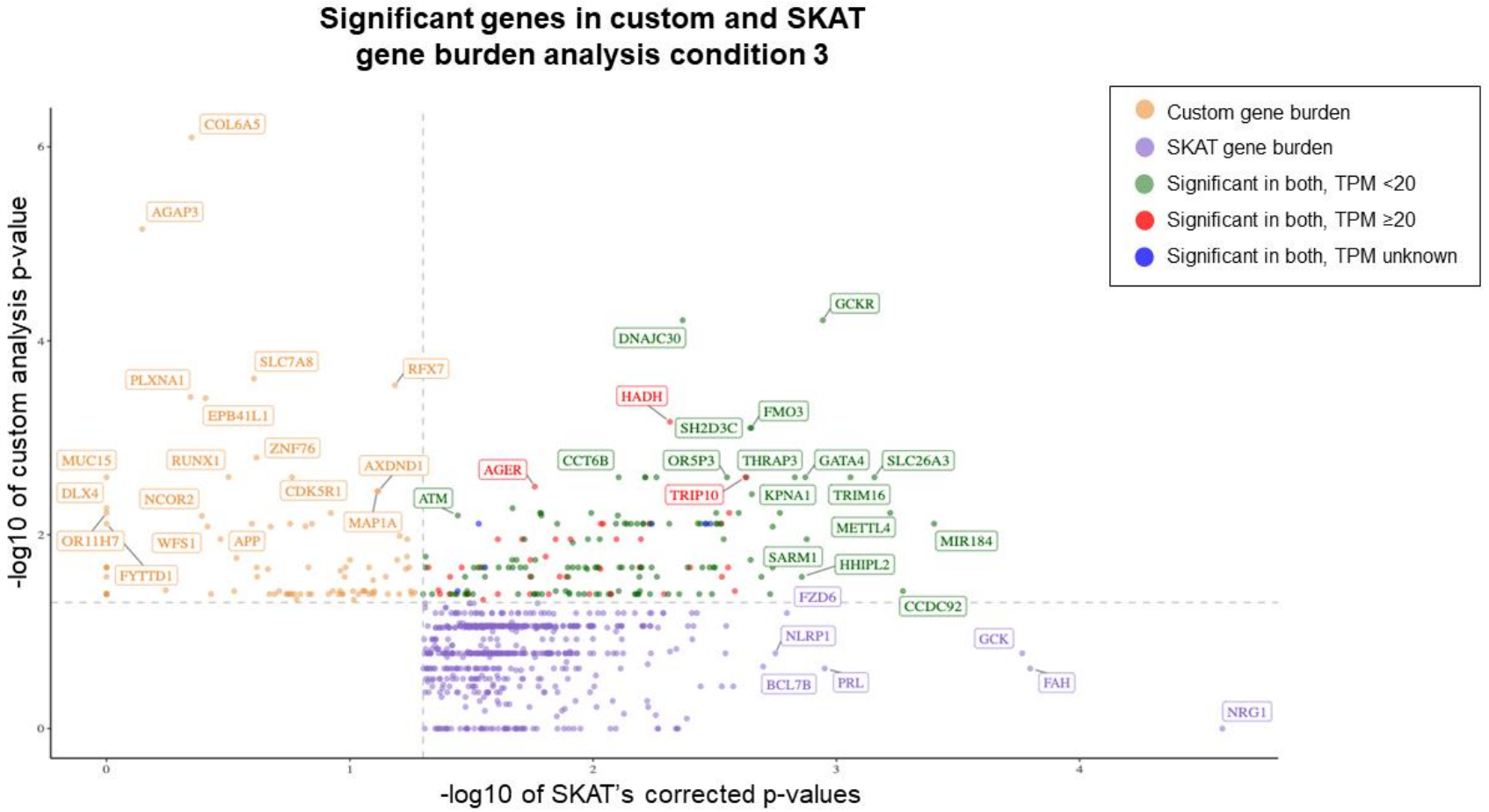


Figure 50 A summary of genes significantly enriched for rare and potentially deleterious variants within a cohort of European FECD cases compared to European controls derived from UCLex exome consortium dataset. Two complementary exome-wide gene burden approaches were applied, including a custom approach (y-axis) and a SKAT gene burden analysis (x-axis). For condition 3, CADD score >20, MAF < 0.0001 (gnomAD exomes MAF, Kaviar MAF) were applied. Candidate genes identified to be significantly enriched for rare variants within the FECD case group by both approaches are highlighted in the top right quadrant of the plot.

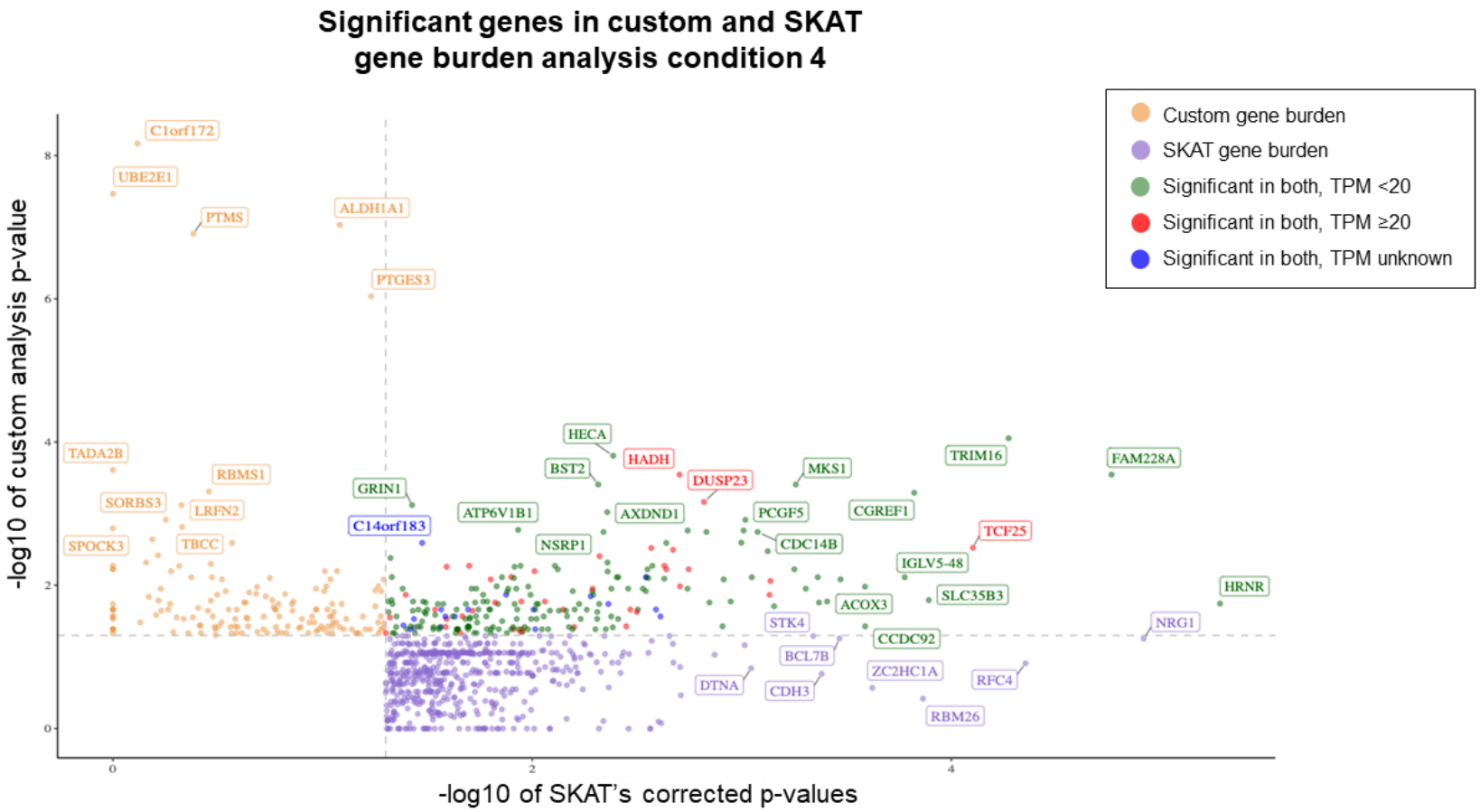


Figure 51 A summary of genes significantly enriched for rare and potentially deleterious variants within a cohort of European FECD cases compared to European controls derived from UCLex exome consortium dataset. Two complementary exome-wide gene burden approaches were applied, including a custom approach (y-axis) and a SKAT gene burden analysis (x-axis). For condition 4, CADD score >10, MAF < 0.001 (gnomAD exomes MAF, Kaviar MAF) were applied. Candidate genes identified to be significantly enriched for rare variants within the FECD case group by both approaches are highlighted in the top right quadrant of the plot.

The gene burden test produced many candidate genes, including *HNRNPM*, *B3GNT7* and *WFS1*, that all warrant further interrogation. However, given the time available I was only able to extensively follow up on one candidate as proof-of-concept that the gene burden approach applied could effectively identify candidate variants in genes that may explain disease. As such, I selected one candidate as an exemplar of this approach, miR-184 based on biological and functional relevance detailed below.

Rare variants in miR-184 were identified to be significantly enriched in the FECD cases cohort compared to the UCLex exome consortium dataset by both the SKAT and custom gene burden approaches under conditions one, two and three respectively. miR-184*,* is a microRNA (miRNA) previously reported to cause EDICT syndrome (OMIM #614303), an autosomal dominant anterior segment dysgenesis characterised by endothelial dystrophy, iris hypoplasia, congenital cataract, and stromal thinning (Iliff et al., 2012; Jun et al., 2002). Given the corneal endothelial cell-specific phenotype that is present in EDICT syndrome it seems biologically plausible that additional variants in the miRNA may also result in other endothelial-cell specific phenotypes such as FECD.

5.2.4.2 MiR-184 variants

A single nucleotide variant in miR-184, +58G>A, MAF in gnomAD exomes 0.000004001, CADD score19.6, (Ensembl transcript ID: ENST00000384962.1) was identified to underlie the statistically association identified by the gene burden approaches ($p= 7.7x10^{-3}$; **Table S8-S11, Figures 48-51**). In total 2 cases and 0 controls were identified to carry this variant by the gene burden test comprising a total of 108 FECD (case) samples and 1,138 UCLex (control) samples. A further FECD sample, derived from a proband of

south Asian ethnicity and hence excluded from the gene burden approach was also retrospectively found to carry this variant. Notably, the variant identified by the gene burden analysis alters the nucleotide located immediately after the variant described to cause EDICT syndrome, +57C>T (Iliff et al., 2012). miRNA are short RNA molecules that play an important role in gene regulation by targeting mRNAs via motif present within 3'UTRs regulating their respective expression (O'Brien, Hayder, Zayed, & Peng, 2018).

Both the variant found here in the gene burden and the causal variant for EDICT syndrome are located in the seed region of miR-184. The seed region is a highly conserved region of the miRNA located from the second to seventh base of the mature miRNA. This region is particularly important for recognition of target mRNA and proper miRNA regulation of protein expression (Lewis, Burge, & Bartel, 2005).

As the variant found in the gene burden analysis was particularly enticing, it was investigated to see if other non-expanded FECD samples that were not included in the gene burden analysis had variants in miR-184. A further miR-184 variant was identified, +73G>T, in a sample presenting with early-onset FECD with polar cataracts. Although this variant does not lie within the seed region, it was located within the mature miRNA sequence and given the phenotypic similarities to EDICT syndrome it made sense to also explore the potential functional impact of this variant alongside those located in the miR-184 seed region.

Clinical data for the three probands harbouring the miR-184, +58G>A variant is present in **Table 30**. Clinical data for these patients were assessed to ensure these patients do not show features of EDICT syndrome and did in fact

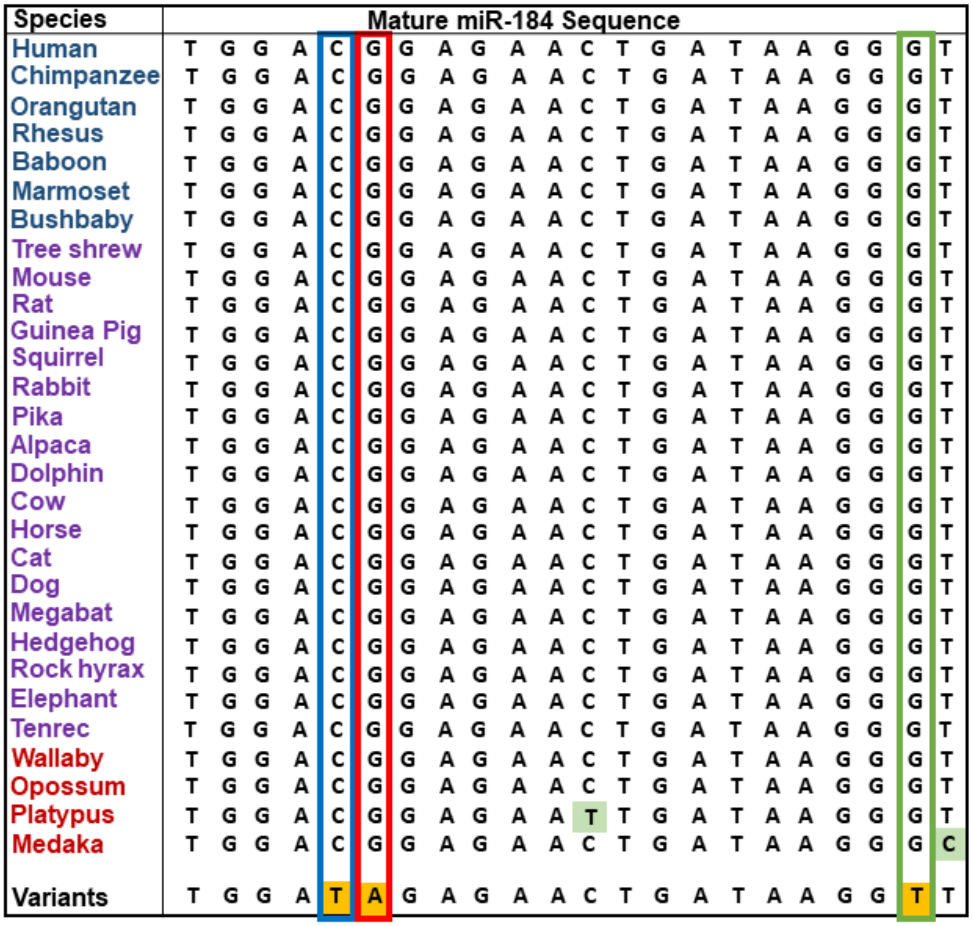
present with an FECD phenotype. None of the three patients displayed features of congenital cataract or iris hypoplasia, key characteristics of EDICT syndrome (Jun et al., 2002). Furthermore, there was no stromal thinning observed in any of the patients and instead showed stromal thickening, a common symptom observed in FECD.

Table 30 Summary of clinical data of three probands with Fuchs endothelial corneal dystrophy (FECD) harbouring miR-184, +58G>A variant.

Age* Sex Ethnicity	Age at DMEK (years)	CCT (µm) Before DMEK	Pre-op BCVA	CCT (µm) After DMEK**	Final BCVA	Congenital cataract or Iris hypoplasia	Ocular Disease	Associated disease	Family history
61 Male White British	60 R 61 L	800 650	6/60 6/18	527 523	6/6 6/6	Nil	divergent squint RE & LE pseudophakia Ocular hypertension	Dementia, learning disability, mild hearing loss, hypertension,	NA
69 Male South Asian	69 R 69 L	NA NA	6/24 6/18	477 470	6/6 6/6	Nil	Nil	Tinnitus	NA
69 Female White British	72 R 71 L	649 637	6/12 6/12	572 550	6/9 6/9	Nil	Nil	Hypertension, ischaemic heart disease	Yes (Not confirmed)

*Age at diagnose, ** at final follow up, RE: right eye, LE: left eye, DMEK: Descemet membrane endothelial keratoplasty, CCT: central corneal thickness, BCVA: best corrected visual acuity, NA: not available.

The mature miR-184 sequence is fully conserved across 28 orthologous sequences that were examined, with the exception of a +64C>T substitution in platypus and a +74T>C substitution in the medaka fish (Iliff et al., 2012) (**Figure 52**). This figure shows the evolutionary conservation across these species, with the EDICT, +57C>T variant highlighted by the blue box, and the variants identified in this study. +58G>A and +73G>T highlighted by the red and green boxes, respectively. All three of these changes were located at positions within the mature miR-sequence that were highly conserved across a diverse range of species.



Species	Mature miR-184 Sequence																				
Human	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Chimpanzee	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Orangutan	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Rhesus	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Baboon	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Marmoset	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Bushbaby	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Tree shrew	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Mouse	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Rat	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Guinea Pig	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Squirrel	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Rabbit	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Pika	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Alpaca	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Dolphin	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Cow	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Horse	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Cat	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Dog	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Megabat	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Hedgehog	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Rock hyrax	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Elephant	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Tenrec	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Wallaby	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Opossum	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	T
Platypus	T	G	G	A	C	G	G	A	G	A	A	T	T	G	A	T	A	A	G	G	G	T
Medaka	T	G	G	A	C	G	G	A	G	A	A	C	T	G	A	T	A	A	G	G	G	C
Variants	T	G	G	A	T	A	G	A	G	A	A	C	T	G	A	T	A	A	G	G	T	T

Figure 52 Evolutionary conservation of miR-184 sequence in 28 nonhuman vertebrates. Blue text: primates; purple text: placental mammals; red text: nonplacental vertebrates. miR-184 variants were highlighted with an orange background. blue box: conservation of EDICT base, +57C>T. Red box conservation of +58G>A. Green box conservation of, +73G>T. Non-conserved bases highlighted with green background: a +64C>T substitution in platypus and a +74T>C substitution in the medaka fish, Figure adapted from (Iliff et al., 2012).

Initially, to evaluate whether these variants could have a potential influence on pathogenicity, in silico tools were applied to predict if the single nucleotide variants discovered in the FECD cohort and previously associated with EDICT syndrome could alter the secondary structure of the miRNA. Secondary miRNA structure predictions were generated using RNAfold web

service by ViennaRNA, Institute of Theoretical Chemistry

(http://rna.tbi.univie.ac.at/), **Figure53.**

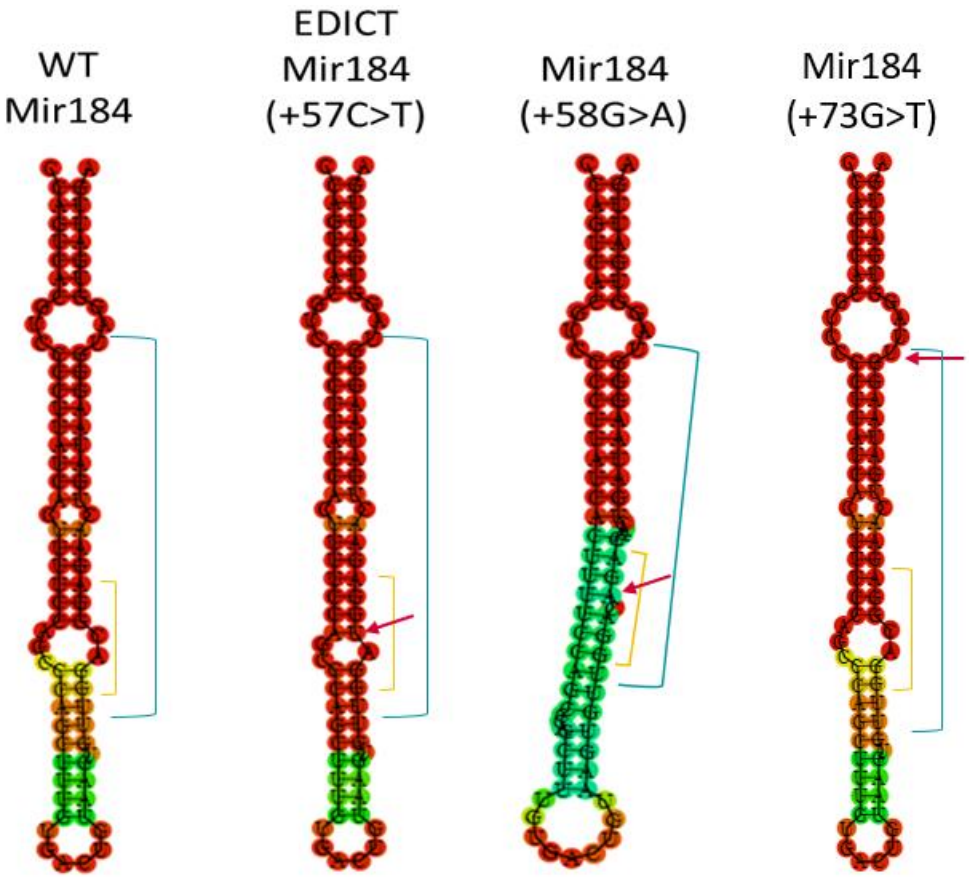


Figure 53 Vienna RNAfold algorithm predicted secondary structure comparing wild-type miR-184, to EDICT associated miR-184(+57C>T), and FECD-associated variants miR-184(+58G>A) and miR-184(+73G>T). Blue bracket: mature miR-184 sequence, orange bracket: seed region, red arrow: substitute base.

5.2.4.3 Investigating effect of miR-184 variants on gene expression

Given the identified miR-184 regions are located within the seed region and mature sequence of miR-184, I was interested to determine if they affect the mRNA targeting capabilities and subsequent ability of miR-184 to regulate gene expression. To achieve this goal, viable target mRNAs first needed to be identified.

To identify target mRNAs for miR-184 three bioinformatic tools were used, DIANA Web Server v5.0 (http://diana.imis.athena-innovation.gr/DianaTools/index.php), miRDB (http://www.mirdb.org/) and Target Scan Human (https://www.targetscan.org/vert_80/). Seed regions of miRNAs function by targeting complementary sequences in mRNA transcripts, usually located within 3'UTR (Peterson et al., 2014). As miRNAs have the potential to target multiple gene transcripts, output genes which were present in all three databases were interrogated based on expression levels in the corneal endothelium and existing published evidence on their respective biological function and overall, four miR-184 gene targets were selected for functional validation to determine if the miR-184 variants identified modulate the miRNA's ability to regulate their respective expression via predicted 3'UTR motifs.

I selected two genes as targets based on the outcomes of the miR-184 target predictions to explore the effect of the miR-184 variants on expression levels, these genes were *SF1* and *EPB41L5*. Additionally, two further genes were selected, based on available published literature, *AKT2* and *INNPL1*. Published evidence suggests that miR-184 directly inhibits *ATK2* and overexpression of miR-184 suppresses cell viability and proliferation (Iliff et al., 2012). Furthermore, the Akt pathway is involved in epithelial–mesenchymal transition, a biological process previously associated with FECD (Iliff et al., 2012). *INPPL1* was also selected to be assessed based on previously published literature. *INNPL1* has been associated with apoptosis and cell death in corneal epithelial cells when miR-184 competes with miR-205, another miRNA. When miR-184 interferes with the ability of miR-205 to suppress INNPL1 levels it results in damping of the Akt signalling pathway via INNPL1 induction leading to increased apoptosis of cells (Yu et al., 2008). Although mir-

205 does not appear to be expressed in the corneal endothelium (TPM 0), it seemed biologically relevant to investigate due to the similarities surrounding cell death, a major phenotype of FECD. Furthermore, it may be that miR-205 is expressed but not captured in the dataset I am exploring as not all miRNAs are captured by RNA-Seq analysis due to their size.

3’UTR sites from the target mRNA genes were initially cloned into a pGEM®-T Easy NheI-HF and SalI-HF restriction enzyme sites incorporated into the amplification primers and used to cleave the inserts from the pGEM®-T Easy Vector before subcloning into the pmirGLO Dual-Luciferase miRNA Target Expression Vector. The pmirGLO Vector is designed to quantitatively evaluate miRNA activity by the insertion of miRNA target sites downstream or 3' of the firefly luciferase gene (*luc2*).

The pmirGLO DNA constructs were co-transfected with miRNA mimics (synthetic miRNAs), designed to imitate WT miR-184 and miR-184 containing the variants described in this study, into HEK293 cells using TransIT®-LT1 Transfection Reagent. To test transfection efficiency HEK293 cells were first transfected with Green Fluorescent Protein (GFP) for 48 hours. When GFP has been transmitted into the cells, the protein emits bright green fluorescence light when excited by UV light. In **Figure 54**, I demonstrate that using TransIT®-LT1 Transfection Reagent I was efficiently able to transfect HEK293 cells.

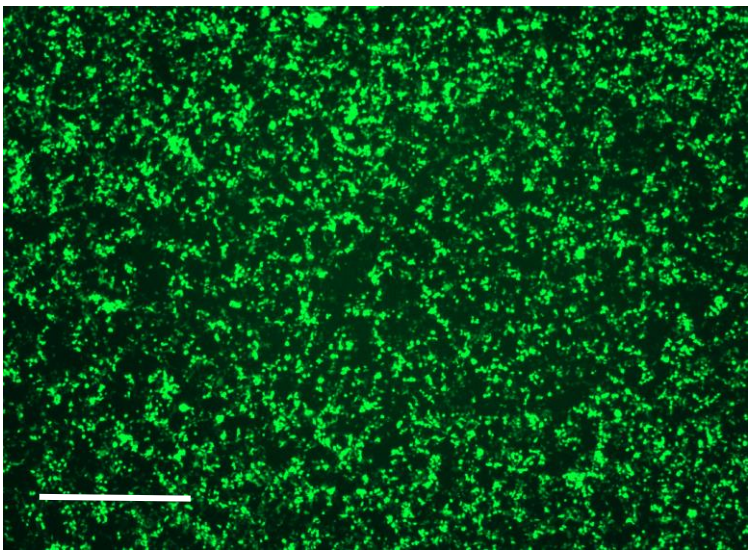
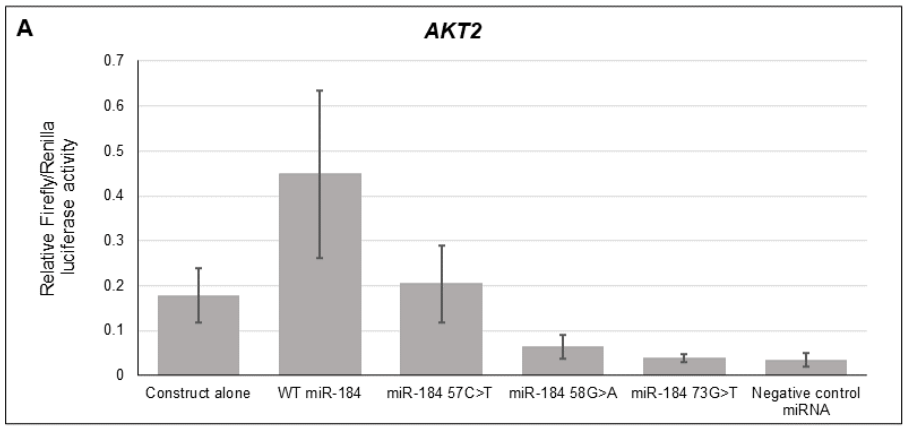
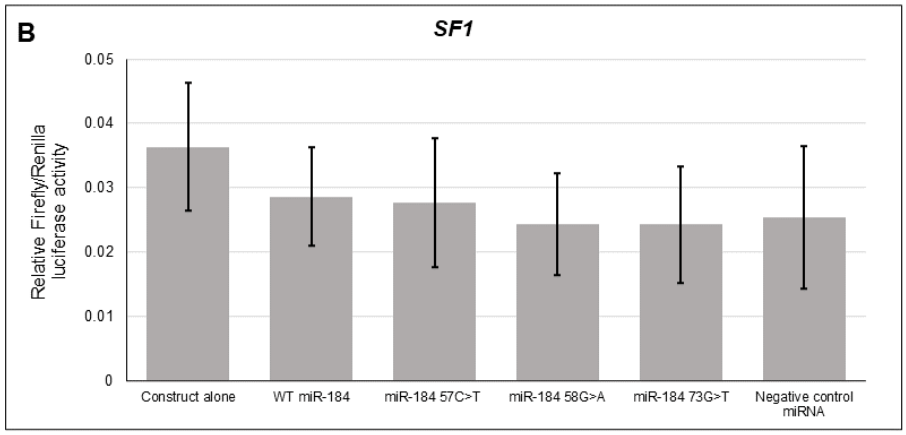
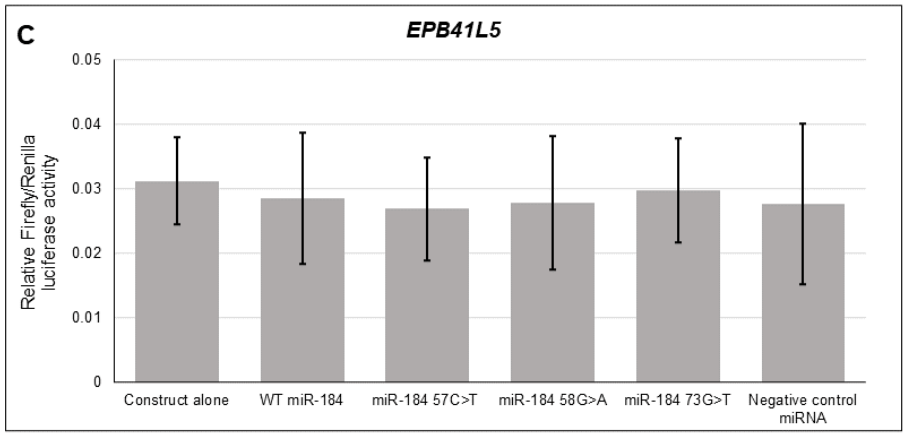


Figure 54 Photograph of the GFP-transfected HEK293 cells using TransIT®-LT1 Transfection Reagent obtained with a fluorescence microscope at 2x optical zoom, Scale bars, 25 µm.

HEK293 cells were then co-transfected with each mRNA target construct and mRNA mimics and incubated at 37°C for 48 hours. Next, the Dual-Glo® Luciferase Assay System was used to measure expression levels for each respective pmirGLO construct, comprising selected mRNA target gene 3'UTR regions fused downstream of *luc2* in the presence or absence of wildtype and mutant miR-184 miRNAs. For each construct and miR-184 condition, 6 replicates were performed to allow for outliers to be removed. **Figure 55** shows the relative firefly/Renilla luciferase activity for the different combinations of pmirGLO constructs and synthetic miRNAs tested in addition to a pmirGLO only and a control miRNA only transfection conditions.


