Investigations of the CEL VNTR by cellular experiments and *in silico* analyses

Rune Haugland Navarsete

Department of Biological Sciences, University of Bergen

Spring 2023

supervised by

Anders Molven[1], Karianne Fjeld[1] and Nathalie Reuter[2]

[1]Gade Laboratory for Pathology, Department of Clinical Medicine, University of Bergen

[2]Computational Biology Unit, Department of Chemistry, University of Bergen

This thesis is submitted in partial fulfilment of the requirements for

the degree of Master of Science

Contents

1 Acknowledgements

The work behind this thesis required about a year of my most sincere attention and effort. However, it would not have been remotely possible for me to complete this task, if not for all the supervision and assistance I received on the way. I will thank everyone now.

First of all, I want to thank my family for encouraging (read ”forcing”) me to take breaks, as I rarely do that on my own volition. Then, I want to thank my friends Benjamin, Elisa, Birgitte and Jørgen. Thank you to Bente, Harsh, Himal, Mahmoud, Monika and Rammah for their assistance in practical endeavours. I want to thank Louise who guided me in my very first days in the lab. Thank you Martin for generously sharing your agar plates and medium. Thank you Miguel for your experienced advice and for the many times you showed me where to find what I needed. Thank you Khadija for advising me on the glycans. Thank you Maria and Sarah for reviewing the Nynorsk version of my abstract.

I want to thank Solrun for guiding me whenever I was in the PCR laboratory. I also want to thank Janniche for spending several weeks of her time to train me in cell culture and Western blotting, and for taking care of my cells when I was sick. I want to thank Reza for helping me with the complicated matter of molecular dynamics, and even worse; using Windows in bioinformatics. My meetings with you were incredibly productive and helpful. Ranveig, thank you. You shared your results, plasmids and lysates, which helped me to finally uncover why my Western blots were not as expected.

Nathalie, thank you for agreeing to be a part of my project. You patiently guided me through my work on the simulations while I was spending most of my time in the wet lab. Also, I am grateful that you gave me the opportunity to access the strongest computer in the nation. Such tools are not often provided to master students at my programme.

Karianne, thank you for all the advice and encouragement, especially when it came to the grueling issues with the Western blots. If you had not encouraged me to do one final test using a different membrane, I would probably never have figured out what was going wrong. Also, thank you for reading my drafts and giving me feedback. It greatly improved this thesis.

Anders, since the very first meeting you have been exceptionally accommodating towards my thoughts and interests. You were always quick to respond and give me meaningful advice. You kept yourself updated and displayed interest in my progress. As a consequence, I rarely felt lost. Further, you were extremely willing and quick to give me in-depth feedback on my thesis drafts. This thesis would not be the same without you.

Rune Haugland Navarsete

June 1st, 2023

Bergen, Norway

2 Abbreviations

AP	Acute Pancreatitis
AP1	Activator Protein 1
AP2	Activator Protein 2
BSA	Bovine Serum Albumin
CD	Circular Dichroism
C/EBPs	CCAAT-Enhancer-Binding-Proteins
CEL	Carboxyl Ester Lipase
CELP	Carboxyl Ester Lipase Pseudogene
CP	Chronic Pancreatitis
EV	Empty Vector
fs	femtoseconds
h3R-DAN-U	human CEL-3R-DAN Unmodified
h3R-DAN-G	human CEL-3R-DAN *O*-glycosylated
HMR	Hydrogen Mass Repartitioning
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
kDa	Kilodaltons
MD	Molecular Dynamics
MIDD	Maternally Inherited Diabetes and Deafness
MODY	Maturity Onset Diabetes of the Young
MOPS	3-(*N*-MOrpholino)PropaneSulfonic acid
MS	Mass Spectrometry
MSA	Multiple Sequence Alignment
MYA	Million Years Ago
NMR	Nuclear Magnetic Resonance
OD	Optical Density
PC	Pancreatic Cancer
PCR	Polymerase Chain Reaction
PDAC	Pancreatic Ductal AdenoCarcinoma
PEST	sequence rich in Pro, Glu, Ser and Thr
PP	Pancreatic Polypeptide
PTM	Post-Translational Modification
PVDF	PolyVinylidene DiFluoride
SAXS	Small-Angle X-ray Scattering
SDS-PAGE	Sodium Dodecyl-Sulphate PolyacrylAmide Gel Electrophoresis
RMSD	Root Mean Square Deviation
RoG	Radius of Gyration
RT	Room Temperature
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
T3cD	Type 3c Diabetes
VNTR	Variable Number of Tandem Repeats
YFP	Yellow Fluorescent Protein
Å	Ångstrøm (= 0.1 nm)

3 Abstract

English

Carboxyl ester lipase (CEL) is a digestive enzyme which is secreted by acinar cells in the pancreas and released into the intestines. In addition to the globular domain responsible for lipase activity, CEL contains a C-terminal region encoded by a Variable Number of Tandem Repeats (VNTR) sequence. Each repeat consists of 11 amino acid residues. The function of the VNTR region is enigmatic, but we know that it is disordered, becomes *O*-glycosylated and is necessary for proper secretion of CEL. The CEL VNTR is also highly polymorphic and at least two variants are confirmed to be pathogenic: CEL-MODY and CEL-HYB. These variants are associated with cellular aggregation and pancreatic disease. A third variant with only three VNTR repeats, CEL-3R, has been discovered in a Danish family, in which it co-segregates with diabetes. However, we do not know if CEL-3R is pathogenic.

In this study, we attempted to investigate the properties of the CEL VNTR with a special emphasis on the Danish CEL-3R variant. To examine if CEL-3R is pathogenic, we expressed it in HEK293 cells and compared its cellular localisation and secretion to the normal CEL protein and to the pathogenic CEL-HYB variant. Further, to better understand the role of the VNTR in general, we examined its evolution and characteristics in vertebrate species by performing a phylogenetic analysis. Finally, we predicted the conformations of the CEL-3R VNTR by molecular simulations.

We found that the CEL-3R variants exhibited levels of aggregation in HEK293 cells comparable to the pathological CEL-HYB variant. Still, CEL-3R displayed increased secretion compared to CEL-HYB. Moreover, we found that the CEL VNTR was present only in mammalian species. The mammalian VNTR sequences were relatively conserved, but exhibited a diverse number of repeats. A short post-VNTR sequence was also well conserved among mammals of the clade Eutheria. We were not able to correlate the number of VNTR repeats with any specific traits of the investigated mammalian species. The molecular simulations of the VNTR region of CEL-3R predicted that it forms unstructured coils and temporary turns, β-bridges and β-sheets. The *O*-glycosylated VNTR structure displayed a similar pattern.

In conclusion, although more transfection experiments are needed, we found that the Danish CEL-3R variant may show signs of pathological aggregation. We also confirmed that the CEL VNTR probably evolved sometime between the origin of Synapsida and the formation of Eutheria and Metatheria. Finally, the typical 'EATPVPPTGDS' repeats in the VNTR may be important for the formation of secondary structures that induce the VNTR to become compact, but still disordered.

Norsk

Karboksylester-lipase (CEL) er eit fordøyingsenzym som vert skilt ut i tarmen frå acinærcellene i bukspyttkjertelen. I tillegg til det globulære domenet som er ansvarleg for lipase-aktiviteten, inneheld CEL ein C-terminal region koda av ein sekvens med eit variabelt tal av tandemrepetisjonar, ein såkalla VNTR region. Kvar repetisjon kodar for 11 aminosyrer. Funksjonen til VNTR-regionen er ikkje kjent, men me veit at han er uordna, blir *O*-glykosylert og er naudsynt for effektiv utskiljing av CEL. VNTR-regionen er også særs variabel (polymorf) og minst to variantar er stadfesta å vera patogene: CEL-MODY og CEL-HYB. Desse variantane er knytt til cellulær aggregering og sjukdom i bukspyttkjertelen. Ein tredje variant med berre tre repetisjonar, CEL-3R, har vorte oppdaga i ein dansk familie der han finst hjå dei medlemmane som har diabetes. Likevel veit me ikkje om CEL-3R er patogen.

I denne studia utforska me eigenskapane til VNTR-regionen til CEL, med særleg merksemd på den danske CEL-3R-varianten. For å undersøkja om CEL-3R er patogen, uttrykte me han i HEK293-celler og samanlikna distribusjon og utskiljing mot det normale CEL-proteinet i tillegg til den patogene CEL-HYB-varianten. Deretter, for å forstå rolla til VNTR-regionen, undersøkte me korleis han kunne vorte utvikla gjennom evolusjonen. Til slutt prøvde me ved å nytte simulasjonar å seie noko om kva konformasjonar VNTR-regionen til CEL-3R kan ha.

Me fann at CEL-3R-variantar viste eit nivå av aggregering i HEK293-celler som kunne minna om den patologiske CEL-HYB-varianten. Likevel hadde CEL-3R auka utskiljing samanlikna med CEL-HYB. Vidare fann me at VNTR-regionen berre var til stades i pattedyrartar. VNTR-sekvensane var konserverte i pattedyr, men talet på repetisjonar varierte sterkt. Ein kort post-VNTR-sekvens var godt konservert i kladen Eutheria (placentale pattedyr). Me fann ingen samanheng mellom talet på repetisjonar i VNTR-regionen og eigenskapar til dyra som vart undersøkte. Dei molekylære simulasjonane av VNTR-regionen til CEL-3R predikerte at han kan danna ustrukturerte kveilar og kortvarige vendingar, β-bruer og β-flak.

Me konkluderte med at den danske CEL-3R-varianten viser teikn på patologisk aggregering, men at me må gjera fleire eksperiment for å få stadfesta dette. Me fann også at VNTR-regionen til CEL sannsynlegvis vart utvikla ein gong mellom fyrste førekomst av kladen Synapsida (pattedyrliknande krypdyr) og forgreininga til Eutheria og Metatheria (pungdyr). Til slutt fann me at den typiske 'EATPVPPTGDS'-repetisjonen i VNTR-regionen kan vera viktig for danninga av sekundære strukturar. Desse strukturane fører til at regionen vert kompakt, men likevel uordna.

4 Introduction

4.1 The Pancreas

4.1.1 Organ structure and function

The pancreas is an organ located behind the lower end of the stomach and along the duodenum (*Walkowska et al.*, 2022). The organ has an angled head, an elongated body and a tail (Figure 4.1A). The branching pancreatic duct transports pancreatic juice to the duodenum via the duodenal papilla. Before opening into the duodenum, the pancreatic duct merges with the common bile duct, mixing the pancreatic juice with bile (*Walkowska et al.*, 2022).

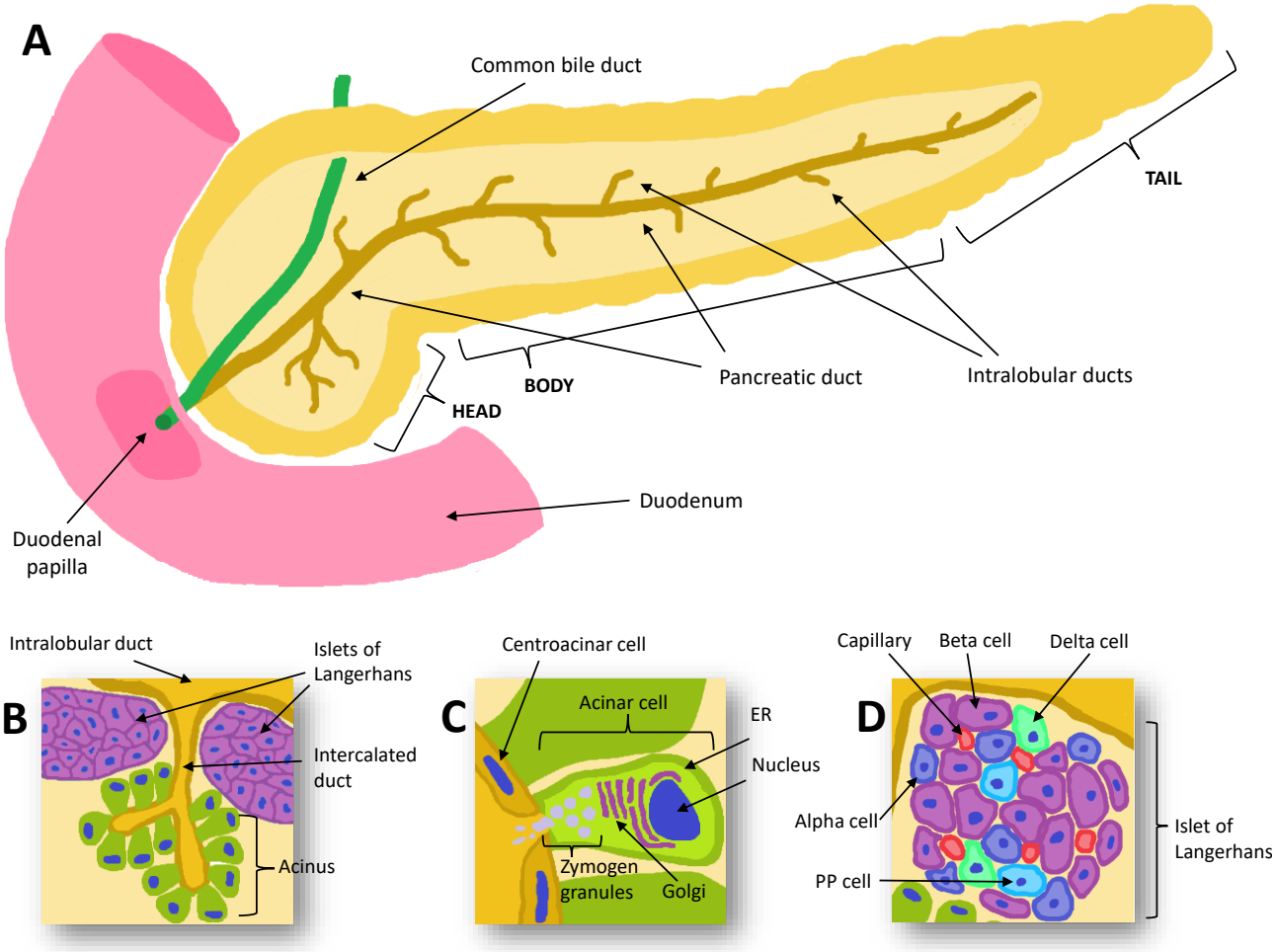


Figure 4.1: **Schematic illustration of the pancreas and its exocrine and endocrine tissues. A.** The pancreas and its position next to the duodenum. **B.** The pancreatic duct branches into intralobular ducts which branch into acini. Islets of Langerhans are dispersed within the acinar tissue. **C.** The acinar cells produce zymogens and secrete them into the ductal system. **D.** Islets of Langerhans are composed of various endocrine cells that secrete hormones into nearby capillaries. The two most common types are alpha and beta cells, producing glucagon and insulin, respectively. (Own illustration)

The pancreas is both an exocrine and an endocrine organ (*Walkowska et al.*, 2022). The exocrine part consists of the ducts, as well as the acinar cells that secrete pancreatic enzymes

(Figure 4.1B). The enzymes are cotranslationally inserted into the rough endoplasmic reticulum (ER) and subsequently transported to, and processed, in the Golgi complex (*Williams*, 2006). The inactivated enzymes (denoted zymogens) are then stored in vacuoles called zymogen granules, from which they are secreted by exocytosis when needed for digestion (*A-Kader and Ghishan*, 2012). Secreted zymogens mix with an alkaline bicarbonate solution produced by the ductal cells to become pancreatic juice with pH $\sim$8.0. The low pH in the duodenum aids in turning the zymogens into active digestive enzymes. Proteases also aid in activating some zymogens by cleaving off specific parts of the polypeptide chain. Bile contributes to emulsification of fat while the pancreatic enzymes break down lipids, proteins, nucleic acids and starch (*Walkowska et al.*, 2022).

The endocrine pancreas consists of Islets of Langerhans which are dispersed within the exocrine tissue (*Walkowska et al.*, 2022). An islet consists of a variety of endocrine cells, where the two most abundant cells are alpha cells and beta cells (Figure 4.1D). Alpha and beta cells regulate blood sugar by releasing the hormones glucagon and insulin, respectively. Delta cells secrete the growth hormone-inhibiting hormone somatostatin. PP (Pancreatic Polypeptide) cells secrete pancreatic polypeptide, which regulates a variety of endocrine and exocrine pancreatic functions (*Walkowska et al.*, 2022).

4.2 Diseases of the pancreas

4.2.1 Pancreatitis

Pancreatitis is a condition of inflammation in the pancreas, either acute or chronic (*Ashraf et al.*, 2021). Acute pancreatitis (AP) is a temporary inflammation from which the pancreas can return to normal. Gallstones are the most common cause of AP, followed by alcohol abuse. Other causes are infection, genetic factors, mechanical trauma, medication, tumors, etc. Genetic mutations involving the pancreatic enzymes are particularly infamous for causing premature enzymatic activation and digestion of the pancreatic tissue. Many people with either type of pancreatitis suffer from abdominal pain attacks. (*Ashraf et al.*, 2021). The prevalence of AP increases with age and is similar between men and women (*A-Kader and Ghishan*, 2012).

Chronic pancreatitis (CP) is characterised by long-lasting, irreversible inflammation and can be caused by recurrent AP, alcohol abuse, genetic factors, autoimmune disease, duct blockage by tumor, etc. (*Luchini et al.*, 2020). CP leads to permanent damage to the pancreas' structure or function with morphological changes such as fibrosis, acinar atrophy and fatty replacement. The loss of pancreatic function may affect exocrine enzyme output, leading to incomplete digestion. Decreased pancreatic function could also affect endocrine blood sugar regulation, leading to diabetes (*Luchini et al.*, 2020).

4.2.2 Pancreatic cancer

Pancreatic cancer (PC) is one of the deadliest types of cancer. The prevalence of PC is higher in men than women, and it increases with age (*Ilic and Ilic*, 2016). Risk factors for PC include alcohol abuse, smoking, obesity, CP, genetic factors, etc. The ABO blood group has also been implicated as a risk factor. Compared to subjects of blood group O, the subjects with blood groups A, B or AB have an odds ratio of developing PC between 1.3 and 2.4. About 65% of patients with cystic neoplasms in the pancreas develop PC. Moreover, development of PC may induce symptoms of new-onset diabetes in patients, due to endocrine dysfunction (*Yadav and Lowenfels*, 2013). Pancreatic adenocarcinomas account for about 85-90% of pancreatic cancer cases, while pancreatic endocrine tumours account for about 5% of cases (*Yadav and Lowenfels*, 2013; *Hidalgo et al.*, 2015).

The most lethal type of PC is pancreatic ductal adenocarcinoma (PDAC) which makes up around 90% of PC cases (*Halbrook et al.*, 2023). PDAC has a 5-year survival rate of 12%. These low rates are due to few recognizable symptoms in the early stages which leads to late discovery and diagnosis. Pancreatic lesions are the origin of about 90% of PDAC cases (*Halbrook et al.*, 2023)

4.2.3 Diabetes mellitus

Diabetes mellitus is a group of diseases with a shared symptom of elevated blood sugar, also called hyperglycemia (*Cole and Florez*, 2020). The symptom is caused by impaired insulin secretion from the beta cells or reduced insulin sensitivity. The most common forms are Type 1 (T1D) and Type 2 (T2D) Diabetes. T1D is an autoimmune disease where the beta cells (Figure 4.1D) are attacked and destroyed by the patient's own immune system. The patient can not produce insulin and becomes unable to lower their blood sugar. Diabetes T1 is caused by both polygenic and environmental factors and usually appears during childhood or adolescence. T2D is a condition where glucose-storing cells become desensitised to the insulin produced by the beta cells. T2D is also partially caused by genetic and environmental factors, but is more dependent on lifestyle factors than T1D. Indeed, T2D usually appears in late adulthood (*Cole and Florez*, 2020).

Other forms of diabetes are monogenic and gestational diabetes (*Schwitzgebel*, 2014). Monogenic diabetes is a term for the 2-5% of diabetes cases which are caused by a single defective gene. The disease manifests either as reduced number of beta cells, or reduced beta cell

functionality, which may lead to similar symptoms as T1 or T2. These similarities often lead to monogenic diabetes being misdiagnosed as T1 or T2. In contrast to T1 and T2, the inheritance pattern of monogenic diabetes is Mendelian and usually dominant. A subgroup of monogenic diabetes is Maturity Onset of the Young (MODY), which is defined as beta cell defects leading to diabetes by the age of 25 years (*Schwitzgebel*, 2014).

Examples of monogenic diabetes are neonatal diabetes and Maternally Inherited Diabetes and Deafness (MIDD) (*Bonnefond et al.*, 2023). Neonatal diabetes is a rare form of diabetes which arises in newborns within their first 6 months of life, but usually passes in infancy. MIDD is a type of diabetes which is inherited through the mitochondria of the mother and is often accompanied with various symptoms such as hearing loss, seizures, and muscle weakness. (*Bonnefond et al.*, 2023).

In contrast to T1D and T2D, some forms of diabetes do not directly affect the beta cells or insulin sensitivity (*Hart et al.*, 2016). Rather, inflammation and disease in the exocrine pancreas may indirectly harm nearby beta cells. As mentioned, exocrine and endocrine pancreatic tissues are highly interspersed (Figure 4.1B). Such diseases are labelled Type 3c Diabetes (T3cD), also sometimes called secondary diabetes of the exocrine pancreas or pancreatic diabetes (*Hart et al.*, 2016; *Wynne et al.*, 2018). Patients with T3cD may exhibit both increased and reduced insulin sensitivity (*Hart et al.*, 2016).

4.3 The pancreatic enzyme CEL

Carboxyl ester lipase (CEL) is a lipase found in pancreatic juice (*Lombardo et al.*, 1978; *Blackberg et al.*, 1980). CEL hydrolyses cholesteryl esters and esters of fat-soluble vitamins (such as vitamins A, D and E) (*Lombardo and Guy*, 1980). The enzyme specificity is broad, as CEL may also hydrolyze amide bonds in lipid-like substrates (*Hui et al.*, 1993) and ceramides (*Nyberg et al.*, 1998). The catalytic activity of CEL is stimulated by bile salts (Figure 11.8) (*Bläckberg and Hernell*, 1983). Bile salts are comprised of planar, steroid backbones which are hydroxylated and conjugated to usually taurine or the amino acid Gly. Bile salts have one hydrophobic face and one hydrophilic face, resulting in the amphiphilic property which makes bile salts into emulsifiers (*Maldonado-Valderrama et al.*, 2011).

In the literature, the CEL protein has been given different names, such as bile salt-activated lipase (BSAL), bile salt-stimulated lipase (BSSL), bile salt-dependent lipase (BSDL), carboxyl ester hydrolase (CEH) and cholesterol esterase. In this thesis, the gene will be referred to as *CEL* and the protein will be denoted CEL.

4.3.1 The *CEL* gene

The *CEL* gene is thought to have originated from a family of carboxylesterase genes during the evolution of early vertebrates (*Wang and Hartsuck*, 1993; *Holmes and Cox*, 2011). In the human reference genome GRCh38, the *CEL* gene is assigned to chromosome 9, more specifically to 9q34.13 (Figure 4.2) (*NCBI*, 2023a). Previously, the *CEL* locus was believed to be at 9q34.3 (*Taylor et al.*, 1991). Human *CEL* consists of 11 exons. The last exon contains a Variable Number of Tandem Repeat (VNTR) region, where each repeat is a 33 bp sequence. The number of VNTR repeats vary between human alleles (*Higuchi et al.*, 2002), as well as between species (*Nilsson et al.*, 1990). The most common human CEL variant has 16 VNTR repeats (*Higuchi et al.*, 2002; *Bengtsson-Ellmark et al.*, 2004; *Fjeld et al.*, 2016).

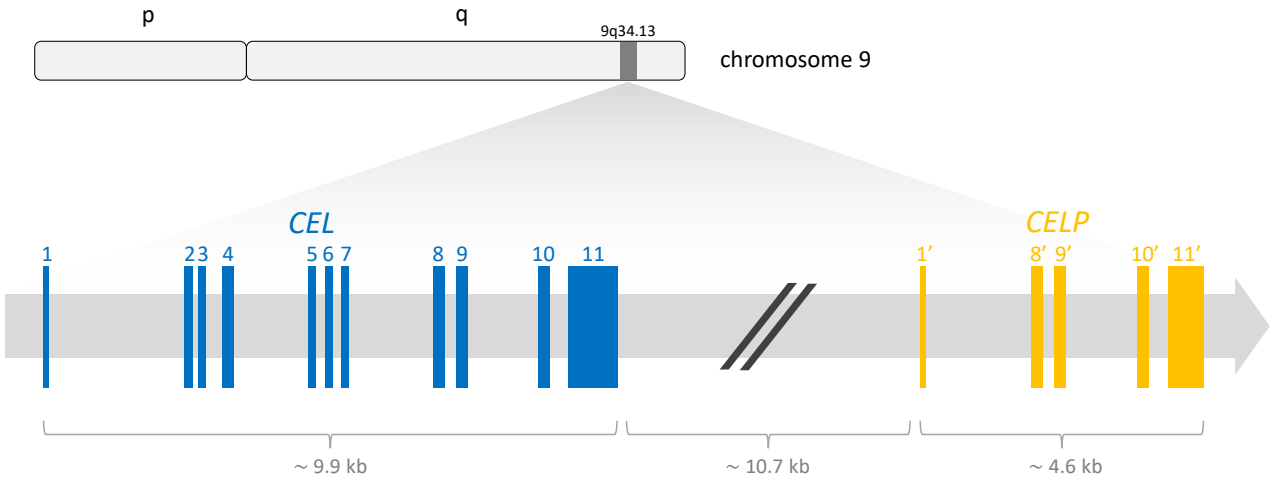


Figure 4.2: **The *CEL* gene and *CELP* pseudogene on chromosome 9.** The *CEL* gene and the neighboring *CELP* pseudogene are located at the q-terminal of human chromosome 9. Exons are shown as coloured boxes. *CELP* exons are numbered with apostrophes to indicate the equivalent exon in *CEL*. (Own illustration). Figure design is partially derived from Fjeld et al.*

* Fjeld et al. A recombined allele of the lipase gene CEL and its pseudogene CELP confers susceptibility to chronic pancreatitis. *Nature Genetics*, 47(5):518–522, mar 2015. ISSN 1546-1718. doi:10.1038/ng.3249

The *CEL* pseudogene (*CELP*) is an inactivated duplicate of *CEL*. The *CEL* gene was most likely duplicated during primate evolution (*Lidberg et al.*, 1992). However, during Hominid evolution, the original *CEL* gene was deactivated and became *CELP* (*Madeyski et al.*, 1999). In humans, *CELP* is also located within 9q34.13 (*NCBI*, 2023b), about 10.7 kb downstream from *CEL*. The pseudogene is missing exons 2-7 that are otherwise found in *CEL* (*Lidberg et al.*, 1992) and also tends to have between 4 and 10 VNTR repeats—otherwise the sequences are fairly identical (*Lidberg et al.*, 1992; *Madeyski et al.*, 1999).