A
AKT2
Relative Firefly/Renilla luciferase activity
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Construct alone
WT miR-184
miR-184 57C>T
miR-184 58G>A
miR-184 73G>T
Negative control miRNA


B
SF1
Relative Firefly/Renilla luciferase activity
0.05
0.04
0.03
0.02
0.01
0
Construct alone
WT miR-184
miR-184 57C>T
miR-184 58G>A
miR-184 73G>T
Negative control miRNA


C
EPB41L5
Relative Firefly/Renilla luciferase activity
0.05
0.04
0.03
0.02
0.01
0
Construct alone
WT miR-184
miR-184 57C>T
miR-184 58G>A
miR-184 73G>T
Negative control miRNA

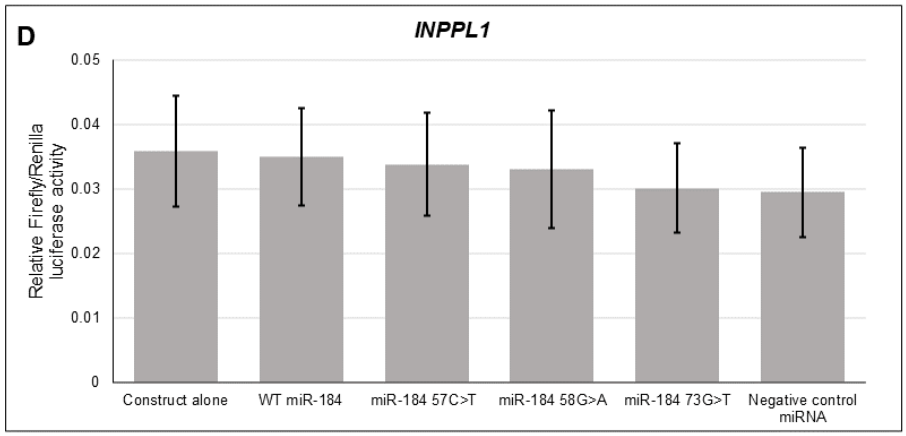
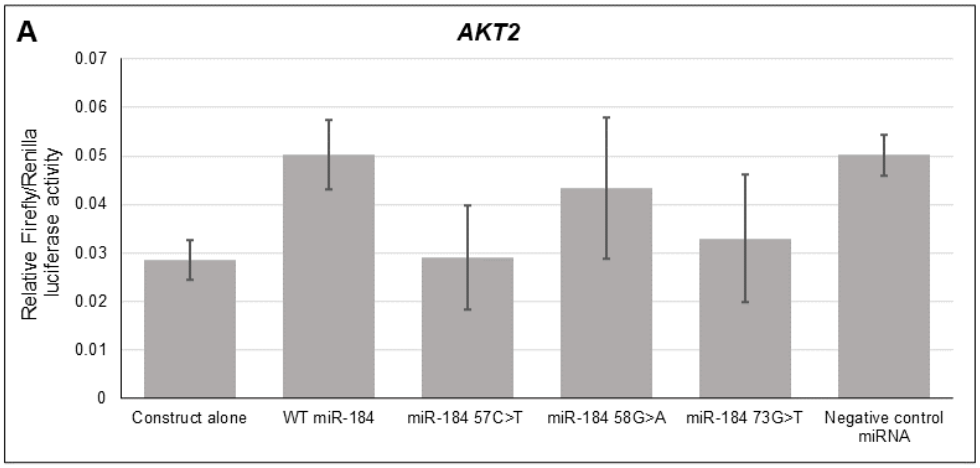
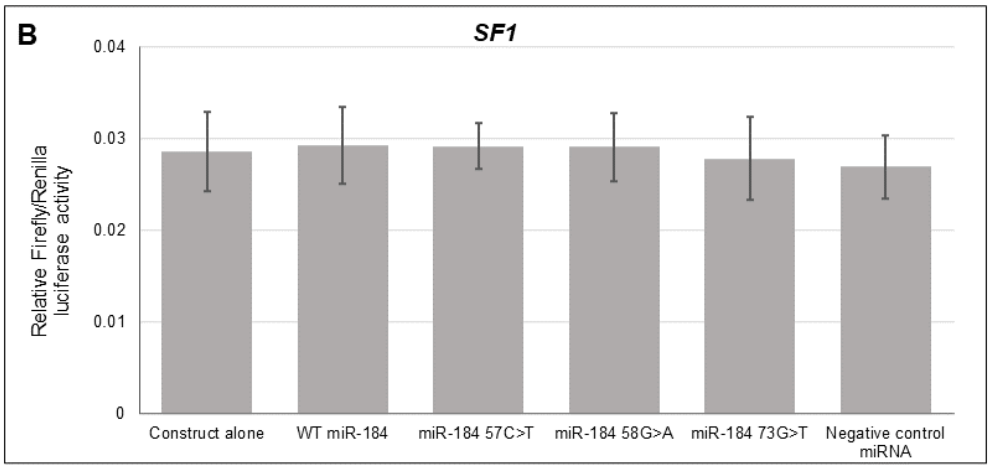
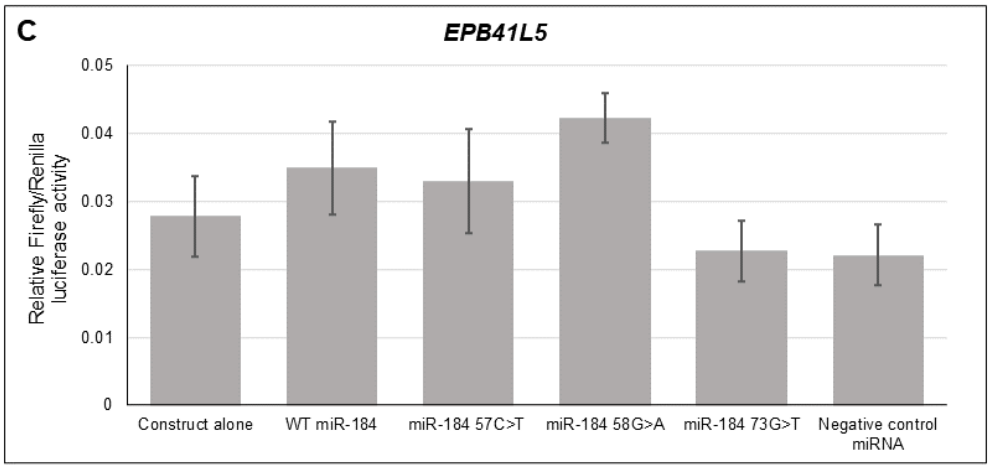

D
INPPL1
Relative Firefly/Renilla luciferase activity
0.05
0.04
0.03
0.02
0.01
0
Construct alone
WT miR-184
miR-184 57C>T
miR-184 58G>A
miR-184 73G>T
Negative control miRNA

Figure 55 Luciferase reporter assay designed to test if miR-184 variants alter the capacity of the microRNA to regulate 3'UTRs regions present within *AKT2*, *SF1*, *EPB41L5*, and *INNPL1*. Target Expression Vector pmirGLO contained the 3′-UTR regions of the four predicted miR-184 primary target genes **(A)** *AKT2*, **(B)** *SF1*, **(C)** *EPB41L5*, **(D)** *INNPL1* were co-transfected into HEK293 cells with either a wild-type miR-184 mimic or a custom designed mimic containing the (+57C>T), (+58G>A) or (+73G>T) variant. Cells were also transfected with the pmirGLO vector-3'UTR mRNA target constructs without any miR-184 mimic and a negative control miRNA mimic for control purposes. Cells were transfected for 48-hours before measuring expression levels using dual-glo luciferase assay.

The luciferase data generated 'noisy' data with large error bars, even after removing outliers. However, for all genes apart from *AKT2*, expression levels were relatively consistent when comparing the effects of adding the WT-miR-184 mimic to the respective mutant version of the miRNA being tested (**Figure 55**) compared to those with the variants in the sequence. For *AKT2*, data was determined to be unreliable given the large amounts of variation observed within the dataset. To overcome this, the experiment was repeated using a transfection time of 24 hours in attempts to remove noise signals from over-confluent and/or dead cells, **Figure 56**.


A
AKT2
Relative Firefly/Renilla luciferase activity
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Construct alone
WT miR-184
miR-184 57C>T
miR-184 58G>A
miR-184 73G>T
Negative control miRNA


B
SF1
Relative Firefly/Renilla luciferase activity
0.04
0.03
0.02
0.01
0
Construct alone
WT miR-184
miR-184 57C>T
miR-184 58G>A
miR-184 73G>T
Negative control miRNA


C
EPB41L5
Relative Firefly/Renilla luciferase activity
0.05
0.04
0.03
0.02
0.01
0
Construct alone
WT miR-184
miR-184 57C>T
miR-184 58G>A
miR-184 73G>T
Negative control miRNA

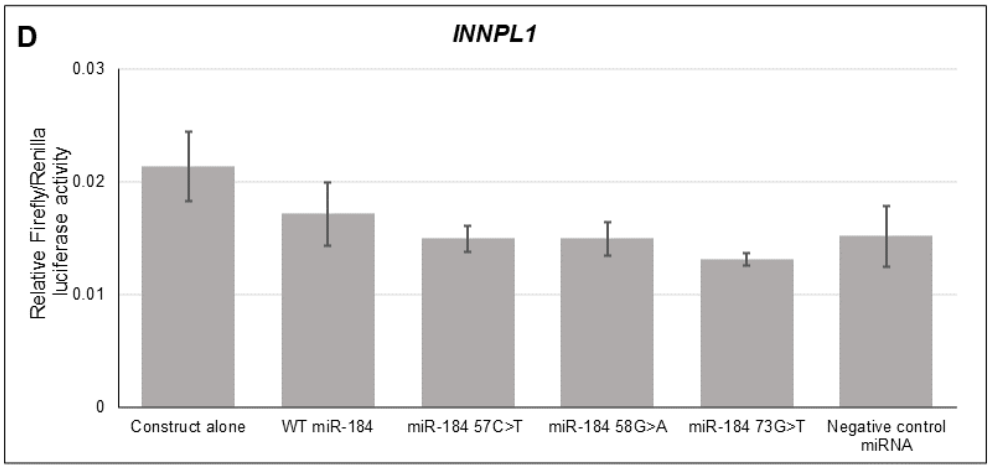

D
INNPL1
Relative Firefly/Renilla luciferase activity
0.03
0.02
0.01
0
Construct alone
WT miR-184
miR-184 57C>T
miR-184 58G>A
miR-184 73G>T
Negative control miRNA

Figure 56 Luciferase reporter assay designed to test if miR-184 variants alter the capacity of the microRNA to regulate 3'UTRs regions present within *AKT2*, *SF1*, *EPB41L5*, and *INNPL1*. Target Expression Vector pmirGLO contained the 3′-UTR regions of the four predicted miR-184 primary target genes **(A)** *AKT2*, **(B)** *SF1*, **(C)** *EPB41L5*, **(D)** *INNPL1* were co-transfected into HEK293 cells with either a wild-type miR-184 mimic or a custom designed mimic containing the (+57C>T), (+58G>A) or (+73G>T) variant. Cells were also transfected with the pmirGLO vector-3'UTR mRNA target constructs without any miR-184 mimic and a negative control miRNA mimic for control purposes. Cells were transfected for 24-hours before measuring expression levels using dual-glo luciferase assay.

The 24-hour transfection improved the data quality slightly, however, replicates from each condition still proved to produce relatively 'noisy' data indicated by the size of the error bars in **Figure 56**. The pmirGLO miRNA Target Expression Vector coupled to the 3′-UTR regions transfected alone were anticipated to have similar expression levels to when co-transfected with the negative control miRNA mimic, given the negative control miRNA is not predicted to bind the respective 3'UTR regions. This control has been extensively tested in human cell lines and tissues and validated to not produce identifiable effects on known miRNA function. This was mostly accurate for *SF1*, *EPB41L5* and *INPPL1*, **Figure 56.B, 56.C and 56.D**, respectively. However, surprisingly for the *AKT2* construct when co-transfected with the negative control miRNA mimics, the relative firefly/Renilla luciferase activity levels were higher than when transfected with the *AKT2*-pmirGLO construct alone. This data however may not be indicative of the true expression levels given the error bars are very large for all *AKT2* conditions, **Figure 56.A**. For *SF1*, there was minimal change between the different transfection conditions suggesting that the miR-184 does not have an effect on this mRNA target gene, **Figure 56.B**. For *EPB41L5*, the relative firefly/Renilla luciferase activity levels did vary between conditions with the EPB41L5-pmirGLO construct co-transfected with the mimic containing the (+58G>A) variant, however, the error bars again

indicate this data cannot be relied upon, **Figure 56.C**. For *INNPL1*, expression levels for all the conditions co-transfected with a miR-184 mimic, WT and those containing variants showed similar expression levels, below the relative luciferase activity of when the *INNPL1*-pmirGLO construct was transfected alone and with the negative control miRNA mimic. This would suggest miR-184 downregulates *INNPL1* expression. The miR-184 mimics containing variants all had slightly lower relative luciferase activity level compared to the WT miR-184 mimic, suggesting all variants have a further dysregulation effect on *INNPL1* compared to WT miR-184. However, again the large error bars do indicate this data may not be reliable, **Figure 56.D**.

5.3 Discussion

5.3.1 Rare variants identified in gene previously associated with FECD

In our cohort three unrelated individuals were identified to harbour the previously early-onset FECD associated *COL8A2* missense mutation p.(Gln455Lys (Biswas, 2001). All three subjects presented with an early-onset phenotype, comparable to the phenotypic presentation of corneal endothelial dystrophy patients previously reported with the same mutation. There has been sufficient studies, utilising knock-in mice models, to support the pathogenicity of mutations in this gene, including p.(Gln455Lys), to be considered causative for the early onset phenotype (Jun et al., 2012; Meng et al., 2013). Two further unrelated cases were also identified to harbour rare *COL8A2* missense variants, p.(Arg434His) and p.(Pro575Leu). The variant p.(Arg434His) had previously been published in literature reporting an individual with typical late-onset FECD (Gottsch et al., 2005). The patient identified in this cohort with this variant, p.(Arg434His), also presented with a late-onset phenotype unlike other reported *COL8A2* mutations, suggesting this variant does not result in an early-

onset phenotype but may still be disease-associated given the associated in silico prediction scores and low frequency in the control population. However, the proband's two daughters were able to be recruited to this study for segregation analysis, one of which carried the variant and one of which was wild-type. Both daughters underwent detailed examination and were found to have no symptoms of FECD. At the time of the examination the daughter who also carried the variant was 43-years of age. Given this variant has previously been associated with a late-onset phenotype, it may be that she is yet to develop symptoms. Further validation work into how p.(Arg434His) could be pathogenic is necessary in determining whether this variant is causative of FECD in this individual. The variant p.(Pro575Leu) had not previously been reported in any literature and familial samples from this individual were not available for segregation analysis. The individual presented with a slightly earlier onset of disease, at 44-years old, in line with the early-onset phenotype, however, again further validation studies are needed before determining if this variant can be considered disease-causing.

Heterozygous missense mutations in *SLC4A11* and *ZEB1* have now been established to be causative of late-onset FECD, although the mechanism behind how these mutations result in the FECD phenotype remains to be elucidated (Chung et al., 2014; Malhotra et al., 2019; Mehta et al., 2008; Riazuddin et al., 2010; Vithana et al., 2008). In this study six missense variants and one splice site variant were identified within *SLC4A11*. Six further synonymous *SLC4A11* variants were also identified. Of these variants, only one, p.(Arg109Leu), was truly suggestive of being potentially pathogenic as it had a notably high CADD score of 31 and found to alter a highly conserved residue located in functional domains of the encoded solute carrier. Future

segregation analysis and/or identification of any of these variants in further patient cohorts would help to resolve if they are in fact causal of FECD, or functionally benign rare polymorphisms.

In *ZEB1* eight heterozygous missense variants, one in-frame deletion variant and one splice site variant were identified along with five synonymous variants. Contradictorily, four of the *ZEB1* variants, p.(Asn78Thr), p.(Lys554Arg), p.(Ser202=) and p.(Ala420=) were seen in together in four unrelated individuals of African American ancestry. These four variants were found to be in close linkage disequilibrium (D' 1.0, R2 >0.9) and all had a MAF of above 5% in the gnomAD African/African American population. It is highly unlikely these variants are disease causing with a frequency as high as 5%, as FECD is thought to affect up to 5% of the population over 40 years of age in Caucasian population with the prevalence being lower in African/ African American populations (Minear et al., 2013). Further analysis, including segregation analysis and functional validation, is needed to establish if the other *ZEB1* variants identified in this study are indeed disease causing in the 139 remaining subjects.

In *LOXHD1* 25 variants (MAF <0.01) in total were detected in our cohort, 14 of which were missense mutations, and two unrelated subjects were identified to share the same variant, p.(Arg524Cys). All missense variants, excluding one, had been observed in one of the control datasets, although these variants were extremely rare (MAF <0.01). Only one of the variants, c.541C>T, p.(Leu181Phe), to the best of my knowledge, had previously been reported in literature . As the frequency for these variants are so rare, it does not rule out the possibility of being disease causing, given FECD is predicted to affect approximately 5% of the Caucasian population (Baratz et al., 2010).

Similar to the study that first proposed *LOXHD1* as a candidate gene for FECD, a high proportion of variants were observed in our cohort with high predicted pathogenic score. Initially, Riazuddin *et al*. suggested there was an enrichment of pathogenic variant observed in the FECD in comparison to the control cohort, however this may be a coincidental finding, as the *LOXHD1* gene appears highly polymorphic (Riazuddin et al., 2012). The initial *LOXHD1* variant identified by Riazuddin *et al*. was a result of linkage analyses identifying a common locus mapping to 18q21.2-q21.32 interestingly, since this finding the *TCF4* CTG18.1 expansion, located on chromosome 18q21.2, has been associated with a high proportion of late-onset FECD (Baratz et al., 2010; Riazuddin et al., 2012; Sundin et al., 2006). Although haplotype analysis performed by the group indicated the signals produced from *LOXHD1* and *TCF4* in these region are independent of one another, these findings have limited evidence and were conducted before the discovery of the *TCF4* CTG18.1 repeat expansion (Riazuddin et al., 2012). Conversely, the pedigrees used to initially identify the 18q21.2-q21.32 region, have not been shown to have been genotyped for the CTG18.1 repeat expansion to eliminate this is where the signal arose from. Furthermore, LOXHD1 is not expressed within the corneal endothelium (TPM 0.1) and hence it is not possible to theorise how mutations inducing either haploinsufficiency (i.e. PTC) or altered function (i.e. missense variants) could exert an effect in a cell type whereby expression is not switched on. This highlights the need to re-evaluate evidence available to support *LOXHD1* variants as being causative of FECD.

In our cohort I have identified one patient with a previously reported nonsense variant, p.(Arg1028Ter), in *AGBL1* and a further three patients with novel missense variants. Some evidence has suggested mutations in *AGBL1*

result in an enrichment of the protein in the nucleus but more importantly, that the AGBL1 protein interacts with TCF4 and the nonsense and a missense variant within *AGBL1* result in reducing the binding affinity of TCF4 (Riazuddin et al., 2013). Although the mechanism of *TCF4* is not fully comprehended in the pathophysiology of FECD, these findings propose a potential mechanism for *AGBL1* mutations to result in a FECD phenotype. Nonetheless, further validation work is required. Initial findings indicated *AGBL1* mutations may account for approximately 1%-2% of the genetic burden for FECD (Riazuddin et al., 2013), however given that only 9 missense and one nonsense variants were identified in our cohort , this would suggest a lower proportion, along with other studies having failed to identify any *AGBL1* mutations in their cohort indicating if *AGBL1* is, in fact causative for FECD, it is extremely rare (Okumura, Hayashi, Nakano, Tashiro, et al., 2019; Skorodumova et al., 2018). Furthermore, the *AGBL1* variant p.(Arg1028Ter) could not explain every account of FECD in the family it was first identified in, proposing FECD in this family may be heterogeneous with multiple causal alleles, however, there is a possibility the family may have a single causal variant that is yet to be identified. Similarly, to *LOXHD1*, *AGBL1* is not expressed in the corneal endothelium (TPM 0), therefore, variants within this gene are unlikely to have a causative effect.

In this study I also discovered two unrelated individuals with rare and predicted deleterious *TCF4* variants. These variants identified were in *cis* missense and nonsense variants, p.(Arg19Ser) and p.(Lys20Ter) in proband BR65 and splice site variant p.(Glu22=) in Proband BR63. Previously *TCF4* has only been associated with FECD through the CTG18.1 expansion and this finding proposes the potential role that rare *TCF4* variants may play in the absence of CTG18.1 expansions in FECD pathogenesis. Both the p.(Lys20Ter)

and p.(Glu22=) have been predicted to result in null alleles; p.(Lys20Ter) through the introduction of a PTC and p.(Glu22=), altering the splice site and is predicted to introduce a short frameshift insertion, followed by a PTC. Furthermore, the rare *TCF4* variants I have identified here are located just upstream of the CTG18.1 locus and in addition are all located within exons that are not included by the vast majority of TCF4 isoforms. This leads to the hypothesis that they would exert loss-of-function and/or regulatory effects on TCF4 functionality, limited to a subset of annotated isoforms. This observation is in keeping with the relatively mild and tissue specific nature of FECD, which is in stark contrast to the severe neurodevelopmental disorder Pitt-Hopkins associated with total TCF4 haploinsufficiency (Sirp et al., 2021). Hence, it can be proposed that the variants identified in this study provide significant insight into which TCF4 isoforms, when dysregulated, may induce CEC-specific disease.

5.3.2 Variants identified in GWAS associated genes

In 2017, a large GWAS study conducted by Afshari *at al*. identified four loci with strong evidence for genome-wide significant association ($P<5x10^{-8}$), with most significant association corresponding to the previously identified *TCF4* locus. Three further signals, on chromosome 1, corresponded with an intronic region within the *KANK4* gene (rs79742895), intergenic region between *LINC00970* and the *ATP1B1* gene (rs1022114) and intronic region within the *LAMC1* gene (rs3768617). As a result, *LAMC1*, *KANK4* and *ATP1B1*, were proposed as likely candidate genes associated with FECD based on their respective locations, however, it does not eliminate the possibility that other genes or transcripts within these regions underlie the associations.

In the CTG18.1 expansion-negative subjects (n=141) only synonymous variants were found in *ATP1B1*. These variants were all predicted to be likely benign as synonymous variants usually do not have a functional impact on a transcript.

In *KANK4*, two missense variants were identified in unrelated individuals within the cohort. Both of the variants were observed in at least one control dataset at a very low frequency. The missense variant p.(Ala588Thr) had a CADD score of 23.3, predicting it to be likely pathogenic. The other variant identified was a splice site variant, c.2231+1del. The cellular function of *KANK4* is not yet established and therefore uncertainty remains as to how mutations can relate to disease, however, previous mutations in the gene have been associated with a rare cause of nephrotic syndrome (Gee et al., 2015). The expression of *KANK4* in corneal tissue, comprised of the endothelium and DM, was minimal; nonetheless, immunostaining revealed *KANK4* protein was localised in the endothelial cytoplasm in both FECD samples and controls however, this may be a result of the antibody being non-specific (Afshari et al., 2017). As the cellular function of *KANK4* is currently unknown it is not possible to confirm if the variants found in this study are the causative for FECD or casual polymorphisms. Future functional analysis is required to confirm if such variants in *KANK4* may cause FECD.

Nine rare (MAF <0.01) *LAMC1* missense variants were identified in six unrelated individuals of our cohort, all variants produced CADD scores predictive of being pathogenic. All variants had been observed in either gnomAD or the UCLex control datasets at low frequencies. However, despite the variants having a very low MAF frequency, there is an apparent enrichment of variants predicted to be pathogenic in the gnomAD control population.

Further investigation through segregation analysis, functional validation and deep phenotyping, is required to explore the implications of these variants and how they may be causative of FECD. Although we cannot currently conclude whether *LAMC1* variants observed in our cohort are causative for FECD, *LAMC1* has been found to be highly expressed in corneal tissue composed of the endothelium and DM and encodes for laminin subunit gamma. Laminins are ECM glycoproteins comprised of laminin alpha, beta and gamma chains, and have a key role in cellular adhesions in basements membrane such as DM, making an ideal candidate gene for causing corneal endothelial disease (Afshari et al., 2017). Prior to the GWAS, *LAMC1* had not been identified as a candidate gene for corneal dystrophies and there has since been only one reported case of an identified *LAMC1* variant in a CTG18.1 expansion negative FECD individual (Wieben et al., 2018). To date there has been no functional analysis published to provide evidence that variants in *LAMC1* may be causative of FECD.

5.3.3 Candidate genes through familial samples

A *COL8A1* missense variant, c.619A>T, p.(Ile207Phe) was identified, in a 35-year-old male presenting with an early-onset FECD phenotype. Despite not previously being associated with FECD, *COL8A1* presented as a strong candidate gene for FECD given that *COL8A1* and *COL8A2* form homotrimers within the DM (Greenhill, Rüger, Hasan, & Davis, 2000) of the cornea and mutations in *COL8A2* has be established to cause early-onset FECD (Aldave et al., 2006). As the proband mentioned here presents with an early-onset phenotype, and the similarities of the genes' functional properties, this variant was very intriguing. However, upon recruitment of familial samples, the variant

was found not to segregate in affected relatives and thus was eliminated for causing disease in this family.

Due to the late-onset nature of FECD, affected first degree relatives are often difficult to recruit as the proband's parents are likely deceased and offspring are not yet knowingly affected, therefore familial cases were limited. However, as I was fortunate to recruit affected familial samples for this proband, I had an excellent opportunity to perform a family-based filtering strategy with the aim to produce a candidate gene list. Two of the proband's affected siblings had exome sequencing performed and shared variants were interrogated in all genes. This led me to identify a novel shared variant, c.658C>T p.(Arg220Cys) in the gene *LYPD3*. LYPD3 has been shown to have an involvement with the adhesion of laminins, a major component of basement membranes such as DM in the cornea (Paret et al., 2005). LYPD3 was also identified to be relatively highly expressed in the endothelium of the cornea. The function of LYPD3 and expression levels within the corneal endothelium suggest that it could be a viable candidate gene for FECD, with dysfunction of LYPD3 potentially inducing defective adhesion of the corneal endothelium to DM. Further validation work is needed to know if this variant and gene could have an impact on the corneal endothelium and be considered to be a candidate gene leading to disease.

5.3.3 Limitations to exome sequencing

A limiting factor of exome sequencing is that only protein coding regions of the genome are sequenced and thus only a small proportion of the genome is interrogated. However, despite covering less than 2% of the entire genome, approximately 85% of Mendelian disease gene mutations are estimated to occur in protein coding regions, therefore exome sequencing can be an excellent genetic technique for investigating monogenic diseases, including

FECD (Rabbani et al., 2014). However, in this study I was only able to confirm causative variants with 1.4% of total cases investigated. Therefore, there is a strong possibility that mutations may be located within the non-coding portion region of the genome. In the future, WGS could be used to explore the non-coding portion of these individuals' genomes to overcome this study limitation.

Furthermore, a major challenge to exome sequencing is the lack of efficient prioritisation of the vast number of variants identified in each proband. Each individual has approximately 27,000 total variants and after filtering for rare variants (MAF <0.01), approximately 1,400 remain. In this study I used the standard approach of filtering by frequency, as any variant with a high frequency in the control population could not be responsible for this rare subset of a disorders affecting approximately 5% of the population, giving we know approximately 80% of FECD is caused by the *TCF4* CTG18.1 expansion. The bioinformatic tool CADD, designed to predict the functional consequence of a mutation, was incorporated into the exome sequencing data to support the assessment of variants rather than as a specific filtering criteria (Kircher et al., 2014; Rentzsch et al., 2019). Variants within genes already previously associated with FECD were first interrogated as a starting point. This led us to the discovery of a number of potential disease-causing variants within a few individuals but could not explain the disease in the majority of the cohort, illustrating that further unsolved genetic heterogeneity remains for FECD.

5.3.4 Gene burden analysis

In attempts to overcome the limitations of exome sequencing through variant prioritisation alone, I applied a gene burden approach to the CTG18.1 expansion-negative cohort with the aim to identify candidate genes by comparing the number of individuals carrying rare, deleterious variants in genes

compared between case and control subjects. The advantage of this approach is where a single variant would usually be underpowered to detect statistical signals between case and control subjects, combining variants across a candidate gene might improve power (M. H. Guo et al., 2018). Furthermore, this approach can be applied to unrelated case subjects thus overcoming limitations where large multiplex families are unavailable or incomplete penetrance, as commonly seen in FECD (M. H. H. Guo et al., 2016).

One limitation to this approach is that all samples should ideally be of the same ancestry to be able to directly compare the frequency or rare and potentially deleterious variants in case subjects compared to control subjects. To overcome this, samples that were not of European ancestry were excluded from this analysis. This introduced a caveat that variants driving disease in non-European populations could be missed as we know that the CTG18.1 expansion is the prevalent cause of disease in European populations and is typically lower in other non-Caucasian ethnic groups meaning a particular ethnic group could possibly have additional genetic causes of disease that would not have been detected (Fautsch et al., 2021). I was unable to replicate the analysis using non-European cases and the analysis would be underpowered due to the low n numbers within these ethnicity groups, **Section 3.2.3**.

Nevertheless, we were able to identify a very interesting variant in miR-184. miR-184, which has previously been associated as the genetic cause for EDICT syndrome (Hughes et al., 2011; Iliff et al., 2012). Given the corneal endothelial cell-specific phenotype is a component of EDICT syndrome it seems biologically plausible that additional variants in the miRNA may also result in other endothelial-cell specific phenotypes such as FECD and therefore this variant was further investigated.

. miRNAs critical regulators of gene expression and miRNA is a multistep process in which both nuclear and subsequent cytoplasmic cleavage events occur by two ribonuclease III endonucleases, DROSHA and DICER1. In brief, miRNA genes are transcribed by RNA polymerase II to produce the miRNA primary transcript (pri-miRNA), which has the characteristic hairpin structure. The pri-miRNA is then cleaved by the microprocessor complex including DROSHA and released into the nucleoplasm as a secondary miRNA precursor (pre-miRNA). Afterwards, pre-miRNAs are exported from the nucleus to the cytoplasm where they it is further processed by the miRNA-induced silencing complex, in which the main component is RNase DICER1. In this process, apical loop of the hairpin-shaped pre-mRNA are removed, generating an ~ 22-bp-long miRNA duplex. One of the strands becomes the mature miRNA, functioning as functions as a guide strand to recognise and silence target mRNAs while the other passenger strand is hydrolysed (J. Liu et al., 2022).

miRNA sequence motifs are highly conserved and crucial for the proper biogenesis of miRNAs. Any mutations leading to the aberrations of the miRNA sequence or structural conformation have the potential to result in impaired miRNA processing, changes in the miRNA level or specificity of miRNA target recognition (Machowska, Galka-Marciniak, & Kozlowski, 2022). **Figure 57** demonstrates the potential effect of different types of genetic variation on the

functionality of miRNA genes.