In humans, the CEL protein is expressed in the mammary glands (*Bläckberg et al.*, 1987), as well as in the acinar cells (*Lombardo et al.*, 1978; *Blackberg et al.*, 1980). Expression of the *CEL* gene in mammary glands only occurs during pregnancy and lactation, unlike the constitutive expression in the pancreas (*Kannius-Janson et al.*, 1998). Furthermore, CEL is secreted into the mother's milk in a broad selection of mammals, including cats and dogs (*Freed et al.*, 1986; *Wang et al.*, 1989; *Wang and Hartsuck*, 1993).

The regulatory elements surrounding the *CEL* gene may affect how its expression is controlled. A TATA box has been detected at about 26 bp upstream from the transcription initiation site of the human *CEL* gene (*Lidberg et al.*, 1992). Genes containing TATA boxes are known to be regulated differently depending on stress factors, as opposed to constitutive expression (*López-Maury et al.*, 2008). Also situated in the 5' region, about 145 bp upstream of *CEL*, is a CCAAT box (*Kumar et al.*, 1997). The CCAAT box is a regulatory element which is targeted by the CCAAT-enhancer-binding-proteins (C/EBPs) family of transcription factors. The C/EBPs transcription factors are known to be important for differentiation, inflammation, liver regeneration and metabolism (*Ramji and Foka*, 2002). The C/EBPs family may both up- and downregulate genes that contain a CCAAT box (*Ramji and Foka*, 2002).

The 1330 bp immediately upstream from the *CEL* transcription initiation site also contains one or several binding sites for the transcription factors AP1 (Activator protein 1), AP2 (Activator protein 2) and SP1 (Specificity protein 1) (*Kumar et al.*, 1997). Collectively, these transcription factors serve various roles in differentiation, apoptosis, cell cycle, growth factors, estrogen receptors, and more (*Kadonaga et al.*, 1987; *Hagen et al.*, 1992; *Hilger-Eversheim et al.*, 2000; *Hess et al.*, 2004). The 5' untranslated region of the *CEL* mRNA is relatively short and there is no indication that it regulates expression (*Kumar et al.*, 1997).

The primate gene duplication event of *CEL* might have had implications on the regulation of its transcription. The new copy seems to have adapted to different types of regulatory elements compared to the original (*Lidberg et al.*, 1992). For instance, the transcriptional regulation of human *CEL* is functionally similar, but mechanistically different to mice (*Kannius-Janson et al.*, 2000).

4.3.2 Structure of the CEL protein

CEL is comprised of a single polypeptide chain (*Baba et al.*, 1991). The chain can be divided into an N-terminal signalling peptide, a catalytic globular domain and a C-terminal VNTR-encoded region (Figure 4.3). The CEL globular domain contains about 14 α-helices surrounding a β-sheet of 11 β-strands (*Wang et al.*, 1997; *Terzyan et al.*, 2000). This distinct fold places CEL in the α/β-hydrolase family of proteins (*Ollis et al.*, 1992; *Wang and Hartsuck*, 1993). Residues Ser217, Asp343 and His458 form the catalytic triad which is crucial for enzymatic activity (*DiPersio et al.*, 1990, 1991; *DiPersio and Hui*, 1993).

The term VNTR is strictly speaking only supposed to be used when referring to repeats in nucleotide sequences. Yet, the term is sometimes used in regards to the repeated protein sequences which are encoded by VNTRs. This is partly due to that there is no separate, widely accepted term for the repeated protein sequences themselves. Further, one may often want to collectively address the repeats in both the DNA and the encoded protein. In these cases, it quickly becomes convoluted to try to use different terms for the two aspects. In this thesis, I will use the term VNTR for both the DNA and its encoded protein sequence.

The numbering of amino acid residues in the CEL protein are incoherent between different sources. This is mainly due to the differing preferences to include or exclude the cleaved signal peptide. See examples in the articles from *Johansson et al.* (2011), *Holmes and Cox* (2011) and *Touvrey et al.* (2019). In this thesis, the amino acids will be numbered as if the signal peptide is present—as is done in the UniProt sequence for human CEL (P19835-1). Figure 4.3 displays both numbering systems for easy comparison. Note also that the length of the signal peptide varies between some sources. In this thesis, the signal peptide is considered to be 23 residues in length.

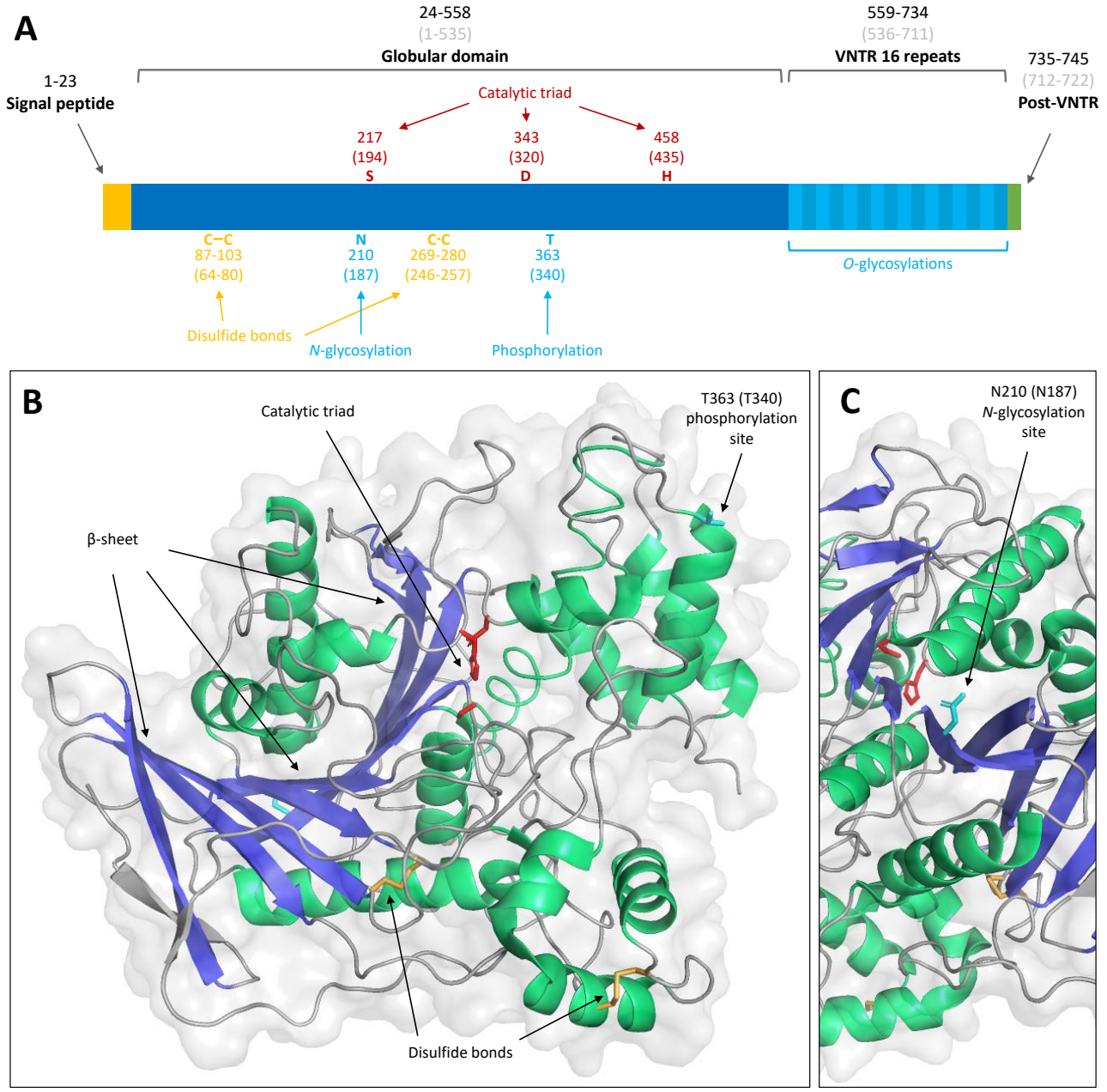


Figure 4.3: **Annotated features of the CEL protein. A.** Annotated domains and regions of a CEL variant with 16 VNTR repeats. Numbers denote the amino acid positions for the fully intact CEL protein (numbers in parentheses denote positions in CEL without signal peptide). Other features are also annotated: catalytic triad (Ser217, Asp343 and His458), disulfide bonds (Cys87-Cys103 and Cys269-Cys280), *N*-glycosylation (Asn210), phosphorylation (Thr363) and *O*-glycosylations (VNTR). **B.** X-ray crystal structure of human CEL globular domain* (PDB accession: 1F6W). The β-sheet is coloured blue. Helices are coloured green. Visualised in PyMOL. **C.** The structure viewed from a different angle than in B. Displays cyan-coloured N210 residue on the edge of the β-sheet. (Own illustration)

* CEL structure from S. Terzyan, C.-S. Wand, D. Downs, B. Hunter, and X. C. Zhang. Crystal structure of the catalytic domain of human bile salt activated lipase. *Protein Science*, 9(9):1783–1790, jan 2000. ISSN 1469-896X. doi:10.1110/PS.9.9.1783.

The CEL polypeptide undergoes several co-translational and post-translational modifications (PTMs) (see annotations in Figure 4.3A). Residue Asn210 is *N*-glycosylated, while the VNTR is particularly abundant in *O*-glycosylations, which are covalently bonded to Thr and Ser residues (*Baba et al.*, 1991). The CEL globular domain has two disulphide bridges—one between Cys87-Cys103, and another between Cys269-Cys280 (*Baba et al.*, 1991). The globular domain also

has a phosphate group attached to Thr363 (*Pasqualini et al.*, 1997; *Verine et al.*, 2001). Human CEL with 16 VNTR-encoded repeats has a predicted protein mass of about 80 kDa. However, the observed size is about 100 kDa (*Lombardo et al.*, 1978), primarily due to the *O*-glycosylations in the VNTR.

The role of the CEL VNTR region has not been clearly identified. It is considered to be a PEST sequence, which is a sequence abundant in residues Pro (P), Glu (E), Ser (S) and Thr (T). PEST sequences are known to target proteins for proteolytic degradation (*Rechsteiner and Rogers*, 1996). Contradictingly, the VNTR may protect CEL from degradation (*Loomes*, 1995) as well as hinder it from dimerizing and forming aggregations (*Loomes and Senior*, 1997). It has been suggested that the *O*-glycosylations on the VNTR prevent CEL from being targeted for degradation (*Loomes*, 1995; *Loomes and Senior*, 1997). Furthermore, the *O*-glycosylations are a prerequisite for CEL secretion (*Bruneau et al.*, 1997).

Indeed, the VNTR has been found to have no effect on the enzymatic activity of human CEL against emulsified lipid substrates (*Hansson et al.*, 1993; *Blackberg et al.*, 1995). In contrast, the VNTR of rat CEL was found to have a small inhibitory effect on the hydrolysation of the water-solubilised substrate *p*-nitrophenyl butyrate at low taurocholate concentrations (*DiPersio et al.*, 1994). The different findings could be due to the studies using different species, substrates, solubilisation, or a combination (*DiPersio et al.*, 1994).

One proposed role of the *O*-glycosylations is to impose a rigid structure to the CEL VNTR, similar to mucins (*Loomes*, 1995; *Johansson et al.*, 2011). Mucins are glycoproteins secreted by a range of epithelial tissues and are responsible for forming mucus layers. Mucins may form disulfide bridges and non-covalent interactions to form viscous or gel-like substances. Similar to the CEL VNTR, mucins are largely composed of *O*-glycosylated PEST sequences (*Strous and Dekker*, 1992).

The *O*-glycosylations on human CEL VNTRs also seem to match with the individual's ABO blood group (*Jellas et al.*, 2018). Such ABO antigen-related *O*-glycosylations could just be a by-product of the blood group, or it may be an advantageous trait. For example, the microbiota of the intestines may interact with glycans on proteins, which could be beneficial to the individual in some way (*Jellas et al.*, 2018).

4.3.3 Folding, modification and secretion of CEL

The folding and modification of CEL is a multi-step process which occurs across several organelles and intracellular localisations. Not all these processes have been studied in detail for CEL, but some pathways are known for secretory proteins in general.

During translation of a signal peptide in the cytosol, the nascent polypeptide is directed into the ER by co-translational translocation. Once the polypeptide is inside the ER, the signal peptide is cleaved off (*Paetzel et al.*, 2002). At the same time as the co-translational translocation, the CEL residue Asn210 is modified with a ”high mannose” *N*-glycosylation (*Strous and Dekker*, 1992; *Abouakil et al.*, 1993). The *N*-glycosylation is important for continuation in the secretory pathway and also increases the enzymatic activity of CEL (*Abouakil et al.*, 1993). Formation of the disulfide bonds Cys87-Cys102 and Cys269-Cys280 are also likely to occur co-translationally in the ER, as is common for other proteins (*Chen et al.*, 1995; *Bulleid*, 2012).

After processing in the ER, CEL is transported to the Golgi complex. The initial *O*-glycosylations of the VNTR is likely performed either in the transition from ER to Golgi or within the *cis*-Golgi (*Strous and Dekker*, 1992). Then, the *O*-glycosylations are expanded during transport to *trans*-Golgi. Also, the *N*-glycosylation is changed from the ”high-mannose” to a ”complex” configuration, probably while CEL is in the Golgi (*Strous and Dekker*, 1992). Finally, residue Thr363 is phosphorylated in the *trans*-Golgi, allowing for CEL to continue its secretory pathway to the zymogen granules (*Verine et al.*, 2001).

4.4 Pathogenic CEL variants

CEL is a highly polymorphic gene (*Johansson et al.*, 2018). Most of the variation is linked to the VNTR, which may expand or contract, hybridize with *CELP* or be subject to single basepair substitutions, deletions and insertions. Several human CEL VNTR variants are associated with pancreatic disease including diabetes (*Johansson et al.*, 2018). The pathogenesis seems to involve pancreatitis which itself induces diabetes and perhaps PC (*Lombardo et al.*, 2018). These variants are listed in Table 4.1 and are further described below.

Table 4.1: **Overview of human CEL variants with a confirmed or possible pathogenic effect compared to normal CEL-WT.** Predicted length, size and pI are derived from CEL variants with a cleaved signal peptide. Values for CEL-DEL5 were estimated from the corresponding literature article. 'Repeats' column displays total number of repeats (number of repeats with atypical residues are shown in parentheses). 'Post-VNTR' column shows whether the post-VNTR sequence is present or if it is absent due to a premature stop codon. 'Genetic explanation' column gives background information on the gene which encodes the protein. For full references, see the end of the thesis.

CEL variant	Length (aa)	Size (kDa)	pI	Repeats		Post-VNTR	Genetic description	Literature
CEL-WT	722	76.3	5.13	16	(0)	Yes	Benign. Most common VNTR variant	*Higuchi et al.* (2002)
CEL-DEL1	649	71.0	9.38	10	(10)	No	Point deletion in repeat 1	*Ræder et al.* (2006)
CEL-DEL4	671	72.9	9.18	12	(9)	No	Point deletion in repeat 4	*Ræder et al.* (2006)
CEL-DEL5	660	71.5	8.99	12	(8)	No	Point deletion in repeat 5	*Pellegrini et al.* (2021)
CEL-HYB1	566	62.2	8.47	3	(3)	No	VNTR exon exchanged with CELP	*Fjeld et al.* (2015)
CEL-3R-DAN	579	63.4	6.85	3	(0)	Yes	Danish variant with 3 repeats	*Torsvik et al.* (2010)

4.4.1 CEL-MODY

CEL-MODY, or MODY-8, is a pathogenic, gain-of-function CEL variant which displays a dominant inheritance pattern. CEL-MODY is caused by single nucleotide deletions and subsequent frameshifts in the *CEL* VNTR (*Ræder et al.*, 2006; *Johansson et al.*, 2011). CEL-DEL1, CEL-DEL4 and CEL-DEL5 are three known variants with deletions in repeat 1, 4 and 5, respectively (*Ræder et al.*, 2006; *Pellegrini et al.*, 2021). These variants are known to cause pancreatic exocrine dysfunction and diabetes with dominant inheritance patterns. The deletions introduce frameshifts and new amino acid residues to the VNTR, like Cys and Arg, as well as premature stop codons (*Ræder et al.*, 2006). Formation of intramolecular disulfide bridges between the Cys residues are believed to be an important factor in intracellular aggregation in acinar cells (*Xiao et al.*, 2016). The pathogenesis possibly begins with cell death in acinar tissue in the exocrine pancreas. The β-cells in the endocrine pancreas are subsequently affected and become dysfunctional (*Kahraman et al.*, 2022).

CEL-MODY is a disease with a history of both false positive and false negative diagnoses (*Jellas et al.*, 2022). Many researchers have assigned CEL-MODY to patients with substitution mutations, which do not fulfill the criterium of frameshift and are unlikely to be pathogenic — false positives. Further, the screening of CEL-MODY is difficult to perform. This is due to the fact that genetic markers and low-fidelity sequencing are often unable detect a single basepair deletion. Thus, real cases may go undetected — false negatives. Proper screening of CEL-MODY deletions typically require additional PCR steps and high-fidelity sequencing (*Jellas et al.*, 2022).

4.4.2 CEL-HYB

CEL-HYB1 is a variant in which the *CEL* gene has recombined with the *CEL* pseudogene (*Fjeld et al.*, 2015). The recombination has occurred between *CEL* intron 10 and *CELP* intron 4'–10' (Figure 4.2). The result is a truncated CEL protein with a globular CEL domain and a *CELP*-encoded C-terminal. *CEL-HYB1* is a genetic risk factor for chronic pancreatitis with an odds ratio of 15.5 in German and French populations. The heterozygous allele frequency for CEL-HYB in these populations were about 1% (*Fjeld et al.*, 2015). Studies in cellular systems and mouse models indicated that the disease mechanism is linked to impaired protein secretion, protein aggregation and ER stress (*Fjeld et al.*, 2015; *Cassidy et al.*, 2020; *Fjeld et al.*, 2022; *Mao et al.*, 2022).

CEL-HYB1 is an ethnic specific variant which is not observed in Asia (*Zou et al.*, 2016). In contrast, *CEL-HYB2* is a different variant which is detected in some Asian populations with an average carrier frequency of 1.7%. *CEL-HYB2* is a result of hybridisation between intron 9 and exon 10 in the *CEL* and *CELP* genes. This hybridisation product results in a stop codon before exon 11, which likely leads to non-sense mediated decay of the *CEL-HYB2* mRNA. Thus, *CEL-HYB2* is barely expressed, if at all, and is unlikely to be proteotoxic (*Zou et al.*, 2016).

4.4.3 VNTR length variants

As previously stated, the *CEL* VNTR is highly polymorphic. Tandem repeats are genetically unstable and may expand or retract in short evolutionary time (*Tompa*, 2003). Human populations have *CEL* VNTRs with varying numbers of repeats (*Higuchi et al.*, 2002; *Bengtsson-Ellmark et al.*, 2004; *Torsvik et al.*, 2010; *Fjeld et al.*, 2016). The *CEL* allele with 16 repeats is the most common and accounts for about 60% of all alleles (*Higuchi et al.*, 2002; *Bengtsson-Ellmark et al.*, 2004). This common variant is known as CEL-16R or CEL-WT. Alleles with other numbers of CEL VNTR repeats become rarer the more the number of repeats deviates from 16 (*Higuchi et al.*, 2002; *Bengtsson-Ellmark et al.*, 2004). This suggests some evolutionary advantage of having 16 repeats in the *CEL* VNTR.

There is little decisive evidence for that *CEL* VNTR lengths which deviate from 16 repeats are pathological. In a study on Chinese patients by *Mao et al.* (2021), they found a statistically significant correlation with homozygosity for short VNTR lengths and patients suffering from idiopathic CP. However, this study and others have failed to find correlations of *CEL* VNTR length with diseases such as PC or alcoholic CP. In general, there is a need for larger cohorts to determine whether the VNTR lengths can induce disease or not (*Fjeld et al.*, 2016; *Dalva et al.*, 2017; *Mao et al.*, 2021).

4.4.4 The CEL-3R variant

CEL-3R is a variant with only 3 repeats in the VNTR (*Torsvik et al.*, 2010). This variant has only been reported in 7 individuals that are all members of the same Danish family (Figure 4.4). The 7 individuals were heterozygous for *CEL-3R* and carried either 16 or 17 VNTR repeats on the other allele. Notably, all except one of the *CEL-3R* carriers had either diabetes, impaired glucose tolerance or impaired fasting glycemia. The non-carriers were all healthy (*Torsvik et al.*, 2010). As with the aforementioned studies on VNTR lengths and pancreatic diseases, more research needs to be performed to confirm that the short *CEL-3R* variant induces diabetes.

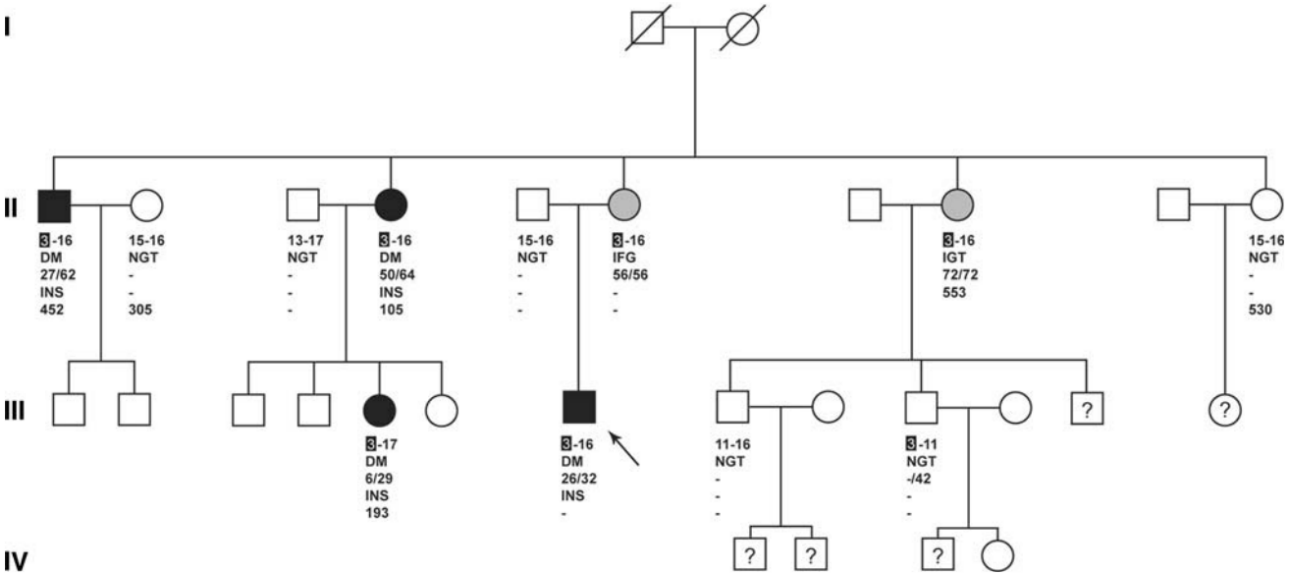


Figure 4.4: **Pedigree of Danish family with the *CEL-3R* variant.** The proband is marked by an arrow. Individuals filled with black colour are diagnosed with diabetes. Individuals filled with grey colour are diagnosed with a pre-diabetic condition. The two numbers separated by ”-” indicate *CEL* VNTR lengths in the respective alleles. Numbers separated by ”/” indicate age of onset and diagnosis, respectively. IFG: impaired fasting glycemia. IGT: impaired glucose tolerance. INS: insulin treatment. Figure is copied from Torsvik et al. Mutations in the VNTR of the carboxyl-ester lipase gene (*CEL*) are a rare cause of monogenic diabetes. *Human Genetics*, 127(1):55–64, jan 2010. ISSN 03406717. doi:10.1007/S00439-009-0740-8. Reproduced with permission from Springer Nature.

4.5 Intrinsically disordered protein regions

Intrinsically disordered proteins (IDPs) are proteins that consist completely or partially of intrinsically disordered regions (IDRs). IDRs are polypeptide segments that do not have any consistent fold or three-dimensional structure. Instead, a disordered polypeptide may assume multiple semi-stable conformations separated by low energy barriers (*Brucale et al.*, 2014). Notably, a well-defined tertiary structure is not an absolute criterion for protein function (*Wright and Dyson*, 1999), despite the previous assumptions of the ”structure-function paradigm” (*Brucale et al.*, 2014). IDRs may therefore play important functional roles and have indeed been shown to be evolutionarily conserved (*Iakoucheva et al.*, 2001). In this thesis, the terms IDPs and IDRs will be used interchangeably whenever both may apply.

The presence of IDRs is frequently seen in polypeptides encoded by tandem repeats (*Tompa*, 2003). It has been proposed that the genetically unstable tandem repeats may rapidly contract or expand into evolutionarily advantageous numbers (*Tompa*, 2003). IDRs are also known to be more common in organisms with higher genomic complexity, such as Eukaryotes (*Dunker et al.*, 2002; *Ward et al.*, 2004).

The formation of some cellular structures may partially depend on IDRs (*Shin and Brangwynne*, 2017). Typically, organelles such as the ER or Golgi are delimited by a lipid bilayer. However, some organelles or cellular structures do not have such a defined border. Instead, they are constituted of a mass held together by a vast number of weak forces. Such masses are denoted ”condensates” and may consist of a mix of proteins, RNAs, DNAs or other compounds. Condensates are highly dynamic structures. They have a liquid-like consistency and may form or dissolve depending on the temperature and the concentration of composites. Proteins with IDRs are common in condensates. The IDRs usually contain mostly polar or charged residues such as Gly, Ser, Gln, Pro, Glu, Lys and Arg. Aromatic residues such as Tyr and Phe are sometimes interspersed within the IDRs (*Shin and Brangwynne*, 2017).

IDPs have been difficult to study for some time. The structure of ordered proteins and domains have traditionally been studied using regularised macromolecule samples and high-resolution techniques, such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) or cryo-electron microscopy (Cryo-EM) (*Schramm et al.*, 2019). In contrast, the numerous conformations of IDPs make them more difficult to crystallize or regularize. Also, the eventual results would tell little about the wider conformational landscape and dynamics of the disordered protein. In addition, IDP samples are often structurally heterogenous anyways (*Gopal et al.*, 2021; *Wang*, 2021), such as the various *O*-glycosylations of CEL.