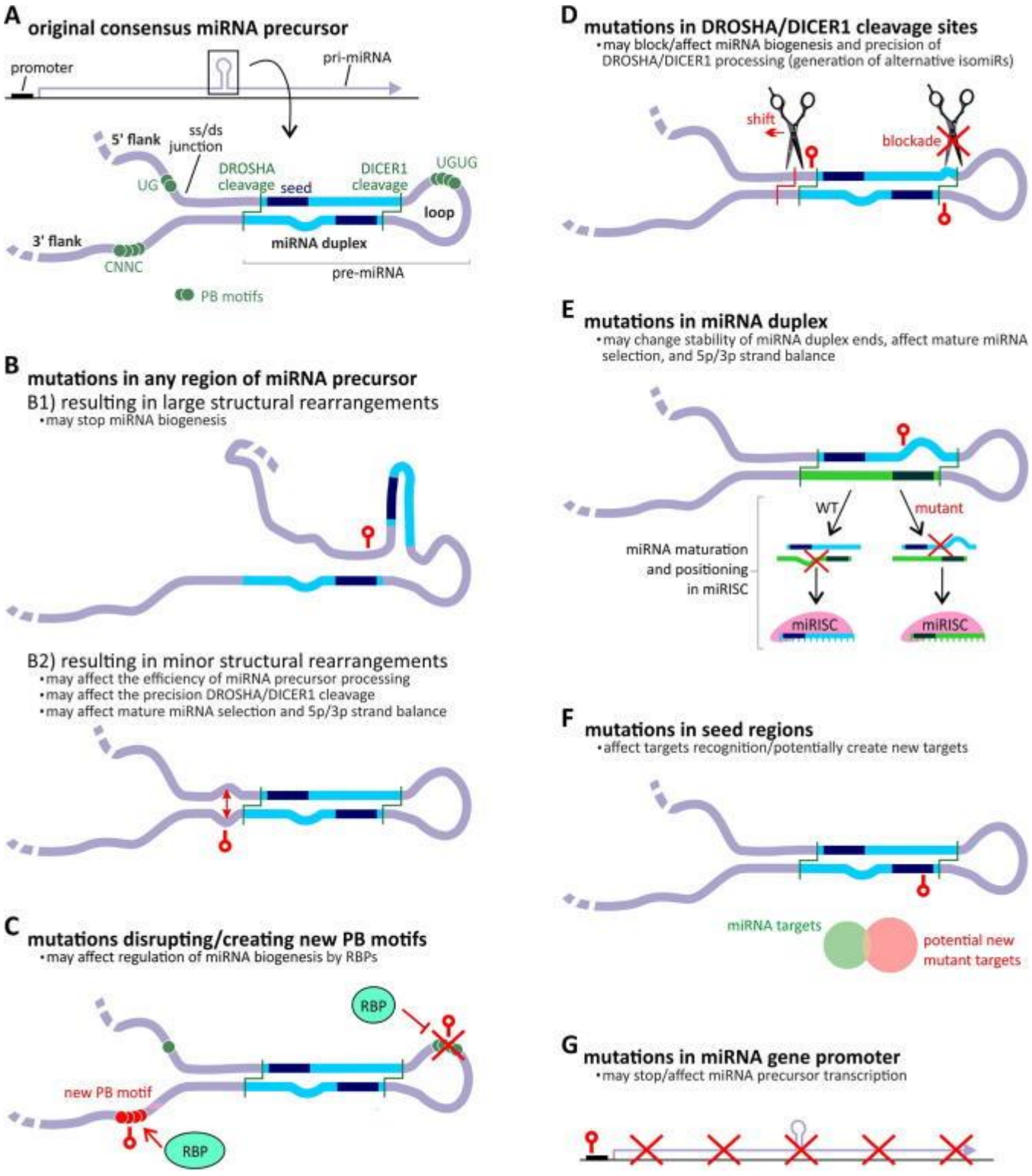


Figure 57 A schematic representation of the potential effects of mutations on the functionality of miRNA genes. A) A schematic representation of the miRNA gene (above) and canonical miRNA precursor (below), with indicated miRNA precursor subregions and functional elements. B-G) Different effects of miRNA gene mutations. Positions of miRNA mutations are indicated as red lollipop symbols. (Machowska et al., 2022)

The miR-184 variant, +58G>A, identified in the gene burden analysis led to the finding of a second variant, +73G>T, through further interrogation of the exome data. Both the variant found to cause EDICT syndrome, +57C>T, and the one identified through this gene burden analysis, +58G>A, are located within the seed region of miR-184. The seed region is a highly conserved region of the miRNA located from the second to seventh base of the mature miRNA. Seed regions are particularly important for recognition of target mRNA and adequate miRNA regulation of protein expression (Lewis, Burge, & Bartel, 2005). **Figure 57.F** suggests mutations in the seed region of miRNA have the potential to affect mRNA target recognition or potentially create new targets (Machowska et al., 2022)

These findings are supportive of the concept that variants within the seed region can result in functional impairment and thus cause disease. Furthermore, the mature sequence of mRNAs is also important, where the variant +73G>T is located, as regulatory proteins can also bind mature miRNA to direct their degradation, preventing their expression this occurs when the RNA duplex is unwound and the single strand mature miRNA is incorporated into the protein complex RISC to function as a guide, directing the silencing of target mRNA (MacFarlane & Murphy, 2010). These findings again support the idea variants found within the miRNAs, including the mature sequence, may lead to functional impairment and therefore disease.

To explore the effect of the miR-184 variants on expression levels, four target mRNA genes were selected based on previous literature and prediction from databases. There is a strong possibility other mRNAs may interact with miR-184 within the cornea. Expression levels of the mRNAs selected were assessed by using the Dual-Glo® Luciferase Assay System, and co-transfecting

with miRNA mimics designed to either imitate the activity of WT miR-184 or when the sequence contained the variants identified. The findings of the study were inconclusive which could be a result of numerous factors. Firstly, the luciferase assay needed further validation work as the data obtained was not the most reliable. This includes the cell seeding density, the transfection time and concentration of the substrates. Unfortunately, due to insufficient time available I was not able to optimise this experiment.

Furthermore, miR-184 has been shown to compete with another miRNA, miR-205 in the epithelia of the cornea (Hughes et al., 2011; Yu et al., 2008). It may be the case that a similar phenomenon is occurring in the endothelium of the cornea with miR-205 or another yet to be identified miRNA. Further investigation is needed to comprehensively investigate the effect of these variants in miR-184 and to establish the mechanism in which they may result in disease. This would include further validating the luciferase experiment and including more potential mRNA targets. Additionally, the idea that miR-184 may compete with other miRNAs in the cornea needs to be explored and the potential miRNAs identified.

Moreover, it is interesting how the +57C>T and +58G>A variants result in two distinct phenotypes despite being one nucleotide apart. The possibility of the three probands which harbour the +58G>A change having EDICT syndrome has been excluded using available clinical data. All three probands do not present with key characteristics of EDICT syndrome including congenital cataracts, iris hypoplasia, keratoconus and stromal thinning. In fact, these patients all displayed thickening of the stroma, a distinct feature of FECD. It can be certain that these three patients with the +58G>A variant do not have EDICT syndrome and do indeed have FECD.

5.4 Conclusion

This section aimed to explore the genetic heterogeneity of non-expanded CTG18.1 FECD through the use of exome sequencing. Interrogation of 141 FECD without a known genetic cause for previously associated FECD genes found 3 patients to harbour the *COL8A2* early-onset disease causing mutation, p.(Glu455Lys). Although other potentially pathogenic variants were identified, the COL8A2 mutations were the only definitive causing variant identified through exome sequencing alone.

To further explore genetic causes of disease in this cohort a gene burden test was performed on the exome data, to seek if there were an enrichment in deleterious variants in cases compared to controls. This study led to the discovery of a novel miR-184 variant, +58G>A, in three unrelated individuals. Although further investigation into whether the miR-184 variant, +58G>A is disease causing, it is a notable candidate gene. Furthermore, the finding of this variant exemplifies the utility of applying a gene-burden style approach to identify novel genetic causes of disease within the unrelated sporadic CTG18.1 expansion-negative FECD cohort studied.

Moreover, this research has highlighted the hypothesis of the possibility that a subset of these cases do not have an underlying genetic cause of disease, as the vast majority of these CTG18.1 expansion-negative cases are sporadic, with no family history. It can be suggested that in some cases environmental factors may play a greater role leading to endothelial failure as a natural ageing process.

6. General discussion and concluding remarks

Since the association between the CTG18.1 expansion and FECD was reported in 2012, significant attempts and progress has been made in identifying potential mechanisms that underlie the disease and how they correlate with phenotypic outcomes. However, much remains unknown with respect to disease mechanisms and molecular consequences associated with CTG18.1 repeat, and when expanded gives rise to a corneal-specific disease phenotype. My thesis aims to further explore how the CTG18.1 expansion correlates with FECD phenotypic outcomes specifically by investigating (1) repeat length of the repeat itself, (2) levels of somatic expansion rates and (3) how genetic modifiers may influence somatic instability.

Furthermore, since the discovery of the CTG18.1 repeat, little research has been undertaken into uncovering the genetic cause of disease in FECD patients which do not harbour a CTG18.1 repeat expansion. The majority of literature which describe FECD-associated genes were conducted prior to the discovery of the CTG18.1 expansion association and were carried out through traditional linkage studies using familial cohorts. In this study I sought to further investigate FECD missing heritability within a large CTG18.1 expansion-negative cohort using exome sequencing to explore the role of previously established and novel rare variants which may contribute to the pathogenesis of FECD.

6.1 Summary of key findings

Firstly, I genotyped a large cohort of 990 unrelated FECD patients for the CTG18.1 repeat using a combination of PCR based approaches (STR and TP-PCR assays). I discovered that almost 80% (n=770) of FECD patients recruited to this study harboured one or more expanded CTG18.1 alleles (≥50 repeats),

in keeping with previous reports of other relatively smaller largely European cohorts investigated (Luther et al., 2016; Mootha et al., 2014; Skorodumova et al., 2018; Zarouchlioti et al., 2018). This strengthens the hypothesis that the CTG18.1 expansion is the most common genetic risk factor for disease within the Caucasian population (OR = 94.59; 95% CI: 60.50-148.74; $p = 6.52 \times 10^{-78}$) (**Section 3**). Overall, a significantly higher prevalence of females, compared to males, was observed in the total cohort (60.8% versus 39.2%; p= 0.00002). This skewing between the prevalence of disease in females and males was more pronounced within the CTG18.1 expansion-negative portion of the cohort (73.2% female versus 26.8% male), indicating involvement of additional CTG18.1 independent genetic and/or environmental factors. Additionally, males were also identified to have a lower trending age-of-recruitment in the CTG18.1 expansion-negative group, suggesting a possible X-linked early-onset form of the disease that may underlie a subset of cases within this group.

In **Section 4,** I further characterised the CTG18.1 locus in CTG18.1 expansion-positive FECD samples (n=630) by applying a targeted high-throughput ultra deep sequencing approach. This allowed me to investigate the dynamic nature of the CTG18.1 expansion and genotyping of the downstream CTC repeat. To the best of my knowledge, this is the first time that a targeted deep sequencing approach has been applied to an FECD patient cohort, on a large scale. My data provides sizable evidence that CTG18.1 is somatically unstable and that these greater levels of instability correlate with increased ePAL length (Alkhateeb, 2018; Hafford-Tear et al., 2019; Wieben, Aleff, et al., 2019). Hence, my thesis substantiates existing evidence that the CTG18.1 repeat is a dynamic unit when expanded, and that we should move away from considering CTG18.1 expanded genotypes as stable entities. Data presented

also suggest that levels of somatic instability may in future prove to modify disease outcome. Characterisation of the allelic structure of CTG18.1 across the cohort has also highlighted that the CTC repeat is more variable on expanded alleles compared to non-expanded alleles. Future work is necessary to understand the implications this variation has on the FECD phenotype.

Additionally, in **Section 4** I explored the frequency and influence of common polymorphisms within DNA repair genes within FECD patients, which have previously been established to have modifying effects on the onset of diseases such as HD and SCAs (Bettencourt et al., 2016; Ciosi et al., 2019; Consortium, 2019; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015). I used a KASP assay to genotype 12 SNPs from DNA repair genes and found a significant enrichment of the minor allele SNP rs1799977 (*MLH1*) in the white British FECD cohort, previously associated with a later residual HD onset of 0.8 years (Consortium, 2019). Through analysis comparing white British FECD samples and an age-matched cohort harbouring repeat expansions in the absence of clinical symptoms of FECD, I found no significant association between the MAF of SNPs between the cohorts investigated. I did however, observe two SNPs, rs1799977 (*MLH1*) and rs1382539 (*MSH3*) which followed trend patterns in the cohort without FECD symptoms where the frequency of these SNPs has been associated with delayed HD motor onset.

Using the somatic instability data acquired from the MiSeq sequencing and the SNP data from the KASP assay I conducted an association analysis to explore the effect these *trans*-acting modifiers had on the CTG18.1 repeat. I identified a significant directional effect for the *MSH3* SNP rs701383 and *FAN1* SNPs rs34017474 and rs3512. The minor allele at these SNPs were all significantly associated with increased levels of somatic instability. This finding

supports the hypothesis that there are underlying modifiers which play a role in variable FECD expressivity which ePAL alone cannot explain. Future studies are necessary in validating the role these *trans*-acting modifiers have on the FECD phenotype. Furthermore, data collected suggests future studies focusing on *trans*-acting modifiers may be crucial in supporting our understanding of the complex nature of FECD, and these may prove to be more significant than *cis*-acting modifiers given interruptions within the CTG18.1 repeat were only identified in 2/630 samples.

In this study I identified three unrelated patients, with an early-onset phenotype, to have the rare, previously reported *COL8A2* disease causing mutation, p.(Glu455Lys) (**Section 5**). Two of these cases were identified through exome sequencing and another through further direct screening of additional early-onset cases (n=12) which did not undergo exome sequencing. This is currently the only established genetic cause for the rare early-onset FECD, however mutations in this gene do not explain of the majority of early-onset FECD (n=20) recruited to this study, suggesting further currently unidentified genetic factors may be responsible for early-onset FECD. The early-onset phenotype attributed to *COL8A2* mutations has also been established to cause definitive characteristics, such as mildly elevated guttae which are associated to an individual CEC, in comparison to the common late-onset FECD where guttae appear sharply raised and located along the borders between CECs. Furthermore, it also presents in a more coarse and distinct distribution, in contrast to a fine, patchy distribution of guttae as seen in the late-onset FECD (Gottsch et al., 2005). As this early-onset FECD phenotype is so rare and has distinct phenotypic characteristics, it leads to the question whether this could be a distinct corneal dystrophy, separate to the typical late-onset

FECD, in which there are additional, currently unknown genetic cause(s), including the *COL8A2* mutations previously established.

Through exome sequencing alone, the *COL8A2* mutations were the only definitive causative variant identified by this study. Although other potentially pathogenic variants were also identified, the majority of the CTG18.1 expansion-negative cohort remained without a known genetic cause. To further utilise the data I had, I collaboratively performed a gene burden test on the exome data which led to the discovery of a novel miR-184 variant, +58G>A, in three unrelated individuals (**Section 5**). This was a particularly enticing candidate variant as previously the cause of EDICT syndrome has been associated with the mutation miR-184 +57C>T and giving the partial phenotypic overlap between FECD and EDICT (Hughes et al., 2011; Iliff et al., 2012). The finding of this variant exemplifies the utility of applying a gene-burden style approach to identify novel genetic causes of disease within the unrelated sporadic CTG18.1 expansion-negative FECD cohort studied. Not only did I identify this miR-184 variant but the gene burden provided a wealth of other candidate genes. Unfortunately, I did not have the availability within my PhD time frame to explore these but many of these genes warrant future follow up experimental work and could potentially genetically solve many other CTG18.1 expansion-negative FECD cases in future. However, as the vast majority of these CTG18.1 expansion-negative cases are sporadic, with no family history, there is the possibility that a subset of these cases do not have an underlying genetic cause of disease. It can be suggested that some patients have an FECD-like phenotype as a result of environmental factors, subsequently leading to the degeneration of the corneal endothelium. Previous studies have proposed that oxidative stress and the accumulation of nuclear DNA damage

can contribute to CEC apoptosis and degeneration and therefore play a critical role in pathogenesis of FECD (Azizi et al., 2011; Jurkunas, Bitar, Funaki, & Azizi, 2010; López-Otín, Blasco, Partridge, Serrano, & Kroemer, 2013). This provides justification that in some cases environmental factors may play a greater role leading to endothelial failure as a natural ageing process.

6.2 Impact of this study on genetic diagnostics and patient care pathways

As a result of the findings of this thesis and the knowledge that new therapeutic treatments are currently under development (Angelbello et al., 2021; Hu et al., 2018; Powers et al., 2022; Zarouchlioti et al., 2018), my work contributed to applying for a new clinical indication under the National genomic test directory commissioned by the NHS, requesting that a new diagnostic STR test to genotype CTG18.1 expansion-mediated FECD. The application for this new test has been accepted and is currently awaiting integration into the service. This test will be in addition to the existing clinical indication, corneal dystrophy; R262, a gene panel. This new clinical indication will allow an efficient, reliable and cost-effective way to genotype patients for this repeat expansion. Having a measure in place to detect disease risk factors at an early stage is particularly important in aiding the advancement of diagnostic and therapeutic approaches for the future.

6.3 Limitations and Future work

In general, tandem repeats in the genome are challenging to genotype due to their polymorphic nature, somatic mosaicism and their amplification can be hindered by bi-allelic skewing. In the first instance, I used traditional methods of genotyping, including an STR and TP-PCR assay to interrogate CTG18.1, and although they are cost effective and relatively high-throughput methods, they have the limitation of only providing crude estimates and mode allele

lengths and therefore did not provide a true reflection of the distribution of allele lengths in this cohort. A further disadvantage to using these methods is the maximum repeat size that can be detected by STR analysis is approximately 120 repeats and although confirmation of larger expansions can be detected using TP-PCR it does not size the largest allele (Warner et al., 1996). I later applied a targeted Illumina MiSeq ultra deep sequencing method to CTG18.1 expansion positive samples. This method is advantageous over traditional STR and TP-PCR methods as it provides sequence level resolution and the ability to quantify levels of somatic instability and the presence and/or absence of variants within the repeat. However, this method comes with its own limitations, namely the length of repeats it can sequence efficiently is similar to the STR assay (approximately 120 repeats).

The recent advancement of long-read sequencing technologies may overcome this challenge given long reads are able to fully encompass and sequence across expanded repeat tracts. A CRISPR-guided non-amplification dependent approach has been used to interrogate a small number of CTG18.1 expansion positive DNA samples and has illustrated that a thousand repeats within blood-derived DNA can be detected (Hafford-Tear et al., 2019; Wieben, Aleff, et al., 2019). However, this method is time consuming, costly and requires large quantities of DNA (minimum of 5ug DNA per sample). More recently, PacBio long read sequencing of mRNA has been performed on RNA extracted from CECs for a small number of FECD patients. This approach demonstrated CTG18.1 expansions can be up to 20 times longer than measured in blood-derived DNA (Wieben et al., 2021). Ultimately, further analysis using corneal endothelial cell–derived DNA is required to provide greater insights into the behaviour of the CTG181.1 expansion in CECs, to potentially explain the tissue-

specific nature of FECD. This would be exceptionally beneficial as the need for alternative FECD treatments is high and there is much interest in developing gene-directed treatment strategies which will rely on accurate genotyping.

As these technologies advance and become more accessible it will provide increasingly accurate genotyping and strengthen our understanding of CTG18.1-mediated FECD pathophysiology; until then, I have utilised the best approaches currently available and provided new insights into the dynamics of the CTG18.1 repeat, regardless of the caveats.

Exome sequencing was employed to explore genetic causes underlying CTG18.1 expansion-negative FECD. A limiting factor of exome sequencing is that only protein coding regions of the genome are sequenced and thus only a small proportion of the genome is interrogated. However, despite covering less than 2% of the entire genome, approximately 85% of Mendelian disease gene mutations are estimated to occur in protein coding regions, therefore WES can be an excellent genetic technique for investigating monogenic diseases, including FECD (Rabbani et al., 2014). In this study I was only able to confirm approximately 1.5% of subjects with causative variants. Therefore, it is a strong possibility that mutations may be located within the non-coding portion region of the genome. In future, WGS could be used to explore the non-coding portion of these individuals' genomes to overcome this study limitation.

A further major challenge to exome sequencing is lack of efficient prioritisation of the vast number of variants identified in each proband. In this study I used the standard approach of filtering by frequency. Any variant with a high frequency in the control population could not be responsible for this rare subset of a disorder affecting approximately 5% of the population, given that we

know approximately 80% of FECD is caused by the CTG18.1 expansion. Where possible I applied a family-based filtering strategy, with the aim of producing a candidate gene list. In addition, RNA-seq data generated from healthy CECs was used to determine the expression levels of a gene in the corneal endothelium to further refine potential candidate variants (Chen et al., 2013). Due to the late-onset nature of FECD, affected first degree relatives are often difficult to recruit as the proband's parents are likely deceased and offspring are not yet knowingly affected, therefore familial cases are always limited. In attempts to overcome this, I applied a gene burden approach, in which I found the miR-184 +58G>A variant. Further work is required to understand how this variant may lead to the pathogenicity of FECD and also how it results in a phenotype that differs from EDICT syndrome, caused by the adjacent miR-184 +57C>T variant.

Additionally since the most common cause of FECD has been associated to the non-coding CTG18.1 repeat expansion, there is also the possibility other non-coding repeat expansions, may be present in the genome resulting in FECD; similar to the findings of the CCTG repeat in *ZNF9* in DM2 discovered after the identification of DMPK CTG expansion causative for DM1 (Liquori et al., 2001). There are now new computational methodologies, such as Expansion Hunter *De novo*, which can be applied to WGS for genome-wide repeat expansion detection in order to explore this hypothesis (Dolzhenko et al., 2020) and large-scale application of long-read sequencing approaches also offer the potential to identify novel genetic causes of disease.

6.4 Concluding remarks

To conclude, the data presented in this thesis has advanced our understanding of the genetic mechanisms underlying FECD and has provided

new areas of interest for further study. Furthermore, it is anticipated that insights gained shed light upon risk prediction factors and will support the development of novel therapeutic strategies to one day treat and/or prevent this sight threatening disease.

References

Afshari, N. A., Igo, R. P., Morris, N. J., Stambolian, D., Sharma, S., Pulagam, V. L., … Iyengar, S. K. (2017). Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14898

Afshari, N. A., Pittard, A. B., Siddiqui, A., & Klintworth, G. K. (2006). Clinical study of Fuchs corneal endothelial dystrophy leading to penetrating keratoplasty: a 30-year experience. *Archives of Ophthalmology (Chicago, Ill. : 1960)*, *124*(6), 777–780. https://doi.org/10.1001/archopht.124.6.777

Agoldberg, R., Raza, S., Walford, E., Feuer, W. J., & Lgoldberg, J. (2014). Fuchs endothelial corneal dystrophy: Clinical characteristics of surgical and nonsurgical patients. *Clinical Ophthalmology*, *8*, 1761–1766. https://doi.org/10.2147/OPTH.S68217

Aldave, A. J., Rayner, S. A., Salem, A. K., Yoo, G. L., Kim, B. T., Saeedian, M., … Yellore, V. S. (2006). No pathogenic mutations identified in the COL8A1 and COL8A2 genes in familial Fuchs corneal dystrophy. *Investigative Ophthalmology & Visual Science*, *47*(9), 3787–3790. https://doi.org/10.1167/IOVS.05-1635

Alkhateeb, M. A. (2018). *Sequence level genotyping at TCF4 CTG repeats associated with late onset Fuchs endothelial corneal dystrophy Mariam Abdulaziz Alkhateeb*. University of Glasgow.

Alonso, M. E., Yescas, P., Rasmussen, A., Ochoa, A., Macías, R., Ruiz, I., & Suástegui, R. (2002). Homozygosity in Hungtington's disease: New ethical dilemma caused by molecular diagnosis. *Clinical Genetics*, *61*(6), 437–442. https://doi.org/10.1034/j.1399-0004.2002.610607.x

American Academy of Ophthalmology. (n.d.). Labeeled anatomy. Retrieved January 25, 2023, from https://www.aao.org/image/anatomy-color-labeled-2

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), 1–12. https://doi.org/10.1186/GB-2010-11-10-R106/COMMENTS

Andrew, S. E., Goldberg, Y. P., Kremer, B., Telenius, H., Theilmann, J., Adam, S., … Kalchman, M. A. (1993). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nature Genetics*, *4*(4), 398–403. https://doi.org/10.1038/ng0893-398

Angelbello, A. J., Benhamou, R. I., Rzuczek, S. G., Choudhary, S., Tang, Z., Chen, J. L., … Disney, M. D. (2021). A small molecule that binds an RNA repeat expansion stimulates its decay via the exosome complex. *Cell Chemical Biology*, *28*(1), 34. https://doi.org/10.1016/J.CHEMBIOL.2020.10.007

Anvret, M., Ahlberg, G., Grandell, U., Hedberg, B., Johnson, K., & Edström, L. (1993). Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. *Human Molecular Genetics*, *2*(9), 1397–1400. https://doi.org/10.1093/hmg/2.9.1397

Ash, P. E. A., Bieniek, K. F., Gendron, T. F., Caulfield, T., Lin, W. L., DeJesus-Hernandez, M., … Petrucelli, L. (2013). Unconventional Translation of C9ORF72 GGGGCC Expansion Generates Insoluble Polypeptides Specific to c9FTD/ALS. *Neuron*, *77*(4), 639–646. https://doi.org/10.1016/j.neuron.2013.02.004

Ashizawa, T., Dubel, J. R., & Harati, Y. (1993). Somatic instability of ctg repeat

in myotonic dystrophy. *Neurology*, *43*(12), 2674–2678. https://doi.org/10.1212/wnl.43.12.2674

Azizi, B., Ziaei, A., Fuchsluger, T., Schmedt, T., Chen, Y., & Jurkunas, U. V. (2011). p53-regulated increase in oxidative-stress--induced apoptosis in Fuchs endothelial corneal dystrophy: a native tissue model. *Investigative Ophthalmology & Visual Science*, *52*(13), 9291–9297. https://doi.org/10.1167/IOVS.11-8312

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., & Shendure, J. (2011, November). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, Vol. 12, pp. 745–755. https://doi.org/10.1038/nrg3031

Bañez-Coronel, M., Ayhan, F., Tarabochia, A. D., Zu, T., Perez, B. A., Tusi, S. K., … Ranum, L. P. W. (2015). RAN Translation in Huntington Disease. *Neuron*, *88*(4), 667–677. https://doi.org/10.1016/j.neuron.2015.10.038

Baratz, K. H., Tosakulwong, N., Ryu, E., Brown, W. L., Branham, K., Chen, W., … Edwards, A. O. (2010). E2-2 protein and Fuchs's corneal dystrophy. *The New England Journal of Medicine*, *363*(11), 1016–1024. https://doi.org/10.1056/NEJMoa1007064

Belzil, V. V., Bauer, P. O., Prudencio, M., Gendron, T. F., Stetler, C. T., Yan, I. K., … Petrucelli, L. (2013). Reduced C9orf72 gene expression in c9FTD/ALS is caused by histone trimethylation, an epigenetic event detectable in blood. *Acta Neuropathologica*, *126*(6), 895–905. https://doi.org/10.1007/s00401-013-1199-1

Bettencourt, C., Hensman-Moss, D., Flower, M., Wiethoff, S., Brice, A., Goizet, C., … Jones, L. (2016). DNA repair pathways underlie a common genetic

mechanism modulating onset in polyglutamine diseases. *Annals of Neurology*, *79*(6), 983. https://doi.org/10.1002/ANA.24656

Bhattacharyya, N., Hafford-Tear, N. J., Sadan, A. N., Szabo, A., Chai, N., Zarouchlioti, C., … Davidson, A. E. (2023). Deciphering novel TCF4-driven molecular origins and mechanisms underlying a common triplet repeat expansion-mediated disease. *BioRxiv*, 2023.03.29.534731. https://doi.org/10.1101/2023.03.29.534731

Biswas, S. (2001). Missense mutations in COL8A2, the gene encoding the alpha2 chain of type VIII collagen, cause two forms of corneal endothelial dystrophy. *Human Molecular Genetics*, *10*(21), 2415–2423. https://doi.org/10.1093/hmg/10.21.2415

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Bonanno, J. A. (2012). Molecular mechanisms underlying the corneal endothelial pump. *Experimental Eye Research*, *95*(1), 2–7. https://doi.org/10.1016/j.exer.2011.06.004

Breschel, T. S., McInnis, M. G., Margolis, R. L., Sirugo, G., Corneliussen, B., Simpson, S. G., … Ross, C. A. (1997). A novel, heritable, expanding CTG repeat in an intron of the SEF2-1 gene on chromosome 18q21.1. *Human Molecular Genetics*, *6*(11), 1855–1863. https://doi.org/10.1093/hmg/6.11.1855

Center, M. E. (n.d.). Corneal Transplants — Moyes Eye Center. Retrieved January 25, 2023, from https://www.moyeseye.com/corneal-transplants

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/S13742-015-0047-8/2707533

Chen, Y., Huang, K., Nakatsu, M. N., Xue, Z., Deng, S. X., & Fan, G. (2013). Identification of novel molecular markers through transcriptomic analysis in human fetal and adult corneal endothelial cells. *Human Molecular Genetics*, *22*(7), 1271–1279. https://doi.org/10.1093/hmg/dds527

Chintalaphani, S. R., Pineda, S. S., Deveson, I. W., & Kumar, K. R. (2021). An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathologica Communications*, *9*(1). https://doi.org/10.1186/S40478-021-01201-X

Chung, D. W. D., Frausto, R. F., Ann, L. B., Jang, M. S., & Aldave, A. J. (2014). Functional impact of ZEB1 mutations associated with posterior polymorphous and fuchs' endothelial corneal dystrophies. *Investigative Ophthalmology and Visual Science*, *55*(10), 6159–6166. https://doi.org/10.1167/iovs.14-15247

Ciosi, M., Cumming, S. A., Chatzi, A., Larson, E., Tottey, W., Lomeikaite, V., … Monckton, D. G. (2021). Approaches to Sequence the HTT CAG Repeat Expansion and Quantify Repeat Length Variation. *Journal of Huntington's Disease*, *10*, 53–74. https://doi.org/10.3233/JHD-200433

Ciosi, M., Cumming, S. A., Mubarak, A., Symeonidi, E., Herzyk, P., McGuinness, D., … Monckton, D. G. (2018). Library preparation and MiSeq sequencing for the genotyping-by-sequencing of the Huntington disease

HTT exon one trinucleotide repeat and the quantification of somatic mosaicism. *Protocol Exchange*. https://doi.org/10.1038/protex.2018.089

Ciosi, M., Maxwell, A., Cumming, S. A., Hensman Moss, D. J., Alshammari, A. M., Flower, M. D., … Monckton, D. G. (2019). A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. *EBioMedicine*, *48*, 568–580. https://doi.org/10.1016/j.ebiom.2019.09.020

Cirulli, E. T., & Goldstein, D. B. (2010, June). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, Vol. 11, pp. 415–425. https://doi.org/10.1038/nrg2779

Consortium, G. M. of H. D. (GeM-H. (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell*, *178*(4), 887-900.e14. https://doi.org/10.1016/j.cell.2019.06.036

Corrales, E., Vásquez, M., Zhang, B., Santamaría-Ulloa, C., Cuenca, P., Krahe, R., … Morales, F. (2019). Analysis of mutational dynamics at the DMPK (CTG)n locus identifies saliva as a suitable DNA sample source for genetic analysis in myotonic dystrophy type 1. *PloS One*, *14*(5), e0216407. https://doi.org/10.1371/journal.pone.0216407

Cross, H. E., Maumenee, A. E., & Cantolino, S. J. (1971). Inheritance of Fuchs' endothelial dystrophy. *Archives of Ophthalmology (Chicago, Ill. : 1960)*, *85*(3), 268–272. https://doi.org/10.1001/archopht.1971.00990050270002

Cumming, S. A., Hamilton, M. J., Robb, Y., Gregory, H., McWilliam, C., Cooper, A., … Monckton, D. G. (2018). De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1. *European Journal of Human Genetics*, *26*(11),

1635. https://doi.org/10.1038/S41431-018-0156-9

Cumming, S. A., Jimenez-Moreno, C., Okkersen, K., Wenninger, S., Daidj, F., Hogarth, F., … Monckton, D. G. (2019). Genetic determinants of disease severity in the myotonic dystrophy type 1 OPTIMISTIC cohort. *Neurology*, *93*(10), e995. https://doi.org/10.1212/WNL.0000000000008056

Davidson, A. E., Liskova, P., Evans, C. J., Dudakova, L., Nosková, L., Pontikos, N., … Hardcastle, A. J. (2016). Autosomal-Dominant Corneal Endothelial Dystrophies CHED1 and PPCD1 Are Allelic Disorders Caused by Non-coding Mutations in the Promoter of OVOL2. *American Journal of Human Genetics*, *98*(1), 75–89. https://doi.org/10.1016/j.ajhg.2015.11.018

Dean, N. L., Tan, S. L., & Ao, A. (2006). Instability in the transmission of the myotonic dystrophy CTG repeat in human oocytes and preimplantation embryos. *Fertility and Sterility*, *86*(1), 98–105. https://doi.org/10.1016/j.fertnstert.2005.12.025

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., … Rademakers, R. (2011). Expanded GGGGCC hexanucleotide repeat in non-coding region of C9ORF72 causes chromosome 9p-linked frontotemporal dementia and amyotrophic lateral sclerosis. *Neuron*, *72*(2), 245. https://doi.org/10.1016/J.NEURON.2011.09.011

DelMonte, D. W., & Kim, T. (2011). Anatomy and physiology of the cornea. *Journal of Cataract and Refractive Surgery*, *37*(3), 588–598. https://doi.org/10.1016/j.jcrs.2010.12.037

Desronvil, T., Logan-Wyatt, D., Abdrabou, W., Triana, M., Jones, R., Taheri, S., … Wiggs, J. L. (2010). Distribution of COL8A2 and COL8A1 gene variants

in Caucasian primary open angle glaucoma patients with thin central corneal thickness. *Molecular Vision*, *16*, 2185–2191.