Studying IDPs favour using solubilised samples and a combination of methods such as small-angle X-ray scattering (SAXS), mass spectroscopy (MS), circular dichroism (CD), certain types of NMR, infrared (IR) spectrophotometry, etc. These techniques will not give high-resolution data about atomic positions, but rather low-resolution data about overall disorder, dihedral angles, secondary structures, etc. Finally, IDPs may be studied *in silico* with computational techniques (*Schramm et al.*, 2019).

4.6 Molecular dynamics

Molecular dynamics (MD) is a computational technique in which molecular systems are simulated to predict movements, conformations and interactions. In general, MD simulations are applications of Newton's laws of motion on individual atoms. The change in the position of an atom is calculated from its velocity and acceleration. The acceleration is the net sum of the differential equations of the forces that act on the atom. The positions of the atoms are updated stepwise with typically a few fs (femtoseconds) between each step (*Leach*, 2001). By sequentially performing these stepwise calculations, an MD simulation may predict the behaviour of a molecular system for a desired amount of simulated time. Increased computational power in recent decades has allowed for simulation of larger systems and in higher quantities.

A force field is a collection of various forces and their weights that are applied to atoms in an MD simulation. Examples of forces defined in a force field are electrostatic forces, van der Waals forces, and numerous bond forces that define ideal distances and angles between covalently bonded atoms. The force fields have historically been adjusted to best replicate the expected behaviour of folded proteins. However, these force fields may have overestimated the overall propensity of proteins to form secondary structures, and have therefore underestimated the disorder. For this reason, MD of IDPs has historically been inaccurate. One solution to this problem has been to increase the weight of the interactions between proteins and water molecules. Thus, proteins become generally more solvated and less compact (*Wang*, 2021). Some of these newer force fields include CHARMM36m (*Huang et al.*, 2016) and Amberff99SB-disp (*Robustelli et al.*, 2018).

4.7 Aim of the project

The overall aim of this project was to obtain new knowledge about the properties of the CEL VNTR and how it may lead to pancreatic disease. The main focus of the investigations was CEL-3R, the short CEL variant of suspected pathology that previously had been identified in a Danish family with diabetes.

The specific objectives were as follows:

- Compare sequences of vertebrate CEL VNTRs by a phylogenetic approach
- Investigate functional properties of the CEL-3R variant by Western blotting
- Predict the conformations of the CEL-3R VNTR with and without *O*-glycosylations by MD simulation

5 Materials

Table 5.1: **Bacterial transformation and plasmid preparation.**

Material	Supplier	Cat. number
Ampicillin	Sigma-Aldrich	A9518
LB Broth	Sigma-Aldrich	L7275-500TAB
OneShot TOP10 Competent bacteria	Invitrogen	C404010
QIAfilter Midi DNA isolation kit	QIAGEN	12143
TE buffer	PanReac Applichem	A0386

Table 5.2: **Plasmids.** Properties of all the applied plasmids in the study. 'Plasmid length (bp)' is the total nucleotide length of the plasmids. 'Protein length' values refer to the number of amino acids in the inserted CEL variants. CEL-TRUNC has a stop codon prior to a V5-His tag, which is not translated.

Plasmids	Plasmid length (bp)	Prot. length (aa)	Supplier	Orig. vector	Vector supplier
CEL-WT	7828	745	Gift from St. Louis lab	pcDNA3	Invitrogen
CEL-3R-USA	7402	602	Gift from St. Louis lab	pcDNA3	Invitrogen
CEL-3R-DAN	7404	602	Gift from St. Louis lab	pcDNA3	Invitrogen
CEL-HYB	7217	589	Gift from St. Louis lab	pcDNA3	Invitrogen
CEL-TRUNC	7154	561	Molven lab group	pcDNA3.1/V5-His B	Invitrogen
EV	5446	-	-	pcDNA3	Invitrogen

Table 5.3: **Primers used for PCR and Sanger sequencing.** Forward and reverse primers are indicated by 'fwd' and 'rev', respectively. All primers were supplied by Merck.

Primer	Sequence (5'-3')	Binding site
T7 (fwd)	TAATACGACTCACTATAGGG	T7 promoter
CF (fwd)	TCACCTTCAACTACCGT	CEL exon 4
BR (rev)	TGGCTCGCCGGATG	CEL exon 6
DF (fwd)	CCGCCGACATCGACTA	CEL exon 8
CR (rev)	CTCCTCCGTGACTTTC	CEL exon 8/9
DR (rev)	TACGGAAAACAGCGGC	CEL exon 11
EF (fwd)	CACACACTGGGAACCCT	CEL exon 11
L11F (fwd)	GTCCCTCACTCATTCTTCTATGGCAAC	CEL intron 7
VNTR-R (rev)	TCCTGCAGCTTAGCCTTGGG	CEL downstream
BGH (rev)	TAGAAGGCACAGTCGAGG	*bgh*-PolyA signal

Table 5.4: **Human DNA samples used in sequencing of the *CEL* gene.** Displays number of VNTR repeats in *CEL* alleles. The Danish family samples had alleles with known repeat numbers*. The Pancreas Biobank samples served as sequencing controls and their repeat numbers were provided by A. Molven (personal correspondence).

Sample ID	Origin	Allele 1	Allele 2
M28-102	Danish family	3	16
M28-111	Danish family	15	16
M28-141	Danish family	3	16
0113-B1	Pancreas Biobank	13	13
0109-B1	Pancreas Biobank	13	15

* Number of VNTR repeats were known from Torsvik et al. Mutations in the VNTR of the carboxyl-ester lipase gene (CEL) are a rare cause of monogenic diabetes. *Human Genetics*, 127(1):55–64, jan 2010. ISSN 03406717. doi:10.1007/S00439-009-0740-8.

Table 5.5: **PCR, restriction digestion, agarose gel electrophoresis and DNA sequencing.**

Material	Supplier	Cat. number
2-log DNA Ladder	New England BioLabs	N3200S
GC Buffer I	Takara	SD1432
BamHI	New England BioLabs	R0136S
Betaine	Sigma	B0300
BigDye Terminator 5X Seq. Buffer	Applied Biosystems	4336697
BigDye Terminator v3.1 Ready reaction mix	Applied Biosystems	4336911
dNTP Mixture	Takara	SD0316
EtBr	Amresco	E406-15ML
ExoProStar	Cytiva	US77705
Gel Loading Buffer	Sigma	G2526-5ML
LA Taq	Takara	RR02AG
MicroAmp Optical 96-well reaction plate	Applied Biosystems	N8010560
Multiscreen 96-well plate	Millipore	MAHVN4510
NEBuffer 3.1	New England BioLabs	B7203S
NuSieve GTG Agarose	Lonza Bioscience	50084
Sephadex G-50	Sigma	G5050-50G
TBE buffer	PanReac AppliChem	A3945,1000

Table 5.6: **Cell culture.**

Material	Stock conc.	Supplier	Cat. number
Cell culture plate, 6-well		Sarstedt	83.3920
Cell scraper		Fisherbrand	08-100-240
cOmplete Mini, EDTA-free Protease inhibitor tablets		Roche Diagnostics GmbH	11836170001
DMEM	1X	Gibco	31966-021
DPBS	1X	Gibco	2235050
Dual Chamber Cell Counting Slides		BioRad	1450011
FBS	10X	Gibco	10270-106
HEK293 Cell line		ATCC	CRL-1573
Lipofectamine 3000 Transfection kit		Invitrogen	L3000-008
Lipofectamine2000		Invitrogen	11668-019
OPTI-MEM	1X	Gibco	31985-062
RIPA buffer	1X	Thermo Scientific	89901
Serological Pipette 10 mL		Thermo Scientific	170356N
Serological Pipette 2 mL		Sarstedt	86.1262.011
Serological Pipette 25 mL		Sarstedt	86.1685.001
Serological Pipette 5 mL		Sarstedt	86.1253.001
T25 cell culture flask		Sarstedt	83.3910.002
T75 EasYFlask		Thermo Scientific	156499
Trypsin-EDTA	1X	Gibco	25300-054

Table 5.7: **SDS-PAGE and western blotting.**

Material	Supplier	Cat. number
Albumin Standard	Thermo Scientific	23209
Blotto Non-Fat Dry Milk	ChemCruz	G0419
Bolt 10% Bis-Tris Plus 1.0mm X 15 well	Invitrogen	NW00105BOX
Bolt 4-12% Bis-Tris plus 1.0mm X 10 well	Invitrogen	NW04120BOX
Chromatography paper	Whatman	3030-866
ECL Prime Western Blotting Detection Reagent A	Amersham	RPN2232V1
ECL Prime Western Blotting Detection Reagent B	Amersham	RPN2232V2
ECL Prime Western Blotting Detection Reagents	Amersham	RPN2232
EMSURE Methanol	Merck	1.06009.2511
NuPAGE Antioxidant	Invitrogen	NP0005
NuPAGE LDS Sample Buffer (4X)	Invitrogen	NP0008
NuPAGE Sample Reducing agent (10X)	Invitrogen	NP0009
NuPAGE Transfer Buffer (20X)	Invitrogen	NP0006-1
PBS Tablets	Gibco	18912-014
Pierce BCA Protein Assay kit	Thermo Scientific	23227
Pierce BCA Protein Assay reagent A	Thermo Scientific	23228
Pierce BCA Protein Assay reagent B	Thermo Scientific	23224
Pierce ECL Plus	Thermo Scientific	32132
Precision Plus Dual Color	Bio-Rad	161-0374
Probumin BSA Powder	Sigma-Aldrich	82-100
PVDF Transfer Membrane 0.45 µm	Thermo Scientific	88518
Restore PLUS Western Blot Stripping Buffer	Thermo Scientific	46430
Restore PLUS Western Blot Stripping Buffer	Thermo Scientific	46430
Trans-Blot Turbo Transfer Pack Midi	Bio-Rad	1704157
Trans-Blot Turbo Transfer Pack Mini	Bio-Rad	1704156
TWEEN 20	Sigma	P1379-500ML
UV 96-well plate	CORNING	3635

Table 5.8: **Antibodies used in western blotting.**

Antibody	Type	Produced in	Dilution	Incubation	Supplier	Cat. number / ref.
anti-CEL	1°	Rabbit	1:5,000	Overnight, 4°C	Gift from St. Louis lab	*Xiao et al.* (2016)
anti-β-Actin	1°	Mouse	1:10,000	1 h, RT	Sigma	A5441-100U
anti-rabbit-HRP	2°	Goat	1:10,000	1 h, RT	Invitrogen	65-6120
anti-mouse-HRP	2°	Goat	1:5,000	1 h, RT	Invitrogen	62-6520

Table 5.9: **Instruments used in the project.** The list is thorough, but not exhaustive.

Instrument	Use	Supplier
ABI 3500xL Genetic Analyzer	Sanger sequencing	Applied Biosystems
AccuSpin Micro 17R	Centrifugation	Fisher scientific
Betzy supercomputer	Running MD simulation	Norwegian Research Infrastructure Services
G:BOX iChemi XR5	Western blot imaging	Syngene
GeneFlash	DNA in agarose gel imaging	Syngene
PowerPac 300	Electrophoresis power supply	Bio-Rad
QIAexpert	DNA spectrophotometry	QIAGEN
Safety Cabinet HERAsafe KS 12	Sterile flow-hood	Kendro Laboratory Products GmbH
Steri-cycle CO2 Incubator HEPA Class 100	Cell incubation	Thermo Electron Corporation
TC20 Automated Cell Counter	Counting cells	Bio-Rad
ThermoMixer C	Sample incubation	Eppendorf
Trans-Blot Turbo	Protein transfer	Bio-Rad
Varioskan Lux	Protein concentration assay	Thermo Scientific
Xcell SureLock Electrophoresis Cell	Chamber for SDS-PAGE	Life Technologies

Table 5.10: **Digital tools.** Programs, websites, databases, scripts and any other application used for bioinformatic or other computational tasks. Version is displayed when applicable. 'Service' column tells whether the tool was run locally or on a server. 'Platform' column tells which operating system, coding language, program, or mode of connection was used for running the tool. For full references, see the end of the thesis.

Tool	Version	Service	Platform	Publisher or developer
BLASTp	2.13.0+	Online	Browser	NCBI *Altschul et al.* (1990, 1997, 2005)
CHARMM-GUI Glycan reader and modeler		Online	Browser	*Jo et al.* (2008, 2011) *Park et al.* (2017, 2019)
CHARMM-GUI Solution Builder		Online	Browser	*Jo et al.* (2008) *Lee et al.* (2016, 2020)
Clustal Omega		Online	Browser	*Sievers et al.* (2011) *Baxevanis et al.* (2020) *Sievers and Higgins* (2018)
ColabFold	1.3+	Online	Browser	*Mirdita et al.* (2022)
DSSP		Software	Linux	*Touw et al.* (2015) *Kabsch and Sander* (1983)
EMBOSS Cons		Online	Browser	*Rice et al.* (2000)
Excel	2210	Software	Windows	Microsoft
ExPASy Compute pI/Mw tool		Online	Browser	Swiss Institute of Bioinformatics
ExPASy Translate tool		Online	Browser	Swiss Institute of Bioinformatics
Fiji		Software	ImageJ2	*Schindelin et al.* (2012)
FinchTV	1.4.0	Software	Windows	Geospiza, Inc
GeneSys	1.2.5.0	Software	Windows	Syngene
GeneTools	4.02.03	Software	Windows	Syngene
ImageJ2		Software	Windows	*Rueden et al.* (2017)
iTOL	6.6	Online	Browser	biobyte solutions *Letunic and Bork* (2021)
IUPred3		Online	Browser	*Erdős and Dosztányi* (2020) *Mészáros et al.* (2018)
Jalview	2.11.2.5	Software	Windows	*Waterhouse et al.* (2009)
Matplotlib	3.7.1	Software	Python 3.7	*Hunter* (2007)
NAMD	3.0a11_64_CUDA	Software	Sigma2/Linux	University of Illinois at Urbana–Champaign: Theoretical and Computational Biophysics Group Parallel Programming Laboratory
phyloT	2	Online	Browser	biobyte solutions
PROPKA3	3	Software	Python 3.7	*Søndergaard et al.* (2011) *Olsson et al.* (2011)
PyMOL	2.5.4	Software	Windows	Schrödinger
Rg-scripts		Software*	VMD Tk	Mohamed shehata
Sigma2		Online	Linux SSH	Norwegian Research Infrastructure Services
SkanIt	5.0.0.42	Software	Windows	Thermo Fisher Scientific Oy
sscache	1.0	Software*	VMD Tk	Andrew Dalke
Taxonomy Database		Online	Browser	*Schoch et al.* (2020)
VEGA ZZ	3.2.3	Software	Windows	*Pedretti et al.* (2021)
VMD	1.9.3	Software	Windows	University of Illinois at Urbana–Champaign: Theoretical and Computational Biophysics Group Parallel Programming Laboratory

* scripts published openly on the web.

6 Methods

6.1 Wet lab methods

6.1.1 Plasmid transformation and bacterial culturing

LB agar plates were pre-heated to 37°C for 45 min. The plasmid samples were added to thawed OneShot TOP10 competent bacteria, followed by incubation on ice for 5 min. The transformed bacteria were inoculated onto the LB agar plates and incubated overnight at 37°C. One colony from each agar plate was inoculated into LB medium with 0.1 mg/mL ampicillin. The pre-cultures were incubated at 37°C in a shaker for 4 hours. Fractions of the pre-cultures were inoculated into more LB medium with 0.1 mg/mL ampicillin. The cultures were incubated overnight at 37°C in a shaker, before being centrifuged at 6000 x g at 4°C for 15 min. The supernatants were discarded and the pellets were used for DNA isolation.

6.1.2 Plasmid isolation

A QIAfilter Midi DNA isolation kit was applied for plasmid preparation according to the instructions from the manufacturer. The bacterial pellets were dissolved in buffer P1. Buffer P2 was mixed in and the solution was incubated for 5 min. Buffer P3 was added, followed by incubation for 10 min. The subsequent lysates were pressed through QIAGEN cartridges and into QIAGEN tips that had been equilibrated with buffer QBT. The QIAGEN tips were washed with buffer QC. DNA was eluted from the QIAGEN tips by using buffer QF. The elutions were mixed with isopropanol and centrifuged at 15,000 x g at 4°C for 30 min. Then, the pellets were submerged in 70% ethanol and centrifuged at 3,000 x g at 4°C for 15 min, followed by air-drying for 75 min and dissolving in 1X TE buffer overnight at room temperature (RT). Finally, the dissolved plasmids were centrifuged at 3,000 x g for 1 min at RT to remove any remaining particulate matter. The supernatants were collected and used as the finished plasmid samples. DNA concentration and purity was measured using a QIAxpert spectrophotometer. The plasmid samples were stored at -20°C when not in use.

6.1.3 Plasmid sequencing

A series of PCR reaction mixes (0.5 µM primer, 1 M betaine, 0.2x sequencing buffer, 0.1x Big Dye v3.1 and 600–1000 ng template DNA) were prepared with varying combinations of plasmids and primers. Sequenced plasmids were: CEL-WT, CEL-3R-USA, CEL-3R-DAN, CEL-HYB, CEL-TRUNC and EV. Applied primers were: T7, CF, BR, DF, CR, DR, EF and BGH (see Table 5.3). The PCR machine was initiated at 96°C for 1 min. The machine subsequently performed 25 cycles of: 96°C for 0.1 sec, 58°C for 0.05 sec and 60°C for 4 min. Next, the PCR reaction products were purified by centrifugation through a Multiscreen 96-well plate with Sephadex-G50 gels at 910 x g for 5 min. The purified reaction products were collected into a MicroAmp Optical 96-well reaction plate and sequenced using an ABI 3500xL capillary sequencer.

6.1.4 Agarose gel electrophoresis of plasmids

One parallel of plasmid samples were left undigested and remained native. Another parallel of plasmid samples were digested and linearised with BamHI at 37°C for 15 min. Agarose gels were prepared (final concentrations: 1% (w/v) agarose and one droplet of EtBr in 1X TBE buffer), placed in electrophoresis trays and submerged in 1X TBE buffer. Samples of 200 ng DNA with loading buffer were loaded into the gels. Electrophoresis was performed at 80 V for up to 2 h. Gels were imaged using a Syngene GeneFlash.

6.1.5 Sequencing of the *CEL* VNTR region in human DNA samples

Long PCR reaction mixes (final concentrations: 0.9X GC-buffer I, 0.36 mM dNTP mix, 0.54 µM L11F primer, 0.54 µM VNTR-R primer, 10 ng template DNA, 0.89 M betaine, 0.022 U/µL LA Taq DNA polymerase) were prepared with human DNA samples (see Table 5.4). The PCR machine was initiated at 94°C for 1 min. The program performed 14 cycles of 94°C for 20 sec and 60°C for 10 min, followed by 20 cycles of 94°C for 20 sec and 62°C for 10 min. The reaction ended with incubation at 72°C for 10 min. PCR products were purified by incubation with ExoProStar at 37°C for 15 min and at 80°C for another 15 min. The prepared PCR reactions were sequenced as explained in Section 6.1.3 'Plasmid sequencing', but with a slightly different PCR mix (final concentrations: 0.5 µM primer, 0.5 M betaine, 0.2x sequence buffer, 0.1x BigDye v3.1, 3 µL purified DNA) and using the primers EF and VNTR-R (see Table 5.3).

6.1.6 Cell culturing

All cell culture operations were performed under sterile conditions in a ventilation hood. Frozen HEK293 cells of passage 14 were thawed in a water bath at 37°C and transferred to DMEM (Dulbecco's Modified Eagle Medium) in a 15 mL Falcon tube. The tube was centrifuged at 13,000 x g for 5 min. The supernatant was decanted and the pellet was dissolved in DMEM with 10% FBS and transferred to a T25 or T75 flask. The cells were grown to 60-90% confluency in a humidified atmosphere at 37°C and with 5% CO_2. Then, the cells were washed with DPBS (Dulbecco's Phosphate Buffered Saline) and incubated in Trypsin-EDTA for 5 min. The cells were split by transferring a portion into a new T25 or T75 flask with DMEM with 10% FBS. Passaging of cells was performed weekly to prevent growth beyond 90% confluency. The HEK293 cells were in use up until passage 30.

6.1.7 Transfection and harvesting of cells

The wells of a 6-well plate were seeded with $5{\cdot}10^5$ HEK293 cells in 2 mL DMEM. The cells were incubated for 24 h at 37°C before being transfected with Lipofectamine2000 or Lipofectamine3000. The Lipofectamine2000 transfection was performed by using 4 µg plasmid and 10 µL Lipofectamine2000 in OPTI-MEM. The mix was incubated for 20 min before being added to the cells. Lipofectamine 3000 transfection kit was applied similar to instructions from manufacturer, but by using 5 µL Lipofectamine3000, 3.5 µg plasmid and 7 µL P3000 reagent. Transfection mixes were added to respective wells 24 h after seeding. The DMEM was renewed 5-6 h after transfection. Then, the transfected cells were incubated for 48 h. The cell media were isolated and centrifuged at 13,300 x g for 5 min at 4°C. The medium supernatants were isolated and further analyzed as the medium fractions. The cells were washed with 2 mL DPBS, followed by lysis in 1X RIPA buffer with 1X protease inhibitor on ice. Lysed cells were harvested and incubated at 4°C for 30 min before centrifuged at 13,300 x g for 15 min at 4°C. Next, the supernatants were isolated and further analyzed as the lysate fractions. The remaining cell pellets were washed and centrifuged twice in DPBS. Finally, each pellet was dissolved in 100 µL loading solution (final concentrations: 2X LDS Sample Buffer, 10% (v/v) Reducing Agent, diluted in dH_2O). All samples were stored at -80°C.

6.1.8 Protein concentration assay

A Pierce BCA Protein Assay kit was applied to determine the total protein concentrations of the cell lysate samples. The assay was performed according to the manufacturer's recommendation. In short, Bovine Serum Albumin (BSA) standards were prepared by serial dilution in DPBS. Lysate samples were diluted 1:5 in DPBS. Two parallels of all standards and samples were added to a UV 96-well plate. A 50:1 mix of reagent A and reagent B was added to each well. The UV 96-well plate

was shaken at 600 RPM for 30 sec, followed by incubation for 5 min at RT. Spectrophotometry was performed by using a Varioskan Lux instrument and analysed with the software SkanIt.

6.1.9 SDS-PAGE

For the soluble lysate fractions, 15 µg protein in 20 µL loading solution (final concentrations: 1X LDS Sample Buffer, 1X Sample Reducing Agent, diluted in 1X RIPA) was loaded to the gel. For the medium fractions, the amount loaded was the same as for the corresponding lysate sample. Lysate and medium fractions were incubated at 56°C for 15 min before loaded to gels. The insoluble pellet fractions were incubated at 95 degrees for 15 min and centrifuged at 13,300 x g for 3 min at 4°C. Then, the pellet samples were loaded in volumes equivalent to 15 µg lysate. Gels were submerged in 1X MOPS (3-(*N*-morpholino)propanesulfonic acid) running buffer and antioxidant was added to inner chamber. Electrophoresis was performed at 80 V for 15 min, followed by 180 V for 1 h and 30 min.

6.1.10 Western blotting

The 0.45 µm polyvinylidene difluoride (PVDF) membranes were activated by incubation in methanol for 1 min, dH_2O for 1 min, then in 1X Transfer Buffer with 10% methanol until use. The 0.2 µm membranes in the Trans-Blot Turbo Transfer Packs were replaced with the activated 0.45 µm membranes. Gels were inserted into the blotting sandwiches and each transfer was run in a Trans-Blot Turbo at setting 'Mixed MW' (7 min, 1.3 A, up to 25 V). The membrane with the transferred lysate samples was cut in two at 50 kDa to separate the β-Actin loading control. All membranes were blocked using 5% (w/v) Non-Fat Dry Milk in PBS-Tween [0.05% (v/v) Tween in Phosphate Buffered Saline (PBS)] for 1 h at RT with 60 RPM rocking. The CEL-specific membranes were incubated in anti-CEL antibody (diluted 1:5,000 in 1% Non-fat Dry Milk in PBS-Tween) overnight at 4°C with rocking at 30 RPM. The loading control membrane was incubated in anti-β-Actin antibody (diluted 1:10,000 in 1% Non-fat Dry Milk in PBS-Tween) for 1 h at RT with 30 RPM rocking. All membranes were washed 3 x 5 min in PBS-Tween with 50 RPM rocking. Membranes were then incubated in corresponding secondary antibodies: anti-rabbit-HRP (diluted 1:5,000) and anti-mouse-HRP (diluted 1:10,000) in 1% Non-fat Dry Milk in PBS-Tween (see Table 5.8). Incubations were performed for 1 h at RT with 30 RPM rocking. All membranes were washed 1 x 10 min and 3 x 5 min in PBS-Tween with 50 RPM rocking. Membranes were exposed to a 1:1 mix of ECL Prime reagents for 5 min and imaged in a Syngene G:BOX iChemi XR5.

6.2 *In silico* methods

6.2.1 CEL sequence search

A BLASTp search was performed (September 26th, 2022) for 5000 sequences using human CEL as query sequence (Swiss-Prot: P19835). The description tables were downloaded for all mammalian hits and non-mammalian, vertebrate hits, respectively.

6.2.2 Filtering the CEL sequences into datasets

Sequences were considered ineligible if they had a description or metadata which contained any of the capital-insensitive strings in Table 11.1. Additionally, sequences with access codes beginning with ”NW” or ”NX” were removed from the vertebrate dataset. These refinements were intended to remove non-CEL sequences as well as CEL sequences that may not correctly represent the VNTR region. Sequences with accession codes XP_035411118.1, XP_035411117.1 and XP_009091641.2 were removed from the vertebrate dataset due to GenBank annotation updates. Sequences were also considered ineligible if the phrases: ”Gnomon”, ”RefSeq” or ”Conceptual translation” were not in the metadata. This was to ensure that the amino acid sequence was based on DNA, and not structural techniques. The non-mammalian, vertebrate hits had 163 eligible sequences from 109 species. The mammalian hits had 216 eligible sequences from 156 species. Eligible sequences in the mammalian collection and the non-mammalian, vertebrate collection were ordered by E-value (or by Total score for entries with equal E-values). An exception was made for *Homo sapiens*: the sequence with accession code AAA63211.1 was selected due to having 16 VNTR repeats.