Doggart, J. H. (1957). Fuchs's epitheltal dystrophy of the cornea. *British Journal of Ophthalmology*, *41*(9), 533–540. https://doi.org/10.1136/bjo.41.9.533

Dolzhenko, E., Bennett, M. F., Richmond, P. A., Trost, B., Chen, S., Van Vugt, J. J. F. A., … Eberle, M. A. (2020). ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biology*, *21*(1). https://doi.org/10.1186/S13059-020-02017-Z

Doyu, M., Sobue, G., Mukai, E., Kachi, T., Yasuda, T., Mitsuma, T., & Takahashi, A. (1992). Severity of X-linked recessive bulbospinal neuronopathy correlates with size of the tandem CAG repeat in androgen receptor gene. *Annals of Neurology*, *32*(5), 707–710. https://doi.org/10.1002/ana.410320517

Du, J., Aleff, R. A., Soragni, E., Kalari, K., Nie, J., Tang, X., … Wieben, E. D. (2015). RNA toxicity and missplicing in the common eye disease fuchs endothelial corneal dystrophy. *The Journal of Biological Chemistry*, *290*(10), 5979–5990. https://doi.org/10.1074/jbc.M114.621607

Eghrari, A. O., Riazuddin, S. A., & Gottsch, J. D. (2015). Overview of the Cornea: Structure, Function, and Development. *Progress in Molecular Biology and Translational Science*, *134*, 7–23. https://doi.org/10.1016/bs.pmbts.2015.04.001

Eghrari, A. O., Vahedi, S., Afshari, N. A., Riazuddin, S. A., & Gottsch, J. D. (2017). CTG18.1 Expansion in TCF4 Among African Americans With Fuchs' Corneal Dystrophy. *Investigative Ophthalmology & Visual Science*,

58(14), 6046–6049. https://doi.org/10.1167/iovs.17-21661

Eghrari, A. O., Vasanth, S., Gapsis, B. C., Bison, H., Jurkunas, U., Riazuddin, S. A., & Gottsch, J. D. (2018). Identification of a Novel TCF4 Isoform in the Human Corneal Endothelium. *Cornea*, *37*(7), 899–903. https://doi.org/10.1097/ICO.0000000000001521

Evans, C., Hardin, J., & Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, *19*(5), 776. https://doi.org/10.1093/BIB/BBX008

Fautsch, M. P., Wieben, E. D., Baratz, K. H., Bhattacharyya, N., Sadan, A. N., Hafford-Tear, N. J., … Davidson, A. E. (2021). TCF4-mediated Fuchs endothelial corneal dystrophy: Insights into a common trinucleotide repeat-associated disease. *Progress in Retinal and Eye Research*, *81*. https://doi.org/10.1016/J.PRETEYERES.2020.100883

Feizi, S. (2018). Corneal endothelial cell dysfunction: etiologies and management. *Therapeutic Advances in Ophthalmology*, *10*, 251584141881580. https://doi.org/10.1177/2515841418815802

Figueroa, K. P., Coon, H., Santos, N., Velazquez, L., Mederos, L. A., & Pulst, S.-M. (2017). Genetic analysis of age at onset variation in spinocerebellar ataxia type 2. *Neurology. Genetics*, *3*(3), e155. https://doi.org/10.1212/NXG.0000000000000155

Flower, M., Lomeikaite, V., Ciosi, M., Cumming, S., Morales, F., Lo, K., … Merkies, I. (2019). MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. *Brain : A Journal of Neurology*, *142*(7), 1876–1886. https://doi.org/10.1093/BRAIN/AWZ115

Foja, S., Luther, M., Hoffmann, K., Rupprecht, A., & Gruenauer-Kloevekorn, C. (2017). CTG18.1 repeat expansion may reduce TCF4 gene expression in corneal endothelial cells of German patients with Fuchs' dystrophy. *Graefe's Archive for Clinical and Experimental Ophthalmology*, *255*(8), 1621–1631. https://doi.org/10.1007/s00417-017-3697-7

Forrest, M. P., Hill, M. J., Quantock, A. J., Martin-Rendon, E., & Blake, D. J. (2014). The emerging roles of TCF4 in disease and development. *Trends in Molecular Medicine*, *20*(6), 322–331. https://doi.org/10.1016/J.MOLMED.2014.01.010

Frausto, R. F., Wang, C., & Aldave, A. J. (2014). Transcriptome Analysis of the Human Corneal Endothelium. *Investigative Ophthalmology & Visual Science*, *55*(12), 7821. https://doi.org/10.1167/IOVS.14-15021

Friedenwald, H., & Friedenwald, J. S. (1925). EPITHELIAL DYSTROPHY OF THE CORNEA. *British Journal of Ophthalmology*, *9*(1), 14–20. https://doi.org/10.1136/bjo.9.1.14

Gee, H. Y., Zhang, F., Ashraf, S., Kohl, S., Sadowski, C. E., Vega-Warner, V., … Hildebrandt, F. (2015). KANK deficiency leads to podocyte dysfunction and nephrotic syndrome. *The Journal of Clinical Investigation*, *125*(6), 2375–2384. https://doi.org/10.1172/JCI79504

Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. (2015). Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell*, *162*(3), 516–526. https://doi.org/10.1016/j.cell.2015.07.003

Geroski, D. H., Matsuda, M., Yee, R. W., & Edelhauser, H. F. (1985). Pump Function of the Human Corneal Endothelium: Effects of Age and Cornea Guttata. *Ophthalmology*, *92*(6), 759–763. https://doi.org/10.1016/S0161-

6420(85)33973-8

Glusman, G., Caballero, J., Mauldin, D. E., Hood, L., & Roach, J. C. (2011). Kaviar: an accessible system for testing SNV novelty. *Bioinformatics (Oxford, England)*, *27*(22), 3216–3217. https://doi.org/10.1093/BIOINFORMATICS/BTR540

Goar, E. L. (1933). Dystrophy of the Corneal Endothelium (Cornea Guttata), with Report of a Histologic Examination. *Transactions of the American Ophthalmological Society*, *31*, 48. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1315418/

Gomes-Pereira, M., & Monckton, D. G. (2006). Chemical modifiers of unstable expanded simple sequence repeats: What goes up, could come down. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *598*(1–2), 15–34. https://doi.org/10.1016/J.MRFMMM.2006.01.011

Goold, R., Flower, M., Moss, D. H., Medway, C., Wood-Kaczmar, A., Andre, R., … Tabrizi, S. J. (2019). FAN1 modifies Huntington's disease progression by stabilizing the expanded HTT CAG repeat. *Human Molecular Genetics*, *28*(4), 650–661. https://doi.org/10.1093/hmg/ddy375

Gottsch, J. D., Sundin, O. H., Liu, S. H., Jun, A. S., Broman, K. W., Stark, W. J., … Magovern, M. (2005). Inheritance of a novel COL8A2 mutation defines a distinct early-onset subtype of fuchs corneal dystrophy. *Investigative Ophthalmology & Visual Science*, *46*(6), 1934–1939. https://doi.org/10.1167/iovs.04-0937

Gould, F. K. (2000). *Comparative PCR analysis of two triplet repeat loci and factors influencing their mutability*. University of Glasgow.

Greenhill, N. S., Rüger, B. M., Hasan, Q., & Davis, P. F. (2000). The α1(VIII) and α2(VIII) collagen chains form two distinct homotrimeric proteins in vivo. *Matrix Biology*, *19*(1), 19–28. https://doi.org/10.1016/S0945-053X(99)00053-0

Greiner, M. A., Terveen, D. C., Vislisel, J. M., Roos, B. R., & Fingert, J. H. (2017). Assessment of a three-generation pedigree with Fuchs endothelial corneal dystrophy with anticipation for expansion of the triplet repeat in the TCF4 gene. *Eye (London, England)*, *31*(8), 1250–1252. https://doi.org/10.1038/eye.2017.60

Guo, M. H. H., Dauber, A., Lippincott, M. F. F., Chan, Y. M., Salem, R. M. M., & Hirschhorn, J. N. N. (2016). Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders. *American Journal of Human Genetics*, *99*(3), 527. https://doi.org/10.1016/J.AJHG.2016.06.031

Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N., & Lippincott, M. F. (2018). Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *American Journal of Human Genetics*, *103*(4), 522–534. https://doi.org/10.1016/J.AJHG.2018.08.016

Hafford-Tear, N. J., Tsai, Y. C., Sadan, A. N., Sanchez-Pintado, B., Zarouchlioti, C., Maher, G. J., … Davidson, A. E. (2019). CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy–associated TCF4 triplet repeat. *Genetics in Medicine*, *21*(9), 2092–2102. https://doi.org/10.1038/s41436-019-0453-x

Han, J. Y., Jang, W., & Park, J. (2022). Intergenerational Influence of Gender and the DM1 Phenotype of the Transmitting Parent in Korean Myotonic

Dystrophy Type 1. *Genes*, *13*(8). https://doi.org/10.3390/GENES13081465

Harper, P. S., Harley, H. G., Reardon, W., & Shaw, D. J. (1992, July). Anticipation in myotonic dystrophy: New light on an old problem. *American Journal of Human Genetics*, Vol. 51, pp. 10–16.

Hu, J., Rong, Z., Gong, X., Zhou, Z., Sharma, V. K., Xing, C., … Mootha, V. V. (2018). Oligonucleotides targeting TCF4 triplet repeat expansion inhibit RNA foci and mis-splicing in Fuchs' dystrophy. *Human Molecular Genetics*, *27*(6), 1015–1026. https://doi.org/10.1093/hmg/ddy018

Hughes, A. E., Bradley, D. T., Campbell, M., Lechner, J., Dash, D. P., Simpson, D. A., & Willoughby, C. E. (2011). Mutation altering the miR-184 seed region causes familial keratoconus with cataract. *American Journal of Human Genetics*, *89*(5), 628–633. https://doi.org/10.1016/J.AJHG.2011.09.014

Iliff, B. W., Riazuddin, S. A., & Gottsch, J. D. (2012). *A Single-Base Substitution in the Seed Region of miR-184 Causes EDICT Syndrome*. https://doi.org/10.1167/iovs.11-8783

Ito, Y. A., & Walter, M. A. (2014). Genomics and anterior segment dysgenesis: A review. *Clinical and Experimental Ophthalmology*, *42*(1), 13–24. https://doi.org/10.1111/CEO.12152

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., … Farh, K. K. H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, *176*(3), 535-548.e24. https://doi.org/10.1016/J.CELL.2018.12.015

Johansson, J., Forsgren, L., Sandgren, O., Brice, A., Holmgren, G., &

Holmberg, M. (1998). Expanded CAG repeats in Swedish spinocerebellar ataxia type 7 (SCA7) patients: effect of CAG repeat length on the clinical manifestation. *Human Molecular Genetics*, *7*(2), 171–176. https://doi.org/10.1093/hmg/7.2.171

Joyce, N. C. (2003). Proliferative capacity of the corneal endothelium. *Progress in Retinal and Eye Research*, *22*(3), 359–389. https://doi.org/10.1016/S1350-9462(02)00065-4

Jun, A. S. (2010). One Hundred Years of Fuchs' Dystrophy. *Ophthalmology*, *117*(5), 859-860.e14. https://doi.org/10.1016/j.ophtha.2010.03.001

Jun, A. S., Broman, K. W., Do, D. V., Akpek, E. K., Stark, W. J., & Gottsch, J. D. (2002). Endothelial dystrophy, iris hypoplasia, congenital cataract, and stromal thinning (EDICT) syndrome maps to chromosome 15q22.1-q25.3. *American Journal of Ophthalmology*, *134*(2), 172–176. https://doi.org/10.1016/S0002-9394(02)01401-0

Jun, A. S., Meng, H., Ramanan, N., Matthaei, M., Chakravarti, S., Bonshek, R., … Kimos, M. (2012). An alpha 2 collagen VIII transgenic knock-in mouse model of Fuchs endothelial corneal dystrophy shows early endothelial cell unfolded protein response and apoptosis. *Human Molecular Genetics*, *21*(2), 384–393. https://doi.org/10.1093/hmg/ddr473

Jurkunas, U. V., Bitar, M. S., Funaki, T., & Azizi, B. (2010). Evidence of Oxidative Stress in the Pathogenesis of Fuchs Endothelial Corneal Dystrophy. *The American Journal of Pathology*, *177*(5), 2278. https://doi.org/10.2353/AJPATH.2010.100279

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., … Daly, M. J. (2020). The mutational constraint spectrum quantified

from variation in 141,456 humans. *Nature 2020 581:7809*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kennedy, L., Evans, E., Chen, C. M., Craven, L., Detloff, P. J., Ennis, M., & Shelbourne, P. F. (2003). Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Human Molecular Genetics*, *12*(24), 3359–3367. https://doi.org/10.1093/HMG/DDG352

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-Based Genome Alignment and Genotyping with HISAT2 andHISAT-genotype. *Nature Biotechnology*, *37*(8), 907. https://doi.org/10.1038/S41587-019-0201-4

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. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kitagawa, K., Kojima, M., Sasaki, H., Shui, Y. B., Chew, S. J., Cheng, H. M., … Sasaki, K. (2002). Prevalence of primary cornea guttata and morphology of corneal endothelium in aging Japanese and Singaporean subjects. *Ophthalmic Research*, *34*(3), 135–138. https://doi.org/10.1159/000063656

Koressaar, T., & Remm, M. (2007). Enhancements and modifications of primer design program Primer3. *Bioinformatics (Oxford, England)*, *23*(10), 1289–1291. https://doi.org/10.1093/bioinformatics/btm091

Krachmer, J. H., Purcell, J. J., Young, C. W., & Bucher, K. D. (1978). Corneal endothelial dystrophy. A study of 64 families. *Archives of Ophthalmology (Chicago, Ill. : 1960)*, *96*(11), 2036–2039.

https://doi.org/10.1001/archopht.1978.03910060424004

Larson, E., Fyfe, I., Morton, A. J., & Monckton, D. G. (2015). Age-, tissue- and length-dependent bidirectional somatic CAG•CTG repeat instability in an allelic series of R6/2 Huntington disease mice. *Neurobiology of Disease*, *76*, 98–111. https://doi.org/10.1016/J.NBD.2015.01.004

Lee, J. M., Correia, K., Loupe, J., Kim, K. H., Barker, D., Hong, E. P., … Myers, R. H. (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell*, *178*(4), 887. https://doi.org/10.1016/J.CELL.2019.06.036

Lee, J. M., Pinto, R. M., Gillis, T., St. Claire, J. C., & Wheeler, V. C. (2011). Quantification of Age-Dependent Somatic CAG Repeat Instability in Hdh CAG Knock-In Mice Reveals Different Expansion Dynamics in Striatum and Liver. *PLoS ONE*, *6*(8). https://doi.org/10.1371/JOURNAL.PONE.0023647

Lee, Y. B., Chen, H. J., Peres, J. N., Gomez-Deza, J., Attig, J., Štalekar, M., … Shaw, C. E. (2013). Hexanucleotide repeats in ALS/FTD form length-dependent RNA Foci, sequester RNA binding proteins, and are neurotoxic. *Cell Reports*, *5*(5), 1178–1186. https://doi.org/10.1016/j.celrep.2013.10.049

Levy, S. G., Moss, J., Sawada, H., Dopping-Hepenstal, P. J. C., & McCartney, A. C. E. (1996). The composition of wide-spaced collagen in normal and diseased Descemet's membrane. *Current Eye Research*, *15*(1), 45–52. https://doi.org/10.3109/02713689609017610

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, *120*(1), 15–20. https://doi.org/10.1016/j.cell.2004.12.035

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930. https://doi.org/10.1093/BIOINFORMATICS/BTT656

Lin, Z. N., Chen, J., & Cui, H. P. (2016, June 18). Characteristics of corneal dystrophies: A review from clinical, histological and genetic perspectives. *International Journal of Ophthalmology*, Vol. 9, pp. 904–913. https://doi.org/10.18240/ijo.2016.06.20

Liquori, C. L., Ricker, K., Moseley, M. L., Jacobsen, J. F., Kress, W., Naylor, S. L., … Ranum, L. P. W. (2001). Myotonic dystrophy type 2 caused by a CCTG expansion in intron I of ZNF9. *Science*, *293*(5531), 864–867. https://doi.org/10.1126/science.1062125

Lisch, W., & Weiss, J. S. (2019, September 1). Clinical and genetic update of corneal dystrophies. *Experimental Eye Research*, Vol. 186. https://doi.org/10.1016/j.exer.2019.107715

Liskova, P., Prescott, Q., Bhattacharya, S. S., & Tuft, S. J. (2007, December). British family with early-onset Fuchs' endothelial corneal dystrophy associated with p.L450W mutation in the COL8A2 gene [8]. *British Journal of Ophthalmology*, Vol. 91, pp. 1717–1718. https://doi.org/10.1136/bjo.2007.115154

Liskova, Petra, Dudakova, L., Evans, C. J., Rojas Lopez, K. E., Pontikos, N., Athanasiou, D., … Hardcastle, A. J. (2018). Ectopic GRHL2 Expression Due to Non-coding Mutations Promotes Cell State Transition and Causes Posterior Polymorphous Corneal Dystrophy 4. *American Journal of Human Genetics*, *102*(3), 447–459. https://doi.org/10.1016/j.ajhg.2018.02.002

Liskova, Petra, Evans, C. J., Davidson, A. E., Zaliova, M., Dudakova, L., Trkova, M., … Hardcastle, A. J. (2016). Heterozygous deletions at the ZEB1 locus verify haploinsufficiency as the mechanism of disease for posterior polymorphous corneal dystrophy type 3. *European Journal of Human Genetics*, *24*(7), 985–991. https://doi.org/10.1038/ejhg.2015.232

Liu, J., Zhou, F., Guan, Y., Meng, F., Zhao, Z., Su, Q., … Wang, X. (2022). The Biogenesis of miRNAs and Their Role in the Development of Amyotrophic Lateral Sclerosis. *Cells 2022, Vol. 11, Page 572*, *11*(3), 572. https://doi.org/10.3390/CELLS11030572

Liu, S., Sadan, A. N., Muthusamy, K., Zarouchlioti, C., Jedlickova, J., Pontikos, N., … Liskova, P. (2023). Phenotype and genotype of concurrent keratoconus and Fuchs endothelial corneal dystrophy. *Acta Ophthalmologica*. https://doi.org/10.1111/AOS.15654

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The Hallmarks of Aging. *Cell*, *153*(6), 1194. https://doi.org/10.1016/J.CELL.2013.05.039

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9

Luther, M., Grünauer-Kloevekorn, C., Weidle, E., Passarge, E., Rupprecht, A., Hoffmann, K., & Foja, S. (2016). [TGC Repeats in Intron 2 of the TCF4 Gene have a Good Predictive Power Regarding to Fuchs Endothelial Corneal Dystrophy]. *Klinische Monatsblatter Fur Augenheilkunde*, *233*(2), 187–194. https://doi.org/10.1055/s-0035-1546138

MacFarlane, L.-A., & Murphy, P. R. (2010). MicroRNA: Biogenesis, Function

and Role in Cancer. *Current Genomics*, *11*(7), 537. https://doi.org/10.2174/138920210793175895

Machowska, M., Galka-Marciniak, P., & Kozlowski, P. (2022). *Consequences of genetic variants in miRNA genes*. https://doi.org/10.1016/j.csbj.2022.11.036

MacKay, C., Déclais, A. C., Lundin, C., Agostinho, A., Deans, A. J., MacArtney, T. J., … Rouse, J. (2010). Identification of KIAA1018/FAN1, a DNA Repair Nuclease Recruited to DNA Damage by Monoubiquitinated FANCD2. *Cell*, *142*(1), 65. https://doi.org/10.1016/J.CELL.2010.06.021

Magovern, M., Beauchamp, G. R., McTigue, J. W., Fine, B. S., & Baumiller, R. C. (1979). Inheritance of Fuchs' Combined Dystrophy. *Ophthalmology*, *86*(10), 1897–1920. https://doi.org/10.1016/S0161-6420(79)35340-4

Malhotra, D., Jung, M., Fecher-Trost, C., Lovatt, M., Peh, G. S. L., Noskov, S., … Casey, J. R. (2019). Defective cell adhesion function of solute transporter, SLC4A11, in endothelial corneal dystrophies. *Human Molecular Genetics*, (780), 1–71. https://doi.org/10.1093/hmg/ddz259

Malter, H. E., Iber, J. C., Willemsen, R., de Graaff, E., Tarleton, J. C., Leisti, J., … Oostra, B. A. (1997). Characterization of the full fragile X syndrome mutation in fetal gametes. *Nature Genetics*, *15*(2), 165–169. https://doi.org/10.1038/ng0297-165

Mankodi, A. (2001). Muscleblind localizes to nuclear foci of aberrant RNA in myotonic dystrophy types 1 and 2. *Human Molecular Genetics*, *10*(19), 2165–2170. https://doi.org/10.1093/hmg/10.19.2165

Massey, T. H., & Jones, L. (2018). The central role of DNA damage and repair in CAG repeat diseases. *Disease Models & Mechanisms*, *11*(1).

https://doi.org/10.1242/dmm.031930

Matsuura, T., Fang, P., Pearson, C. E., Jayakar, P., Ashizawa, T., Roa, B. B., & Nelson, D. L. (2006). Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: Repeat purity as a disease modifier? *American Journal of Human Genetics*, *78*(1), 125–129. https://doi.org/10.1086/498654

Matthaei, M., Hribek, A., Clahsen, T., Bachmann, B., Cursiefen, C., & Jun, A. S. (2019). Fuchs Endothelial Corneal Dystrophy: Clinical, Genetic, Pathophysiologic, and Therapeutic Aspects. *Annual Review of Vision Science*, *5*, 151–175. https://doi.org/10.1146/annurev-vision-091718-014852

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297. https://doi.org/10.1101/GR.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/S13059-016-0974-4/TABLES/8

McMurray, C. T. (2010). Expansions in simple DNA repeats underlie ~20 severe neuromuscular and neurodegenerative disorders. *Nature Publishing Group*, *11*. https://doi.org/10.1038/nrg2828

Mehta, J. S., Vithana, E. N., Tan, D. T. H., Yong, V. H. K., Yam, G. H. F., Law, R. W. K., … Aung, T. (2008). Analysis of the posterior polymorphous corneal dystrophy 3 gene, TCF8, in late-onset fuchs endothelial corneal

dystrophy. *Investigative Ophthalmology and Visual Science*, *49*(1), 184–188. https://doi.org/10.1167/iovs.07-0847

Meng, H., Matthaei, M., Ramanan, N., Grebe, R., Chakravarti, S., Speck, C. L., … Jun, A. S. (2013). L450W and Q455K Col8a2 knock-in mouse models of fuchs endothelial corneal dystrophy show distinct phenotypes and evidence for altered autophagy. *Investigative Ophthalmology and Visual Science*, *54*(3), 1887–1897. https://doi.org/10.1167/iovs.12-11021

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., … Marshall, D. (n.d.). *Using Tablet for visual exploration of second-generation sequencing data*. https://doi.org/10.1093/bib/bbs012

Minear, M. A., Li, Y. J., Rimmler, J., Balajonda, E., Watson, S., Rand Allingham, R., … Gregory, S. G. (2013). Genetic screen of African Americans with Fuchs endothelial corneal dystrophy. *Molecular Vision*, *19*, 2508. Retrieved from /pmc/articles/PMC3859630/

Mirkin, S. M. (2007, June 21). Expandable DNA repeats and human disease. *Nature*, Vol. 447, pp. 932–940. https://doi.org/10.1038/nature05977

Miyajima, T., Vasanth, S., Melangath, G., Deshpande, N., Chen, Y., zhu, shan, … Jurkunas, U. V. (2019). NQO1 downregulation generates genotoxic DNA adducts in in vitro FECD model. *Investigative Ophthalmology & Visual Science*, *60*(9), 2177–2177.