6.2.3 Annotating for VNTR properties

The highest scoring sequence of each species was reviewed and, if VNTR was present, annotated. The reviewed non-mammalian, vertebrate sequences will be referred to as the non-mammalian, vertebrate dataset. The reviewed and annotated mammalian sequences will be referred to as the mammalian dataset. The sequences in the non-mammalian, vertebrate dataset were not annotated for VNTR properties as no VNTRs were found. Each sequence in the mammalian dataset was annotated for a ’globular domain’, a ’VNTR’ and a ’post-VNTR’, if possible. The VNTRs were further annotated for individual repeats. The consensus sequence of all the repeats in each VNTR was computed using EMBOSS Cons with normal settings. The mammalian dataset was appended with the VNTR annotations, such as: number of repeats, repeat lengths, consensus repeat sequence, post-VNTR sequence and VNTR interpretability. Clustal Omega and Jalview were used for aligning and analysing sequences, respectively.

6.2.4 Phylogenetic tree

The data from the sequence reviews and annotations were combined with phylogenetic trees to make figures. Small trees of up to 9 species were constructed in Newick format by the tool phyloT, which itself acquired taxonomies from the NCBI Taxonomy Database. The phylogenetic trees constructed by phyloT were merged into larger trees using a Python script written by the author of the thesis. Then, the large phylogenetic trees were visualised in online tool iTOL.

6.2.5 Generating files for MD

A protein structure of the CEL-3R VNTR region was constructed using ColabFold prediction with standard settings. Then, the pKa values of the structure were predicted using PROPKA3. Glycans were modeled using CHARMM-GUI Glycan reader and modeler. Next, MD simulation files were generated using CHARMM-GUI Solution Builder, with a periodic simulation box solvated with TIP3P water molecules and 6 x K^+ ions for system neutralisation. The following parameters were activated: 'automatic PME FFT generation', 'CHARMM36m force field', 'Hydrogen Mass Repartitioning' (HMR), 'WYF parameter for cation-pi interaction' and 'NAMD'. The 'NPT ensemble' was set to 310 K (equivalent to 37°C). The subsequent MD simulation files were downloaded.

6.2.6 MD equilibration and production

The equilibration and production input files generated by CHARMM-GUI were modified. The 'timestep' was set to 4 fs, 'LangevinPistonPeriod' was set to 200, 'LangevinPistonDecay' was set to 100, 'pairlistdist' was set to 14, 'margin' was set to 4 and 'dcdfreq' was set to 25000. For the production input file, the 'CUDASOAintegrate' term was activated to allow for computations by a Graphics Processing Unit. Files were uploaded to the Betzy supercomputer from the Sigma2 service. Finally, the MD simulations were run using the program NAMD.

6.2.7 MD analysis

The subsequent simulation trajectories were stripped of water molecules using VEGA ZZ. Then, the trajectories were analyzed using VMD. The simulation energy outputs were inspected using the 'NAMD Plot' tool in VMD. The secondary structures in the trajectories were assigned and visualised in VMD with the aid of the 'sscache' script. Further, the secondary structures were separately assigned and calculated using the tool DSSP. Also, the root mean square deviation (RMSD) and radius of gyration (RoG) values of the trajectories were analysed using VMD and 'Rg-scripts', respectively.

7 Results

7.1 Phylogenetic analysis

As explained previously, the number of VNTR repeats in *CEL* varies greatly in human populations. The most common allele has 16 VNTR repeats, while the Danish family with diabetes carried a variant with only 3 repeats. Notably, this is the same number of repeats found in the mouse *Cel* gene (*Mackay and Lawn*, 1995). We therefore decided to start the project with a phylogenetic analysis of the CEL protein with the intention that this might shed light on the evolution and function of the VNTR region. Moreover, such information could be useful in the interpretation of the possible pathogenic effect of the CEL-3R variant.

7.1.1 Vertebrate CEL and origin of VNTR

For the phylogenetic analysis, CEL sequences were acquired from various species by a BLASTp search performed with a human CEL query sequence. This resulted in 5000 sequence hits that were segregated into mammalian species and non-mammalian, vertebrate species. The non-vertebrate species were excluded for two reasons: CEL had previously been described as a vertebrate protein in literature (*Wang and Hartsuck*, 1993; *Holmes and Cox*, 2011) and the non-vertebrate sequence hits had been assigned with questionable species. In an effort to isolate the most trustworthy sequences and to limit the analysis, a large number of the sequences were filtered away based on their metadata (Figure 7.1).

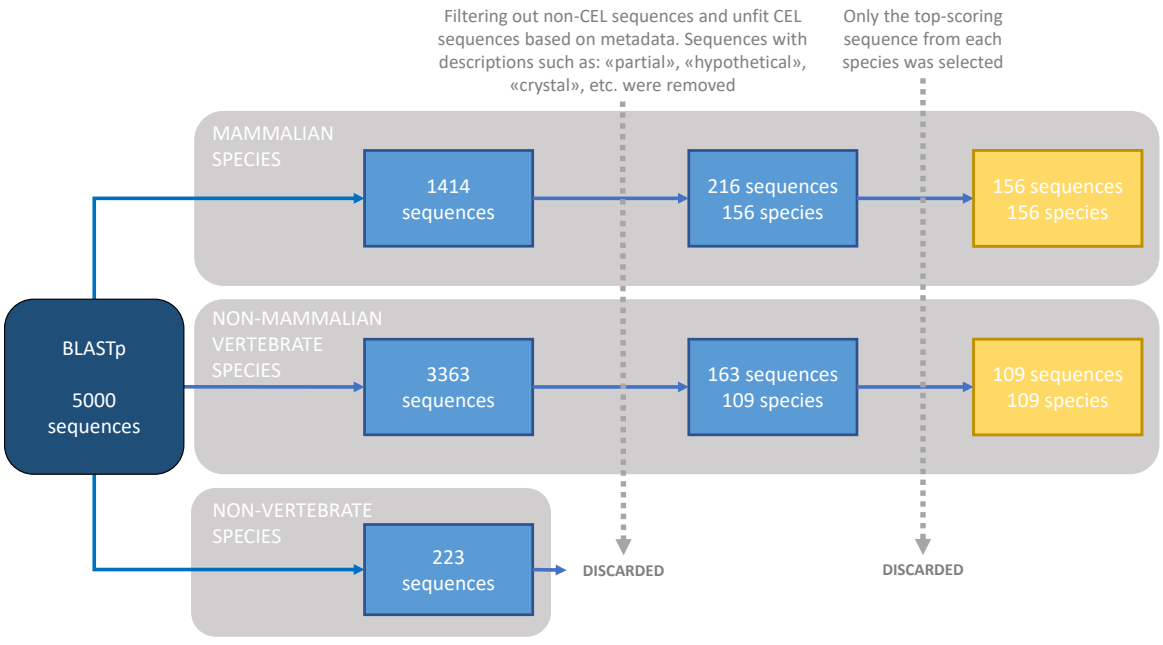
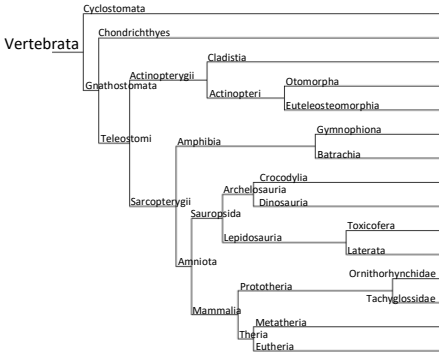


Figure 7.1: **Filtering of 5000 protein sequence hits from the BLASTp search.** The search was performed using the canonical CEL sequence from Swiss-Prot (P19835).

The CEL sequences of fifteen vertebrate species were compared. The species were selected based on phylogenetic diversity (representing major animal clades), as well as for having easily interpretable sequences. The sequences were reviewed for the presence of a VNTR and the resulting data were combined with a phylogenetic tree in order to display the evolutionary relationship of the species (Figure 7.2). We found that the VNTR is exclusively present in mammalian CEL sequences, more specifically in mammals of the clade Theria. The 13 other non-therian species exhibited CEL sequences without VNTRs. However, the metatherian species Australian echidna did display a CEL sequence with a proline-rich, C-terminal extension which resembled the PEST sequence of the VNTR.

In addition to the species in Figure 7.2, the CEL sequences of the 163 eligible non-mammalian, vertebrate species were manually reviewed without finding a VNTR in any of them (data not shown). However, a CEL sequence from Tanaka's snailfish *Liparis tanakae* contained a C-terminal region of 29 consecutive lysine repeats (GenBank: TNN38450.1). A few other vertebrate sequences also displayed repeats of large regions within the CEL protein. However, none of these were remotely recognised as the VNTR typically seen in the C-terminal "tail" of CEL.



Scientific Name	Common Name	VNTR	Protein length	Accession
Petromyzon marinus	Sea lamprey	**No**	568	XP_032832354.1
Rhincodon typus	Whale shark	**No**	562	XP_048469723.1 *
Erpetoichthys calabaricus	Reedfish	**No**	557	XP_028664415.1 *
Danio rerio	Zebrafish	**No**	550	AAI24754.1
Salmo salar	Atlantic salmon	**No**	555	XP_013984108.1 *
Rhinatrema bivittatum	Two-lined caecilian	**No**	557	XP_029467137.1
Rana temporaria	Common frog	**No**	552	XP_040180373.1
Alligator sinensis	Chinese alligator	**No**	559	XP_006029883.1
Gallus gallus	Chicken	**No**	556	NP_001013015.1 *
Python bivittatus	Burmese python	**No**	559	XP_007420625.1
Zootoca vivipara	Common lizard	**No**	557	XP_034958083.1
Ornithorhynchus anatinus	Platypus	**No**	557	XP_007670968.2
Tachyglossus aculeatus	Australian echidna	**No** **	639	XP_038601598.1
Sarcophilus harrisii	Tasmanian devil	**Yes**	631	XP_003757493.1
Homo sapiens	Human	**Yes**	742	AAA63211.1 *

Figure 7.2: **Phylogenetic analysis showing the unique presence of CEL VNTR in mammals.** All branches are labelled with corresponding clades. Important clades have different shades. Vertebrate species were selected from a BLASTp search with a query sequence from human CEL (P19835. Taxonomy was acquired from NCBI Taxonomy Database by online tool phyloT. Phylogenetic tree was visualised in iTOL.

*Alternative, eligible sequences were available for this species. **Non-VNTR, 82 aa, proline-rich extension on C-terminal.

The sequences from Figure 7.2 were also subjected to a multiple sequence alignment (MSA). The alignment was performed on the CEL sequences excluding any C-terminal extensions beyond the globular domain (Figure 7.3). Note that the annotated residue positions are in reference to the human CEL structure in Figures 4.3 and 11.8. Overall, the alignment displayed many matching residues and few internal gaps. In total, there were 170 residues with complete conservation across all 15 species. Some of the least conserved regions were the signal peptide at 1-23 and a region at 352-411 which overlaps with the CEL ”small domain” (see Figure 11.8).

The alignment also displayed the conservation of residues that are known to be important in the CEL globular domain (Figure 7.3). The residues Ser217, Asp343 and His458 which constitute the catalytic triad are completely conserved, while many of their immediate neighbors are also well conserved. The same holds for the Cys residues which form the disulfide bridges Cys87-Cys103 and Cys269-Cys280. The *N*-glycosylated Asp210 did not display a particularly high conservation value. However, all species did have at least one Asp residue in close proximity to position 210. The phosphorylated Thr363 residue was even less conserved. Some species displayed the OH-containing Ser or Tyr residues instead of Thr. Other species, like the Atlantic salmon and Australian echidna, displayed Pro and Ala residues, while the closest OH-containing residues were at least 2 positions away.

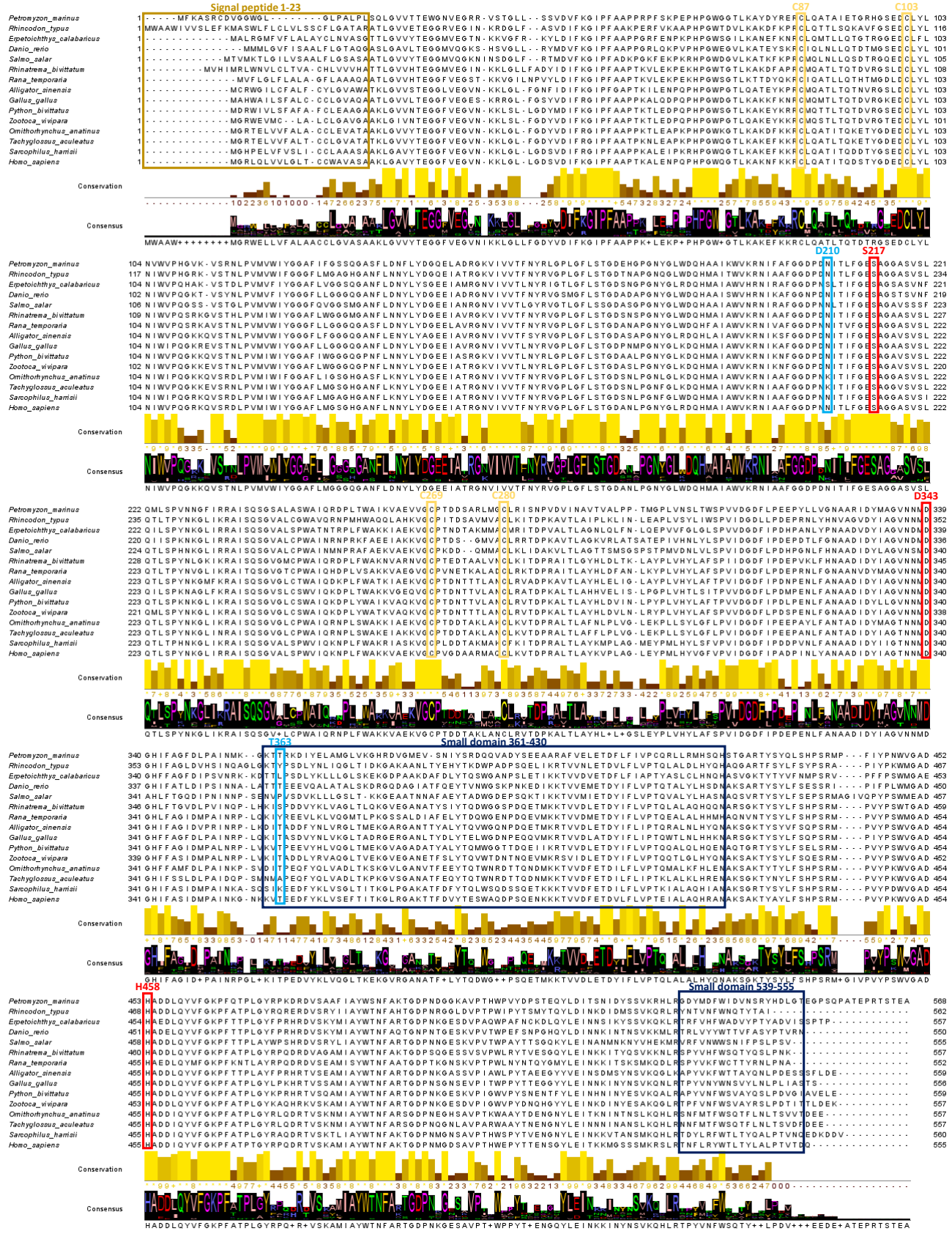


Figure 7.3: **MSA of CEL globular domain sequences from the 15 vertebrate species.** Annotations are according to a human CEL reference (see Figures 4.3 and 11.8). The C-terminal, post-globular domain sequences were removed from the species *Tachyglossus aculeatus*, *Sarcophilus harrisii* and *Homo sapiens*. **Consensus** shows the residue propensities[♦♦]. **Conservation** scores show the similarities of the residues[♦♦♦]. The symbols '+' and '*' denote fully conserved properties and residues, respectively.

[♦♦] as defined by J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, may 1982. ISSN 0022-2836. doi:10.1016/0022-2836(82)90515. [♦♦♦] as defined by C. D. Livingstone and G. J. Barton. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Computer applications in the biosciences : CABIOS*, 9(6):745–756, 1993. ISSN 0266-7061. doi:10.1093/BIOINFORMATICS/9.6.745.

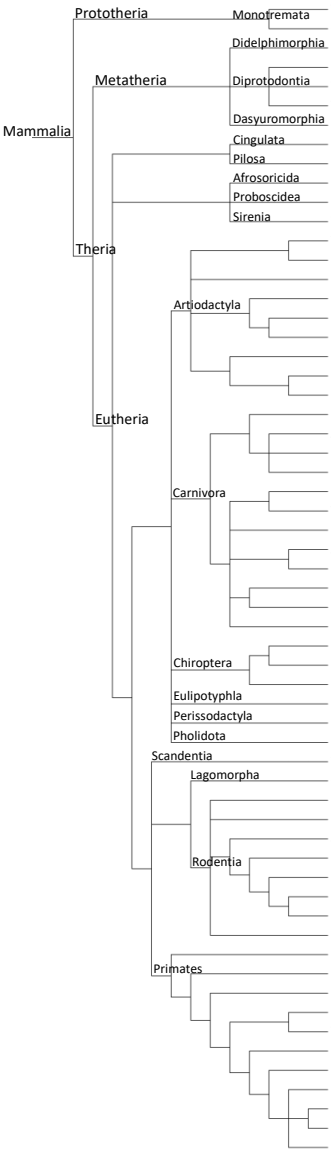
7.1.2 Diversity of VNTR in Mammalia

We then decided to analyse the diversity of the CEL VNTR in Mammalia, and the identified 156 mammalian VNTR sequences were manually annotated. Next, these annotations were collected into a dataset of VNTR traits of the mammalian CEL protein. From this dataset, the annotations for 60 species were selected and juxtaposed against a mammalian phylogenetic tree to show evolutionary relationships (Figure 7.4). These 60 species were selected for phylogenetic diversity, as well as easily interpretable VNTRs.

The number of repeats was highly variable across Theria (Figure 7.4). The lowest number of repeats was found in the Agile gracile mouse opossum with only one repeat, while the highest number was found in the Western gorilla with 39 repeats. Evolutionarily close species tended to have a similar number of repeats. Species with 10 or more repeats were mainly present in orders Carnivora and Primates. Other species with 10 or more repeats were the Tasmanian devil, African savannah elephant and Woodchuck. As already discussed in the vertebrate analysis, the VNTR was not present in Prototheria.

The mammalian phylogenetic tree of CEL VNTR sequences showed that the repeat lengths varied between Metatheria and Eutheria (Figure 7.4). The metatherians were annotated with repeats of 5 residues, as opposed to 11 in the eutherians. Despite the difference in repeat length, both Metatheria and Eutheria displayed residues with similar properties. The eutherian consensus repeats usually contained residues Glu, Pro, Thr, Ser, Asp, Gly, Ala and Val. The order of appearance of these residues was well conserved.

The consensus repeats in Metatheria also contained the residues Pro, Asp, Val and Ala, but less Thr and Ser (Figure 7.4). Also, the metatherians displayed a couple of Lys residues. The residue orders between Metatheria and Eutheria were less correlated. The greatest similarity in the order was that the Val and Pro residues were usually positioned next to each other.



Species	Common name	Consensus repeat	Repeat length	N repeats	Post-VNTR	Accesion number
T. aculeatus *	Austral. echidna	-	-	-	-	XP_038601598.1
O. anatinus	Platypus	-	-	-	-	XP_007670968.2
G. agilis *	Agile g. m. opossum	D Q A P V	5	1	PTAADAEVTSALPPVT…	XP_044517241.1
P. cinereus	Koala	E D K V P	5	3	LIAMAADS	XP_020837662.1
T. vulpecula	Common brushtail	P E D K V	5	3	PVYMAADS	XP_036607804.1
V. ursinus	Common wombat	g E n k v	5	2	PIATAADS	XP_027725445.1
S. harrisii *	Tasmanian devil	P E G S V	5	10	PIAPVADSI	XP_003757493.1
D. novemcinctus	Nine-b. armadillo	E t S p L P t v D G q	11	2	-	XP_004485292.2
C. didactylus	South. two-toed sloth	E A N P V P P A D D S	11	3	EEPQMPIVIGF	XP_037653412.1
E. telfairi	Small Mdg. hedgehog	E S A P x P P K G D S	11	3	EGQMPIVIGF	XP_004713286.1
L. africana	Afrc. svnh. elephant	E A A P V P P T D D P	11	14	KEAQMPLGGHNGGNP…	XP_003407452.1
T. m. latirostris	Florida manatee	E A A P V P P T x D S	11	3	EEAQMPIVIGF	XP_004375850.1
V. pacos	Alpaca	a S t s L p P M a D S	11	2	EVAQMPIAIGF	XP_006216679.1
C. dromedarius	Arabian camel	a S T s L p P M a D S	11	2	EVARMPIAIGF	XP_010973713.1
S. scrofa	Pig	E A T P A P P T G D S	11	4	EVAQMPIVIGF	XP_003353742.2
C. elaphus	Red deer	E A S x V P P A D N S	11	3	EVAQMPVVIGF	XP_043772433.1
O. aries	Sheep	E A S S V P P A D N S	11	3	EVAQMPVVIGF	XP_004005621.5
B. taurus	Cow	G A S x V P P A D N S	11	3	EVAQMPVVIGF	NP_001013601.1 **
B. a. scammoni	Minke whale	E p T P a p P t G D S	11	4	EVAQMPVIIGF	XP_007194497.1
O. orca	Killer whale	E A T P A P T A G D S	11	4	EVAQMPLVIGF	XP_033282762.1
D. leucas	Beluga whale	E A T P A P P A G D S	11	5	EVAQMPIVIGF	XP_022434031.1
H. hyaena	Striped hyena	D A A P A P P S G D S	11	6	EGAQMPAVIGF	XP_039085574.1
A. jubatus	Cheetah	D A A P A P P S G D S	11	7	EGVQMPAVIDF	XP_026891324.1
F. catus	Domestic cat	D A A P A P P S G D S	11	11	EGVQMPAVIGF	XP_011278010.1
P. leo	Lion	D A A P A P P S G D S	11	12	EGVQMPAVIVF	XP_042768715.1
C. l. familiaris	Dog	N A P P A P P S G D S	11	13	EGAQMPIVIGF	XP_548401.1
V. vulpes	Red fox	D A A P A P P S G D S	11	14	EGAQMPVVIGF	XP_025865362.1
C. ursinus	Northern fur seal	D A T P t P P S G D S	11	4	EGAQMPIVIGF	XP_025743751.1
A. melanoleuca	Giant panda	D A A P a P P S G D S	11	18	EGAQMPAVIGF	XP_034520642.1
U. arctos	Brown bear	D A A P A P P S G D S	11	5	EGAQMPAVIGF	XP_026370049.2
M. putorius furo	Domestic ferret	E A A P A P P S G D S	11	7	EGAQMPVVIGF	XP_004757098.1
G. g. luscus	Wolverine	D A A P A P P S G D S	11	5	EVAQMPVVIGF	KAI5763052.1
H. grypus	Gray seal	D A A P A P P S G D S	11	9	EGAQMPIVIGF	XP_035963565.1
M. lucifugus	Little brown bat	x A V P V P P S D D S	11	3	EGAQMSAAIGF	XP_006104472.1 **
D. rotundus	Com. vampire bat	g A T P V P P A G g S	11	2	EGAPMPAVIGF	XP_024418293.2
R. aegyptiacus	Egyptian rousette	D V T P A P P T D G P	11	7	EGAQMPVVIGF	XP_036073986.1
S. etruscus	Etruscan shrew	E a A P L P P T D D S	11	4	EETQVPLAIGF	XP_049630017.1
E. caballus	Horse	E A A P x P P S G D S	11	3	EGAWMPAVIGF	XP_001498324.1
M. javanica	Malay. pangolin	E A A P A P P A E D P	11	3	EGAQMPAVIGF	XP_036863713.1 **
T. chinensis	Chin. tree shrew	E A A P v P P t D D S	11	4	EGAPMPKVFEFCARIL…	XP_027622430.1 **
O. cuniculus *	Rabbit	E d V P v P P t D D A	11	2	VAAQMPMAIGF	NP_001095181.3
C. porcellus	Dom. guinea pig	E A V P V P P A D D S	11	3	LAPQMPAAIGF	XP_003470795.1
C. canadensis	American beaver	E A A P V P P A D D S	11	3	QGAQMPVVIGF	XP_020026272.1
J. jaculus	Lesser Egpt. jerboa	Q A x P V P P A D D S	11	3	METQMPVAIGF	XP_004654104.2
C. griseus	Chinese hamster	Q A A P V P P T D D S	11	8	VEAQMPVIIGF	XP_027278863.1 **
M. unguiculatus	Mongolian gerbil	Q A A P V P P x D D S	11	3	VEAQMPAVIGF	XP_021481963.1
R. norvegicus *	Norway rat	Q x x P V P P T D D S	11	4	QAGDSVEAQMPVAIGF	NP_058693.2 **
M. musculus	House mouse	Q V x P x P P S D D S	11	3	VEAQMPATIGF	NP_034015.1 **
M. monax *	Woodchuck	R P P P F P P A N D S	11	17	PSPPIPTRQ	XP_046321629.1**
L. catta	Ring-tail. lemur	E A T P V P P T G D S	11	4	VGAQMPTLIGF	XP_045419623.1
C. syrichta	Philip. tarsier	E A a P V P P T G D S	11	4	EGAQMPTVIGF	XP_008047713.1
C. jacchus	White-t. marmoset	E A x P v P P T G D S	11	8	KGAQMPAVIGF	XP_017832605.1 **
M. mulatta	Rhesus monkey	E A T P V P P T G D S	11	9	KGAQMPVVISF	XP_014971947.2 **
T. francoisi	Francois' langur	E A A P V P P T G D S	11	13	KGAQMSAVISF	XP_033051039.1 **
H. moloch	Silvery gibbon	E A A P V P P T G D S	11	5	KEVQMPAVISF	XP_032021402.1 **
P. abelii	Sumatran orangutan	G A P P V P P T G D S	11	19	KEVQMPAVISF	XP_024108043.1
G. gorilla	Western gorilla	G A P P V P P T G D S	11	39	KEAQMPAVISF	AAF71700.1 **
P. troglodytes	Chimpanzee	G A P P V P P T G D S	11	10	KEAQMPAVISF	XP_024201717.1 **
P. paniscus	Pygmy chimpanzee	G A P P V P P T G D S	11	11	KEAQMPAVISF	XP_008974297.2
H. sapiens	Human	G A P P V P P T G D S	11	16	KEAQMPAVIRF	AAA63211.1 ***

Figure 7.4: **Phylogenetic analysis of the CEL VNTR in mammalian species.** Prototheria are coloured blue, Metatheria are coloured green and Eutheria are coloured grey. All taxological orders are labelled and shaded differently. Mammalian species were selected from a BLASTp search with query sequence from human CEL (Swiss-Prot: P19835). Assigning the consensus residues of the repeats was performed by EMBOSS Cons. Capital letter indicates that the residue was in the majority. Lower case letter indicate that the residue was not in the majority. The letter 'x' indicates no consensus residue. 'Repeat length' column displays the number of amino acid residues in each VNTR repeat. 'N repeats' column shows the number of annotated repeats in the VNTR. 'Post-VNTR' column shows the annotated sequence after the VNTR. Taxonomy was acquired from NCBI Taxonomy Database by online tool phyloT. Phylogenetic tree was visualised in iTOL.