Mok, J. W., Kim, H. S., & Joo, C. K. (2009). Q455V mutation in COL8A2 is associated with Fuchs' corneal dystrophy in Korean patients. *Eye*, *23*(4), 895–903. https://doi.org/10.1038/eye.2008.116

Monckton, D. G., Wong, L. J. C., Ashizawa, T., & Caskey, C. T. (1995). Somatic

mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Human Molecular Genetics*, *4*(1), 1–8. https://doi.org/10.1093/HMG/4.1.1

Mootha, V. V., Gong, X., Ku, H. C., & Xing, C. (2014). Association and Familial Segregation of CTG18.1 Trinucleotide Repeat Expansion of TCF4 Gene in Fuchs' Endothelial Corneal Dystrophy. *Investigative Ophthalmology & Visual Science*, *55*(1), 33. https://doi.org/10.1167/IOVS.13-12611

Morales, F., Couto, J. M., Higham, C. F., Hogg, G., Cuenca, P., Braida, C., … Monckton, D. G. (2012). Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Human Molecular Genetics*, *21*(16), 3558–3567. https://doi.org/10.1093/HMG/DDS185

Morales, F., Vásquez, M., Santamaría, C., Cuenca, P., Corrales, E., & Monckton, D. G. (2016). A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA Repair*, *40*, 57–66. https://doi.org/10.1016/J.DNAREP.2016.01.001

Musova, Z., Mazanec, R., Krepelova, A., Ehler, E., Vales, J., Jaklova, R., … Sedlacek, Z. (2009). Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. *American Journal of Medical Genetics. Part A*, *149A*(7), 1365–1374. https://doi.org/10.1002/AJMG.A.32987

Nakano, M., Okumura, N., Nakagawa, H., Koizumi, N., Ikeda, Y., Ueno, M., … Baratz, K. H. (2015). Trinucleotide Repeat Expansion in the TCF4 Gene in Fuchs' Endothelial Corneal Dystrophy in Japanese. *Investigative*

Ophthalmology & Visual Science, *56*(8), 4865–4869. https://doi.org/10.1167/iovs.15-17082

Nanda, G. G., Padhy, B., Samal, S., Das, S., & Alone, D. P. (2014). Genetic association of TCF4 intronic polymorphisms, CTG18.1 and rs17089887, with Fuchs' endothelial corneal dystrophy in an Indian population. *Investigative Ophthalmology & Visual Science*, *55*(11), 7674–7680. https://doi.org/10.1167/iovs.14-15297

O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology*, *9*(AUG), 402. https://doi.org/10.3389/FENDO.2018.00402/BIBTEX

Okumura, N., Hayashi, R., Nakano, M., Tashiro, K., Yoshii, K., Aleff, R., … Koizumi, N. (2019). Association of rs613872 and Trinucleotide Repeat Expansion in the TCF4 Gene of German Patients with Fuchs Endothelial Corneal Dystrophy. *Cornea*, *38*(7), 799–805. https://doi.org/10.1097/ICO.0000000000001952

Okumura, N., Hayashi, R., Nakano, M., Yoshii, K., Tashiro, K., Sato, T., … Koizumi, N. (2019). Effect of trinucleotide repeat expansion on the expression of TCF4 mRNA in fuchs' endothelial corneal dystrophy. *Investigative Ophthalmology and Visual Science*, *60*(2), 779–786. https://doi.org/10.1167/iovs.18-25760

Okumura, N., Puangsricharern, V., Jindasak, R., Koizumi, N., Komori, Y., Ryousuke, H., … Suphapeetiporn, K. (2019). Trinucleotide repeat expansion in the transcription factor 4 (TCF4) gene in Thai patients with Fuchs endothelial corneal dystrophy. *Eye (Basingstoke)*.

https://doi.org/10.1038/s41433-019-0595-8

Ong Tone, S., Kocaba, V., Böhm, M., Wylegala, A., White, T. L., & Jurkunas, U. V. (2021). Fuchs Endothelial Corneal Dystrophy: The Vicious Cycle of FuchsPathogenesis. *Progress in Retinal and Eye Research*, *80*, 100863. https://doi.org/10.1016/J.PRETEYERES.2020.100863

Paret, C., Bourouba, M., Beer, A., Miyazaki, K., Schnölzer, M., Fiedler, S., & Zöller, M. (2005). Ly6 family member C4.4A binds laminins 1 and 5, associates with galectin-3 and supports cell migration. *International Journal of Cancer*, *115*(5), 724–733. https://doi.org/10.1002/ijc.20977

Paulson, H. (2018). Repeat expansion diseases. *Handbook of Clinical Neurology*, *147*, 105–123. https://doi.org/10.1016/B978-0-444-63233-3.00009-9

Peterson, S. M., Thompson, J. A., Ufkin, M. L., Sathyanarayana, P., Liaw, L., & Congdon, C. B. (2014). Common features of microRNA target prediction tools. *Frontiers in Genetics*, *5*(FEB). https://doi.org/10.3389/FGENE.2014.00023

Pontikos, N., Yu, J., Moghul, I., Withington, L., Blanco-Kelly, F., Vulliamy, T., … Van Heyningen, V. (2017). Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. *Bioinformatics (Oxford, England)*, *33*(15), 2421–2423. https://doi.org/10.1093/BIOINFORMATICS/BTX147

Powers, A., Rinkoski, T. A., Cheung, K., Schehr, H., Osgood, N., Livelo, C., … Fautsch, M. P. (2022). GeneTAC^{TM} small molecules reduce toxic nuclear foci and restore normal splicing in corneal endothelial cells derived from patients with Fuchs endothelial corneal dystrophy (FECD) harboring repeat

expansions in transcription factor 4 (TCF4). *Investigative Ophthalmology & Visual Science*, *63*(7), 2753 – A0242-2753 – A0242.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rabbani, B., Tekin, M., & Mahdieh, N. (2014, January). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, Vol. 59, pp. 5–15. https://doi.org/10.1038/jhg.2013.114

Ranen, N. G., Stine, O. C., Abbott, M. H., Sherr, M., Codori, A. M., Franz, M. L., … Ross, C. A. (1995). Anticipation and instability of IT-15 (CAG)(N) repeats in parent-offspring pairs with Huntington disease. *American Journal of Human Genetics*, *57*(3), 593–602.

Rao, B. S., Ansar, S., Arokiasamy, T., Sudhir, R. R., Umashankar, V., Rajagopal, R., & Soumittra, N. (2018). Analysis of candidate genes *ZEB1* and *LOXHD1* in late-onset Fuchs' endothelial corneal dystrophy in an Indian cohort. *Ophthalmic Genetics*, *39*(4), 443–449. https://doi.org/10.1080/13816810.2018.1474367

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Riazuddin, S. A., Parker, D. S., McGlumphy, E. J., Oh, E. C., Iliff, B. W., Schmedt, T., … Gottsch, J. D. (2012). Mutations in LOXHD1, a recessive-deafness locus, cause dominant late-onset Fuchs corneal dystrophy.

American Journal of Human Genetics, *90*(3), 533–539. https://doi.org/10.1016/j.ajhg.2012.01.013

Riazuddin, S. A., Vasanth, S., Katsanis, N., & Gottsch, J. D. (2013). Mutations in AGBL1 cause dominant late-onset Fuchs corneal dystrophy and alter protein-protein interaction with TCF4. *American Journal of Human Genetics*, *93*(4), 758–764. https://doi.org/10.1016/j.ajhg.2013.08.010

Riazuddin, S. A., Zaghloul, N. A., Al-Saif, A., Davey, L., Diplas, B. H., Meadows, D. N., … Katsanis, N. (2010). Missense Mutations in TCF8 Cause Late-Onset Fuchs Corneal Dystrophy and Interact with FCD4 on Chromosome 9p. *American Journal of Human Genetics*, *86*(1), 45–53. https://doi.org/10.1016/j.ajhg.2009.12.001

Rifé, M., Badenas, C., Quintó, L., Puigoriol, E., Tazón, B., Rodriguez-Revenga, L., … Milà, M. (2004). Analysis of CGG variation through 642 meioses in Fragile X families. *Molecular Human Reproduction*, *10*(10), 773–776. https://doi.org/10.1093/molehr/gah102

Saade, J. S., Xing, C., Gong, X., Zhou, Z., & Mootha, V. V. (2018). Instability of TCF4 Triplet Repeat Expansion With Parent-Child Transmission in Fuchs' Endothelial Corneal Dystrophy. *Investigative Ophthalmology & Visual Science*, *59*(10), 4065–4070. https://doi.org/10.1167/iovs.18-24119

Santoro, M., Masciullo, M., Pietrobono, R., Conte, G., Modoni, A., Bianchi, M. L. E., … Silvestri, G. (2013). Molecular, clinical, and muscle studies in myotonic dystrophy type 1 (DM1) associated with novel variant CCG expansions. *Journal of Neurology*, *260*(5), 1245–1257. https://doi.org/10.1007/S00415-012-6779-9

Schmid, E., Lisch, W., Philipp, W., Lechner, S., Göttinger, W., Schlötzer-

Schrehardt, U., … Janecke, A. R. (2006). A new, X-linked endothelial corneal dystrophy. *American Journal of Ophthalmology*, *141*(3). https://doi.org/10.1016/J.AJO.2005.10.020

Schmidt, M. H. M., & Pearson, C. E. (2016). Disease-associated repeat instability and mismatch repair. *DNA Repair*, *38*, 117–126. https://doi.org/10.1016/J.DNAREP.2015.11.008

Semagn, K., Babu, R., Hearne, S., & Olsen, M. (2014). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement. *Molecular Breeding*, *33*(1), 1–14. https://doi.org/10.1007/S11032-013-9917-X/FIGURES/4

Sepp, M., Kannike, K., Eesmaa, A., Urb, M., & Timmusk, T. (2011). Functional Diversity of Human Basic Helix-Loop-Helix Transcription Factor TCF4 Isoforms Generated by Alternative 5′ Exon Usage and Splicing. *PLoS ONE*, *6*(7). https://doi.org/10.1371/JOURNAL.PONE.0022138

Sirp, A., Leite, K., Tuvikene, J., Nurm, K., Sepp, M., & Timmusk, T. (2020). The Fuchs corneal dystrophy-associated CTG repeat expansion in the TCF4 gene affects transcription from its alternative promoters. *Scientific Reports*, *10*(1). https://doi.org/10.1038/S41598-020-75437-3

Sirp, A., Roots, K., Nurm, K., Tuvikene, J., Sepp, M., & Timmusk, T. (2021). Functional consequences of TCF4 missense substitutions associated with Pitt-Hopkins syndrome, mild intellectual disability, and schizophrenia. *The Journal of Biological Chemistry*, *297*(6). https://doi.org/10.1016/J.JBC.2021.101381

Skorodumova, L. O., Belodedova, A. V, Antonova, O. P., Sharova, E. I.,

Akopian, T. A., Selezneva, O. V, … Malyugin, B. E. (2018). CTG18.1 Expansion is the Best Classifier of Late-Onset Fuchs' Corneal Dystrophy Among 10 Biomarkers in a Cohort From the European Part of Russia. *Investigative Ophthalmology & Visual Science*, *59*(11), 4748–4754. https://doi.org/10.1167/iovs.18-24590

Sobczak, K., & Krzyzosiak, W. J. (2004). Patterns of CAG repeat interruptions in SCA1 and SCA2 genes in relation to repeat instability. *Human Mutation*, *24*(3), 236–247. https://doi.org/10.1002/humu.20075

Soliman, A. Z., Xing, C., Radwan, S. H., Gong, X., & Mootha, V. V. (2015). Correlation of Severity of Fuchs Endothelial Corneal Dystrophy With Triplet Repeat Expansion in TCF4. *JAMA Ophthalmology*, *133*(12), 1386–1391. https://doi.org/10.1001/jamaophthalmol.2015.3430

Soragni, E., Petrosyan, L., Rinkoski, T. A., Wieben, E. D., Baratz, K. H., Fautsch, M. P., & Gottesfeld, J. M. (2018). Repeat-Associated Non-ATG (RAN) Translation in Fuchs' Endothelial Corneal Dystrophy. *Investigative Ophthalmology & Visual Science*, *59*(5), 1888–1896. https://doi.org/10.1167/iovs.17-23265

Squitieri, F., Gellera, C., Cannella, M., Mariotti, C., Cislaghi, G., Rubinsztein, D. C., … Donato, S. Di. (2003). Homozygosity for CAG mutation in Huntington disease is associated with a more severe clinical course. *Brain : A Journal of Neurology*, *126*(Pt 4), 946–955. https://doi.org/10.1093/brain/awg077

Sridhar, M. S. (2018, February 1). Anatomy of cornea and ocular surface. *Indian Journal of Ophthalmology*, Vol. 66, pp. 190–194. https://doi.org/10.4103/ijo.IJO_646_17

Stolle, C. A., Frackelton, E. C., McCallum, J., Farmer, J. M., Tsou, A., Wilson,

R. B., & Lynch, D. R. (2008). Novel, complex interruptions of the GAA repeat in small, expanded alleles of two affected siblings with late-onset Friedreich ataxia. *Movement Disorders*, *23*(9), 1303–1306. https://doi.org/10.1002/mds.22012

Sundin, O. H., Broman, K. W., Chang, H. H., Vito, E. C. L., Stark, W. J., & Gottsch, J. D. (2006). A common locus for late-onset Fuchs corneal dystrophy maps to 18q21.2-q21.32. *Investigative Ophthalmology & Visual Science*, *47*(9), 3919–3926. https://doi.org/10.1167/iovs.05-1619

Sutherland, G. R., Kremer, E., Lynch, M., Pritchard, M., Yu, S., Richards, R. I., & Haan, E. A. (1991). Hereditary unstable DNA: a new explanation for some old genetic questions? *The Lancet*, *338*(8762), 289–292. https://doi.org/10.1016/0140-6736(91)90426-P

Suzuki, T., Kinoshita, Y., Tachibana, M., Matsushima, Y., Kobayashi, Y., Adachi, W., … Kinoshita, S. (2001). Expression of sex steroid hormone receptors in human cornea. *Current Eye Research*, *22*(1), 28–33. https://doi.org/10.1076/CEYR.22.1.28.6980

Sznajder, Ł. J., Thomas, J. D., Carrell, E. M., Reid, T., McFarland, K. N., Cleary, J. D., … Swanson, M. S. (2018). Intron retention induced by microsatellite expansions as a disease biomarker. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), 4234–4239. https://doi.org/10.1073/pnas.1716617115

Taneja, K. L., McCurrach, M., Schalling, M., Housman, D., & Singer, R. H. (1995). Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues. *Journal of Cell Biology*, *128*(6), 995–1002. https://doi.org/10.1083/jcb.128.6.995

Tang, H., Zhang, W., Yan, X. M., Wang, L. P., Dong, H., Shou, T., … Guo, Q. (2016). Analysis of SLC4A11, ZEB1, LOXHD1, COL8A2 and TCF4 gene sequences in a multi-generational family with late-onset Fuchs corneal dystrophy. *International Journal of Molecular Medicine*, *37*(6), 1487–1500. https://doi.org/10.3892/ijmm.2016.2570

Tassone, F., Iwahashi, C., & Hagerman, P. J. (2004). FMR1 RNA within the intranuclear inclusions of fragile X-associated tremor/ataxia syndrome (FXTAS). *RNA Biology*, *1*(2), 103–105. https://doi.org/10.4161/rna.1.2.1035

Telenius, H., Kremer, B., Goldberg, Y. P., Theilmann, J., Andrew, S. E., Zeisler, J., … Hayden, M. R. (1994). Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. *Nature Genetics 1994 6:4*, *6*(4), 409–414. https://doi.org/10.1038/ng0494-409

Trang, H., Stanley, S. Y., Thorner, P., Faghfoury, H., Schulze, A., Hawkins, C., … Yoon, G. (2015). Massive CAG repeat expansion and somatic instability in maternally transmitted infantile spinocerebellar ataxia type 7. *JAMA Neurology*, *72*(2), 219–223. https://doi.org/10.1001/JAMANEUROL.2014.1902

Tseng-Rogenski, S. S., Munakata, K., Choi, D. Y., Martin, P. K., Mehta, S., Koi, M., … Carethers, J. M. (2020). *The Human DNA Mismatch Repair Protein MSH3 Contains Nuclear Localization and Export Signals That Enable Nuclear-Cytosolic Shuttling in Response to Inflammation*. https://doi.org/10.1128/MCB.00029-20

Tuft, S. J., & Coster, D. J. (1990). The corneal endothelium. *Eye (Basingstoke)*, *4*(3), 389–424. https://doi.org/10.1038/eye.1990.53

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm,

M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, *40*(15). https://doi.org/10.1093/nar/gks596

Van den Bogerd, B., Dhubhghaill, S. N., Koppen, C., Tassignon, M. J., & Zakaria, N. (2018). A review of the evidence for in vivo corneal endothelial regeneration. *Survey of Ophthalmology*, *63*(2), 149–165. https://doi.org/10.1016/J.SURVOPHTHAL.2017.07.004

Vasanth, S., Eghrari, A. O., Gapsis, B. C., Wang, J., Haller, N. F., Stark, W. J., … Gottsch, J. D. (2015). Expansion of CTG18.1 trinucleotide repeat in TCF4 is a potent driver of fuchs' corneal dystrophy. *Investigative Ophthalmology and Visual Science*, *56*(8), 4531–4536. https://doi.org/10.1167/iovs.14-16122

Vedana, G., Villarreal, G., & Jun, A. S. (2016, February 18). Fuchs endothelial corneal dystrophy: Current perspectives. *Clinical Ophthalmology*, Vol. 10, pp. 321–330. https://doi.org/10.2147/OPTH.S83467

Vilas, G. L., Loganathan, S. K., Liu, J., Riau, A. K., Young, J. D., Mehta, J. S., … Casey, J. R. (2013). Transmembrane water-flux through SLC4A11: A route defective in genetic corneal diseases. *Human Molecular Genetics*, *22*(22), 4579–4590. https://doi.org/10.1093/hmg/ddt307

Vithana, E. N., Morgan, P. E., Ramprasad, V., Tan, D. T. H., Yong, V. H. K., Venkataraman, D., … Aung, T. (2008). SLC4A11 mutations in Fuchs endothelial corneal dystrophy. *Human Molecular Genetics*, *17*(5), 656–666. https://doi.org/10.1093/hmg/ddm337

Vithana, E. N., Morgan, P., Sundaresan, P., Ebenezer, N. D., Tan, D. T. H., Mohamed, M. D., … Aung, T. (2006). *Mutations in sodium-borate cotransporter SLC4A11 cause recessive congenital hereditary endothelial*

dystrophy (CHED2). https://doi.org/10.1038/ng1824

Walkow, T., Anders, N., & Klebe, S. (2000). Endothelial cell loss after phacoemulsification: Relation to preoperative and intraoperative parameters. *Journal of Cataract and Refractive Surgery*, *26*(5), 727–732. https://doi.org/10.1016/S0886-3350(99)00462-9

Warner, J. P., Barron, L. H., Goudie, D., Kelly, K., Dow, D., Fitzpatrick, D. R., & Brock, D. J. H. (1996). A general method for the detection of large GAG repeat expansions by fluorescent PCR. *Journal of Medical Genetics*, *33*(12), 1022–1026. https://doi.org/10.1136/jmg.33.12.1022

Wieben, E. D., Aleff, R. A., Basu, S., Sarangi, V., Bowman, B., McLaughlin, I. J., … Fautsch, M. P. (2019). Amplification-free long-read sequencing of TCF4 expanded trinucleotide repeats in Fuchs Endothelial Corneal Dystrophy. *PloS One*, *14*(7), e0219446. https://doi.org/10.1371/journal.pone.0219446

Wieben, E. D., Aleff, R. A., Rinkoski, T. A., Baratz, K. H., Basu, S., Patel, S. V., … Fautsch, M. P. (2021). Comparison of TCF4 repeat expansion length in corneal endothelium and leukocytes of patients with Fuchs endothelial corneal dystrophy. *PLOS ONE*, *16*(12), e0260837. https://doi.org/10.1371/JOURNAL.PONE.0260837

Wieben, E. D., Aleff, R. A., Tang, X., Kalari, K. R., Maguire, L. J., Patel, S. V., … Fautsch, M. P. (2018). Gene expression in the corneal endothelium of fuchs endothelial corneal dystrophy patients with and without expansion of a trinucleotide repeat in TCF4. *PLoS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0200005

Wieben, E. D., Aleff, R. A., Tosakulwong, N., Butz, M. L., Highsmith, W. E., Edwards, A. O., & Baratz, K. H. (2012). A common trinucleotide repeat

expansion within the transcription factor 4 (TCF4, E2-2) gene predicts Fuchs corneal dystrophy. *PloS One*, *7*(11), e49083. https://doi.org/10.1371/journal.pone.0049083

Wieben, E. D., Baratz, K. H., Aleff, R. A., Kalari, K. R., Tang, X., Maguire, L. J., … Fautsch, M. P. (2019). *Gene Expression and Missplicing in the Corneal Endothelium of Patients With a TCF4 Trinucleotide Repeat Expansion Without Fuchs' Endothelial Corneal Dystrophy*. https://doi.org/10.1167/iovs.19-27689

Willoughby, C. E., Ponzin, D., Ferrari, S., Lobo, A., Landau, K., & Omidi, Y. (2010). Anatomy and physiology of the human eye: effects of mucopolysaccharidoses disease on structure and function - a review. *Clinical & Experimental Ophthalmology*, *38*, 2–11. https://doi.org/10.1111/j.1442-9071.2010.02363.x

Wilson, S. E., & Hong, J. W. (2000, July). Bowman's layer structure and function: Critical or dispensable to corneal function? A Hypothesis. *Cornea*, Vol. 19, pp. 417–420. https://doi.org/10.1097/00003226-200007000-00001

Wojciechowska, M., & Krzyzosiak, W. J. (2011). Cellular toxicity of expanded RNA repeats: focus on RNA foci. *Human Molecular Genetics*, *20*(19), 3811–3821. https://doi.org/10.1093/HMG/DDR299

Wong, L. J. C., Ashizawa, T., Monckton, D. G., Caskey, C. T., & Richards, C. S. (1995). Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent. *American Journal of Human Genetics*, *56*(1), 114. Retrieved from /pmc/articles/PMC1801291/?report=abstract

Woo, J. H., Ang, M., Htoon, H. M., & Tan, D. (2019). Descemet Membrane Endothelial Keratoplasty Versus Descemet Stripping Automated

Endothelial Keratoplasty and Penetrating Keratoplasty. *American Journal of Ophthalmology*, *207*, 288–303. https://doi.org/10.1016/j.ajo.2019.06.012

Wright, A. F., & Dhillon, B. (2010). Major Progress in Fuchs's Corneal Dystrophy. *New England Journal of Medicine*, *363*(11), 1072–1075. https://doi.org/10.1056/NEJMe1007495

Wright, G. E. B., Collins, J. A., Kay, C., McDonald, C., Dolzhenko, E., Xia, Q., … Hayden, M. R. (2019). Length of Uninterrupted CAG, Independent of Polyglutamine Size, Results in Increased Somatic Instability, Hastening Onset of Huntington Disease. *American Journal of Human Genetics*, *104*(6), 1116–1126. https://doi.org/10.1016/j.ajhg.2019.04.007

Wu, H. T., Zhong, H. T., Li, G. W., Shen, J. X., Ye, Q. Q., Zhang, M. L., & Liu, J. (2020). Oncogenic functions of the EMT-related transcription factor ZEB1 in breast cancer. *Journal of Translational Medicine*, *18*(1), 1–10. https://doi.org/10.1186/S12967-020-02240-Z/FIGURES/3

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, *89*(1), 82–93. https://doi.org/10.1016/J.AJHG.2011.05.029

Xing, C., Gong, X., Hussain, I., Khor, C.-C., Tan, D. T. H., Aung, T., … Mootha, V. V. (2014). Transethnic replication of association of CTG18.1 repeat expansion of TCF4 gene with Fuchs' corneal dystrophy in Chinese implies common causal variant. *Investigative Ophthalmology & Visual Science*, *55*(11), 7073–7078. https://doi.org/10.1167/iovs.14-15390

Yu, J., Ryan, D. G., Getsios, S., Oliveira-Fernandes, M., Fatima, A., & Lavker, R. M. (2008). MicroRNA-184 antagonizes microRNA-205 to maintain

SHIP2 levels in epithelia. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(49), 19300. https://doi.org/10.1073/PNAS.0803992105

Zarouchlioti, C., Sanchez-Pintado, B., Hafford Tear, N. J., Klein, P., Liskova, P., Dulla, K., … Davidson, A. E. (2018). Antisense Therapy for a Common Corneal Dystrophy Ameliorates TCF4 Repeat Expansion-Mediated Toxicity. *American Journal of Human Genetics*, *102*(4), 528–539. https://doi.org/10.1016/j.ajhg.2018.02.010

Zavala, J., López Jaime, G. R., Rodríguez Barrientos, C. A., & Valdez-Garcia, J. (2013). Corneal endothelium: Developmental strategies for regeneration. *Eye (Basingstoke)*, Vol. 27, pp. 579–588. https://doi.org/10.1038/eye.2013.15

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(Database issue), D754. https://doi.org/10.1093/NAR/GKX1098

Zhang, D., Dey, R., & Lee, S. (2020). Fast and robust ancestry prediction using principal component analysis. *Bioinformatics*, *36*(11), 3439. https://doi.org/10.1093/BIOINFORMATICS/BTAA152

Zu, T., Cleary, J. D., Liu, Y., Bañez-Coronel, M., Bubenik, J. L., Ayhan, F., … Ranum, L. P. W. (2017). RAN Translation Regulated by Muscleblind Proteins in Myotonic Dystrophy Type 2. *Neuron*, *95*(6), 1292-1305.e5. https://doi.org/10.1016/j.neuron.2017.08.039

Zu, T., Gibbens, B., Doty, N. S., Gomes-Pereira, M., Huguet, A., Stone, M. D., … Ranum, L. P. W. (2011). Non-ATG-initiated translation directed by microsatellite expansions. *Proceedings of the National Academy of*

Sciences of the United States of America, *108*(1), 260–265. https://doi.org/10.1073/pnas.1013343108

Zuallaert, J., Godin, F., Kim, M., Soete, A., Saeys, Y., & De Neve, W. (2018). SpliceRover: interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics (Oxford, England)*, *34*(24), 4180–4188. https://doi.org/10.1093/BIOINFORMATICS/BTY497

Supplementary data

Table S1 List of primers used for PCR amplification and Sanger sequencing

Gene	Target	Forward primer	Reverse primer	Amplimer Size (bp)
COL8A2	c.1363C>A, p.(Gln455Lys)	GTGACCAGGGGCCTAGTG	CCTGCGATGCCAGTCTCAT	448
	c.1301G>A, p.(Arg434His)			
	c.1724C>T, p.(Pro575Leu)	TGGAGAGGGGAGAGCAGG	CGTTGTTCTTGTACAGGGCC	449
	c.660G>A, p.(Gly220=)	GAAAACCAGGTGCCCAAGG	GACTCCCACACCGTCTACTC	329
	c.22C>T, p.(Leu8=)	CCCGCGACTTTGAAAATTGC	GAGGCTCTCACCCAGGTAC	390
	c.297A>G, p.(Lys99=)	CCATTCTTCCTCTCCCGTGT	CTGGTCCCCTCGTATTCCTG	372
	c.96C>T, p.(Ala32=)	TCTGATCTTTTGGTGACCCC	GAATGAGGAGCTGTGGAGGG	395
SLC4A11	c.326G>T, p.(Arg109Leu)	GCTAGGGAATGCTGGAGACT	GGAGAAAAGCGGGGAGGG	489
	c.1121G>A, p.(Arg374Gln); c.1120C>T, p.(Arg374Trp)	CAGAGGTACAGGGTAGAGGC	CAAGGCCTGGAAAGCAGAG	214
	c.1018G>C, p.(Val340Leu); c.1012G>A, p.(Gly338Ser)	GGGAGAGCACCTTCACCTG	GGGCTGGAGGAGAGGACA	386

	c.530A>G, p.(Asn177Ser)	GACCCTGACAACAATGAGCC	GAGGACACAGTGCACAGTTG	387
	c.1800C>T, p.(Thr600=)	TGGGCATTACTTGGACGACT	CATGAGCACAGCCTTTGACC	289
	c.2265C>T, p.(His755=)	CTCTGCTGTGGGGCTCTG	GCCTCACTCCTCCCTATGTC	297
	c.2421C>T, p.(Pro807=)	TGGGACATAGGGAGGAGTGA	CCCTCCGGATGTAGTGTGTC	385
	c.2739C>T, p.(Asp913=)	TGGGCTGGGATGGGTGTC	ACCCCTACAATGCCCAGATG	296
	c.2073C>T, p.(Leu691=)	AATCGAGAGTGAGTTGGGGC	GTGTTGATGATGGCGAGGAG	399
	c.1851C>G, p.(Thr617=)	ATCCCTTGTCAGCCTGTCAG	AGGATCTCTCGCACGCAG	299
	c.173-8C>G	TTTAACAGCCAGGCCCTCTT	AGTCACACCTGCCCAGTC	250
ZEB1	c.698C>T, p.(Thr233Met)	TCAAATTCTGTCCCCACTATCAC	TCTCCCAATTAGTGTATGCCAA	481
	c.998T>C, p.(Ile333Thr)	GAGAAGCCATATGAATGCCCA	CAGAACAACAGCTTGCACCA	396
	c.3099G>C, p.(Glu1033Asp); c.3093G>A, p.(Arg1031=)	GCTTCTCACACTCTGGGTCT	TTCAGCCCTGTCATCCTTCA	395
	c.2254A>G, p.(Thr752Ala); c.2064A>C, p.(Pro688=)	ACTCCCCAGTTTTACCAGTG	GTTGGCTCTACGGGACTGAT	394
	c.2563C>A, p.(Gln855Lys);	CCAAGTGCCAACCCCATAAA	AGCAAACAACCAACTGAAGACA	400

	c.2522A>C, p.(Gln841Pro)			
	c.1260G>A, p.(Ala420=) c.1661A>G, p.(Lys554Arg)	GTTCTCCTCAGGGCATGGT	GCTCTTCTGCACTTGGTTGT	678
	c.105_107del, p.(Asp35del)	ACTCTCTCTCTGCCTTGATTTTC	ACACTGTCTGGTCTGTTGGC	499
	c.794-7T>G	ACCGCTTGTTTTAGGGAAATGA	TTTTGCCGTATCTGTGGTCG	286
	c.233A>C, p.(Asn78Thr)	GCTGACTGTGAAGGTGTACC	AGAGTATTCATTCGGGGTTACAA	420
	c.2706T>C, p.(Asn902=)	TCAGTGTGCTTGCTTTGGTC	GATTGAGATTGCGTGCCACT	491
	c.606T>C, p.(Ser202=)	GGGACTCAGTGGAAACTTTGG	TGAACTCTCAGTCATTGCACT	299
AGBL1	c.484G>A, p.(Val162Met); c.490A>T, p.(Ile164Phe); c.477C>T, p.(Asn159=)	TCTCAAAGAGGTGTGGCTGT	CAGAGTGAATCCCCATGCTG	295
	c.668C>T, p.(Pro223Leu); c.671C>T, p.(Thr224Met)	GCTTGGAGAGTGTTATTAGCTGT	ACCAAGCAAAGCAGAAACAGT	492
	c.1994A>G, p.(Tyr665Cys)	CCCTCCCACCTCTCCCTTAT	CCTCTGCACCCCATCAGG	297
	c.2884A>G, p.(Lys962Glu)	GCCATTCACAATAAATCAGCTGG	GCCTGTGATTCTGCTCAGTT	400

	c.3082C>T, p.(Arg1028Ter); c.3149A>G, p.(Asn1050Ser)	CCATGTTCTGTTTGGGCCTC	ATTTTCTCTGTGTAAAGACCCCT	232
	c.1204A>G, p.(Arg402Gly); c.1290A>C, p.(Pro430=)	CTGCCTCCTCAAAACAGCAT	TTGAAGTCCACCACAGAGCT	421
	c.2622G>A, p.(Leu874=)	GAAACCAATGACCTGACCTGG	AATGAGTGGGGCATGGTCTT	265
	c.2471G>A, p.(Ser824Asn)	GCTTAAAGGACAGATCTACAGCT	AAGGTGGAGGCAGAAGGAAG	330
LOXHD1	c.3338G>T, p.(Cys1113Phe)	AACTTCATGGGGTCCTGCTC	TTTTCTCCACTTGCCACTGG	383
	c.569T>G, p.(Leu190Arg); c.541C>T, p.(Leu181Phe)	TCTCTGCTCAGGTCCTTGATG	AAGGAAGAACTGGGGCTGAG	291
	c.703A>C, p.(Lys235Gln)	ATCTCAGGAGGAAGTTGCCC	GAGTGGATGCAGATGGACCT	388
	c.1690C>T, p.(Arg564Cys)	CACTGGGAAGCACAAGGAC	AGGTAGGCTGTTCTTCCCAC	290
	c.911G>A, p.(Arg304Gln)	TGAGCTGATGAATCCCTGAAGT	AGCCTTCCCATGGTGATGAG	275
	c.1570C>T, p.(Arg524Cys)	AATAGCCTTGGCTTCTCTGC	TAAGGGGCCTGAAGATGCAA	353
	c.2469C>A, p.(Asn823Lys)	CACCCTAAGCCTCACCTTGT	GGCCTTGAGTGGGAGCTAC	