*some difficulty in annotating VNTR. **alternative, eligible sequences for this species were available. ***16 repeat sequence was prioritised over alternative, eligible sequences

7.1.3 The post-VNTR sequence in Eutheria

Characteristic of the human CEL protein is also a short sequence of 11 amino acids immediately following the VNTR region (see Figure 4.3). We denote this the "post-VNTR sequence". A comparison of the mammalian post-VNTR sequences can also be seen in Figure 7.4.

The metatherian post-VNTR sequences were somewhat similar to the VNTR repeats of eutheria (Figure 7.4). In contrast, the eutherian post-VNTR sequences were not similar to any sequence seen anywhere else. The post-VNTR in Eutheria tended to begin with the charged residues Glu or Lys, followed by mostly hydrophobic residues such as Ala, Val, Met and Ile, before almost always ending with the Phe residue.

In order to better understand the conservation of the CEL post-VNTR in Eutheria, an MSA was performed. From the 156 annotated CEL sequences in the mammalian dataset, 134 eutherian post-VNTR sequences were selected for having a length of 11 residues. The sequences were aligned and analysed (Figure 7.5). The alignment of the 134 post-VNTR sequences showed a correlation between residue conservation, hydrophobicity and proximity to the C-terminal (Figure 7.5). The consensus of the alignment yielded by Jalview was 'EGAQMPAVIGF', while EMBOSS Cons computed a similar sequence 'EGAQMPVVIGF'. The aligned sequences displayed decent conservation of the residues 'AQMP' and 'IGF'. The Pro and Ile residues showed even higher conservation. The C-terminal Phe residue was completely conserved as it was present in 100% of the 134 aligned sequences. The hydrophobicity and conservation of the residues tended to increase towards the C-terminal. Residues Ser or Thr were rare in the post-VNTR sequences.

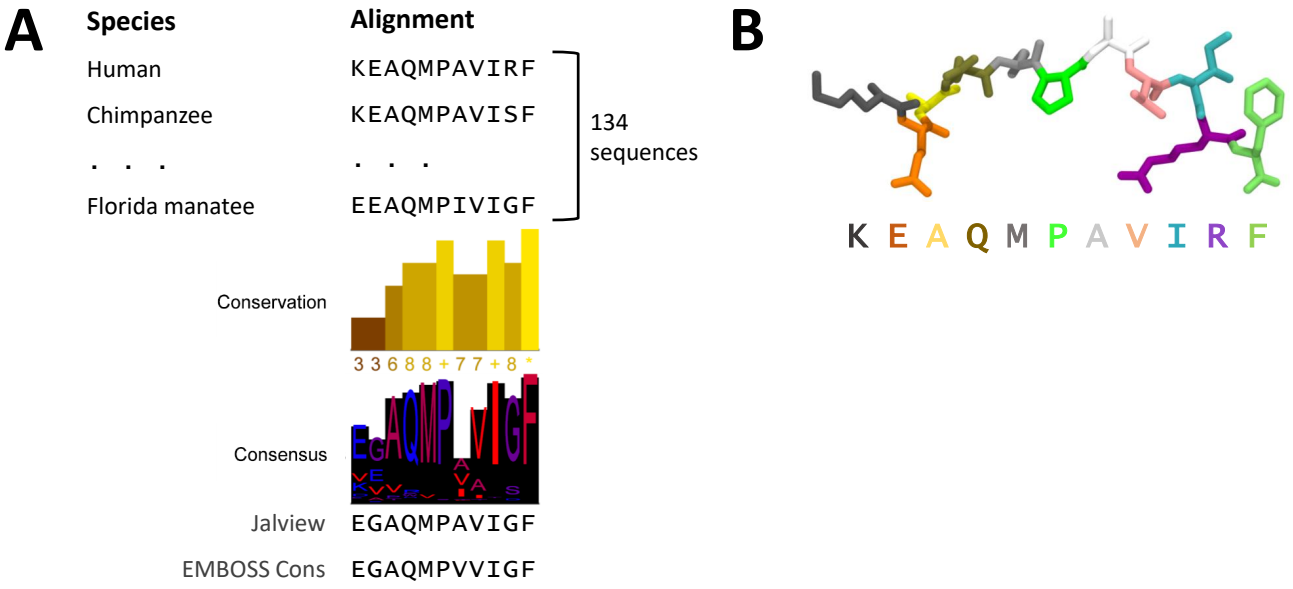


Figure 7.5: **MSA of CEL post-VNTR sequences in 134 eutherian species. A.** The CEL post-VNTR sequences of 134 eutherian species were aligned by Clustal Omega. Conservation levels and consensus histograms were produced in Jalview. **Conservation** scores show the similarities of the residues♦♦. The symbols '+' and '*' denote fully conserved properties and residues, respectively. **Consensus** shows the residue propensities♦♦♦. Hydrophobic and hydrophilic residues are coloured red and blue, respectively. **B.** A structural model of the human CEL post-VNTR peptide. Residues are colour-coded with the 'KEAQMPAVIRF' sequence. Visualised in VMD.

♦♦ as defined by C. D. Livingstone and G. J. Barton. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Computer applications in the biosciences : CABIOS*, 9(6):745–756, 1993. ISSN 0266-7061. doi:10.1093/BIOINFORMATICS/9.6.745. ♦♦♦ as defined by J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, may 1982. ISSN 0022-2836. doi:10.1016/0022-2836(82)90515.

7.2 Cellular analysis

The CEL-3R variant was suspected to be pathological, as it was correlated with the Danish family members who suffer from Diabetes (see Section 4.4.4: 'The CEL-3R variant'). As seen in other variants with confirmed pathology, the CEL protein tends to be involved in cellular aggregation and ER stress. Thus, we wanted to examine how the CEL-3R variant localises within human cells and how it undergoes the secretion pathway.

In order to study the cellular localisation and secretion of CEL-3R, we would express CEL-3R in cell culture, fractionate the cell compartments, and detect the amounts of CEL-3R in the samples using Western blotting. We would use a CEL-3R-USA variant as a model for CEL-3R. CEL-3R-USA is an artificially designed variant in which the VNTR repeats match with the first three repeats of CEL-WT (Figure 7.7). The CEL-3R-USA variant would be compared against other variants with previously confirmed properties: CEL-WT, CEL-HYB and CEL-TRUNC.

7.2.1 Sequencing

In order to study the functional properties of the CEL-3R variant, we first desired to confirm that the *CEL-3R-USA* variant was an exact nucleotide match with the *CEL-3R* gene seen in the Danish family. Thus, we needed to sequence some of the Danish family members. We Sanger sequenced the DNA samples from two members who were confirmed to be heterozygous for *CEL-3R*.

The sequencing of the *CEL* VNTRs in the two Danish individuals yielded two identical *CEL-3R* VNTR sequences. However, the Danish *CEL-3R* VNTR sequences were not matching with the artificial *CEL-3R-USA* variant (Figure 7.6A). The nucleotide sequence of the *CEL-3R* variant found in the Danish family will hereby be denoted *CEL-3R-DAN*. The mismatches all appeared in the second and third repeats. Further, two of these mismatches would lead to the encoding of different amino acids (Figure 7.6B). In contrast to their VNTRs, the globular domain of both genes were found to be exact matches with the exons of the *CEL-WT* reference sequence ENSG00000170835 (Solrun J. Steine, personal correspondence).

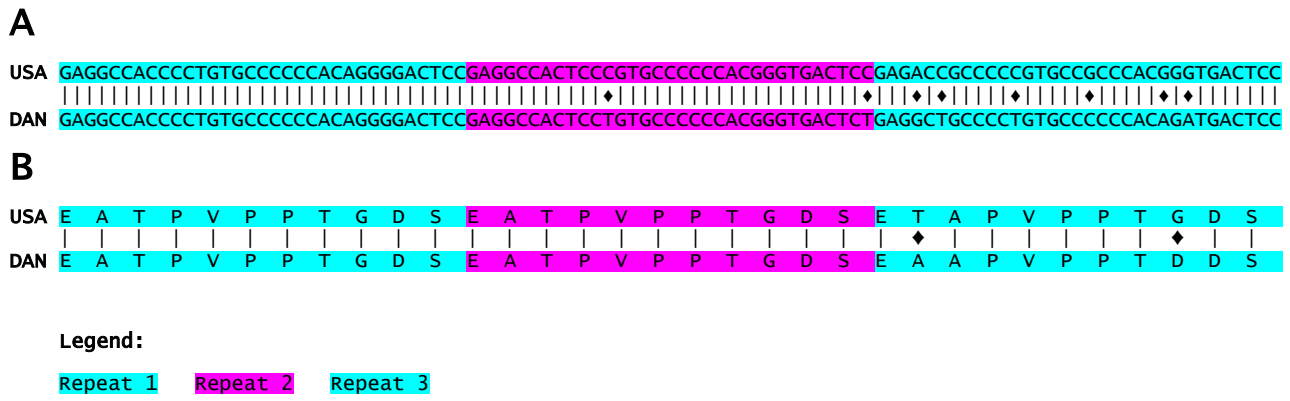


Figure 7.6: **Alignment of *CEL-3R* VNTR sequences.** DNA sequences (**A**) and corresponding amino acids (**B**) of the VNTR regions of the *CEL-3R-USA* and *CEL-3R-DAN* plasmids. Pairwise alignment was performed manually. Mismatches are marked with ♦.

Compared to the reference sequence of *CEL-WT*, the first and last repeats of *CEL-3R-DAN* were exact matches (Figure 7.7). In contrast, the second repeat of *CEL-3R-DAN* had no exact match with any of the reference repeats. However, the early part of the second repeat matched with the second repeat of *CEL-WT*, while the latter part matched with the 15th repeat of *CEL-WT*. In short, the VNTR of *CEL-3R-DAN* was not contiguously matching with *CEL-WT*.

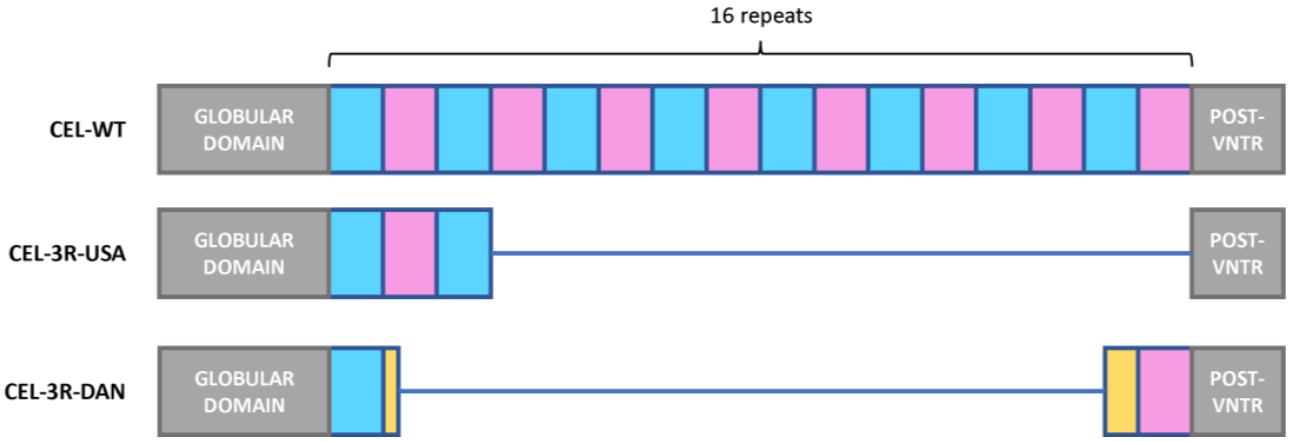


Figure 7.7: **Alignments of *CEL-3R-USA* and *CEL-3R-DAN* against reference *CEL-WT*.** Illustration of matching segments of CEL-WT, *CEL-3R-USA* and *CEL-3R-DAN*. Coloured boxes represent individual repeats. Repeats of *CEL-3R-USA* and *CEL-3R-DAN* are aligned with matching repeats in *CEL-WT*. The globular domain and post-VNTR are not drawn to scale.

Plasmids inserted with all the aforementioned variants were acquired: CEL-WT, CEL-3R-USA, CEL-3R-DAN, CEL-HYB and CEL-TRUNC. An empty vector (EV) plasmid was also included as a negative control. The sequenced plasmids (Figures 11.1-11.6) were exact matches with their respective VNTR reference sequences. Further, the CEL insertions all had matching non-VNTR sequences, except for CEL-TRUNC which contained the silent mutation A1497T. Despite not being a confirmed human variant, we chose to include the CEL-3R-USA variant in the cellular experiments to compare against CEL-3R-DAN. The differences in properties between the two variants could tell about the function of their mismatching amino acids.

7.2.2 Transformation and purification of plasmids

In order to increase the quantity of the plasmids, they were transformed into competent bacteria which were cultured on agar plates with ampicillin. Then, the ampicillin-resistant colonies were selected for further culturing in liquid agar with ampicillin. The bacteria cultures were harvested, centrifuged and isolated as pellets. DNA was then isolated from the bacterial pellets using the QIAfilter Midi DNA isolation kit. The resulting DNA samples were measured for optical density (OD) to determine concentration and purity.

Overall, the plasmid samples displayed decent concentrations and purity (Table 7.1). The measured concentrations were between 700 and 1000 ng/µL. All plasmids displayed similar ratios of OD. The A260/A280 ratios were all at either 1.88 or 1.89—slightly higher than the optimal value at 1.80. Meanwhile, some of the A260/A230 ratios were slightly higher than the optimal value at about 2.0–2.2. The high A260/A280 and A260/A230 ratios indicated that the plasmid samples were low in contaminants such as proteins, phenol and salts. However, the high A260/A280 ratios could also indicate that there was RNA contamination in the plasmid samples.

Table 7.1: **Concentration and absorbance ratios in the isolated plasmid samples.** TE buffer was used as both the plasmid solvent and a blank for the measurements.

Plasmids	Conc. (ng/µL)	A260/A280	A260/A230
CEL-WT	981.9	1.88	2.31
CEL-3R-USA	965.9	1.88	2.30
CEL-3R-DAN	913.4	1.89	2.35
CEL-HYB	760.4	1.88	2.35
CEL-TRUNC	777.9	1.89	2.18
EV	793.6	1.88	2.36

As a final plasmid quality check, an agarose gel electrophoresis was performed. Both native plasmids and plasmids linearised using BamHI restriction digestion were loaded on the gels. The native plasmids were expected to display multiple migration bands due to different uncoiled, coiled and supercoiled configurations. In contrast, the linearised plasmids were expected to only display one band.

As expected the gel image displayed multiple bands for the native plasmids, and only one band for the linearised plasmids (Figure 7.8). The native, uncoiled bands were positioned entirely above the 10 kb mark, while the native, supercoiled bands were positioned either somewhat higher or lower than the expected size (see Table 5.2). Most plasmids mainly displayed the supercoiled configuration, except for CEL-3R-DAN and CEL-TRUNC, whose main configurations were less

coiled. In contrast, the linearised plasmid bands were positioned roughly at expected sizes (see Table 5.2). Based on the OD analysis and the agarose gel electrophoresis, all plasmids were assumed to be of good quality.

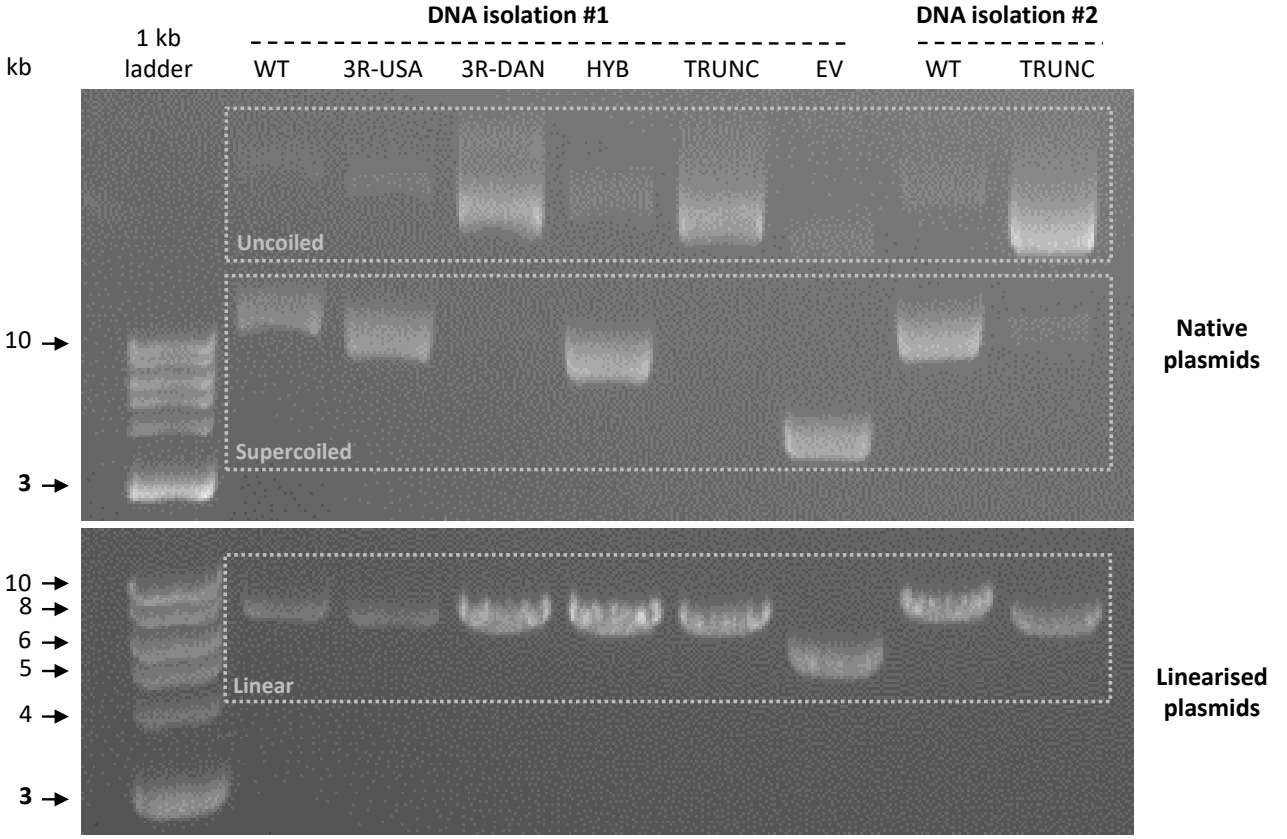


Figure 7.8: **Agarose gel electrophoresis of native and linearised CEL plasmids.** The native plasmid bands are annotated for uncoiled (nicked) bands and supercoiled bands. A 1% (w/v) agarose gel was prepared and the wells were loaded with 200 ng of the plasmids CEL-WT, CEL-3R-USA, CEL-3R-DAN, CEL-HYB, CEL-TRUNC and EV. Plasmids were linearised using BamHI restriction digestion at 37°C for 15 min. Gels were imaged in a Syngene GeneFlash.

7.2.3 Transfection optimisation

To ensure a high transient transfection efficiency for HEK293 cells, a plasmid expressing Yellow Fluorescent Protein (YFP) was used for transfection optimisation. The cells were seeded, then 24 h later transfected using various mixtures of the YFP plasmid prepared in Lipofectamine2000 or Lipofectamine3000 reagents. Cells were observed 24 h and 48 h after transfection using a UV microscope (Figure 7.9).

A

				24 h post transfection		48 h post transfection	
Well	Lipofectamine3000 (µL)	DNA (µg)	P3000 (µL)	Confluency	Fluorescence	Confluency	Fluorescence
1	5	2.5	5	60	+	80	++++
2	7	2.5	5	50	++	70	++++
3	5	3.5	7	50	+++	70	+++++
4	7	3.5	7	50	++	60	+++
5	10 (Lipofectamine2000)	4	-	40-50	++	70	++++
6	-	-	-	90-100	0	100	0

B

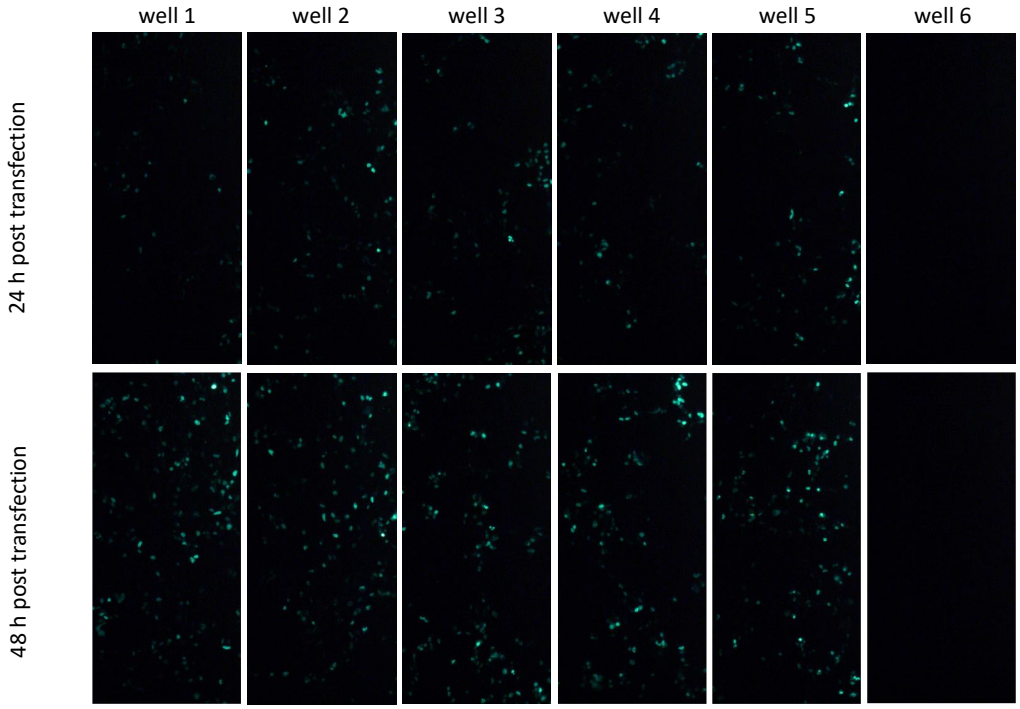


Figure 7.9: **Optimisation of transfecting HEK293 cells with YFP plasmid. A.** Six wells were seeded with $5{\cdot}10^5$ cells and incubated for 24 h. The wells were transfected with YFP plasmids using Lipofectamine3000 with varying volumes of reagents (wells 1–4) or Lipofectamine2000 (well 5). Well 6 was left untransfected. Medium was changed 6 h post transfection. The displayed scores for Confluency and Fluorescence were based on observations in the microscope. **B.** Images from the wells at 24 h and 48 h post transfection. Displayed in 10X zoom.

Under the microscope, the transfected cells were labelled green, and the number of green cells markedly increased from 24 h to 48 h in all wells (Figure 7.9 A-B). Some differences in transfected cells were observed between the wells at 24 h, but less so after 48 h. The highest transfection efficiency was observed in well 4 (not obvious in Figure 7.9B). From these results, the Lipofectamine3000 and Lipofectamine2000 setups from the respective wells 4 and 5 were applied in later transfections.

7.2.4 Cell fractionation and Western blotting – optimisation

We wanted to perform Western blotting on the CEL-3R variants to investigate their cellular localisation. The CEL-3Rs could be compared against the already characterised normal CEL-WT and the pathogenic, ER-aggregating CEL-HYB (*Cassidy et al.*, 2020; *Tjora et al.*, 2020; *Fjeld et al.*, 2022).

As previously mentioned, CEL is translated into the ER and undergoes modifications as it moves to the Golgi, then to the zymogen granules, before it is secreted out of the cell. These cellular compartments in the secretion pathway of CEL can be tracked by fractioning the cell samples. The dense microsomes (remnants of the ER) are a predominant constituent of the insoluble pellet fraction after centrifugation of lysed cells. Meanwhile, the soluble lysate fraction is mainly associated with the less dense Golgi compartments and other associated membranes (*Torsvik et al.*, 2014). Finally, the medium fraction is associated with the fully secreted protein. Thus, we could approximately track the localisation of the CEL variants as they underwent the secretion pathway (Figure 7.10).

Furthermore, the CEL proteins with different PTMs would be distinguishable in SDS-PAGE due to mass differences (*Gravdal et al.*, 2021). Thus, we expected the fully *O*-glycosylated CEL variant proteins to be associated with high kDa bands which should be prominent in the medium fractions. Meanwhile, less mature CEL proteins were expected to be associated with lower kDa bands, which should be prominent in the pellet fractions.

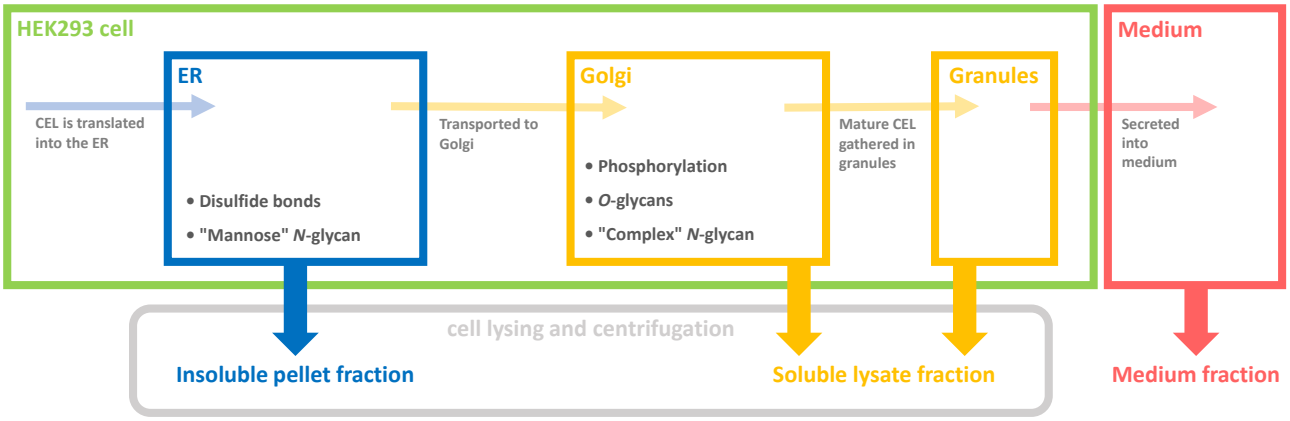


Figure 7.10: **Model of CEL secretion pathway and ideal cell fractioning.** The illustration is a simplified model for the secretion pathway of CEL in HEK293 cells. It also indicated the ”ideal” separation of cellular compartments into fractions: insoluble pellet, soluble lysate and medium fraction. ER and Golgi are annotated with the modifications that are performed on the CEL protein within their compartments. (Own illustration)

We wanted to investigate the functional properties of the CEL-3R proteins in a cellular system. To do so, plasmids expressing for CEL-WT, CEL-3R-USA, CEL-3R-DAN, CEL-HYB, CEL-TRUNC or EV were transiently transfected into HEK293 cells. Cells were harvested 48 h post-transfection and separated into 3 fractions; (1) the insoluble pellet, (2) the soluble lysate and (3) the medium. Protein samples from the three fractions were separated using SDS-PAGE, transferred to a 0.20 µm PVDF membrane, blocked in BSA and targeted using a primary anti-CEL antibody. The constitutively expressed β-Actin proteins in the lysate samples were targeted using a primary anti-β-Actin antibody. Protein bands were imaged using secondary HRP-antibodies and chemiluminescence.

For the first Western blot experiments, the results were as shown in Figure 7.11. The bands generally appeared straight and clearly defined. However, compared to previous work with Western blots on CEL-WT and CEL-TRUNC, the CEL bands were weak (see *Gravdal et al.*, 2021). Indeed, the CEL-TRUNC bands appeared completely invisible.

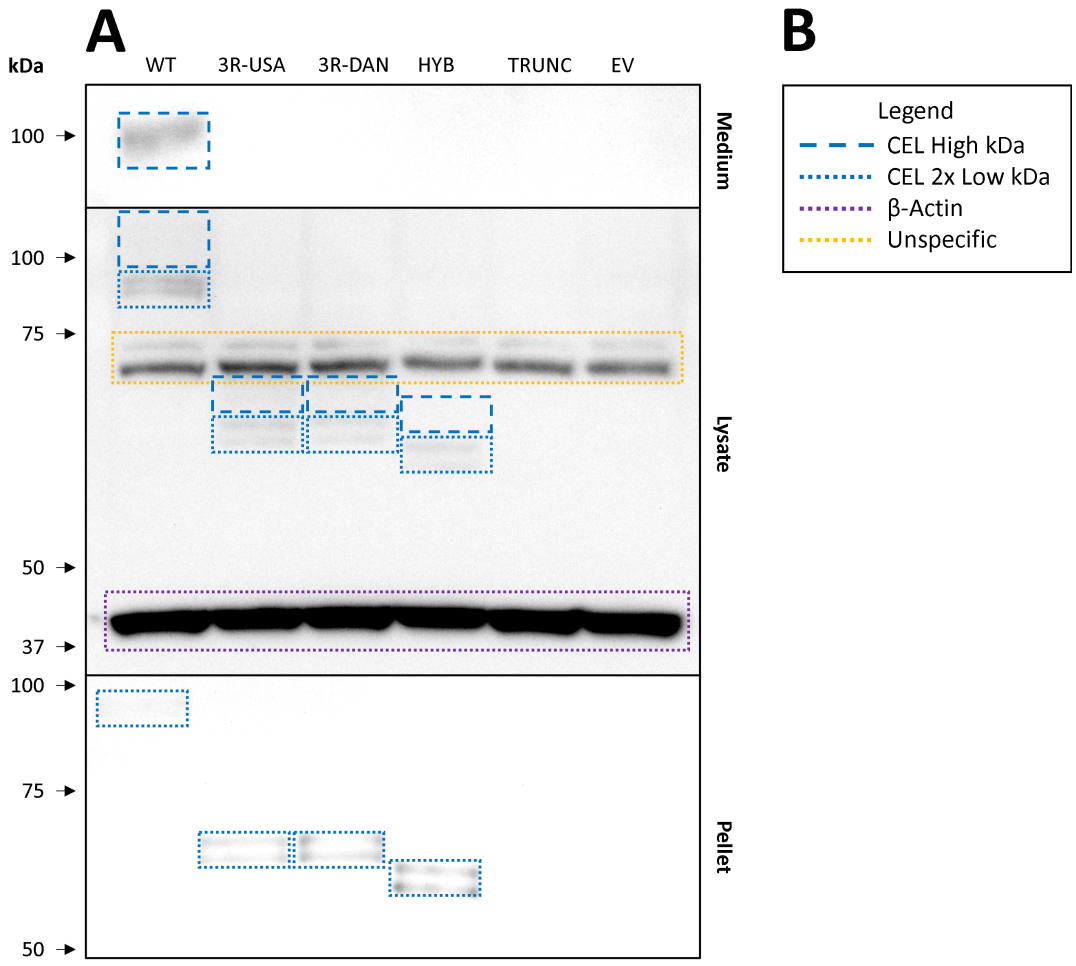


Figure 7.11: **Western blot of fractionated CEL variants.** Cell fractions were loaded in exclusive gels. SDS-PAGE was performed using 1.0 mm 4-12% Bis-Tris gels. Transfer was performed using the Trans-Blot Turbo system from Bio-Rad. Membranes were incubated in ECL Prime and imaged using a G:BOX iChemi XR5. **A.** The imaged membranes with annotations for observed bands. **B.** Legend describing the annotations.

To improve the strength of the CEL signals in comparison to the unspecific bands, we tested a different ECL kit and milk as a blocking buffer instead of BSA. A new round of transfection, SDS-PAGE and Western blotting was performed. To see if the amount of loaded protein had an effect, two gels were loaded with 4 µg and 15 µg protein, respectively. The ECL Plus kit was used for imaging the CEL membranes (Figure 7.12). For the membrane loaded with 4 µg protein, the result was weak signals and a comparatively high background noise. For the 15 µg membrane, both the CEL signals and the unspecific bands were weaker than seen previously. Therefore, neither low amount of protein nor the new ECL kit were used further in Western blots. In contrast, the use of milk as a blocking buffer was continued.

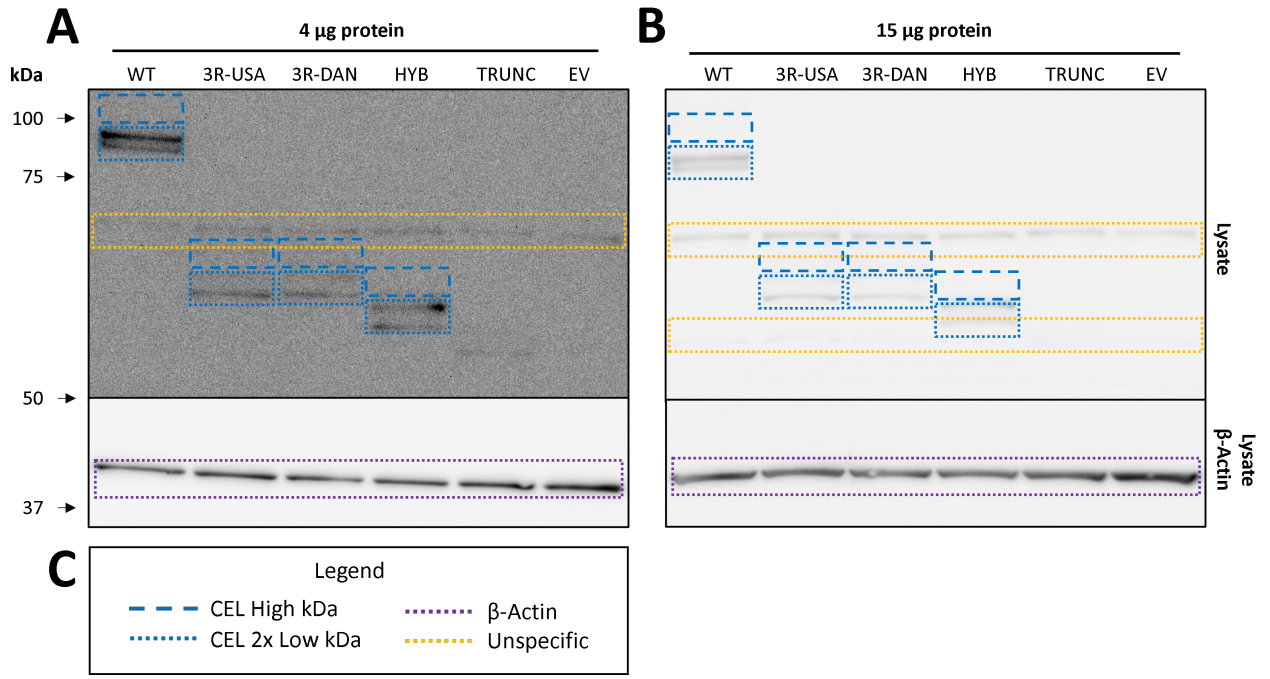


Figure 7.12: **Western blot of fractionated CEL variants loaded as 4 µg and 15 µg.** SDS-PAGE was performed using 1.0 mm 4-12% Bis-Tris gels. Transfer was performed using the Trans-Blot Turbo system from Bio-Rad. Lysate CEL-specific membranes were incubated in ECL Plus. Lysate β-Actin membranes were incubated in ECL Prime. The membranes were imaged using a G:BOX iChemi XR5. Observed bands are annotated. **A.** The membranes prepared with 4 µg protein. **B.** The membranes prepared with 15 µg protein. **C.** Legend describing the band annotations.

Up until now, the Western blot transfer was performed using the Trans-Blot Turbo system. In contrast, the lab group had previously applied the traditional wet transfer method, resulting in visible CEL-TRUNC bands (see *Gravdal et al.*, 2021). We believed that the shorter and more intense transfer used in the Trans-Blot Turbo system was not properly transferring the CEL proteins to the membrane. To test this, a new transfer using the wet method was performed. Independently prepared plasmids and lysates were applied as positive controls.

The wet transfer results showed that there was something about the Trans-Blot Turbo method that was at fault for the weak CEL-TRUNC bands seen in previous attempts. The imaging of wet transfer membranes finally displayed visible CEL-TRUNC bands (Figure 7.13). There was also satisfactory little background noise.

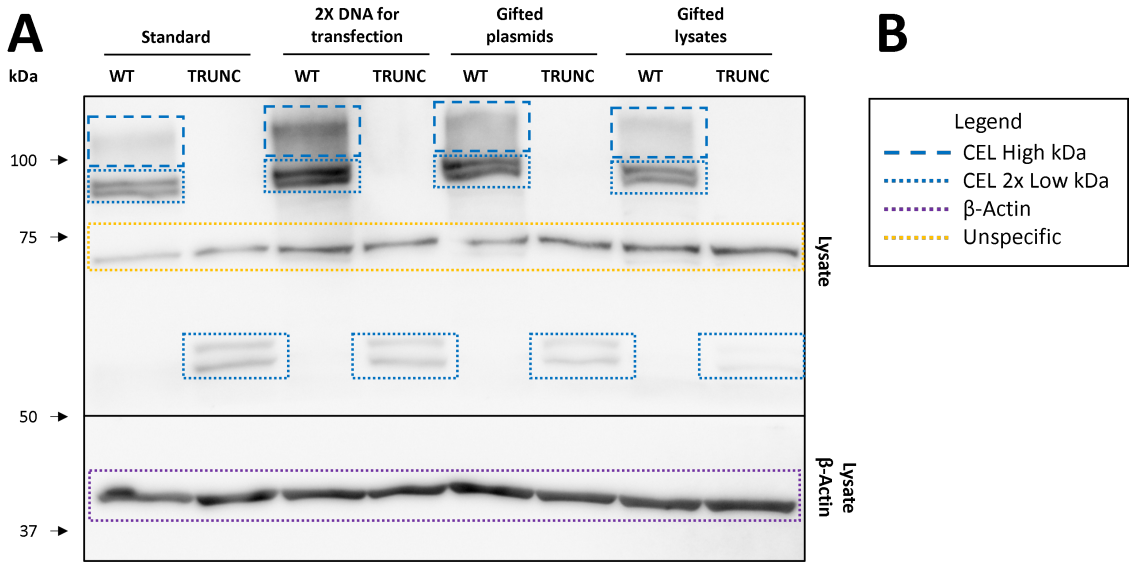


Figure 7.13: **Western blot of fractionated CEL variants by wet transfer.** SDS-PAGE was performed using a 1.0 mm 4-12% Bis-Tris gel. Transfer was performed using the wet transfer system. Membranes were incubated in ECL Prime and imaged using a G:BOX iChemi XR5. **A.** The imaged membranes with annotations for observed bands. **B.** Legend describing the band annotations.

In order to reveal why the CEL-TRUNC bands only were observed after attempting the wet transfer method, the Trans-Blot Turbo transfer was attempted again with modified setups to imitate the wet transfer. The previously used samples were run once again in SDS-PAGE. Two parallel transfers were performed using the Trans-Blot Turbo system: (1) one where the built-in 0.2 µm PVDF membrane was replaced with a 0.45 µm PVDF membrane, and (2) the other where a less intense, 30 min transfer time was applied. The result of the modified Trans-Blot Turbo transfers finally showed that a 0.45 µm PVDF membrane was necessary to get visible CEL-TRUNC bands (Figure 7.14).

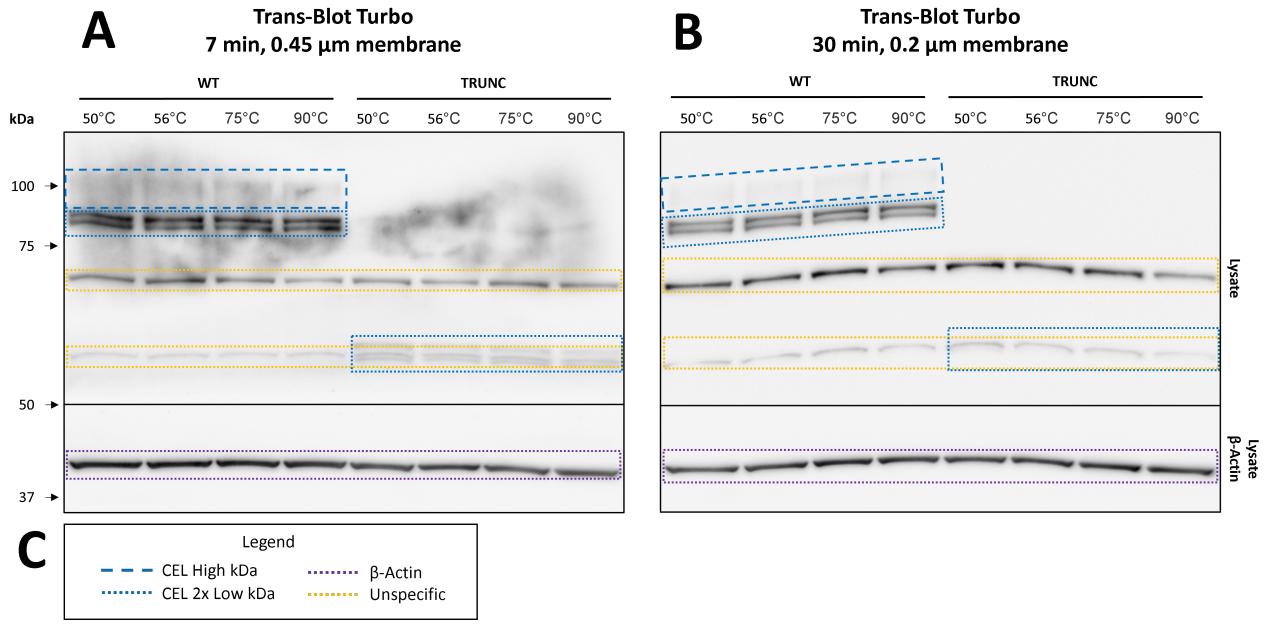


Figure 7.14: **Western blot of fractionated CEL variants.** The SDS-PAGE was performed using 1.0 mm 4-12% Bis-Tris gels. Transfer was performed using the Trans-Blot Turbo system from Bio-Rad. Membranes were incubated in ECL Prime and imaged using a G:BOX iChemi XR5. Observed bands are annotated. **A.** The 0.2 µm PVDF membrane in the Turbo-Blot sandwich was replaced with an activated 0.45 µm PVDF membrane. Transfer was run at setting 'Mixed MW'. **B.** Transfer was performed for 30 min using the 'Standard SD' setting. **C.** Legend describing the band annotations.

7.2.5 Cell fractionation and Western blotting – final results

As the method for protein transfer was finally resolved, a final round of transfection, SDS-PAGE and Western blotting was performed. The cells were transfected with 4 µg plasmid using Lipofectamine2000. Loading samples of 15 µg protein were separated by SDS-PAGE and then transferred to 0.45 µm PVDF membranes in a Trans-Blot Turbo for 25 V for 7 min. The lysate membrane was cut into a CEL-specific membrane and a β-Actin-specific membrane. All membranes were blocked using milk. Then, the membranes were incubated in antibodies to target CEL and β-Actin, respectively. Finally, the membranes were imaged using ECL Prime.

The final Western blot on fractions from transfected HEK293 cells resulted in all the CEL variants appearing in all the membranes (Figure 7.15). The variants all displayed three bands in total: two lower kDa bands in close proximity and one smear with higher kDa. The exception is CEL-TRUNC, which did not display the higher kDa smear band. These three CEL-specific bands displayed different strengths that varied between the pellet, lysate and medium fractions. The bands in the pellet fraction displayed only the two lower kDa CEL bands, while the medium fraction predominantly displayed the higher kDa smear bands. The lysate fraction displayed both the lower and higher kDa bands. The signal quantification analysis included all the CEL-specific bands that were visible in each membrane.

In addition to the CEL-specific bands, two unspecific bands appeared in the lysate fraction (Figure 7.15). The strongest unspecific band appeared at about 60-70 kDa and has been reported before (*Gravdal et al.*, 2021). The other, weaker unspecific band appeared somewhat higher than 50 kDa and overlapped with the CEL-TRUNC bands. These bands were assigned as unspecific due to appearing in the EV negative control.

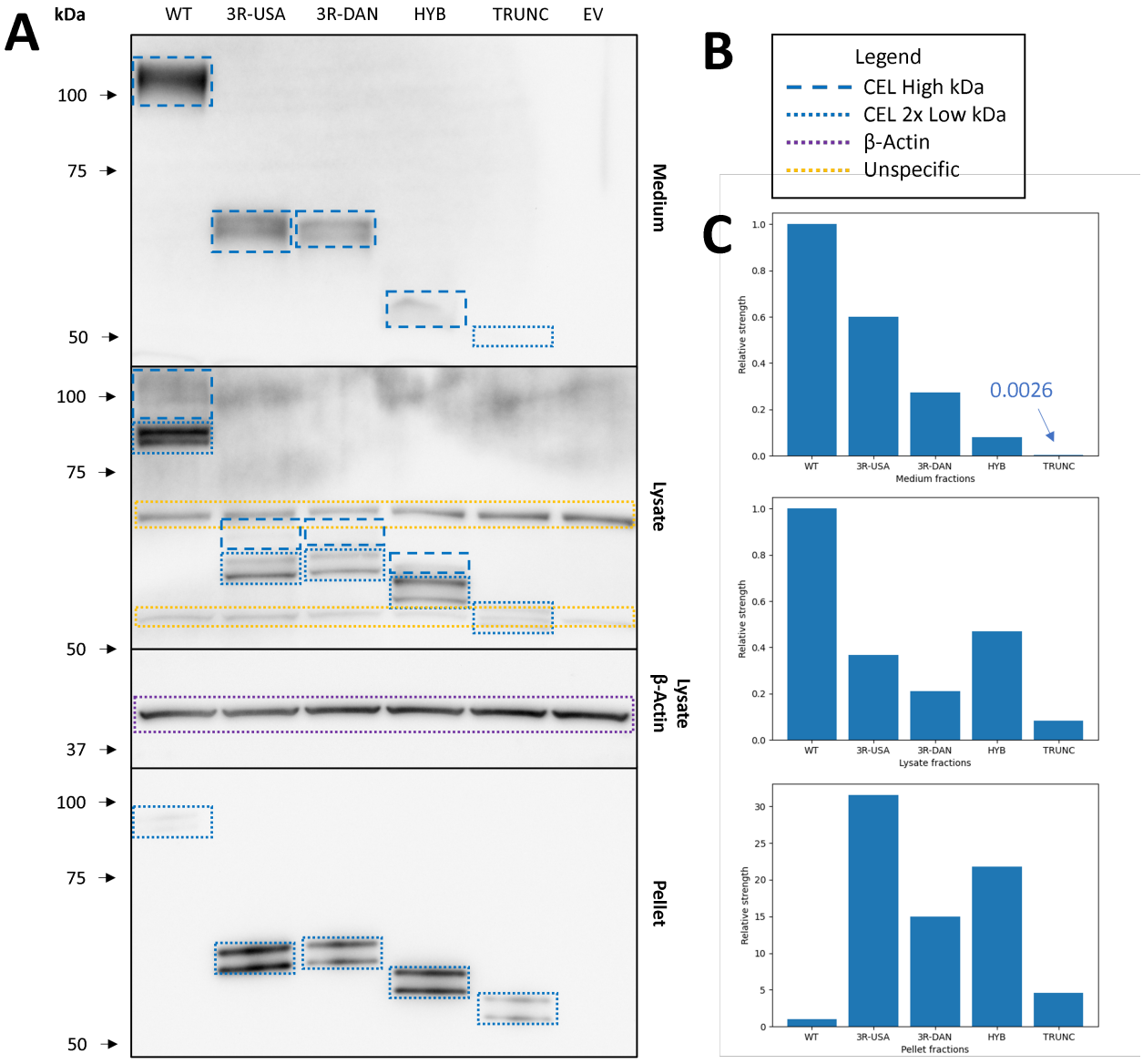


Figure 7.15: **Western blot of fractionated CEL variants.** SDS-PAGE was performed using a 1.0 mm 4-12% Bis-Tris gel. Transfer was performed using the Trans-Blot Turbo system from Bio-Rad. However, the 0.2 µm PVDF membranes were replaced with 0.45 µm PVDF membranes. Membranes were incubated in ECL Prime and imaged using a G:BOX iChemi XR5. **A.** The membranes of the medium, lysate and pellet fractions. Observed bands are annotated. **B.** Legend describing the annotations on the bands. **C.** Quantified signal strengths of CEL variants displayed relative to CEL-WT.

The CEL-3R-DAN bands displayed some patterns across the different fractions (Figure 7.15). First, CEL-3R-DAN always displayed bands that were about half as strong as the CEL-3R-USA bands. Secondly, the two CEL-3R variants displayed strengths comparable to CEL-HYB in the pellet and lysate fractions, but not in the medium where they were both much stronger. The CEL-3R variants displayed relatively strong bands in both the pellet and in the medium fractions, unlike CEL-WT, CEL-HYB and CEL-TRUNC.

7.3 Molecular dynamics

The CEL VNTR is predicted to be disordered by bioinformatical tools (Figure 7.16). Indeed, the disorder of the CEL VNTR has been mentioned in the literature since at least the 1990's (*Reue et al.*, 1991; *Chen et al.*, 1995). Despite this, the functional relationship between the disordered VNTR, the *O*-glycosylations and the rest of the CEL protein have never been well understood.

As previously explained, IDPs have historically been difficult to examine using traditional structural techniques. We theorised that the *in silico* technique MD simulation could aid in predicting the physical properties of the CEL VNTR. MD simulations produce trajectories which display the movements of atoms and molecules over time. These trajectories can be analysed to investigate the conformations of the VNTR and the roles of various residues like Pro, Glu, Ser, Thr, Gly and Val. Further, the VNTR could be simulated with and without the *O*-glycosylations, which could reveal to us how this type of modification affects the physical properties of the protein.

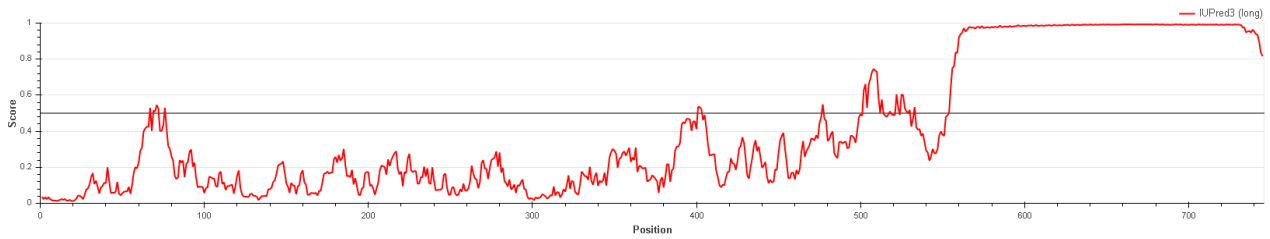


Figure 7.16: **Predicted disorder of the CEL-WT protein.** X-axis displays residue numbers. Y-axis displays the disorder score, where 0.5 is considered the threshold for disorder. Predicted by IUPred3.

7.3.1 Preparations for molecular dynamics

The short VNTR of CEL-3R-DAN was selected as the structure for the MD simulations due to it being the main topic of the thesis, as well as requiring less computational power to simulate than for example the larger CEL-WT. Further, the post-VNTR sequence 'KEAQMPAVIRF' was included. Thus, the total simulated protein region was a 44-residue sequence consisting of three VNTR repeats with 11 residues each, followed by the 11-residue post-VNTR sequence. See Figure 7.17 for the complete simulated sequence.

Figure 7.17: **The simulated CEL-3R-DAN VNTR sequence.**

The simulation system was prepared twice, once with and once without the addition of *O*-glycosylations. The simulated structure without modification will hereby be denoted as h3R-DAN-U (human CEL-3R-DAN Unmodified), while the *O*-glycosylated structure will be denoted as h3R-DAN-G (human CEL-3R-DAN *O*-glycosylated).

As for how the *O*-glycosylations of the h3R-DAN-G structure were designed, there was little pre-existing knowledge to take from. In a study by *Jellas et al.* (2018), there was found two large glycan peaks at 708 and 953 m/z in the MS of CEL from a patient with blood type A. These glycans were somewhat arbitrarily selected as the glycans for the h3R-DAN-G structure and denoted glycan A and glycan B, respectively (Figure 7.18A). The α/β conformations of the glycans were assumed from common depictions on the web. Using the band sizes we observed in the Western blot of lysates from HEK293 cells, the total mass of the CEL-3R-DAN *O*-glycosylations were estimated to be around 3-4 kDa (Figure 7.18B). Thus, we added a total glycan mass of 3.2 kDa to the h3R-DAN-G structure by dispersing 2x glycan A and 3x glycan B on various Thr and Ser residues (Figure 7.18C).