	c.1708G>A, p.(Asp570Asn); c.1742T>C, p.(Val581Ala)	GGTCAGCCCAGATGAGAACT	TTAACAGGGCAGGGAAGACG	356
	c.274G>A, p.(Val92Ile)	GGAATGGGATCTTGTTGCTCA	GACAGGGAAAGATTTGGGCC	355
	c.2027A>G, p.(Asp676Gly)	CCTCCTTCCAATCTCAGCCA	TGCTCGTGTTTTCTTGAAGGG	378
	c.815C>T, p.(Thr272Met)	TCAGTGCCTTATCTCCTTTCCA	TTTGCCTTGAACCTGCTCTG	207
	c.2080G>T, p.(Asp694Tyr)	CTTCCTCCCTCTGCCTTGG	ACATGGTCTTGGGAAGGAGA	250
	c.1042+7G>A	TGTCTGTCTGTCTGTCCCAC	GAGAGGAGGGAAGGAGGGTA	227
	c.1147C>T, p.(Arg383Ter)	TGTCTGTCTGTCTGTCCCAC	GAGAGGAGGGAAGGAGGGTA	227
	c.1008C>T, p.(Gly336=)	TTGTCACCGTCTTCACTGGG	CCACCCAAAATGCCCAATGA	383
	c.2887T>C, p.(Leu963=)	CATCCCATCCCTGTTCCCTG	TCATACCCTGCTCTCTTGCC	362
	c.1887C>T, p.(Ser629=)	TGGGGTAGCCACTGTCTAAC	GGTAGTAGGGCTGGGTCTTC	286
	c.231C>T, p.(Leu77=)	TCTGAGCCTGCAATCTGTCT	CAAGAGAGGAGCTGAGGGAG	360
	c.228G>A, p.(Lys76=)	GGCTGTCCTCTTTCCTCCTT	CCCCTTGGAAATTTCTGCTGA	257
	c.93G>A, p.(Val31=)	TCCAATCTTTCCCTATCCACCC	GGGAGGGAAGGAAGATGGAG	279
	c.2535C>T, p.(His845=)	TGTACCCCTGACTCCTCTGA	CAGAGTCAATGTGCTGCCAG	379

	c.2370C>T, p.(Asp790=)	CCTCACTCGGGTTTCCTTCA	CACCACAGCCTCCTCCATAC	234
	c.966G>C, p.(Gly322=)	TGCAGCTCACATTGAACACT	AGGGGTTGAATCAGGGAAGG	339
TCF4	c.57G>T, p.(Arg19Ser); c.58A>T, p.(Lys20Ter); c.66G>A, c.(Glu22=)	GTGTGAGTGAGAGGGAACGA	GTCGGGCAGGCTAGGATG	261
	c.26T>C, p.(Ile9Thr)	GCCTGTGATTGATTAGTTTTGGC	GTGGCAACCCTGTAAGTTTG	492
	c.944C>T, p.(Ala315Val)	GCCTGGTTTTCATATTCTGCCT	TCCAGTGACTGTTATGCAAGAA	250
	c.51A>G, p.(Gln17=)	CTCGGCCATCCCAGGAAG	AAGGACCAGAGGCTACTTCC	244
	c.1419G>T, p.(Pro473=)	TGGCCTCTGGAAATAGCTGT	GCTTCTTGAGGGATGAACACC	296
KANK4	c.1762G>A, p.(Ala588Thr)	TCCAATCTCCCAGGGAAGG	TCAGGCCCCTATATTGATGGT	309
	c.2231+1del	GCCCATCTCACCTGCGAG	CACAGTGTGGGCATGAAACA	250
	c.797A>G, p.(Asp266Gly)	AGGATGCCGAGCTCACTTT	GCTTCTCTGGCATTGTGTTCA	382
	c.1550G>A, p.(Arg517Lys)	AATCAGAGCCCAGCAGAACG	GGTGCTCCTTCCCTGGGA	250
	c.1229C>T, p.(Thr410Met); c.1230G>A, p.(Thr410=)	TGAAACAGCAGGTCTCGGC	AGCTTTCAGACTCCATGCTG	300

	c.33T>C, p.(Ser11=)	CATCTCAGCATAATTTCGAGGC	TCCCTTCTCGATGTCATCCA	255
	c.912G>A, p.(Glu304=)	CCAGAGAAGCAGAGGTGTTG	GTTTCAGGCTGGAGATGCTG	231
LAMC1	c.95G>C, p.(Cys32Ser)	CTTGCCTTCGCCGTGACC	GAGTCCCACACGTGTTGGT	250
	c.1669C>T, p.(Arg557Trp)	TTCTGTATACAAGGTGTGGTCTT	AGTATCTCGCCTGTCCACTC	396
	c.2701G>A, p.(Val901Met)	GGCAACACAGGTCTAAAGAATCT	CAGCACTGCCCCTGAAGG	395
	c.1240G>A, p.(Gly414Ser)	ATCAGGAGATTGCATACTGGTT	AGCTAGTCCTCAGTCTTGTTGA	378
	c.4456A>G, p.(Met1486Val); c.4467A>C, p.(Ala1489=)	TGCCAAGAATGAACCAAGCT	ACACTGTTCACTTTCTGCCA	612
	c.3009C>T, p.(Cys1003=); c.3106C>T, p.(Arg1036Trp)	AAGTGTGTCCCAAAGGTTGC	AGTGCACAGTAAGTTGCCAC	479
	c.4303G>A, p.(Ala1435Thr)	ACCCTGACTTCCATTGTTCAT	GCTCTGAGGGTCTGGAAAGA	389
	c.1836C>T, p.(Gly612=)	CAGTCTGACCTGCTGTGTG	AGGCTCCCTTTTACAAAATCACC	400
	c.282C>T, p.(Ala94=)	GATGAGAGGGAGCCATCGG	TGTCACCGCTTACCCAGG	432
	c.2553G>A, p.(Lys851=)	ACATCAGATTGTCTCATCCCCA	GACAAAAGCGGTTCACAATGT	394
	c.3780C>G, p.(Ala1260=);	CCTGACCTGAAGTGATCTGC	GAATCCTCCCACCTCAGCC	626

	c.3796G>A, p.(Glu1266Lys)			
	c.757C>G, p.(Leu253Val); c.768G>C, p.(Leu256=)	TCAGTCCCTCCACTTCTCTT	CACCTCTATGCAAAAGCTCCA	234
	c.4797C>T, p.(Gly1599=)	TCTATGTACTTTCTGACCCTCCA	ATGGACAGCAGCAGAGGAG	396
	c.882G>A, p.(Glu294=); c.894C>T, p.(Asn298=); c.999G>A, p.(Ala333=)	GGACTTTGCAGCTGTTCCAT	AGACAACTGACAGACTGACGT	249
ATP1B1	c.222G>A, p.(Pro74=)	TCCGTGGGAAGATTAAACTTTCA	ACCCATGCTATCTCTGAAGGA	249
	c.321G>A, p.(Arg107=)	TGTCTTCGTTTCTGCCTTCC	GCCACAGACATCTACAATGAGT	236
COL8A1	c.619A>T, p.(Ile207Phe)	GGAATGCCAGGGAAGCCA	CCGAAGCCCTTGTCTCCTTT	248
LYPD3	c.658C>T p.(Arg220Cys)	TCTGTGTCTCCTTCCTGCAG	GTGACAGATGTGGTTGAGGC	243
mir184	+58G>A; +73G>T	GCACAGAGGGGCTTTGAATT	ACAAAACACAAAGGCTACCCC	384

Table S2 List of primers used for PCR amplification of miR-184 mRNA target genes, with restriction enzymes NheI and SalI tagged to the forward and reverse primers, respectively, to enable cloning.

Gene	Target region	Forward primer tagged with NheI sequence (GCTAGC)	Reverse primer tagged with SalI sequence (GTCGACG)	Amplimer Size (bp)
AKT2	3TUR	CACTGTGATCCATGAGCTGC	GCTAGTACAGGAGGAGCTGG	371
SF1	3TUR	AAGGAGAGGGGAGCAAATGG	TCAAACCCCTACACACTGCA	398
EPB41L5	3TUR	aatgtcctcctccaaacccc	ACCAGAGGCAGGCTTGTTAT	478
INNPL1	3TUR	GGTTACTCTGGTGCTGTCCT	CACAGACCAGGAGACAGTGA	358

Table S3 List of MiSeq primer sequences

Forward MiSeq primer name	P5 adaptor	common part before index	index 1	seq primer binding site	spacer	Forward locus specific PCR primer
S502	AATGATACGGCGACCACCGAGATCT	ACAC	CTCTCTAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AATATAT	AATCCAAACCGCCTTCCAAGTG
S503	AATGATACGGCGACCACCGAGATCT	ACAC	TATCCTCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ATACTT	AATCCAAACCGCCTTCCAAGTG
S505	AATGATACGGCGACCACCGAGATCT	ACAC	GTAAGGAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CT	AATCCAAACCGCCTTCCAAGTG
S506	AATGATACGGCGACCACCGAGATCT	ACAC	ACTGCATA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCC	AATCCAAACCGCCTTCCAAGTG
S507	AATGATACGGCGACCACCGAGATCT	ACAC	AAGGAGTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGAT	AATCCAAACCGCCTTCCAAGTG
S508	AATGATACGGCGACCACCGAGATCT	ACAC	CTAAGCCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	G	AATCCAAACCGCCTTCCAAGTG
S510	AATGATACGGCGACCACCGAGATCT	ACAC	CGTCTAAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TATTA	AATCCAAACCGCCTTCCAAGTG
S511	AATGATACGGCGACCACCGAGATCT	ACAC	TCTCTCCG	ACACTCTTTCCCTACACGACGCTCTTCCGATCT		AATCCAAACCGCCTTCCAAGTG
S513	AATGATACGGCGACCACCGAGATCT	ACAC	TCGACTAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCT		AATCCAAACCGCCTTCCAAGTG
S515	AATGATACGGCGACCACCGAGATCT	ACAC	TTCTAGCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCTT	AATCCAAACCGCCTTCCAAGTG
S516	AATGATACGGCGACCACCGAGATCT	ACAC	CCTAGAGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	C	AATCCAAACCGCCTTCCAAGTG
S517	AATGATACGGCGACCACCGAGATCT	ACAC	GCGTAAGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TG	AATCCAAACCGCCTTCCAAGTG
S518	AATGATACGGCGACCACCGAGATCT	ACAC	CTATTAAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAATATT	AATCCAAACCGCCTTCCAAGTG

S520	AATGATACGGCGACCACCGAGATCT	ACAC	AAGGCTAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CATATA	AATCCAAACCGCCTTCCAAGTG
S521	AATGATACGGCGACCACCGAGATCT	ACAC	GAGCCTTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ATC	AATCCAAACCGCCTTCCAAGTG
S522	AATGATACGGCGACCACCGAGATCT	ACAC	TTATGCGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ATACT	AATCCAAACCGCCTTCCAAGTG
Reverse MiSeq primer name	**P7**		**Index 2**	**seq primer binding site**	**spacer**	**Reverse locus specific PCR primer**
N701	CAAGCAGAAGACGGCATACGAGAT		TCGCCTTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	CAAAACTTCCGAAAGCCATTTCTCC
N702	CAAGCAGAAGACGGCATACGAGAT		CTAGTACG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	CAAAACTTCCGAAAGCCATTTCTCC
N703	CAAGCAGAAGACGGCATACGAGAT		TTCTGCCT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	CAAAACTTCCGAAAGCCATTTCTCC
N704	CAAGCAGAAGACGGCATACGAGAT		GCTCAGGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	CAAAACTTCCGAAAGCCATTTCTCC
N705	CAAGCAGAAGACGGCATACGAGAT		AGGAGTCC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	CAAAACTTCCGAAAGCCATTTCTCC
N706	CAAGCAGAAGACGGCATACGAGAT		CATGCCTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	CAAAACTTCCGAAAGCCATTTCTCC
N707	CAAGCAGAAGACGGCATACGAGAT		GTAGAGAG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	CAAAACTTCCGAAAGCCATTTCTCC
N710	CAAGCAGAAGACGGCATACGAGAT		CAGCCTCG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		CAAAACTTCCGAAAGCCATTTCTCC
N711	CAAGCAGAAGACGGCATACGAGAT		TGCCTCTT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	CAAAACTTCCGAAAGCCATTTCTCC
N712	CAAGCAGAAGACGGCATACGAGAT		TCCTCTAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	CAAAACTTCCGAAAGCCATTTCTCC
N714	CAAGCAGAAGACGGCATACGAGAT		TCATGAGC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	CAAAACTTCCGAAAGCCATTTCTCC

N715	CAAGCAGAAGACGGCATACGAG AT		CCTGAGA T	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	TA	CAAAACTTCCGAAAGCCATTTCT CC
N716	CAAGCAGAAGACGGCATACGAG AT		TAGCGAG T	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	AT	CAAAACTTCCGAAAGCCATTTCT CC
N718	CAAGCAGAAGACGGCATACGAG AT		GTAGCTC C	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	G	CAAAACTTCCGAAAGCCATTTCT CC
N719	CAAGCAGAAGACGGCATACGAG AT		TACTACGC	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	TA	CAAAACTTCCGAAAGCCATTTCT CC
N720	CAAGCAGAAGACGGCATACGAG AT		AGGCTCC G	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC		CAAAACTTCCGAAAGCCATTTCT CC
N721	CAAGCAGAAGACGGCATACGAG AT		GCAGCGT A	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	G	CAAAACTTCCGAAAGCCATTTCT CC
N722	CAAGCAGAAGACGGCATACGAG AT		CTGCGCA T	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	G	CAAAACTTCCGAAAGCCATTTCT CC
N723	CAAGCAGAAGACGGCATACGAG AT		GAGCGCT A	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	G	CAAAACTTCCGAAAGCCATTTCT CC
N724	CAAGCAGAAGACGGCATACGAG AT		CGCTCAG T	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	G	CAAAACTTCCGAAAGCCATTTCT CC
N726	CAAGCAGAAGACGGCATACGAG AT		GTCTTAG G	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	AT	CAAAACTTCCGAAAGCCATTTCT CC
N727	CAAGCAGAAGACGGCATACGAG AT		ACTGATC G	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	TA	CAAAACTTCCGAAAGCCATTTCT CC
N728	CAAGCAGAAGACGGCATACGAG AT		TAGCTGC A	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	AT	CAAAACTTCCGAAAGCCATTTCT CC
N729	CAAGCAGAAGACGGCATACGAG AT		GACGTCG A	GTGACTGGAGTTCAGACGTGTGCTCTTCCG ATC	G	CAAAACTTCCGAAAGCCATTTCT CC

Table S4 Summary of patient demographics included in the MiSeq assay, including Short Tandem Repeat genotype, MiSeq determined estimated progenitor allele length and somatic expansion scores.

Sample Identifier	Age	Sex	STR genotype of expanded allele	ePAL of expanded allele	Somatic expansion scores for reads larger than ePAL from ePAL to the end	Somatic expansion scores for larger than 116 from ePAL to the end
AS1-N701-A-S502	74	F	76	76	0.966382101	0.106324178
AS1-N701-A-S503	69	M	51	52	0.724392112	0.001372136
AS1-N701-A-S505	49	F	96	97	0.969861036	0.714612284
AS1-N701-A-S507	61	F	74	106	0.973546564	0.76080683
AS1-N701-A-S508	78	F	69	70	0.894022524	0.002562807
AS1-N701-A-S510	69	F	68	68	0.878623889	0.003865481
AS1-N701-C-S513	52	M	79	80	0.917406682	0.029261342
AS1-N701-C-S516	59	F	76	74	0.860309385	0.002699336
AS1-N701-C-S517	77	M	99	98	0.954159203	0.576317119
AS1-N701-C-S518	85	M	89	89	0.977050153	0.789758301
AS1-N701-C-S520	61	F	88	88	0.970028338	0.561893649
AS1-N701-C-S521	76	M	58	57	0.827670602	0.002951339
AS1-N702-A-S502	68	M	74	74	0.938938718	0.019153449
AS1-N702-A-S503	39	M	93	93	0.977509352	0.771758237
AS1-N702-A-S505	59	F	67	65	0.870923525	0.001199383
AS1-N702-A-S506	67	F	84	85	0.960193408	0.241711947
AS1-N702-A-S507	88	F	56	59	0.815559157	0.002377093
AS1-N702-A-S510	81	F	107	105	0.947038114	0.465636002
AS1-N702-A-S511	75	F	81	83	0.970677492	0.47053955
AS1-N702-C-S513	85	F	80	78	0.958346701	0.318682494

AS1-N702-C-S515	57	M	72	68	0.834826151	0.002186332
AS1-N702-C-S516	60	F	89	82	0.981840194	0.347025251
AS1-N702-C-S518	67	F	69	63	0.96024495	0.001679095
AS1-N702-C-S520	81	M	81	79	0.961879226	0.150571229
AS1-N702-C-S521	75	F	83	82	0.980148253	0.382514216
AS1-N702-C-S522	79	F	68	67	0.879313908	0.002118206
AS1-N703-A-S502	85	F	73	74	0.91226173	0.006928152
AS1-N703-A-S505	70	F	90	92	0.949894406	0.567518324
AS1-N703-A-S506	76	F	86	87	0.952543258	0.53931508
AS1-N703-A-S507	80	F	70	71	0.904845865	0.001933796
AS1-N703-A-S508	74	F	65	65	0.844256447	0.002106479
AS1-N703-A-S510	64	M	77	74	0.943584623	0.003328341
AS1-N703-C-S515	77	M	89	87	0.966166667	0.509666667
AS1-N703-C-S516	73	M	93	91	0.977194393	0.670726103
AS1-N703-C-S517	57	F	107	108	0.965876035	0.775453122
AS1-N703-C-S518	76	F	84	84	0.974167366	0.46904562
AS1-N703-C-S520	46	F	96	95	0.974860724	0.638788301
AS1-N703-C-S522	76	F	77	77	0.959314877	0.376865229
AS1-N704-A-S502	69	F	86	89	0.979940681	0.616921636
AS1-N704-A-S503	73	M	89	90	0.920438027	0.478825468
AS1-N704-A-S505	68	M	73	71	0.947706888	0.011423253
AS1-N704-A-S506	84	M	87	86	0.978646131	0.479436392
AS1-N704-A-S508	70	M	95	85	0.972197459	0.232142857
AS1-N704-A-S510	74	F	79	78	0.97388372	0.213213213
AS1-N704-C-S513	63	M	92	91	0.956026712	0.62172025
AS1-N704-C-S515	77	F	91	81	0.980347969	0.104963618
AS1-N704-C-S516	90	M	80	80	0.96386453	0.436498337
AS1-N704-C-S517	86	F	72	69	0.901458789	0.002552881

AS1-N704-C-S518	65	M	95	94	0.939556673	0.470978062
AS1-N704-C-S521	69	M	86	87	0.95900706	0.452483977
AS1-N704-C-S522	75	F	96	95	0.953759121	0.303025334
AS1-N705-A-S502	88	M	78	79	0.964984387	0.379647175
AS1-N705-A-S503	72	F	83	86	0.973239687	0.568234353
AS1-N705-A-S506	65	M	67	65	0.89242781	0.001883753
AS1-N705-A-S508	88	F	78	87	0.95031398	0.633469208
AS1-N705-A-S511	86	F	78	79	0.941294919	0.371385791
AS1-N705-C-S513	76	F	X	108	0.926022628	0.532637076
AS1-N705-C-S515	76	F	93	92	0.932802979	0.529197698
AS1-N705-C-S516	72	M	80	81	0.971044271	0.127637311
AS1-N705-C-S517	51	M	99	102	0.979317203	0.763274776
AS1-N705-C-S518	81	M	82	82	0.984872588	0.493531831
AS1-N705-C-S520	69	F	83	83	0.943081707	0.388034779
AS1-N705-C-S521	82	F	61	61	0.83302982	0.002503984
AS1-N705-C-S522	54	M	84	80	0.940062316	0.028889253
AS1-N706-A-S502	66	F	101	79	0.992500457	0.187671483
AS1-N706-A-S503	82	F	87	90	0.969769572	0.622507223
AS1-N706-A-S505	79	F	81	84	0.967743875	0.421794679
AS1-N706-A-S506	59	M	70	69	0.866254424	0.001516743
AS1-N706-A-S507	89	F	80	81	0.959400797	0.604682607
AS1-N706-A-S508	58	M	82	83	0.959491975	0.126549508
AS1-N706-C-S513	68	F	76	74	0.895747056	0.003421081
AS1-N706-C-S515	66	M	76	77	0.961298684	0.128192523
AS1-N706-C-S516	60	M	99	100	0.925432032	0.303442645
AS1-N706-C-S517	45	F	90	89	0.969746125	0.188552449
AS1-N706-C-S518	69	M	96	95	0.960359599	0.601484252
AS1-N706-C-S520	69	M	78	77	0.963351717	0.143813378

AS1-N706-C-S521	75	F	101	99	0.940568475	0.521963824
AS1-N706-C-S522	78	F	70	69	0.883307373	0.002855659
AS1-N707-A-S502	74	M	96	98	0.931037401	0.5010154
AS1-N707-A-S503	79	F	50	51	0.79464101	0.004655493
AS1-N707-A-S505	69	M	66	56	0.976143025	0.002198124
AS1-N707-A-S506	64	F	84	86	0.971237676	0.307300579
AS1-N707-A-S507	76	F	79	81	0.96456146	0.368179616
AS1-N707-A-S508	68	F	87	88	0.958993126	0.36404157
AS1-N707-A-S510	68	M	75	73	0.917750678	0.003279133
AS1-N707-A-S511	85	F	88	87	0.981714601	0.599222028
AS1-N707-C-S513	61	M	73	70	0.907571962	0.003733989
AS1-N707-C-S515	75	M	75	75	0.949988649	0.330374574
AS1-N707-C-S516	57	M	87	87	0.97430744	0.310287073
AS1-N707-C-S517	78	M	93	94	0.979650203	0.693103069
AS1-N707-C-S518	70	M	71	70	0.913265557	0.001866313
AS1-N707-C-S521	72	F	78	74	0.955740144	0.055138854
AS1-N707-C-S522	55	M	63	63	0.64863109	0.004640371
AS1-N710-A-S502	73	F	81	82	0.973423275	0.201480993
AS1-N710-A-S503	54	F	71	71	0.863578054	0.001874877
AS1-N710-A-S506	67	F	80	83	0.959953199	0.274052013
AS1-N710-A-S507	79	M	87	91	0.93596268	0.314249364
AS1-N710-A-S508	80	M	85	88	0.895662144	0.207529142
AS1-N710-A-S510	62	M	62	59	0.788386084	0.001046299
AS1-N710-A-S511	73	M	86	86	0.955977557	0.642857143
AS1-N710-C-S513	49	M	105	84	0.959138571	0.281210326
AS1-N710-C-S515	80	M	85	83	0.952341178	0.279418268
AS1-N710-C-S517	77	F	63	61	0.801603666	0.00137457
AS1-N710-C-S518	77	F	62	69	0.893173187	0.002419853

AS1-N710-C-S520	66	F	74	71	0.952601836	0.403880704
AS1-N710-C-S521	74	F	95	95	0.952565248	0.465317059
AS1-N710-C-S522	64	F	91	91	0.905880547	0.415937356
AS1-N711-A-S502	77	M	X	98	0.96124031	0.618001723
AS1-N711-A-S503	77	M	86	86	0.945119983	0.573872807
AS1-N711-A-S505	79	F	83	79	0.98263949	0.166637797
AS1-N711-A-S506	69	F	84	86	0.973905847	0.5938854
AS1-N711-A-S507	78	F	77	77	0.958293159	0.119904077
AS1-N711-A-S508	56	M	107	110	0.95079508	0.731473147
AS1-N711-A-S510	76	M	83	83	0.946153846	0.517769827
AS1-N711-A-S511	84	M	75	73	0.935438399	0.017394982
AS1-N711-C-S513	59	M	88	90	0.964264396	0.643201082
AS1-N711-C-S516	71	F	81	87	0.94570229	0.477217114
AS1-N711-C-S517	76	M	87	85	0.959933441	0.52156588
AS1-N711-C-S518	50	M	88	88	0.957953916	0.183361047
AS1-N711-C-S520	61	F	79	78	0.929645866	0.296845453
AS1-N711-C-S521	69	F	62	60	0.869016984	0.001415337
AS1-N711-C-S522	63	M	66	66	0.786682712	0.004492579
AS1-N712-A-S503	69	M	105	106	0.958705973	0.570497957
AS1-N712-A-S505	65	F	70	69	0.876165422	0.002166428
AS1-N712-A-S506	71	M	76	75	0.913073493	0.005321444
AS1-N712-A-S507	71	F	84	84	0.971865443	0.456636086
AS1-N712-A-S508	73	M	81	82	0.937649165	0.253729117
AS1-N712-A-S510	86	F	X	103	0.939021808	0.520643659
AS1-N712-A-S511	86	F	73	75	0.954359957	0.094528192
AS1-N712-C-S515	77	M	105	103	0.938034961	0.359975889
AS1-N712-C-S516	67	F	79	77	0.943556701	0.037293814
AS1-N712-C-S518	65	F	77	77	0.880877356	0.008124807

AS1-N712-C-S521	45	F	78	81	0.955051611	0.128495464
AS1-N712-C-S522	77	M	91	92	0.925232824	0.471182569
AS1-N714-A-S502	78	F	69	67	0.895110642	0.001819569
AS1-N714-A-S507	85	F	X	91	0.943185121	0.469776228
AS1-N714-A-S508	73	F	82	82	0.961168089	0.564785094
AS1-N714-A-S510	71	M	82	81	0.95007638	0.340321902
AS1-N714-A-S511	67	F	87	88	0.961899432	0.572283451
AS1-N714-C-S515	69	F	90	89	0.974494202	0.566740686
AS1-N714-C-S516	53	M	103	104	0.965433143	0.710360302
AS1-N714-C-S517	68	M	86	81	0.949735305	0.044970802
AS1-N714-C-S518	65	M	75	71	0.885461937	0.002032805
AS1-N714-C-S520	68	F	75	79	0.95708476	0.050406678
AS1-N714-C-S521	68	M	86	88	0.97631597	0.639263468
AS1-N714-C-S522	71	F	71	68	0.87944835	0.003244997
AS1-N715-A-S503	61	M	78	77	0.959103763	0.067761584
AS1-N715-A-S505	76	M	97	99	0.956441576	0.38050242
AS1-N715-A-S506	72	F	84	86	0.956830652	0.359285781
AS1-N715-A-S507	82	F	85	84	0.968266254	0.50374097
AS1-N715-C-S513	59	F	83	81	0.960236432	0.300913487
AS1-N715-C-S515	84	F	79	78	0.963663514	0.237137427
AS1-N715-C-S516	70	M	107	107	0.962900708	0.685196302
AS1-N715-C-S517	84	F	86	85	0.956408215	0.422894444
AS1-N715-C-S518	69	M	85	85	0.978787593	0.248279438
AS1-N716-B-S502	66	F	98	100	0.962839596	0.435828042
AS1-N716-B-S503	78	F	93	94	0.968629091	0.590661926
AS1-N716-B-S505	89	M	77	77	0.957735934	0.084715307
AS1-N716-B-S507	UNKNOWN	M	72	73	0.946769157	0.063319721
AS1-N716-B-S508	84	F	95	97	0.930240157	0.363339324

AS1-N716-B-S510	60	M	70	69	0.842561438	0.00209954
AS1-N716-B-S511	83	F	91	92	0.858830478	0.249261666
AS1-N716-D-S513	69	F	69	68	0.868089178	0.002667208
AS1-N716-D-S515	71	F	78	78	0.963973747	0.124031874
AS1-N716-D-S516	65	M	81	80	0.941457677	0.129946504
AS1-N716-D-S517	66	M	82	81	0.947617043	0.377578224
AS1-N716-D-S518	70	F	86	86	0.931572343	0.550103842
AS1-N716-D-S520	69	M	92	93	0.969538332	0.594687654
AS1-N716-D-S521	71	F	86	86	0.973957854	0.534002024
AS1-N716-D-S522	72	M	94	93	0.966274348	0.621033794
AS1-N718-B-S505	69	F	X	95	0.969504511	0.521649277
AS1-N718-B-S506	69	M	79	75	0.94165836	0.017692499
AS1-N718-B-S507	72	M	77	77	0.93474691	0.054276742
AS1-N718-B-S508	67	F	86	86	0.939407474	0.341223761
AS1-N718-B-S510	73	M	93	94	0.944928981	0.499875405
AS1-N718-B-S511	86	F	81	79	0.968807399	0.182837624
AS1-N718-D-S513	91	F	94	93	0.948414291	0.482938579
AS1-N718-D-S515	88	M	X	84	0.95658289	0.578303675
AS1-N718-D-S516	71	M	77	76	0.943368579	0.036909429
AS1-N718-D-S517	57	F	75	74	0.905974593	0.004362307
AS1-N718-D-S518	55	F	69	67	0.87477162	0.00182704
AS1-N718-D-S520	62	F	90	90	0.973020114	0.639018778
AS1-N718-D-S521	82	F	77	77	0.96128591	0.120070354
AS1-N718-D-S522	88	F	82	82	0.972727094	0.462205018
AS1-N719-B-S502	66	F	58	58	0.830729997	0.001027167
AS1-N719-B-S503	87	M	74	76	0.969154103	0.310477512
AS1-N719-B-S505	69	F	70	69	0.884363118	0.002519011
AS1-N719-B-S507	86	F	76	74	0.95524055	0.013459336