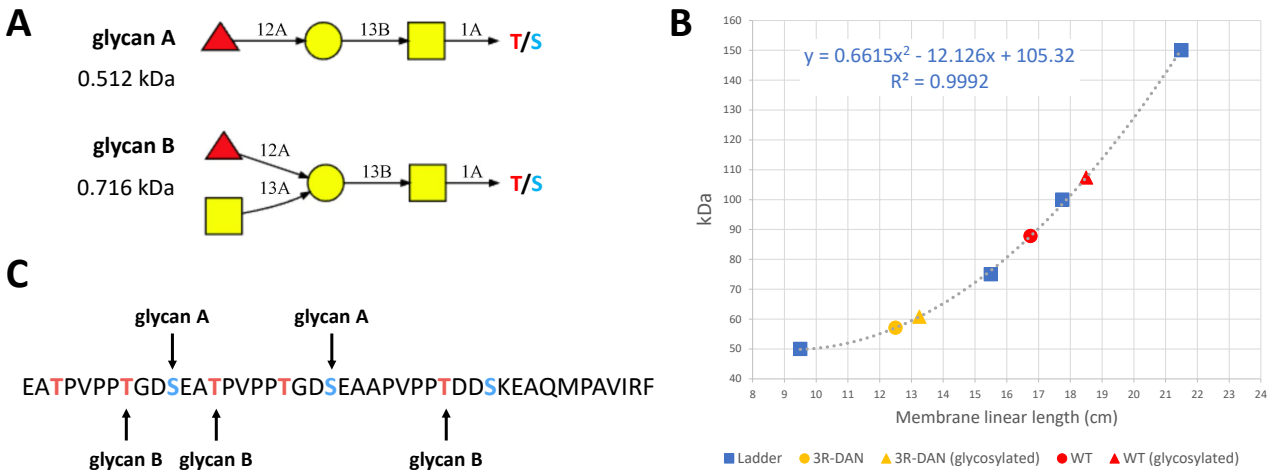
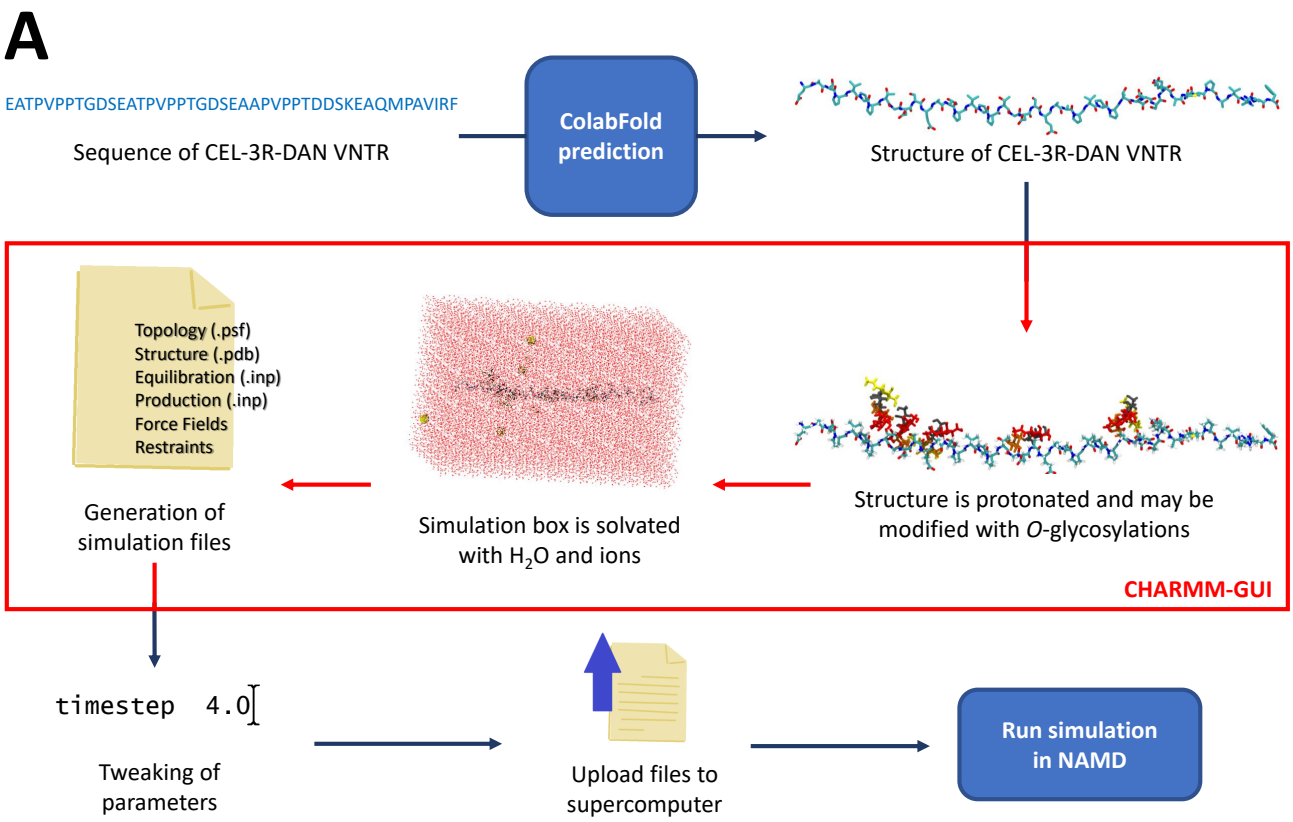


Figure 7.18: **Determining the *O*-glycosylations for the h3R-DAN-G structure. A.** The two glycans prepared in CHARMM-GUI Glycan reader and modeler, displayed in SNFG format (*Varki et al.*, 2015). Full reference at the end of the thesis. **B.** The predicted mass increase of *O*-glycosylated CEL-3R-DAN in the Western blot. The 2nd degree polynomial regression curve was calculated from the ladder (blue squares). The yellow and red indicators show the observed CEL-3R-DAN and CEL-WT band positions fitted to the curve. **C.** The sequence of the simulated structure. Thr and Ser are displayed in red and blue, respectively. The applied *O*-glycosylations are annotated at their respective residues.



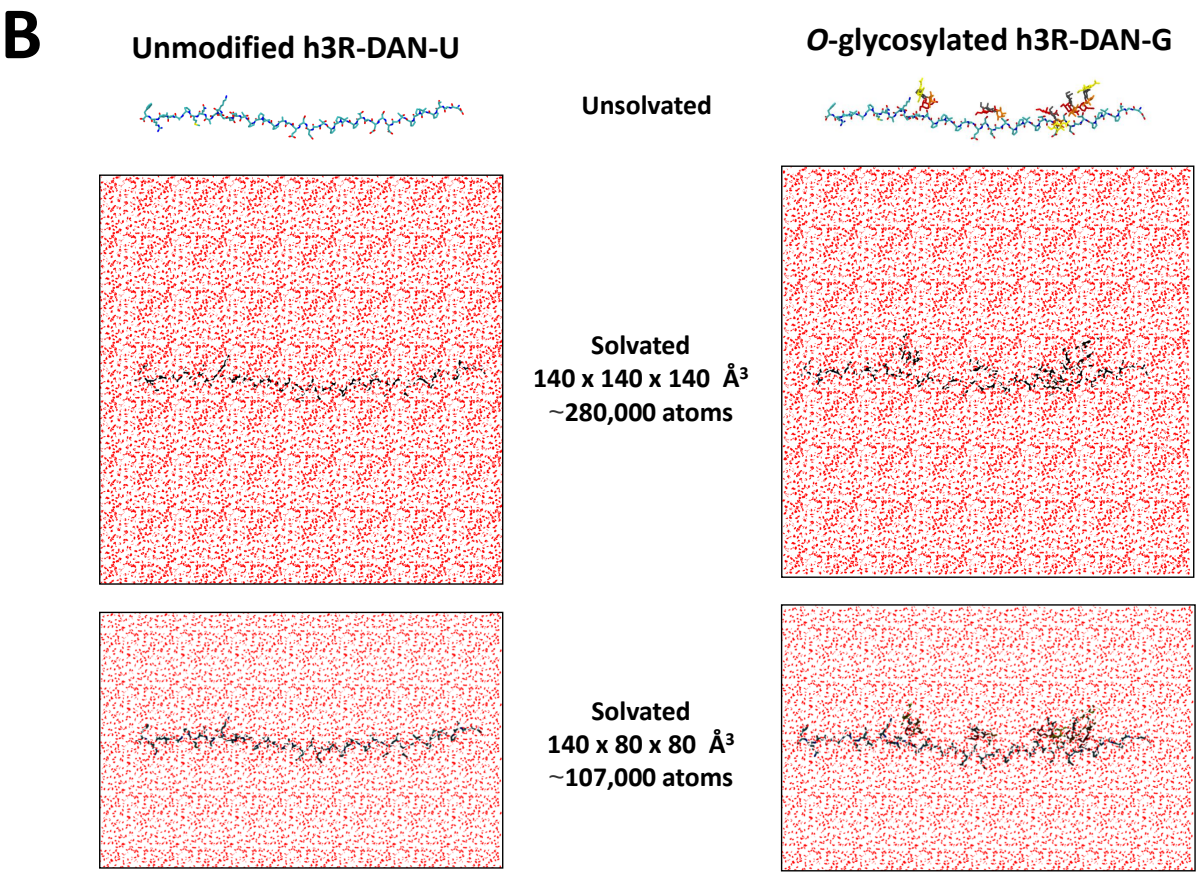


Figure 7.19: **Pipeline and setup for preparing for MD simulation. A.** The pipeline from the protein sequence to running of the simulation. **B.** The unmodified h3R-DAN-U and *O*-glycosylated h3R-DAN-G structures were both solvated in two differently sized simulation boxes.

To run the MD simulations, a series of preparatory steps was performed (Figure 7.19A). A linear, unprotonated protein structure was generated from the amino acid sequence using ColabFold. Then, the CHARMM-GUI Solution builder was applied to generate the simulation files. The structure was protonated assuming pH 7. Then, one version of the structure was *O*-glycosylated, while the other was left unmodified. The structure was solvated inside the simulation box. Na^+ ions were distributed in the solvent to neutralize the net charge of the system. CHARMM36m was selected as the force field due to its reported ability to predict IDPs (*Huang et al.*, 2016; *Wang*, 2021).

7.3.2 Simulations in large water boxes

As a starting point for the MD simulations, the structures h3R-DAN-U and h3R-DAN-G were solvated. The length of the linearised structures were about 120 Å. To prevent the protein from self-interacting across the simulation box, the box was defined with approximate dimensions of 140 x 140 x 140 $Å^3$. The added water atoms constituted more than 99% of the approximate 280,000 atoms in the systems. The h3R-DAN-U and h3R-DAN-G systems were each simulated until they had formed stable secondary structures. The runtimes for all the simulations can be viewed in Table 11.2.

The conformation and compactness of the simulated trajectories were analysed by plotting the RMSD and RoG values. The RMSD value indicates the distance between the atoms' starting positions and every subsequent frame. In other words, the lower the RMSD, the closer the structure is to its starting conformation. The RoG value indicates how compact the structure is. The lower the RoG, the closer the atoms are to the proteins' center of mass.

Figure 7.20 displays the calculated RMSD and RoG of the large water box simulations of h3R-DAN-U and h3R-DAN-G. In both trajectories, the RMSD and RoG values were highly negatively correlated. After the first 100 ns, both trajectories generally displayed one of two RoG patterns: (1) volatile variation within the approximate range 14-22 Å, and (2) a less volatile plateau at approximately 11 Å. The systems both entered the second pattern after about 200 ns. However, the h3R-DAN-G system displayed a slightly longer and more straight plateau. The h3R-DAN-U system eventually reached a somewhat stable conformation at about 800 ns.

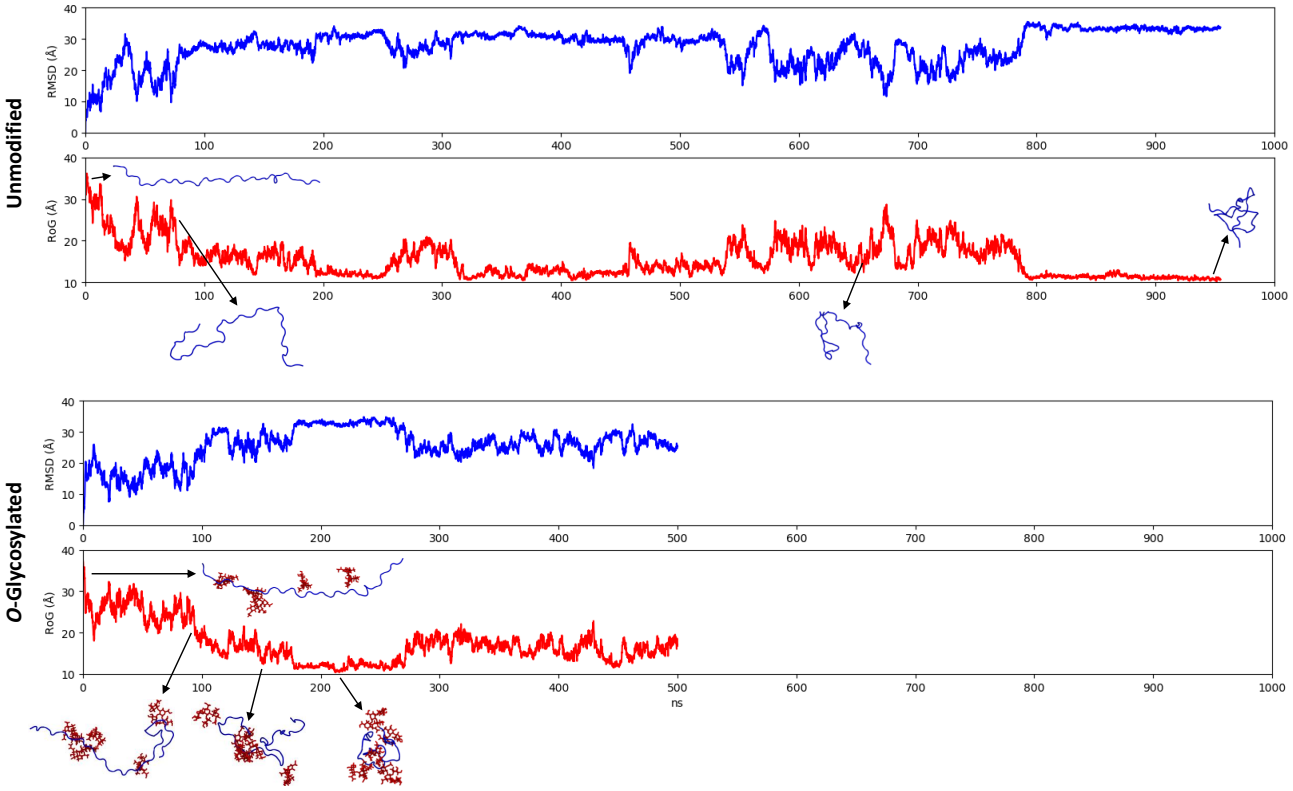
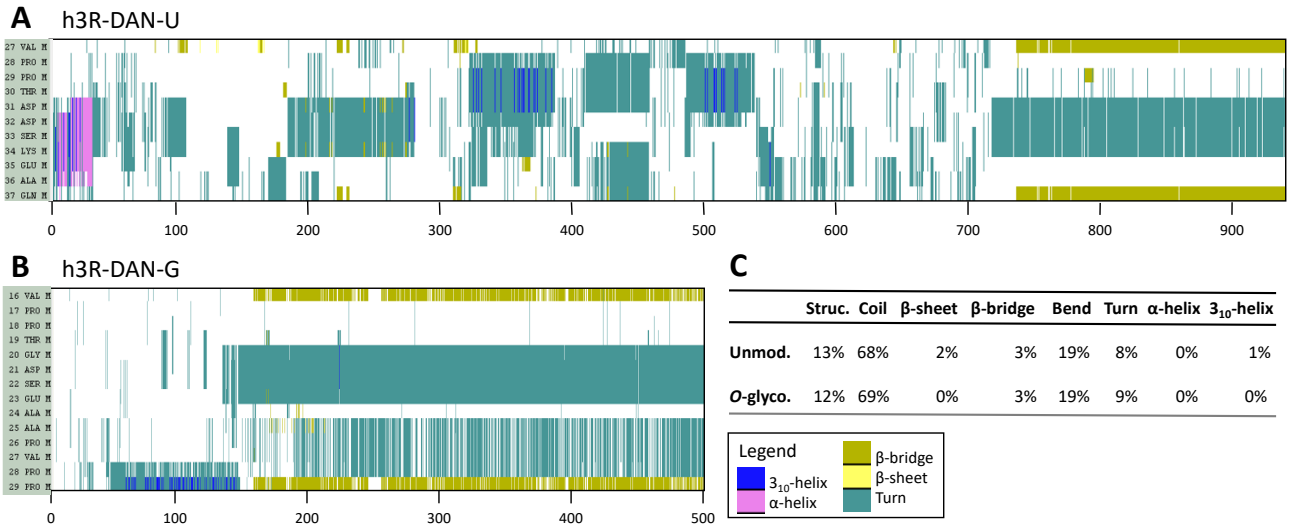


Figure 7.20: **RMSD and RoG of trajectories from h3R-DAN-U and h-3R-DAN-G in large water boxes.** RMSD calculation was performed on the 'C CAN' backbone of the trajectory aligned to frame 0. The RoG was calculated only for the protein mass and excluded the glycan mass. Plotted using Matplotlib. The RoG plots are annotated with simulation snapshots to visualise how RoG values relate to structure compactness. Proteins were visualised in VMD.

The h3R-DAN-U and h3R-DAN-G systems displayed different patterns of secondary structures (Figure 7.21). The h3R-DAN-U structure displayed mostly unstable turns, β-bridges and 3_{10}-helices in a wide range of its residues. A seemingly stable turn at Asp32-Glu35 and a β-bridge between Val27 and Glu37 formed after about 750 ns. In contrast, the h3R-DAN-G system formed stable secondary structures after only 140 ns. The modified system formed a turn at Gly20-Glu23 and a β-bridge between Val16 and Pro29. Residues Ala25-Pro23 were also alternatingly assigned as a turn. Despite the more stable secondary structures of h3R-DAN-G, both trajectories displayed about the same total prevalence of secondary structures at about 12-13%.



	Struc.	Coil	β-sheet	β-bridge	Bend	Turn	α-helix	3_{10}-helix
Unmod.	13%	68%	2%	3%	19%	8%	0%	1%
***O*-glyco.**	12%	69%	0%	3%	19%	9%	0%	0%

Figure 7.21: **Secondary structures of h3R-DAN-U and h3R-DAN-G in large water boxes.** X-axes display the full simulated time (ns) in each trajectory. Y-axes display only the residues with the most stable secondary structures within each trajectory. Visualised in VMD. **A.** h3R-DAN-U trajectory. **B.** h3R-DAN-G trajectory. **C.** The assigned % of secondary structures for all residues in each trajectory. The first 100 ns of each trajectory were excluded from the calculations. Secondary structure calculation was performed by DSSP.

7.3.3 Replica simulations in smaller water boxes

In order to validate the results of the simulations in large water boxes, the h3R-DAN-U and h3R-DAN-G systems were each simulated in three additional replicas. However, the dimensions of the simulation boxes were decreased to about 140 x 80 x 80 $Å^3$ in order to increase the run speed. This resulted in about 107,000 atoms, which was 60% less compared to the previous systems (see Figure 7.19B). The structures had quickly become compacted in the previous simulations, so protein self-interaction across the smaller simulation box was not a worry. Each replica simulation was run for 500 ns. The replicas were assigned randomly generated seeds in order to give each simulation a different velocity profile for the thermostat, which would result in unique trajectories (Figure 7.22).

Figure 7.22 shows the RMSD and RoG values of the h3R-DAN-U and h3R-DAN-G replica trajectories in the smaller water boxes. Again, the RMSD and RoG values displayed a strong, negative correlation. There was no obvious difference between the RoG pattern of the two systems. The h3R-DAN-U replicas 1 and 3 generally showed trajectories with higher and more volatile RoG values than the replicas of the h3R-DAN-G system. In contrast, replica 2 of the h3R-DAN-U displayed the least volatile and lowest RoG values of all the trajectories.

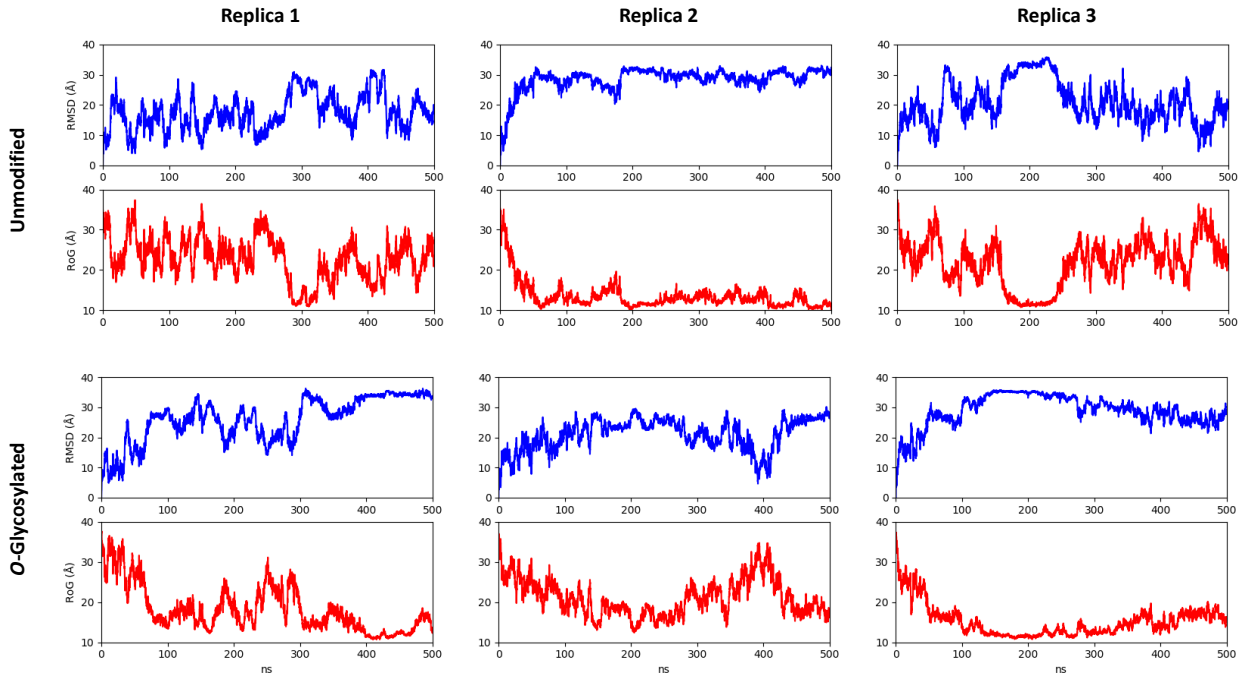


Figure 7.22: **RMSD and RoG of replica trajectories of h3R-DAN-U and h3R-DAN-G.** The RMSD calculation was performed on 'C CAN' backbone of the trajectory aligned to frame 0. The RoG was calculated only for protein mass and excluded the glycan mass.

A histogram plot was constructed in order to more easily discern the differences in RoG values between the h3R-DAN-U and h3R-DAN-G trajectories (Figure 7.23). The trajectories both displayed similar prevalence peaks at very low RoG values around 10-15 Å, although the h3R-DAN-U system was slightly more prevalent in the low range at 10-12 Å. Both systems displayed one more broad peak each, but in different ranges. The h3R-DAN-U system had a very broad peak in between about 20-27 Å. The h3R-DAN-G system had a less broad peak at about 16-18 Å. Overall, the h3R-DAN-U system displayed a broader prevalence of RoG values, while the h3R-DAN-G system displayed a more narrow prevalence around 12-18 Å.

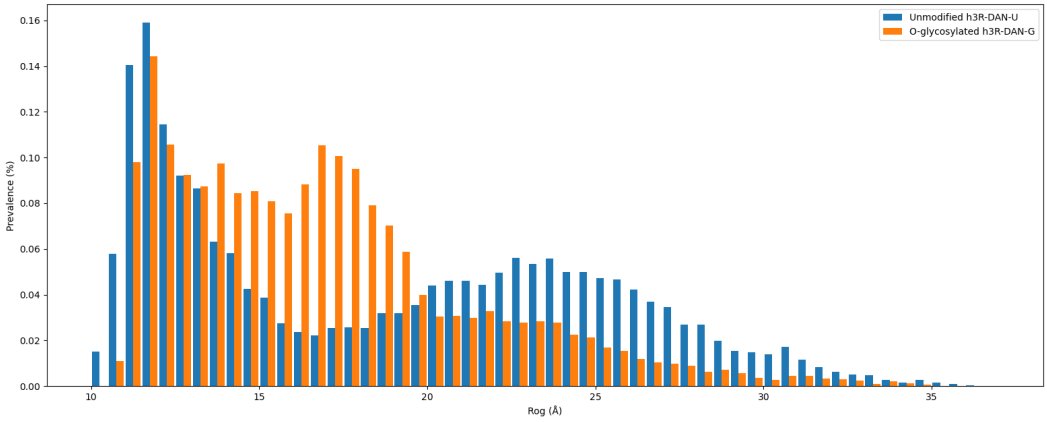


Figure 7.23: **Prevalence of RoG values in replica trajectories of h3R-DAN-U and h3R-DAN-G.** The last 400 ns of the replica trajectories were appended together. X-axis displays the binned RoG values. Y-axis displays the prevalence of the binned RoG values in the appended trajectories. The RoG values were calculated only for protein mass and excluded any glycan mass. Visualised in Matplotlib.

The secondary structures of the replica trajectories were analysed (Figure 7.24). The most common form of secondary structure in the trajectories was the turn. The turns usually consisted of 4 residues (also denoted as $i \rightarrow i \pm 3$). The second most common secondary structure was the β-bridge. A β-bridge is when a single H-bond connects the backbone of two protein strands. The most stable turns and β-bridges usually occurred together, with the β-bridges appearing about 3-4 residues away from the turn. In some cases, multiple H-bonds connected two strands, yielding a β-sheet. The least stable secondary structures in the simulations were the α-helices and 3_{10}-helices. These two secondary structures displayed short lifespans.

As with the RMSD and RoG results, the h3R-DAN-U and h3R-DAN-G trajectories displayed some mixed patterns of secondary structure (Figure 7.24). The h3R-DAN-U system exhibited mostly unstable and volatile secondary structures, with the exception of replica 2 which displayed the most stable secondary structure of all the simulations. The h3R-DAN-G systems generally displayed larger secondary structures, like β-sheets and some double turns with 8 residues. Still, both structures exhibited a similar total prevalence of secondary structures (Table 7.2). Although, h3R-DAN-G displayed β-sheets more often than h3R-DAN-U.