AS1-N719-B-S508	74	F	90	88	0.972910333	0.559913313
AS1-N719-B-S511	81	F	75	75	0.90579656	0.020234451
AS1-N719-D-S513	67	M	89	88	0.974154523	0.537805884
AS1-N719-D-S515	76	M	78	77	0.905159332	0.218512898
AS1-N719-D-S516	80	M	74	73	0.901498697	0.008217492
AS1-N719-D-S517	75	M	92	91	0.863154043	0.284394438
AS1-N719-D-S518	83	M	80	78	0.892109365	0.100581371
AS1-N719-D-S520	78	F	90	88	0.974826623	0.541002432
AS1-N719-D-S521	61	M	83	83	0.943815154	0.15770116
AS1-N719-D-S522	61	M	86	88	0.960470659	0.421789449
AS1-N720-B-S502	66	M	65	65	0.86396158	0.00197981
AS1-N720-B-S503	68	F	80	78	0.966685896	0.070377848
AS1-N720-B-S505	72	F	53	53	0.891131832	0.00114076
AS1-N720-B-S506	70	M	83	83	0.94588063	0.286488466
AS1-N720-B-S507	71	F	83	85	0.950828064	0.364619091
AS1-N720-B-S510	67	M	76	77	0.948701004	0.040147041
AS1-N720-B-S511	83	F	79	79	0.977883634	0.238759539
AS1-N720-D-S513	84	F	102	101	0.930652174	0.360869565
AS1-N720-D-S515	69	M	84	85	0.97440904	0.211416227
AS1-N720-D-S516	91	M	X	87	0.964629869	0.606165774
AS1-N720-D-S517	65	M	98	98	0.930947287	0.540750827
AS1-N720-D-S518	89	F	92	94	0.927967986	0.294353046
AS1-N720-D-S520	62	F	89	90	0.958892146	0.571081031
AS1-N720-D-S521	65	F	90	90	0.94950838	0.514083768
AS1-N720-D-S522	38	F	83	83	0.967560541	0.301421346
AS1-N721-B-S502	74	M	52	52	0.750593824	0.001543943
AS1-N721-B-S503	78	F	107	107	0.946904469	0.577490775
AS1-N721-B-S505	70	M	97	96	0.955475947	0.379426817

AS1-N721-B-S506	70	F	83	83	0.966426041	0.39358457
AS1-N721-B-S507	74	M	X	97	0.93048554	0.290130008
AS1-N721-B-S508	74	M	94	95	0.947896282	0.342547293
AS1-N721-B-S511	70	M	79	80	0.966827586	0.143678161
AS1-N721-D-S513	76	F	91	91	0.979979913	0.724204888
AS1-N721-D-S515	70	F	X	104	0.938116776	0.605263158
AS1-N721-D-S516	62	M	75	75	0.881764534	0.002553049
AS1-N721-D-S517	72	M	59	59	0.721972333	0.001951965
AS1-N721-D-S518	64	F	77	76	0.947794942	0.034836722
AS1-N721-D-S520	60	F	X	102	0.96120108	0.614934773
AS1-N721-D-S521	91	F	89	89	0.951741584	0.3238957
AS1-N721-D-S522	61	M	81	80	0.972510728	0.104826171
AS1-N722-B-S502	60	F	92	91	0.952546131	0.445572879
AS1-N722-B-S503	77	F	103	103	0.939734121	0.481093058
AS1-N722-B-S505	68	F	103	104	0.960314136	0.665759162
AS1-N722-B-S507	70	M	74	71	0.897483999	0.008018833
AS1-N722-B-S510	67	M	79	80	0.943001077	0.351403587
AS1-N722-D-S513	73	F	86	85	0.970358062	0.356078473
AS1-N722-D-S515	68	F	83	84	0.963462532	0.428656331
AS1-N722-D-S516	59	F	65	64	0.732209909	0.001785878
AS1-N722-D-S517	59	F	86	87	0.975539313	0.431027971
AS1-N722-D-S518	74	F	85	84	0.96137261	0.370999584
AS1-N722-D-S520	61	M	85	85	0.96289417	0.47279328
AS1-N722-D-S521	67	F	89	77	0.959257853	0.127136115
AS1-N722-D-S522	79	F	86	88	0.973906423	0.646940753
AS1-N723-B-S502	80	F	84	82	0.932038835	0.384573894
AS1-N723-B-S503	65	F	109	108	0.919618529	0.473773842
AS1-N723-B-S505	78	M	97	95	0.934319834	0.280373832

AS1-N723-B-S506	33	F	90	85	0.952874859	0.028635851
AS1-N723-B-S507	69	M	80	81	0.936703208	0.238317444
AS1-N723-B-S508	25	M	84	83	0.865009971	0.005471187
AS1-N723-B-S511	64	F	101	101	0.935448874	0.38224392
AS1-N723-D-S513	55	M	71	69	0.840608763	0.002618337
AS1-N723-D-S515	54	M	86	86	0.971434049	0.293663727
AS1-N723-D-S516	63	M	81	76	0.93367484	0.019627029
AS1-N723-D-S517	71	M	81	77	0.94020777	0.029595484
AS1-N723-D-S518	82	M	83	81	0.977870306	0.61353998
AS1-N723-D-S522	59	M	91	89	0.961654894	0.552855701
AS1-N724-B-S502	77	F	85	88	0.978303583	0.638267631
AS1-N724-B-S505	65	F	91	91	0.96345795	0.505207393
AS1-N724-B-S506	68	M	76	75	0.936557847	0.022657912
AS1-N724-B-S507	71	M	79	80	0.963349632	0.064226246
AS1-N724-B-S508	69	M	85	82	0.943372464	0.028336257
AS1-N724-B-S510	77	M	68	65	0.877962244	0.001204981
AS1-N724-B-S511	67	F	98	98	0.924073247	0.239392586
AS1-N724-D-S513	75	F	94	92	0.89901662	0.40709644
AS1-N724-D-S515	65	F	81	78	0.951899974	0.016142201
AS1-N724-D-S516	61	F	75	74	0.911066421	0.003519356
AS1-N724-D-S517	74	F	82	82	0.962489428	0.167305109
AS1-N724-D-S520	67	F	67	60	0.949968886	0.001991288
AS1-N724-D-S521	63	M	82	84	0.976399191	0.176789327
AS1-N724-D-S522	73	F	95	94	0.864598025	0.403102962
AS1-N726-B-S502	85	F	85	85	0.933430255	0.458719106
AS1-N726-B-S503	59	M	84	85	0.965344943	0.335927662
AS1-N726-B-S505	87	M	101	99	0.908249275	0.243540141
AS1-N726-B-S506	80	M	91	89	0.90651037	0.364869252

AS1-N726-B-S507	79	M	81	79	0.957727018	0.326255772
AS1-N726-B-S508	63	F	92	91	0.98156906	0.543918351
AS1-N726-B-S510	95	M	79	78	0.953049743	0.48500425
AS1-N726-B-S511	74	F	103	102	0.946056991	0.487342611
AS1-N726-D-S513	83	F	83	84	0.96348864	0.390992527
AS1-N726-D-S515	50	M	75	74	0.849268001	0.010706105
AS1-N726-D-S516	73	M	102	99	0.970277481	0.307630736
AS1-N726-D-S517	71	M	61	60	0.844844394	0.002849644
AS1-N726-D-S518	76	M	84	85	0.970948592	0.568523431
AS1-N726-D-S521	68	F	103	103	0.939672925	0.548273773
AS1-N726-D-S522	53	F	83	84	0.959019438	0.23036606
AS1-N727-B-S502	54	F	70	67	0.809430473	0.001331361
AS1-N727-B-S503	66	F	92	91	0.973776158	0.556599166
AS1-N727-B-S506	86	M	87	87	0.970949789	0.606486861
AS1-N727-B-S507	70	F	99	98	0.960486891	0.388389513
AS1-N727-B-S508	72	M	73	68	0.870513267	0.002339822
AS1-N727-B-S510	91	F	93	90	0.964456308	0.541444226
AS1-N727-B-S511	83	M	86	78	0.971706298	0.093539192
AS1-N727-D-S513	76	M	92	91	0.923673329	0.516712612
AS1-N727-D-S515	43	F	94	95	0.970836512	0.43120124
AS1-N727-D-S516	58	F	81	78	0.930994392	0.032006565
AS1-N727-D-S517	69	M	72	69	0.827862468	0.012684989
AS1-N727-D-S518	68	M	83	83	0.931462802	0.216531192
AS1-N727-D-S520	54	M	97	96	0.959187364	0.596636122
AS1-N727-D-S521	75	M	72	67	0.90026705	0.001848809
AS1-N728-B-S502	59	M	89	93	0.936541256	0.581714122
AS1-N728-B-S503	56	M	82	80	0.944766651	0.063566508
AS1-N728-B-S505	59	F	86	88	0.965875682	0.337043259

AS1-N728-B-S506	77	F	55	55	0.696040724	0.003280543
AS1-N728-B-S507	69	F	96	96	0.975278785	0.668211743
AS1-N728-B-S508	54	F	93	92	0.979004467	0.664262923
AS1-N728-B-S510	67	F	85	83	0.965001949	0.239138909
AS1-N728-B-S511	71	M	90	92	0.948860105	0.552395396
AS1-N728-D-S513	81	F	82	82	0.943863202	0.479637755
AS1-N728-D-S515	76	M	79	79	0.967320261	0.512233953
AS1-N728-D-S517	69	f	93	94	0.939825168	0.649454496
AS1-N728-D-S518	75	f	80	80	0.950068485	0.206481136
AS1-N728-D-S520	63	F	92	97	0.967433408	0.705800775
AS1-N728-D-S521	70	F	87	88	0.976629274	0.554420406
AS1-N728-D-S522	77	F	75	76	0.923583439	0.016668287
AS1-N729-B-S502	77	M	67	65	0.821851353	0.001833712
AS1-N729-B-S503	79	M	76	72	0.921614468	0.007917953
AS1-N729-B-S505	68	F	63	60	0.880745189	0.000939658
AS1-N729-B-S506	72	M	60	60	0.853636085	0.001585353
AS1-N729-B-S507	69	M	89	88	0.9755802	0.603885135
AS1-N729-D-S513	75	F	100	101	0.935655738	0.510245902
AS1-N729-D-S515	72	M	86	85	0.98093693	0.322449807
AS1-N729-D-S516	70	F	98	85	0.985370066	0.105448447
AS1-N729-D-S517	68	M	82	83	0.974024604	0.24268321
AS1-N729-D-S518	71	M	84	85	0.970869005	0.439957573
AS2-N701-A-S502	44	F	94	93	0.930357547	0.537961077
AS2-N701-A-S503	79	F	89	88	0.927991253	0.521555764
AS2-N701-A-S505	58	F	73	70	0.887833169	0.001357354
AS2-N701-A-S506	67	M	85	81	0.947101406	0.039310509
AS2-N701-A-S507	69	F	71	71	0.924849176	0.211153864
AS2-N701-A-S508	88	M	88	88	0.934197887	0.51560999

AS2-N701-A-S510	82	F	82	81	0.956552372	0.4348677
AS2-N701-A-S511	64	M	82	81	0.955923567	0.099981802
AS2-N701-C-S513	74	M	85	84	0.948730922	0.206762796
AS2-N701-C-S515	67	F	69	69	0.883685124	0.000811825
AS2-N701-C-S516	70	F	84	86	0.939189189	0.507475561
AS2-N701-C-S517	66	M	81	78	0.931350017	0.009405856
AS2-N701-C-S520	65	F	91	92	0.946049321	0.223553095
AS2-N701-C-S522	71	F	51	50	0.838605486	0.001230454
AS2-N702-A-S502	84	M	76	72	0.915616889	0.003165158
AS2-N702-A-S503	72	F	77	76	0.928343869	0.039747155
AS2-N702-A-S505	78	F	77	76	0.945414222	0.150665331
AS2-N702-A-S507	72	M	90	89	0.935723536	0.434735192
AS2-N702-A-S508	70	M	X	76	0.963066162	0.094509681
AS2-N702-A-S510	80	M	95	94	0.903738318	0.388785047
AS2-N702-A-S511	63	F	79	75	0.922003508	0.008302475
AS2-N702-C-S513	52	F	79	77	0.93697318	0.008572797
AS2-N702-C-S515	77	F	91	90	0.958026989	0.391717078
AS2-N702-C-S516	52	F	90	89	0.96063376	0.454844607
AS2-N702-C-S517	71	M	90	91	0.93350807	0.394537178
AS2-N702-C-S518	84	M	52	52	0.717976538	0.000406862
AS2-N702-C-S520	68	M	88	87	0.949222011	0.381100569
AS2-N702-C-S521	51	M	86	65	0.883495146	0.000950506
AS2-N702-C-S522	74	F	83	83	0.952503103	0.224079437
AS2-N703-A-S502	69	F	56	54	0.823047914	0.000613909
AS2-N703-A-S503	68	M	103	103	0.946027489	0.417029836
AS2-N703-A-S505	60	F	107	109	0.92325856	0.638577332
AS2-N703-A-S506	71	M	81	80	0.948358629	0.254812588
AS2-N703-A-S507	67	F	79	78	0.972635907	0.097898584

AS2-N703-A-S510	65	F	55	53	0.939148681	0.00069944
AS2-N703-A-S511	67	F	83	81	0.973958997	0.140965236
AS2-N703-C-S513	87	M	54	53	0.857355627	0.000533933
AS2-N703-C-S515	68	F	82	82	0.94598443	0.042790835
AS2-N703-C-S516	87	M	96	97	0.947717444	0.387890437
AS2-N703-C-S517	40	F	95	95	0.972268058	0.554505775
AS2-N703-C-S518	72	M	87	85	0.947517297	0.314915332
AS2-N703-C-S520	65	F	84	84	0.954178626	0.230187405
AS2-N703-C-S521	70	M	85	87	0.968154582	0.443008147
AS2-N703-C-S522	71	F	86	82	0.974633057	0.215645607
AS2-N704-A-S502	51	M	99	99	0.971867271	0.638911789
AS2-N704-A-S503	62	F	91	91	0.956645121	0.627857481
AS2-N704-A-S505	76	M	81	82	0.956021695	0.219499445
AS2-N704-A-S506	60	F	93	92	0.960784314	0.56754902
AS2-N704-A-S507	75	F	78	74	0.926109556	0.006397441
AS2-N704-A-S508	64	F	94	95	0.9138322	0.515563801
AS2-N704-A-S510	74	F	71	69	0.915519477	0.001790234
AS2-N704-A-S511	72	F	92	92	0.942937325	0.609396113
AS2-N704-C-S513	70	F	65	63	0.903578493	0.000791225
AS2-N704-C-S515	45	M	76	76	0.86488324	0.00133018
AS2-N704-C-S516	65	F	X	101	0.941299062	0.521361584
AS2-N704-C-S517	55	M	93	92	0.939834025	0.547079477
AS2-N704-C-S518	67	F	66	64	0.889988999	0.000433377
AS2-N704-C-S520	57	M	80	76	0.950383109	0.004971905
AS2-N704-C-S521	73	F	94	95	0.938508226	0.362402172
AS2-N704-C-S522	73	M	99	98	0.947249264	0.35680326
AS2-N705-A-S502	65	F	84	85	0.947100646	0.254451764
AS2-N705-A-S503	53	F	85	82	0.961722488	0.074417184

AS2-N705-A-S505	68	F	92	92	0.960334029	0.545581072
AS2-N705-A-S506	61	F	89	86	0.964773736	0.142768412
AS2-N705-A-S507	44	F	86	82	0.966783689	0.084211
AS2-N705-A-S508	73	M	76	72	0.942399563	0.004140316
AS2-N705-A-S510	65	F	84	84	0.954325405	0.224689539
AS2-N705-C-S513	70	F	89	88	0.955782598	0.517879575
AS2-N705-C-S515	70	F	96	95	0.967523095	0.601039261
AS2-N705-C-S516	90	F	87	88	0.964156558	0.328770886
AS2-N705-C-S517	56	F	73	68	0.89352518	0.001111838
AS2-N705-C-S518	61	F	73	72	0.925400641	0.000721154
AS2-N705-C-S521	74	F	79	82	0.947302447	0.199709186
AS2-N705-C-S522	71	M	91	88	0.952245092	0.51423137
AS2-N706-A-S502	69	F	60	61	0.846275458	0.000650981
AS2-N706-A-S503	61	F	73	71	0.91556695	0.001597041
AS2-N706-A-S505	75	F	91	89	0.916330275	0.407706422
AS2-N706-A-S506	64	F	101	99	0.881944444	0.189484127
AS2-N706-A-S507	71	F	66	66	0.897490516	0.137642253
AS2-N706-A-S508	59	F	100	102	0.939588101	0.501144165
AS2-N706-A-S510	71	F	88	103	0.937956204	0.502689205
AS2-N706-C-S513	78	M	74	69	0.897844874	0.001333618
AS2-N706-C-S516	81	F	58	56	0.876708917	0.000387447
AS2-N706-C-S517	61	M	98	96	0.974194149	0.659400037
AS2-N706-C-S518	76	F	106	107	0.934195635	0.436216653
AS2-N706-C-S520	53	M	94	93	0.925295508	0.530496454
AS2-N706-C-S521	69	F	83	82	0.967777655	0.195105747
AS2-N707-A-S502	71	M	84	80	0.941784431	0.092680159
AS2-N707-A-S503	46	F	87	87	0.962330991	0.162071205
AS2-N707-A-S505	75	F	58	58	0.844383536	0.001370446

AS2-N707-A-S506	69	F	61	61	0.758641912	0.001128321
AS2-N707-A-S507	48	F	96	96	0.973606033	0.623800274
AS2-N707-A-S508	56	F	105	111	0.932814457	0.673734121
AS2-N707-A-S510	70	F	78	78	0.957643171	0.147745134
AS2-N707-A-S511	87	F	74	75	0.930123541	0.203769715
AS2-N707-C-S513	46	F	73	70	0.845369726	0.001392564
AS2-N707-C-S515	46	F	83	80	0.939761789	0.022821399
AS2-N707-C-S516	72	F	79	77	0.949081927	0.084486504
AS2-N707-C-S517	62	F	73	69	0.86981224	0.001623289
AS2-N707-C-S518	64	F	94	93	0.924886009	0.481521478
AS2-N707-C-S520	69	M	89	89	0.955183429	0.480011616
AS2-N707-C-S521	72	M	59	57	0.885770607	0.000622504
AS2-N707-C-S522	74	M	X	110	0.932767402	0.682173175
AS2-N710-A-S502	54	F	93	95	0.957621925	0.496453901
AS2-N710-A-S503	73	F	86	86	0.963997231	0.297792138
AS2-N710-A-S505	61	M	83	77	0.949804026	0.020978087
AS2-N710-A-S506	76	M	77	76	0.919401148	0.008900738
AS2-N710-A-S507	85	F	84	83	0.906141482	0.263121341
AS2-N710-A-S508	76	F	67	67	0.8036482	0.00122835
AS2-N710-A-S510	66	F	102	102	0.935002103	0.59991586
AS2-N710-A-S511	72	F	57	56	0.852460807	0.000682799
AS2-N710-C-S513	80	M	94	93	0.964071856	0.707869974
AS2-N710-C-S515	71	F	87	88	0.955089025	0.412968376
AS2-N710-C-S516	60	M	76	77	0.940528634	0.103176443
AS2-N710-C-S517	67	M	76	75	0.920199552	0.003038453
AS2-N710-C-S518	71	F	67	67	0.933605745	0.045522945
AS2-N710-C-S520	69	F	68	65	0.880925136	0.001733763
AS2-N710-C-S521	78	M	76	76	0.940468772	0.025451757

AS2-N710-C-S522	62	F	85	87	0.957124376	0.555165404
AS2-N711-A-S502	75	F	91	90	0.947535462	0.500485783
AS2-N711-A-S503	73	F	90	89	0.95857461	0.641128434
AS2-N711-A-S505	65	M	80	80	0.926396834	0.032260688
AS2-N711-A-S506	76	F	69	65	0.888228129	0.000573189
AS2-N711-A-S507	62	M	83	84	0.948946623	0.272528105
AS2-N711-A-S508	66	F	58	58	0.755865136	0.000499161
AS2-N711-A-S510	68	F	88	88	0.967189884	0.580717244
AS2-N711-A-S511	64	M	70	68	0.882579403	0.001572024
AS2-N711-C-S515	76	M	54	54	0.72960373	0.001165501
AS2-N712-A-S502	65	F	65	65	0.948398577	0.003558719
AS2-N712-A-S503	76	F	65	65	0.839284812	0.000858629
AS2-N712-A-S506	68	M	92	92	0.923208416	0.52991453
AS2-N712-A-S507	56	M	111	112	0.942744324	0.764067127
AS2-N712-A-S508	63	F	93	94	0.933368588	0.569539926
AS2-N712-A-S510	81	F	76	78	0.9217317	0.019994304
AS2-N712-C-S517	82	M	65	65	0.857604288	0.001305752
AS2-N712-C-S518	79	M	53	53	0.661157025	0.016528926
AS2-N714-A-S502	59	F	92	91	0.907309721	0.445571008
AS2-N714-A-S503	62	M	75	73	0.860014892	0.000856292
AS2-N714-A-S505	85	M	90	88	0.864955501	0.405391461
AS2-N714-A-S506	67	F	82	82	0.952527161	0.214643363
AS2-N714-A-S507	71	F	62	61	0.849361443	0.043888101
AS2-N714-A-S508	60	F	83	82	0.949653846	0.166846154
AS2-N714-A-S510	69	F	83	83	0.953891164	0.303101229
AS2-N714-A-S511	62	M	50	50	0.656596919	0.000403687
AS2-N715-A-S502	52	F	83	83	0.958487624	0.17053206
AS2-N715-A-S503	65	F	76	73	0.893459651	0.002533957

AS2-N715-A-S505	74	F	99	99	0.913150147	0.353287537
AS2-N715-A-S506	63	F	56	55	0.769983793	0.000324149
AS2-N715-A-S507	64	F	93	92	0.966747279	0.53030532
AS2-N716-B-S502	71	F	81	80	0.965554532	0.087034886
AS2-N716-B-S503	53	F	77	76	0.929381736	0.037719796
AS2-N716-B-S505	72	M	99	99	0.886346301	0.250572082
AS2-N716-B-S506	71	M	89	88	0.933587085	0.384214901
AS2-N716-B-S507	54	F	90	90	0.944033986	0.388991504
AS2-N716-B-S508	77	M	63	61	0.741542809	0.000815151
AS2-N716-B-S510	86	F	83	78	0.953425194	0.039367092
AS2-N716-B-S511	71	F	89	88	0.954551603	0.547725326
AS2-N718-B-S502	67	M	81	83	0.947854291	0.114520958
AS2-N718-B-S503	61	M	91	90	0.958085612	0.543400713
AS2-N718-B-S505	58	F	78	79	0.943399627	0.208916662
AS2-N718-B-S506	69	F	82	83	0.935253867	0.348053591
AS2-N718-B-S507	64	M	90	89	0.954261821	0.555033773
AS2-N718-B-S508	69	M	67	63	0.89965292	0.00116374
AS2-N718-B-S510	62	M	71	68	0.847637482	0.00082956
AS2-N718-B-S511	88	F	53	53	0.762224711	0.00933184
AS2-N718-D-S513	81	F	95	95	0.856492027	0.225512528
AS2-N718-D-S515	79	F	89	89	0.933107705	0.356957993
AS2-N718-D-S516	75	F	84	84	0.969992898	0.111860795
AS2-N718-D-S517	72	F	63	63	0.776346856	0.001609788
AS2-N718-D-S518	82	M	61	61	0.853636028	0.00036914
AS2-N718-D-S520	80	F	42	42	0.289852465	0.003418496
AS2-N719-B-S502	62	F	103	104	0.967660484	0.608623871
AS2-N719-B-S503	85	F	74	72	0.939715395	0.011556705
AS2-N719-B-S505	70	M	71	69	0.852304275	0.00079321

AS2-N719-B-S506	82	M	71	69	0.861266294	0.000850134
AS2-N719-B-S507	62	M	99	101	0.922776573	0.453362256
AS2-N719-B-S508	82	F	91	90	0.924842226	0.388028304
AS2-N719-B-S510	70	F	76	75	0.949403782	0.060083009
AS2-N719-B-S511	70	F	90	90	0.837137569	0.297347316
AS2-N719-D-S513	58	M	69	69	0.835896849	0.001772543
AS2-N719-D-S515	69	F	79	79	0.921134556	0.010377032
AS2-N719-D-S516	75	F	71	71	0.868834805	0.00107689
AS2-N719-D-S517	88	M	79	79	0.957687723	0.084028605
AS2-N719-D-S518	73	M	82	82	0.931776557	0.14514652
AS2-N719-D-S520	68	M	82	82	0.944562397	0.159471573
AS2-N719-D-S521	86	F	89	89	0.959648058	0.479368932
AS2-N719-D-S522	66	M	70	69	0.865258053	0.001731902
AS2-N720-B-S502	74	F	88	87	0.945415481	0.457464723
AS2-N720-B-S503	68	M	74	75	0.899958392	0.002158423
AS2-N720-B-S505	77	M	80	76	0.955245109	0.025696135
AS2-N720-B-S506	69	M	62	60	0.779802098	0.000594556
AS2-N720-B-S507	65	F	69	69	0.89345417	0.228446657
AS2-N720-B-S508	67	F	98	99	0.89093702	0.464669739
AS2-N720-B-S510	65	F	59	59	0.767968181	0.000672231
AS2-N720-B-S511	72	M	87	87	0.981134314	0.588277171
AS2-N720-D-S513	77	F	80	80	0.92555332	0.092555332
AS2-N720-D-S515	87	F	56	56	0.797947629	0.000353857
AS2-N720-D-S516	78	F	64	64	0.776103714	0.001751927
AS2-N720-D-S518	78	F	74	74	0.900900901	0.033333333
AS2-N720-D-S520	83	F	57	57	0.794612795	0.001224365
AS2-N720-D-S521	87	M	74	74	0.838541667	0.019097222
AS2-N720-D-S522	68	F	94	94	0.670846395	0.23092999

AS2-N721-B-S502	69	M	91	90	0.914864475	0.355680507
AS2-N721-B-S503	65	M	81	78	0.963921097	0.089841756
AS2-N721-B-S505	64	F	93	94	0.950933388	0.529327757
AS2-N721-B-S506	78	F	106	105	0.958465945	0.640356461
AS2-N721-B-S507	69	M	83	82	0.950406966	0.247113383
AS2-N721-B-S508	68	F	88	87	0.969411434	0.443733446
AS2-N721-B-S510	70	M	87	87	0.942331641	0.416514992
AS2-N721-B-S511	86	F	78	78	0.903945965	0.219480981
AS2-N721-D-S513	74	F	81	81	0.953414634	0.135176152
AS2-N721-D-S515	88	F	X	90	0.935169988	0.25123095
AS2-N721-D-S516	66	M	71	70	0.901955641	0.00332255
AS2-N721-D-S517	31	F	86	88	0.961333661	0.13664291
AS2-N721-D-S518	61	F	100	101	0.917485809	0.351100651
AS2-N721-D-S520	73	M	94	93	0.939314922	0.475524476
AS2-N722-B-S502	78	F	65	65	0.831849057	0.001735849
AS2-N722-B-S503	71	F	110	111	0.923597294	0.578989256
AS2-N722-B-S505	68	F	64	62	0.851296236	0.000656322
AS2-N722-B-S506	71	M	83	81	0.962840663	0.144859409
AS2-N722-B-S507	58	F	55	55	0.721285324	0.001937101
AS2-N722-B-S508	72	F	85	85	0.926829268	0.449805198
AS2-N722-B-S510	86	M	75	75	0.88662889	0.004186401
AS2-N722-B-S511	67	F	71	69	0.866192725	0.001217315
AS2-N722-D-S513	65	F	93	92	0.927161036	0.462483545
AS2-N722-D-S515	78	F	64	62	0.783322328	0.000377323
AS2-N722-D-S516	60	F	103	103	0.956677665	0.740926696
AS2-N722-D-S520	72	M	80	79	0.949012806	0.084950105
AS2-N723-B-S502	71	F	89	89	0.970428325	0.597997613
AS2-N723-B-S503	82	M	83	83	0.932581913	0.379554391

AS2-N723-B-S505	66	F	72	72	0.887208194	0.002560743
AS2-N723-B-S506	80	F	86	86	0.945870708	0.509743273
AS2-N723-B-S507	81	M	82	75	0.969183922	0.028217621
AS2-N723-B-S508	73	F	104	105	0.948329448	0.473193473
AS2-N723-B-S510	64	M	83	82	0.955226077	0.153305618
AS2-N723-B-S511	67	F	86	86	0.960494881	0.513566553
AS2-N723-D-S513	67	F	84	82	0.945912043	0.127840909
AS2-N723-D-S516	62	F	86	85	0.947856133	0.370975342
AS2-N723-D-S518	78	F	86	87	0.914990421	0.168821839
AS2-N723-D-S520	89	F	87	85	0.963593059	0.509794391
AS2-N724-B-S502	74	M	57	55	0.908605256	0.175750794
AS2-N724-B-S503	68	F	76	72	0.897733026	0.002418856
AS2-N724-B-S505	55	M	93	94	0.955519684	0.606672047
AS2-N724-B-S506	77	F	88	82	0.960030261	0.281679486
AS2-N724-B-S507	80	F	86	86	0.949196326	0.454793341
AS2-N724-B-S508	60	F	63	63	0.804335033	0.001422126
AS2-N724-B-S510	79	F	95	95	0.935697115	0.493389423
AS2-N724-B-S511	76	F	81	80	0.971490607	0.237373737
AS2-N726-B-S502	81	M	57	58	0.861391606	0.000788892
AS2-N726-B-S503	72	F	65	62	0.862404416	0.001085924
AS2-N726-B-S505	73	M	100	101	0.931755757	0.485315867
AS2-N726-B-S506	68	F	78	78	0.946444182	0.050110557
AS2-N726-B-S507	70	F	76	76	0.942110263	0.06521406
AS2-N726-B-S508	82	F	72	71	0.878993224	0.001075616
AS2-N726-B-S510	77	F	97	96	0.967231387	0.40637539
AS2-N726-B-S511	70	F	84	84	0.966467805	0.327967496
AS2-N727-B-S502	74	F	88	86	0.964104135	0.400980092
AS2-N727-B-S503	68	F	103	106	0.936492428	0.388373229

AS2-N727-B-S505	77	F	82	81	0.970883337	0.231884058
AS2-N727-B-S506	78	F	61	61	0.846234548	0.002052639
AS2-N727-B-S507	57	M	86	87	0.965653833	0.351625525
AS2-N727-B-S508	54	F	83	85	0.95074642	0.232118073
AS2-N727-B-S510	64	F	84	83	0.923194764	0.373233062
AS2-N727-B-S511	67	F	62	60	0.743329755	0.002027914
AS2-N727-D-S513	78	M	81	79	0.933595103	0.217432729
AS2-N727-D-S518	62	F	72	71	0.845229932	0.001267585
AS2-N728-B-S503	84	F	51	49	0.79087148	0.000467103
AS2-N728-B-S505	82	F	90	89	0.894496593	0.196412498
AS2-N728-B-S507	61	F	65	63	0.854756028	0.001898614
AS2-N728-B-S508	73	M	80	78	0.898054105	0.001993355
AS2-N728-B-S510	73	M	102	99	0.955194623	0.292355083
AS2-N728-B-S511	68	F	61	61	0.904641584	0.097815929
AS2-N728-D-S513	58	F	86	85	0.959143281	0.322612568
AS2-N728-D-S515	71	F	74	71	0.911448913	0.002193573
AS2-N728-D-S516	64	F	76	77	0.93669715	0.04398107
AS2-N728-D-S517	69	F	84	81	0.938692227	0.024200437
AS2-N729-B-S502	62	F	69	61	0.887125841	0.000996293
AS2-N729-B-S503	62	F	89	91	0.978735687	0.453045319
AS2-N729-B-S505	53	F	85	84	0.943219722	0.019246252
AS2-N729-B-S506	65	F	89	91	0.94074358	0.584668455
AS2-N729-B-S507	73	F	X	98	0.933842987	0.579535431
AS2-N729-D-S513	85	F	62	61	0.872481423	0.000779632
STR: short tandem repeat assay, ePAL; estimated progenitor allele length, F: female, M: male, X: expanded allele confirmed by triplet primed PCR						

Table S5 Linear regression models of relationships between CTG18.1 allele length and age for allele lengths <79.

Explanatory variable for proportion of reads greater than ePAL (<79)	AIC	Adjusted R^2
CTG + Age	-609.9149	0.678
CTG + Age + (CTG x Age)	-609.2692	0.679
CTG + Age + (CTG x Age) + CTG^2 + Age^2	-608.0847	0.681
CTG + Age + (CTG x Age) + CTG^2 + Age^2 + (CTG x Age^2) + (Age x CTG^2)	-607.6873	0.683
CTG + Age + (CTG x Age) + CTG^2 + Age^2 + (CTG x Age^2) + (Age x CTG^2) + (CTG^2 x Age^2)	-608.7557	0.687
AIC: Akaike information criterion		

Table S6 Linear regression models of relationships between CTG18.1 allele length and age for allele lengths >80.

Explanatory variable for proportion of reads greater than 166 from ePAL (>80)	AIC	Adjusted R^2
CTG + Age	-293.9071	0.3277
CTG + Age + (CTG x Age)	-329.003	0.4053
CTG + Age + (CTG x Age) + CTG^2 + Age^2	-372.2355	0.4901
CTG + Age + (CTG x Age) + CTG^2 + Age^2 + (CTG x Age^2) + (Age x CTG^2)	-369.8443	0.4893
CTG + Age + (CTG x Age) + CTG^2 + Age^2 + (CTG x Age^2) + (Age x CTG^2) + (CTG^2 x Age^2)	-368.596	0.4888
AIC: Akaike information criterion		

Table S7Summary of rare synonymous or missense variants predicted not to be deleterious identified in FECD-associated genes and GWAS-hit genes from a total of 141 FECD cases analysed by exome sequencing. This table includes variants with a synonymous effect or missense variants with a CADD score <10. MAF < 0.01 in publicly available gnomAD genomes, exomes and Kaviar was used to determine rarity.

Subject ID	Functional change	Genomic co-ordinates (Hg19)	Change	In silco predictions			GnomAD Frequency (genomes)	UCLex Frequency (AC/AN)	UCLex Frequency without FECD cases	Variant interpretation
				CADD	DANN	Reveal				
COL8A2 (ENST00000397799.2) **247.16TPM**										
BR68	S	1-36564622-C-T	c.660G>A, p.(Gly220=)	9.399	0.534	0	0.0003946 (90/228066)	0.0010799 (11/10186)	0.000981932 (10/10184)	Likely Benign
CZ42	S	1-36565748-G-A	c.96C>T, p.(Ala32=)	3.79	0.866	0	0.0007739 (215/277826)	0.0005612 (6/10692)	0.000467727 (5/10690)	Likely Benign
CZ36; BR85	S	1-36565822-G-A	c.22C>T, p.(Leu8=)	7.071	0.742	0	0.003617 (758/209580)	0.0013693 (14/10224)	0.001174168 (12/10220)	Likely Benign
BR78	S	1-36564985-T-C	c.297A>G, p.(Lys99=)	9.831	0.417	0	0.0002552 (50/195904)	0.0001134 (1/8820)	0 (0/8818)	Likely Benign
SLC4A11 (ENST00000380059) 2938.99TPM										
BR6	MS	20-3212035-C-G	c.1018G>C, p.(Val340Leu)	5.134	0.757	0.156	0 (0/0)	0.0000890 (1/11240)	0 (0/11238)	Likely Benign
BR47	MS	20-3212041-C-T	c.1012G>A, p.(Gly338Ser)	0.065	0.774	0.156	0.0001521 (43/282718)	0.0000890 (1/11242)	0 (0/11240)	Likely Benign
CZ29	S	20-3210241-G-A	c.1800C>T, p.(Thr600=)	0.048	0.609	0	0.00007246 (18/248426)	0.0000898 (1/11134)	0 (0/11132)	Likely Benign
BR52	S	20-3209540-G-A	c.2265C>T, p.(His755=)	2.707	0.508	0	0.00008505 (24/282182)	0.0000893 (1/11202)	0 (0/11200)	Likely Benign

CZ9	S	20-3210190-G-C	c.1851C>G, p.(Thr617=)	9.204	0.592	0	0.003115 (874/280594)	0.0054435 (61/11206)	0.00535523 (60/11204)	Likely Benign
BR70	S	20-3209254-G-A	c.2421C>T, p.(Pro807=)	11.28	0.899	0	0.001727 (487/282058)	0.0003590 (4/11142)	0.0002693 (3/11140)	Likely Benign
BR38; BR63	S	20-3208451-G-A	c.2739C>T, p.(Asp913=)	8.915	0.856	0	0.002799 (786/280774)	0.0005560 (6/10792)	0.000370782 (4/10788)	Likely Benign
BR82	S	20-3209815-G-A	c.2073C>T, p.(Leu691=)	6.501	0.814	0	0.0001424 (40/280942)	0.0001865 (2/10726)	9.32488E-05 (1/10724)	Likely Benign
ZEB1 (ENST00000361642) **4.11TPM**										
BR6;BR32; BR63;BR38+	MS	10-31750140-A-C	c.233A>C, p.(Asn78Thr)	7.849	0.957	0.187	0.00511 (1441/281942)	0.0016077 (18/11196)	0.001161959 (13/11188)	Likely Benign
CZ7	S	10-31812962-T-C	c.2706T>C, p.(Asn902=)	10.21	0.632	0	0.0005280 (149/282214)	0.0012186 (13/10668)	0.00112507 (12/10666)	Likely Benign
BR6;BR32; BR63;BR38+	S	10-31799722-T-C	c.606T>C, p.(Ser202=)	11.32	0.702	0	0.005698 (1610/282548)	0.0017531 (19/10838)	0.001292705 (14/10830)	Likely Benign
BR6;BR32; BR63;BR38+	S	10-31809520-G-A	c.1260G>A, p.(Ala420=)	8.982	0.651	0	0.005941 (1675/281926)	0.0017879 (20/11186)	0.001341922 (15/11178)	Likely Benign
BR21; BR34; BR66	S	10-31810324-A-C	c.2064A>C, p.(Pro688=)	4.663	0.416	0	0.006258 (1766/282180)	0.0025714 (29/11278)	0.0023066 (26/11272)	Likely Benign
BR59	S	10-31815907-G-A	c.3093G>A, p.(Arg1031=)	12.1	0.554	0	0.0003836 (107/278908)	0.0006237 (7/11224)	0.000534664 (6/11222)	Likely Benign
AGBL1 (ENST00000635782) **0.00TPM**										
BR58	MS	15-86807744-A-G	c.1204A>G, p.(Arg402Gly)	6.295	0.855	0.048	0.001493 (372/249108)	0.0007719 (8/10364)	0.000675545 (7/10362)	Likely Benign

CZ5	MS	15-87066094-G-A	c.2471G>A, p.(Ser824Asn)	1.39	0.222	0.021	0.0001889 (53/280616)	0.0003763 (4/10630)	0.000282273 (3/10628)	Likely Benign
BR56;CZ30	MS	15-87531283-A-G	c.3149A>G, p.(Asn1050Ser)	6.049	0.805	0.037	0.003699 (1028/277938)	0.0063588 (57/8964)	0.006138393 (55/8960)	Likely Benign
BR26;CZ40;CZ33	MS	15-86790997-G-A	c.484G>A, p.(Val162Met)	8.172	0.77	0.046	0.006270 (1756/280080)	0.0073934 (78/10550)	0.007111701 (75/10546)	Likely Benign
BR4; BR45	S	15-86790990-C-T	c.477C>T, p.(Asn159=)	0.541	0.82	0	0.007628 (2135/279902)	0.0044558 (47/10548)	0.00426783 (45/10544)	Likely Benign
BR5	S	15-87089307-G-A	c.2622G>A, p.(Leu874=)	9.275	0.764	0	0.002671 (748/280056)	0.0007564 (8/10576)	0.000662001 (7/10574)	Likely Benign
BR11; BR21; BR66	S	15-86807830-A-C	c.1290A>C, p.(Pro430=)	0.575	0.611	0	0.004406 (1235/280316)	0.0008864 (9/10154)	0.000591191 (6/10149)	Likely Benign
LOXHD1 (ENSP00000300591.6/ENST00000536736.5 *) **0.01TPM**										
CZ12	S	18-44181306-G-A	c.1008C>T, p.(Gly336=)*	8.165	0.648	0	0.00006321 (12/189834)	0.0002791 (3/10748)	0.000186116 (2/10746)	Likely Benign
BR41	S	18-44063671-A-G	c.2887T>C, p.(Leu963=)	3.086	0.546	0	0.0001696 (32/188706)	0.0001920 (2/10418)	9.60061E-05 (1/10416)	Likely Benign
BR21	S	18-44157753-G-A	c.1887C>T, p.(Ser629=)*	0.03	0.69	0	0.0002480 (47/189522)	0.0005654 (6/10612)	0.00038059 (4/10510)	Likely Benign
CZ31; CZ50	S	18-44229132-G-A	c.231C>T, p.(Leu77=)*	10.19	0.729	0	0.0002075 (39/187956)	0.0003851 (4/10386)	0.000192641 (2/10382)	Likely Benign
BR26	S	18-44125338-C-T	c.228G>A, p.(Lys76=)	13.78	0.709	0	0.0002607 (50/191788)	0.0004679 (5/10686)	0.000374392 (4/10684)	Likely Benign
CZ29	S	18-44126946-C-T	c.93G>A, p.(Val31=)	11.54	0.828	0	0.001202 (228/189758)	0.0022422 (24/10704)	0.002149131 (23/10702)	Likely Benign

BR2	S	18-44085811-G-A	c.2535C>T, p.(His845=)	2.209	0.682	0	0.001742 (327/187688)	0.0013451 (14/10408)	0.001249279 (13/10406)	Likely Benign
BR31; BR63; BR47	S	18-44146287-G-A	c.2370C>T, p.(Asp790=)*	10.54	0.733	0	0.009132 (1739/190426)	0.0046236 (50/10814)	0.004348631 (47/10808)	Likely Benign
BR31; BR32; BR52	S	18-44181348-C-G	c.966G>C, p.(Gly322=)*	10.32	0.699	0	0.007262 (1379/189888)	0.0018660 (20/10718)	0.001587005 (17/10712)	Likely Benign
TCF4 (ENST00000566286*/ENST00000544241**/ENST00000354452***/ ENST00000568169****)										
CZ44	S	18-52901846-C-G	c.1419G>T, p.(Pro473=)***	0.438	0.6296	0	0.002203 (623/282806)	0.0070247 (79/11246)	0.006937033 (78/11244)	Likely Benign
BR12	S	18-53331939-T-C	c.51A>G, p.(Gln17=)****	0.575	0.2517	0	0.003278 (520/158650)	0.0067370 (23/3414)	0.006447831 (22/3412)	Likely Benign
KANK4 0.77TPM0										
BR21	MS	1-62739226-C-T	c.1550G>A, p.(Arg517Lys)	0.407	0.746	0.01	0.003996 (1129/282536)	0.0010636 (12/11282)	0.000975177 (11/11280)	Likely Benign
BR5	MS	1-62739979-T-C	c.797A>G, p.(Asp266Gly)	0.725	0.798	0.027	0.0001154 (29/251294)	0.0000888 (1/11264)	0 (0/11262)	Likely Benign
BR69	MS	1-62739547-G-A	c.1229C>T, p.(Thr410Met)	1.805	0.941	0.048	0.002687 (760/282876)	0.0006190 (7/11308)	0.000530692 (6/11306)	Likely Benign
CZ1	S	1-62739546-C-T	c.1230G>A, p.(Thr410=)	0.067	0.337	0	0.000003976 (1/251482)	0.0001769 (2/11308)	8.84486E-05 (1/11306)	Likely Benign
BR30	S	1-62740743-A-G	c.33T>C, p.(Ser11=)	6.799	0.604	0	0 (0/0)	0.0000907 (1/11024)	0 (0/11022)	Likely Benign
BR81	S	1-62739864-C-T	c.912G>A, p.(Glu304=)	6.65	0.394	0	0.0007287 (206/282688)	0.0008908 (10/11226)	0.000801853 (9/11224)	Likely Benign

***LAMC1 (*ENST00000258341) 13.71TPM**										
BR46	S	1-183086817-C-T	c.1836C>T, p.(Gly612=)	12.63	0.671	0	0.000007071 (2/282852)	0.0000899 (1/11124)	0 (0/11122)	Likely Benign
BR46	S	1-182993133-C-T	c.282C>T, p.(Ala94=)	15.21	0.929	0	0.0004018 (112/278758)	0.0006364 (7/11000)	0.000545554 (6/10998)	Likely Benign
BR77	S	1-183093917-G-A	c.2553G>A, p.(Lys851=)	9.5	0.6	0	0.001966 (556/282828)	0.0028455 (32/11246)	0.002757026 (31/11244)	Likely Benign
BR81	S	1-183096425-C-T	c.3009C>T, p.(Cys1003=)	12.28	0.661	0	0.002139 (605/282860)	0.0041593 (47/11300)	0.004071517 (46/11298)	Likely Benign
BR56	S	1-183102616-C-G	c.3780C>G, p.(Ala1260=)	0.753	0.502	0	0.001928 (545/282640)	0.0025958 (29/11172)	0.002506714 (28/11170)	Likely Benign
BR66;BR69	S	1-183077455-G-C	c.768G>C, p.(Leu256=)	9.948	0.765	0	0.002033 (575/282776)	0.0007866 (8/10170)	0.000590203 (6/10166)	Likely Benign
BR63;BR69	S	1-183106956-A-C	c.4467A>C , p.(Ala1489=)	10.84	0.786	0	0.009249 (2516/272018)	0.0015788 (17/10768)	0.001393534 (15/10764)	Likely Benign
BR69	S	1-183111892-C-T	c.4797C>T, p.(Gly1599=)	11.44	0.875	0	0.002694 (759/281774)	0.001479805 (17/11488)	0.001393 (16/11486)	Likely Benign
BR31	S	1-183079650-G-A	c.882G>A, p.(Glu294=)	12.64	0.47	0	0.009199 (2598/282432)	0.006460388 (76/11764)	0.006376467 (75/11762)	Likely Benign
BR31	S	1-183079662-C-T	c.894C>T, p.(Asn298=)	7.024	0.613	0	0.009224 (2605/282424)	0.0061976 (68/10972)	0.006107566 (67/10970)	Likely Benign
BR31	S	1-183079767-G-A	c.999G>A, p.(Ala333=)	7.901	0.753	0	0.009264 (2616/282390)	0.0062995 (70/11112)	0.006210621 (69/11110)	Likely Benign
***ATP1B1 (*ENST00000367816) 77TPM**										
BR2	S	1-169080732-G-A	c.222G>A, p.(Pro74=)	5.201	0.645	0	0.001269 (358/282112)	0.0014342 (16/11156)	0.001344809 (15/11154)	Likely Benign

BR67, BR88	S	1-169094216-G-A	c.321G>A, p.(Arg107=)	11.51	0.589	0	0.006824 (1927/282402)	0.0088057 (96/10902)	0.008625436 (94/10898)	Likely Benign

MS: missense, S: synonymous variant; TPM: transcripts per million, CADD: Combined Annotation Dependent Depletion, FECD: Fuchs endothelial corneal dystrophy, MAF: minor allele frequency, gnomAD: The Genome Aggregation Database, AC/AN: allele count/allele number, +: homozygous.

Table S8Top 50 genes significantly enriched for rare and potentially deleterious variants within a cohort of European FECD cases compared to European controls derived from UCLex exome consortium dataset identified in a SKAT and custom gene burden analysis (P-value < 0.05) using Condition 1. CADD score >20, MAF < 0.01 (gnomAD exomes MAF, Kaviar MAF) were applied.

Gene	P-value in Custom gene burden	P-value in SKAT gene burden	Number of variants in cases	Number of variants in controls
HMOX2	9.94E-05	3.52E-06	5	0
OR11L1	0.017364533	4.98E-05	5	2
FAM228A	0.000684697	6.77E-05	3	0
PRELID2	0.000792283	8.28E-05	4	2
RBM26	0.003009863	8.82E-05	7	11
NR4A1	0.006334701	0.000120119	8	10
UTP11L	0.02718592	0.000138312	3	2
METTL4	0.003187147	0.000152282	4	2
LRRC3	0.002554743	0.000356067	13	30
HNRNPM	0.000758653	0.000365618	10	20
MIR184	0.007704388	0.000407162	2	0
HFM1	0.016071161	0.000414485	6	15
PAICS	0.007704388	0.000530946	2	0
SH2D3C	0.002271611	0.000550713	7	15
N6AMT1	0.01818638	0.000669587	3	1
ADH4	0.000792283	0.000764656	5	2
MIER1	0.000792283	0.000826321	4	1
TMTC1	0.002271611	0.000911485	7	11
NDUFA2	0.007704388	0.000957069	2	0
COX6A1	0.00595999	0.000974613	3	1
LGALS4	0.011127697	0.000993921	3	1

SLBP	0.000284578	0.001045067	4	1
FBLN2	0.001528068	0.001087943	10	16
XYLB	0.006334701	0.001105905	6	11
KPNA1	0.002554743	0.001128982	3	1
FIBIN	0.00595999	0.001138593	3	1
HSD17B3	0.001600582	0.001201082	5	5
OAZ1	0.008196309	0.001411949	6	14
SLC2A4RG	0.003187147	0.001440683	4	2
OVCH2	0.038114445	0.001505917	3	1
SLC23A2	0.011127697	0.001519727	3	0
TUBGCP5	0.001680682	0.001522533	7	12
GATA4	0.00595999	0.001663229	3	1
CARHSP1	0.00532887	0.001724145	4	4
UTP6	0.013032732	0.001811986	6	15
RPL8	0.007704388	0.001826233	2	0
PDIA6	0.02718592	0.001896574	3	2
ITGB3BP	0.007704388	0.002029111	2	0
PEX19	0.011127697	0.002159955	3	1
CBLL1	0.007704388	0.002179655	3	1
MEGF11	0.044596593	0.002193042	6	13
SMO	0.003554585	0.00230079	10	35
HADH	0.003554585	0.002351903	13	34
TMEM184A	0.022583824	0.002351983	4	3
BSCL2	0.00595999	0.002406612	4	1
VLDLR	0.004827961	0.002412697	8	15
INSC	0.03851929	0.002569833	6	12
NREP	0.007704388	0.002798474	2	0
CEP95	0.002554743	0.002821578	3	1

HEPACAM	0.040932795	0.002865797	5	8

Table S9 Top 50 genes significantly enriched for rare and potentially deleterious variants within a cohort of European FECD cases compared to European controls derived from UCLex exome consortium dataset identified in a SKAT and custom gene burden analysis (P-value < 0.05) using Condition 2. CADD score >20, MAF < 0.001 (gnomAD exomes MAF, Kaviar MAF) were applied.

Gene	P-value in Custom gene burden	P-value in SKAT gene burden	Number of variants in cases	Number of variants in controls
FAM228A	0.000684697	7.92E-05	3	0
HFM1	0.006334701	8.32E-05	6	12
NRG1	0.046881895	9.53E-05	5	16
TCF25	0.039523381	0.000317008	5	15
MIR184	0.007704388	0.000482491	2	0
CCDC92	0.038114445	0.000498823	3	0
PAICS	0.007704388	0.000504433	2	0
METTL4	0.00595999	0.000538633	3	0
MORC1	0.038114445	0.000709383	3	6
ADH4	0.000792283	0.000760998	5	2
HADH	6.13E-05	0.000854527	9	18
NDUFA2	0.007704388	0.000969202	2	0
COX6A1	0.00595999	0.000977966	3	1
N6AMT1	0.01818638	0.001101963	3	1
LGALS4	0.00595999	0.001171963	3	1
ZNF226	0.001716119	0.001178344	4	2
SARM1	0.011127697	0.001183983	3	1
NR4A1	0.04634194	0.001237498	6	8

CCT6B	0.003187147	0.001268009	5	1
SMO	0.003554585	0.001358351	7	24
SLC23A2	0.011127697	0.001402979	3	0
ZP2	0.01818638	0.001533721	4	3
KPNA1	0.002554743	0.001572182	3	1
GATA4	0.002554743	0.001812445	3	1
FMO3	0.005503139	0.001877934	3	1
ITGB3BP	0.007704388	0.002011892	2	0
C1orf56	0.002554743	0.002093885	3	0
TRIM16	0.000792283	0.002132211	5	4
PEX19	0.011127697	0.002206219	3	1
TMEM184A	0.022583824	0.002345327	4	3
PDIA6	0.02718592	0.002375701	3	2
BSCL2	0.00595999	0.00275514	4	1
AGER	0.003187147	0.002854187	4	3
CCER1	0.007704388	0.002867532	4	3
TRIP10	0.00595999	0.002914439	3	2
SIRT4	0.00532887	0.003082927	4	5
C9orf117	0.02718592	0.003101821	3	3
S100PBP	0.007704388	0.003176591	2	0
MIER1	0.00595999	0.00320643	3	1
PCGF5	0.007704388	0.003233984	2	0
OXER1	0.00595999	0.003372825	3	2
ASPN	0.007704388	0.003402694	2	0
NPRL3	0.011127697	0.003450609	3	2
HIST2H3D	0.011127697	0.00373566	2	0
NEDD4L	0.017364533	0.003798939	6	9
B3GNTL1	0.011127697	0.003845358	3	2

TTC17	0.022583824	0.003864218	4	5
PDE12	0.021743769	0.003952612	2	0
AHSA2	0.021743769	0.003991696	2	0
SLC29A1	0.021743769	0.004037383	2	0

Table S10 Top 50 genes significantly enriched for rare and potentially deleterious variants within a cohort of European FECD cases compared to European controls derived from UCLex exome consortium dataset identified in a SKAT and custom gene burden analysis (P-value < 0.05) using Condition 3. CADD score >20, MAF < 0.0001 (gnomAD exomes MAF, Kaviar MAF) were applied.

Gene	P-value in Custom gene burden	P-value in SKAT gene burden	Number of variants in cases	Number of variants in controls
MIR184	0.007704388	0.000396635	2	0
CCDC92	0.038114445	0.000532772	3	0
METTL4	0.00595999	0.000600907	3	0
SLC26A3	0.002554743	0.000697929	3	1
TRIM16	0.002554743	0.000874909	4	1
GCKR	6.13E-05	0.001136779	3	0
SARM1	0.011127697	0.00132453	3	1
GATA4	0.002554743	0.001341618	3	1
HHIPL2	0.02718592	0.001387732	3	1
KPNA1	0.002554743	0.001482673	3	1
LGALS4	0.00595999	0.001707294	3	1
RHOBTB2	0.008252425	0.001827243	4	5
RSPO3	0.021743769	0.001828646	2	1
INPP5A	0.040932795	0.001872863	2	0
GDAP1	0.02718592	0.002032915	3	4

ARHGEF10L	0.003820386	0.002226388	5	8
FMO3	0.000792283	0.002235548	4	4
HLCS	0.01818638	0.002251361	3	4
SH2D3C	0.000792283	0.002258956	4	2
THRAP3	0.002554743	0.002343782	3	1
TRIP10	0.002554743	0.002366373	3	1
ZDHHC8	0.00595999	0.002520082	3	1
ZNF521	0.038114445	0.0026151	3	2
BSCL2	0.00595999	0.002758408	4	1
PFN1	0.021743769	0.002788803	2	0
OR5P3	0.002554743	0.002816598	3	1
NPRL3	0.011127697	0.002905459	3	1
STK11IP	0.02718592	0.002956139	4	3
S100PBP	0.007704388	0.002969467	2	0
UTP14C	0.007704388	0.002983908	2	0
UGT3A2	0.007704388	0.00307603	2	0
ZNF652	0.00595999	0.003120193	3	2
CCER1	0.007704388	0.003125466	2	0
OR52W1	0.021743769	0.003155286	2	0
ZNF557	0.021743769	0.003200743	2	0
C7orf31	0.040932795	0.003247261	2	1
SEPN1	0.007704388	0.003295645	2	0
FIBIN	0.021743769	0.003324381	2	0
C1QTNF4	0.007704388	0.00341234	2	0
TMEM8C	0.007704388	0.003465298	2	0
CPNE4	0.021743769	0.003555405	2	0
ZNF429	0.007704388	0.003621589	2	0
CCDC170	0.040932795	0.003661003	2	0

TMC3	0.007646013	0.003710104	5	12
SLC29A1	0.021743769	0.003997301	2	0
ZBTB7A	0.021743769	0.004088597	2	0
DNAJC30	6.13E-05	0.004283595	2	0
MICU3	0.040932795	0.004426469	2	0
HADH	0.000684697	0.00483034	7	15
MBLAC2	0.007704388	0.004862438	2	0

Table S11 Top 50 genes significantly enriched for rare and potentially deleterious variants within a cohort of European FECD cases compared to European controls derived from UCLex exome consortium dataset identified in a SKAT and custom gene burden analysis (P-value < 0.05) using Condition 4. CADD score >10, MAF < 0.001 (gnomAD exomes MAF, Kaviar MAF) were applied.

Gene	P-value in Custom gene burden	P-value in SKAT gene burden	Number of variants in cases	Number of variants in controls
HRNR	0.017958938	5.23E-06	9	18
FAM228A	0.000284578	1.72E-05	4	0
TRIM16	8.84E-05	5.33E-05	8	9
TCF25	0.002985476	7.89E-05	9	19
SLC35B3	0.016071161	0.000128107	5	15
CGREF1	0.000509993	0.000150425	8	4
IGLV5-48	0.007704388	0.000166881	2	0
CCDC92	0.037332493	0.000257782	4	1
ACOX3	0.010418071	0.000258151	6	11
ANXA10	0.008252425	0.000337727	4	4
PLEKHA1	0.016810458	0.000391475	5	6

DNAJB5	0.017364533	0.000428295	5	5
MIR184	0.007704388	0.000433724	2	0
METTL4	0.011127697	0.000495124	3	0
MKS1	0.000390858	0.000552237	10	20
OR7E36P	0.00595999	0.000561272	3	1
CA14	0.019576972	0.000700808	8	14
NDUFS1	0.008682227	0.000729661	8	20
AAMP	0.013528436	0.000737106	5	7
ARHGEF10L	0.003335327	0.000751667	11	30
CDC14B	0.001798106	0.000839913	5	4
TK1	0.007704388	0.000895361	23	101
PCGF5	0.00121739	0.000960799	6	8
HTR1F	0.001716119	0.000979462	4	3
OR5P3	0.002539515	0.001004725	6	6
CBLL1	0.016810458	0.001129399	4	5
TSPYL4	0.008252425	0.001219279	5	3
MYF5	0.037332493	0.001234568	3	1
CAGE1	0.017364533	0.001426646	7	14
LAYN	0.001798106	0.00147366	6	7
DUSP23	0.000684697	0.001513843	3	0
SMIM4	0.00595999	0.001791865	3	2
ADH4	0.001716119	0.001810976	5	3
C16orf82	0.011127697	0.001814781	2	1
MSTO1	0.010304153	0.001970539	5	9
HADH	0.000284578	0.001981303	13	34
PITPNA	0.003187147	0.002127056	4	4
THY1	0.002554743	0.002290483	3	1
SLBP	0.00595999	0.002291518	3	2

FYTTD1	0.00532887	0.002341545	4	5
PDCD11	0.00807255	0.0023425	12	36
RP11-21L19.1	0.02718592	0.002441022	2	0
C19orf24	0.021743769	0.002565103	2	0
ZNF793	0.00532887	0.002669756	4	4
SMARCA5	0.003009863	0.002703279	7	16
TMPRSS7	0.00788579	0.002789653	7	16
PSMB2	0.00595999	0.002814298	3	2
HIST4H4	0.007704388	0.002857439	2	0
CCER1	0.007704388	0.002857441	2	0
SIRT4	0.012053477	0.002866935	4	5

Table S12 Details of who conducted each piece of work within collaborative projects

MiSeq expermiment	
MiSeq library preparation	Amanda sadan @ UCL
MiSeq library quantification by quibit	Amanda sadan @ UCL
MiSeq library quantification by Bioanalyser	Dr. Marc Ciosi @ University of Glasgow
MiSeq sequencing	Glasgow Polyomics @Univerisrty of Glasgow
Preparation of MiSeq data; normalisation, cutadapt, trimming, etc.	Amanda sadan @ UCL
Processing repeat genotype (RGT)	Dr. Vilija Lomeikaite and Dr. Marc Ciosi @ University of Glasgow
Alignment of MiSeq reads	Amanda sadan @ UCL
Interpretation of MiSeq data; calling ePAL, genotype-phenotype association analysis	Amanda sadan @ UCL
Quantifying CTG18.1 somatic expansion levels	Dr. Marc Ciosi @ University of Glasgow
Interpretation of CTG18.1 somatic expansion levels	Amanda sadan @ UCL
Kompetitive Allele Specific (KASP) assay	
KASP library preparation	Amanda sadan @ UCL
KASP genotyping	Outsourced to the LGC group Twickenham, UK
KASP genotype interpretation and analysis	Amanda sadan @ UCL
Exome data analysis	
Predicting ancestry of patients using a genome-wide SNP array	Anita Szabo @UCL
Generation of transcript per million reads mapped (TPM) gene expression levels using RNA-Seq data	Dr. Nathaniel Hafford Tear @UCL
Library preparation, exome capture and sequencing	Outsourced to Novogene
Alignment, variant calling, and annotation of exome data and variant calling	Dr. Nikolas Pontikos and Anita Szabo @ UCL

Exome sequencing data interpretation	Amanda sadan @ UCL
PCA ancestry prediction	Dr Cian Murphy @UCL
Producing gene burden analysis data	Dr Cian Murphy, Dr. Nikolas Pontikos and Anita Szabo @UCL
Interpretation of Gene burden data	Amanda sadan @ UCL