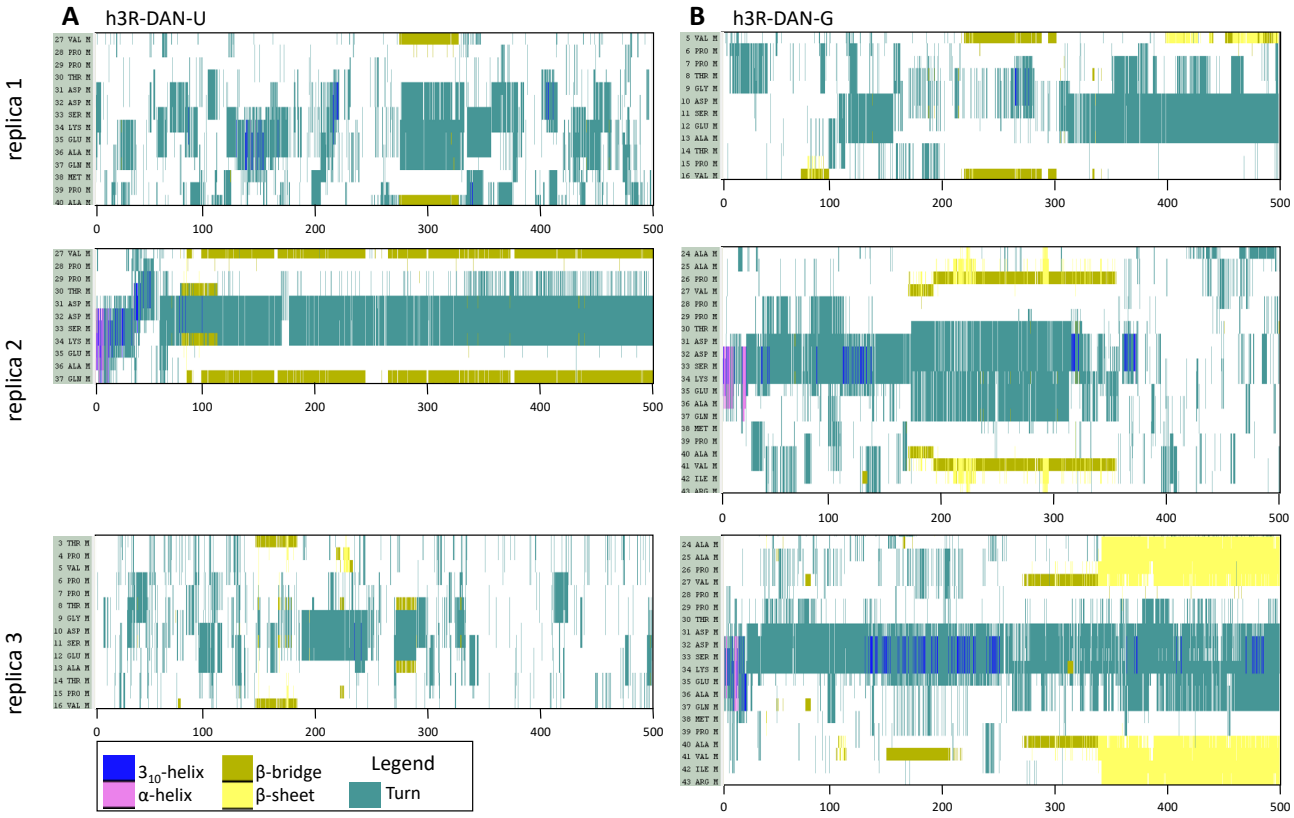


Figure 7.24: **Secondary structure in the replica simulations of h3R-DAN-U and h3R-DAN-G.** X-axes display the full 500 ns of simulated time in every trajectory. Y-axes display only the residues with the most stable secondary structures within each trajectory. Visualised in VMD. **A.** The three replica trajectories of the h3R-DAN-U structure. **B.** The three replica trajectories of the h3R-DAN-G structure.

Table 7.2: **Assigned secondary structures in replica trajectories of h3R-DAN-U and h3R-DAN-G.** Values show the prevalence of secondary structure in the last 400 ns of the trajectories. 'Structure' is the combination of 'α-Helix', 'β-Sheet', 'β-Bridge' and 'Turn'. Secondary structure assignments were calculated using DSSP.

	Structure	**Coil**	**β-sheet**	**β-bridge**	**Bend**	**Turn**	**α-Helix**	**3_{10}-Helix**
h3R-DAN-U								
Replicate 1	6%	75%	0%	1%	19%	4%	0%	0%
Replicate 2	17%	61%	2%	6%	23%	8%	0%	0%
Replicate 3	6%	77%	0%	1%	16%	5%	0%	1%
h3R-DAN-G								
Replicate 1	11%	72%	3%	1%	17%	6%	0%	1%
Replicate 2	9%	74%	1%	2%	17%	6%	0%	0%
Replicate 3	17%	63%	7%	2%	18%	9%	-	2%

Figure 7.25A displays some examples of observed secondary structures in the h3R-DAN-U trajectories. The h3R-DAN-U structure usually formed turns at the 'GDSE' sequences. The 'GDSE' sequence consists of the flexible Gly residue, as well as the polar or acidic residues Ser, Asp and Glu. The turns rarely formed backbone-backbone H-bonds. Instead, the turns sometimes displayed sidechain-backbone or sidechain-sidechain H-bonds. The 'TDDS' sequence sometimes formed two consecutive turns together with 'KEAQ'. Here, the Ser33 and Lys34 sidechains frequently formed H-bonds with the Asp31 and Asp32 residues, which might have helped stabilize the turns. Further, the Val27 and Ala40 residues on opposite strands formed a β-bridge. These two strands were aligned anti-parallel by a turn.

Figure 7.25B shows some of the secondary structures observed in h3R-DAN-G. Turns often formed around 'TGDS' and similar sequences, as was seen in the h3R-DAN-U system. Sometimes, the *O*-glycans formed H-bonds with the surrounding residues. The double-turn which was observed in the h3R-DAN-U system also appeared here. The h3R-DAN-G system formed β-sheets more often than the h3R-DAN-U system (Table 7.2). The β-sheets were small, consisting of only two, short β-strands. The residues within these β-strands were typically hydrophobic.

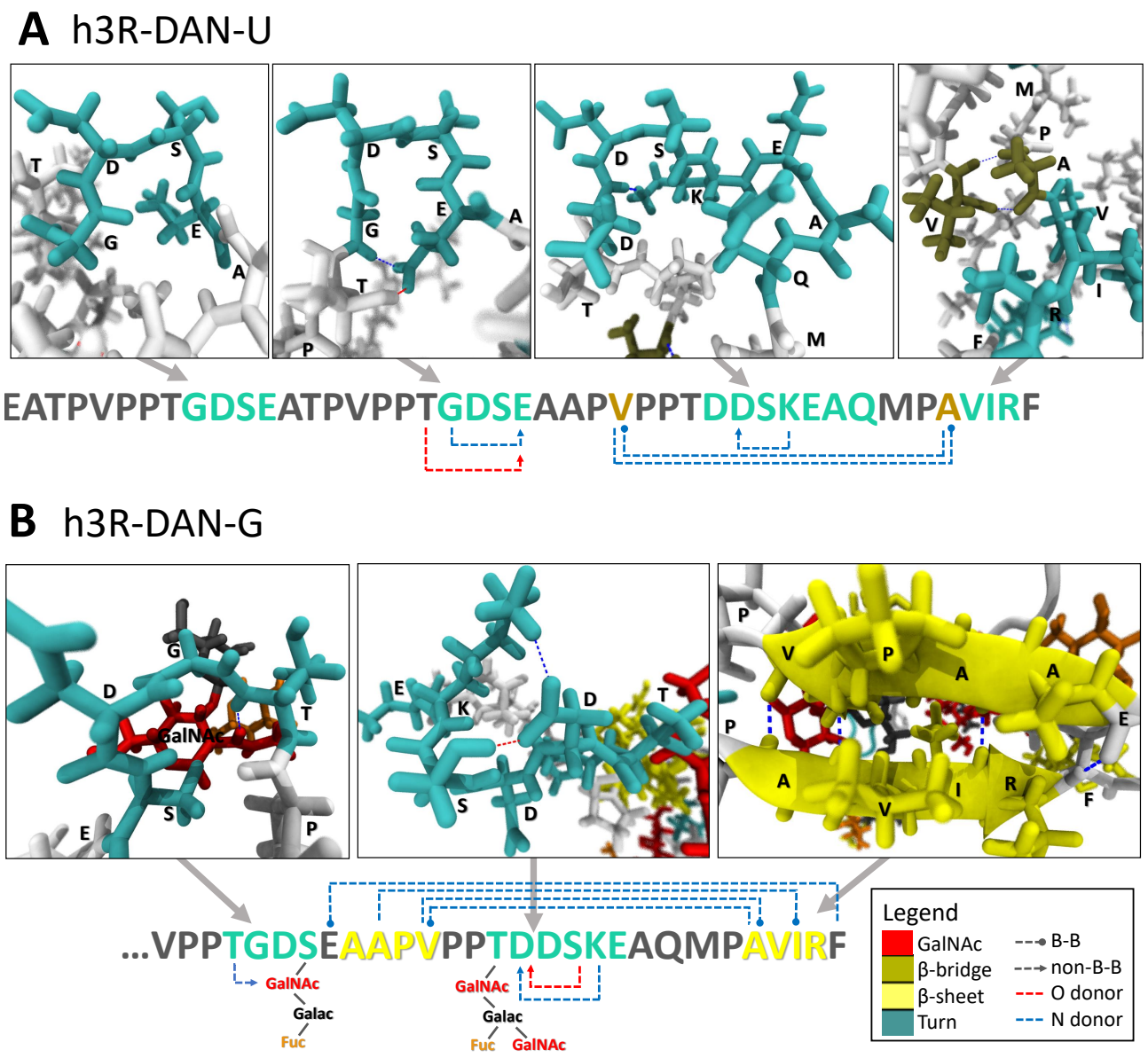


Figure 7.25: **Examples of secondary structures in the replica trajectories for h3R-DAN-U and h3R-DAN-G.** Panels display snapshots of secondary structures. Sequences and models are colour-coded: red is the sugar molecule GalNAc, grey is unstructured coil, blue is turn, brown is β-bridge and yellow is β-sheet. The H-bonds are illustrated as arrows. Red and blue arrows indicate oxygen and nitrogen as H-bond donors, respectively. The arrows point to the acceptor residue. Round arrowheads indicate backbone-backbone H-bonds. Sharp arrow heads indicate non-B-B H-bonds. **A** and **B** display examples of the secondary structures of h3R-DAN-U and h3R-DAN-G, respectively.

8 Discussion

The CEL protein was isolated by different laboratories in the 1960's and 70's (*Morgan et al.*, 1968; *Erlanson*, 1970; *Hernell and Olivecrona*, 1974). Since then, researchers have studied different aspects of CEL, such as; enzymatic activity, bile salt-activation, expression in the mammary glands and pancreas, intracellular localisation, structure of the *CEL* gene, pathology of *CEL* variants and interspecies variation. In this discussion, I will consider previous work focused on the pathological variants and their cellular distribution, as well as CEL sequence variation among species. Important for this thesis, there is not much literature regarding the physical properties of the disordered CEL VNTR. Many researchers have discussed the properties of this region, but few experiments have actually been done.

The results of this thesis were the outcomes of three different approaches, utilizing both *in vivo* and *in silico* techniques. They will be discussed one by one. Finally, there will be a joint discussion of the implications of my studies regarding our understanding of the properties of the CEL VNTR.

8.1 Cellular properties of the CEL VNTR

8.1.1 Differences between the constructs CEL-3R-USA and CEL-3R-DAN

We first tested whether the CEL construct with three repeats, made by a collaborating group in USA and available when this thesis work started, had identical sequence to the CEL-3R allele carried by the Danish family. We found that the Danish variant displayed two mismatching residues compared to the artificially designed American variant: Gly/Asp and Thr/Ala (see Figure 7.6). We decided to test both variants and denoted the respective constructs CEL-3R-DAN and CEL-3R-USA.

The Gly/Asp mismatch is unlikely to be functionally important, as both Gly and Asp residues regularly appear at the ends of the VNTR repeats in the CEL protein (see Figure 7.4). Indeed, in the MD simulations of the CEL-3R-DAN VNTR, both the 'GDS' and the 'DDS' sequence were observed to form structural turns (Figure 7.25).

We believe that the Thr/Ala mismatch is more likely to affect the VNTR structure and function, as the loss of a Thr residue means that one less site is available for *O*-glycosylation. Indeed, when comparing CEL-3R-USA and CEL-3R-DAN in the Western blot medium fraction, it seems like the *O*-glycosylated smear band is narrower for CEL-3R-DAN, suggesting that this variant has a reduced mass of added *O*-glycans. Further, the CEL-3R-DAN abundance (band strength) was about half the abundance of CEL-3R-USA in all fractions. PEST sequences are known to target proteins for

degradation (*Rechsteiner and Rogers*, 1996), while *O*-glycosylations are thought to protect against this degradation (*Loomes*, 1995; *Loomes and Senior*, 1997). This may indicate that the CEL-3R-DAN VNTR is less *O*-glycosylated and more frequently targeted for degradation than CEL-3R-USA.

The CEL-3R-USA bands observed in the Western blots (Figure 7.15) are somewhat consistent with the work in a previous master thesis by *Gravdal* (2016). Although difficult to judge by eye, it may seem like the signal strength ratios between CEL-WT and CEL-3R-USA in the medium and lysate fractions are similar to what was observed in my thesis. However, for the pellet fractions in my thesis, the CEL-WT bands are much weaker than CEL-3R-USA. This band strength difference is not as apparent in the previous thesis by *Gravdal* (2016).

Since the CEL-3R-USA and CEL-3R-DAN variants displayed a fixed ratio of their band strengths when compared against each other in all samples, they will be collectively referred to as CEL-3R in the further comparisons against CEL-WT, CEL-HYB and CEL-TRUNC.

8.1.2 Cellular aggregation and secretion of CEL-3R

The CEL-3R variants displayed cellular properties that somewhat contradictingly implied that they both aggregate in the ER, but also that they are properly secreted. The pellet fraction presented CEL-3R bands that were much stronger than CEL-WT and about as strong as CEL-HYB (Figure 7.15). However, the CEL-3R variants also displayed relatively strong bands in the medium fraction. The interpretation of this is uncertain. It may be that since the CEL-3R variants have a VNTR length similar to CEL-HYB, they are all equally impeded from being transported to the Golgi, which results in ER aggregation. In contrast, the residue compositions of the CEL-3R VNTRs are more similar to CEL-WT, which perhaps ensures better *O*-glycosylation and secretion once the CEL protein has reached the Golgi.

Regardless if secreted well or not, the potential ER aggregation of the CEL-3R variants is highly interesting. As mentioned before, the confirmed pathogenic variants CEL-MODY and CEL-HYB are both associated with ER aggregation and stress (see Sections 4.4.1 ’CEL-MODY’ and 4.4.2 ’CEL-HYB’). The ER stress caused by the CEL-3R variants could lead to acinar cell death, which could lead to pancreatitis, which could lead to damage of insulin producing beta cells. In other words, the diabetes seen in the Danish family carrying *CEL-3R* may be a type of T3cD (see Section 4.2.3 ’Diabetes mellitus’). However, the potential ER stress induced by the CEL-3R variants would have to be confirmed by detection of ER stress markers.

8.1.3 Reflections on the cellular analysis

The cellular analysis of CEL-3R variants was somewhat successful, but also incomplete due to time constraints. The Western blots indicated that the CEL-3R variants could aggregate in the ER. For further verification, the final Western blotting protocol should be employed in three, independent replications in order to calculate p-values of the findings. Furthermore, the potential ER stress induced by CEL-3R variants should be confirmed by testing of ER stress markers. This work could be done by immunofluorescent staining and microscopy of cells expressing CEL-3R, as well as by Western blotting. The final western blotting protocol could also be further optimised. The CEL-TRUNC were detectable in the Western blot membranes only after switching to a 0.45 PVDF µm membrane. However, sometimes the membrane appeared dirty, likely as a result of activation in a different buffer than that used in the blotting sandwich (see lysate membrane in Figure 7.15). The protocol for 0.45 µm PVDF could be further optimised by trying different Turbo-Blot packages and transfer settings.

It is uncertain why CEL-TRUNC only appeared in the Western blot imaging after switching to a membrane with larger pore sizes. The CEL protein is relatively large. However, CEL-TRUNC is the smallest CEL variant tested in this thesis. Its mass (predicted to be around 59.3 kDa) is about half of the fully *O*-glycosylated CEL-WT (around 100 kDa, see Table 4.1 and Figure 7.15). It would be reasonable to think that too small pore sizes would disproportionally affect the larger CEL variants. Yet, the band intensities of the larger variants became only marginally stronger after switching the membrane. Meanwhile, the small CEL-TRUNC variant went from being completely undetectable to fairly detectable.

It is worth noting that the apparent kDa sizes of the bands observed in the Western blots do not necessarily reflect the real sizes of the migrating CEL proteins. The apparent size of IDPs is usually 1.2–1.8 times higher in Western blots than what is predicted from their protein sequence or observed in MS (*Schramm et al.*, 2019). There are at least two possible factors behind this effect: (1) the low hydrophobicity of IDP sequences make weaker connections to the SDS reagents, meaning less negative charge and slower migration. (2) Proline-rich sequences enforce a rigid structure on the IDP which could further slow the migration (*Schramm et al.*, 2019).

Another factor that could affect the migrating CEL bands in SDS-PAGE is PTMs. Disulfide bridges are broken by the reducing agent and should therefore not affect the migration. However, all the other modifications (phosphorylation, *N*-glycosylation and *O*-glycosylations) add mass to the CEL protein and slows its migration. They also branch off the protein backbone, meaning that the molecule is not completely linearised.

8.2 The CEL protein in phylogenetically diverse species

8.2.1 When did the CEL VNTR first originate?

The results of the vertebrate phylogenetic analysis of CEL showed that the CEL VNTR is only present in species of the mammalian lineage and not in any other vertebrates (Figure 7.2). A more limited analysis by *Holmes and Cox* (2011) also identified the CEL VNTR in mammalian species only. This implies that the CEL VNTR originated after the clade Synapsida (ancestors of Mammalia) branched away from Sauropsida (ancestors of all birds and living reptiles), which happened around 320 million years ago (MYA) (*Angielczyk and Kammerer*, 2018). Further, as the VNTR was identified in both metatherian and eutherian species (Figure 7.4), it must have originated before these clades branched away from each other around 100 MYA (*Zachos*, 2020).

The estimated range of when the CEL VNTR originated could be greatly reduced if the presence of a VNTR in prototherian mammals could be determined. The CEL sequence of the prototherian Australian echidna displayed a C-terminal, PEST sequence which could be interpreted as a decayed VNTR. Only two sequences in the whole clade of Prototheria were present in the mammalian CEL VNTR dataset. Thus, more sequences should be collected and reviewed to form a dataset representative of prototherians. This animal lineage branched off from Theria at about 160 MYA (*Zachos*, 2020). A confirmation or rejection of the presence of the VNTR in Prototheria would move its origin to about 320–160 MYA or 160–100 MYA, respectively. We propose Figure 8.1 as a model for the evolution of the CEL protein.

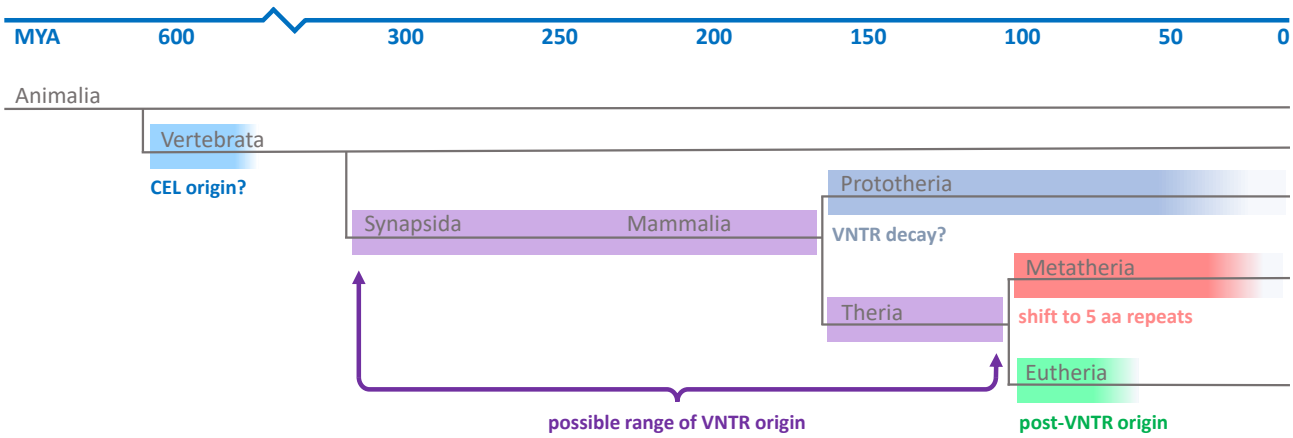


Figure 8.1: **Proposed model of the evolution of the CEL VNTR region.** A phylogenetic tree showing the animal lineages leading to the three main mammalian clades Prototheria, Metatheria and Eutheria. The tree is annotated with the proposed evolutionary events that have impacted the structure of the CEL protein and its VNTR region. Time is displayed horizontally in Million Years Ago (MYA). The coloured boxes represent the estimated ranges of time.

Determining if the VNTR is present in all mammals could have interesting implications on the effect of the VNTR on the CEL protein. Mammalia is defined by the ability to feed offspring with milk and the CEL protein is present in the milk of many mammals (*Freed et al.*, 1986; *Wang et al.*, 1989; *Wang and Hartsuck*, 1993). The VNTR could therefore be an evolutionary adaptation for secretion of the enzyme in milk, assuming the presence of CEL in milk confers some advantage. One such advantage could be to aid in enzymatically digesting the milk fats for the offspring. The role of the mucin-like VNTR could be to make the CEL protein more soluble in the milk. Or rather, the *O*-glycosylations on the CEL VNTR in the ingested milk is thought to maybe interact with the microbiota in the intestines (*Jellas et al.*, 2018).

8.2.2 Characterisation of the VNTR in Mammalia

The number of CEL VNTR repeats observed in the different species varied considerably (Figure 7.4). The smallest number of repeats was found in the Agile gracile mouse opossum with 1 repeat (although technically not a VNTR). The largest number of repeats was found in the Western gorilla with 39 repeats. Closely related species tended to have similar repeat numbers, but not always. Diet did not seem to correlate with VNTR length, as both carnivores and herbivores displayed low and high repeat numbers. Animal size might be a contributing factor, as larger animals tended to display higher repeat numbers, but with many exceptions such as the large species in the Artiodactyla order. In summary, no clear connection was found between the number of CEL VNTR repeats and the morphology or physiology of the species.

The mammalian phylogenetic analysis showed that the eutherian and metatherian VNTRs shared similar residue types and positions. (Figure 7.4). However, the VNTR repeat lengths were usually 5 residues in Metatheria and 11 in Eutheria. Either of these lengths, or a different length entirely, may have been present in the CEL VNTR of the common ancestors of metatherians and eutherian. The metatherian post-VNTRs sometimes contained sequences very similar to the 11-residue Eutherian repeats, which indicates that a sequence similar to the 11-residue repeat evolved first and later was reduced to 5 residues in Metatheria. Indeed, *Holmes and Cox* (2011) identified the metatherian Opossum (not to be confused with the Agile gracile mouse opossum in this thesis) with one VNTR repeat which was fairly identical to the 11 aa repeats in Eutheria.

The annotations in the mammalian phylogenetic analysis (Figure 7.4) is largely in agreement with the work of *Holmes and Cox* (2011). One important exception is the number of VNTR repeats in the human CEL gene. We deliberately chose a sequence of 16 repeats for the phylogenetic analysis, as this is the most common CEL allele variant (*Higuchi et al.*, 2002; *Bengtsson-Ellmark et al.*, 2004; *Fjeld et al.*, 2016). Meanwhile, *Holmes and Cox* (2011) used a rare allele of 17 repeats, probably because this CEL variant was one of the first sequenced and is annotated as the canon sequence in some databases (see CEL entries in GenBank and UniProt). Other results which differed from their work, such as number of repeats for the Rhesus monkey, may stem from the use of different sequences or different interpretations and annotations of the VNTR repeats. For example, *Holmes and Cox* (2011) annotate the beginning and end of the VNTR repeats differently than we do. Their consensus sequence 'PVPPTGDSEAA' would be annotated as 'EAAPVPPTGDS' in this thesis.

8.2.3 The role of the post-VNTR in Eutheria

The eutherian post-VNTR was found to be a well-conserved sequence of hydrophobic residues (Figure 7.5). The post-VNTR sequence probably does not have the same properties or function as the preceding VNTR repeats, as it does not share a similar residue composition and is unlikely to be *O*-glycosylated. Instead, the conserved, hydrophobic residues (especially the 100% conserved Phe residue) may imply that the post-VNTR is important for a potential hydrophobic binding partner.

One such binding partner may be the CEL globular domain itself. In an X-ray crystallography study on bovine CEL, *Chen et al.* (1998) discovered that the post-VNTR was embedded into the active site of the globular domain. They hypothesised that the post-VNTR blocks substrates from reaching the active site, until bile salts or other emulsifiers induce conformational changes which displace the post-VNTR from the globular domain (*Chen et al.*, 1998).

There is not much evidence to support that the post-VNTR, or the VNTR itself, plays a significant role in the CEL enzymatic activity. As previously mentioned, *DiPersio et al.* (1994) found a slightly lower K_m and slightly higher V_{max} in rat CEL variants with truncated VNTRs when bile salt levels were below saturation. However, *Hansson et al.* (1993) did not find such effects for human CEL, although they used a different experimental setup. Further, *Gravdal et al.* (2021) and *Cassidy et al.* (2020) have compared CEL variants with very different VNTRs and only found small differences in enzymatic activity. Still, it is possible that the post-VNTR is a preventative measure against pancreatic self-digestion by the CEL protein.

8.2.4 Reflections on the phylogenetic analyses

The phylogenetic analysis of CEL was successful in collecting sequences from a broad range of vertebrate species and in mapping the diversity of the VNTR in Mammalia. We also gained strong evidence in support of the mammalian origin of the VNTR and for the conservation of the eutherian VNTR and post-VNTR sequences. However, we failed in finding any association between the number of VNTR repeats and characteristics of the species.

Further, the whole phylogenetic work could have been optimised for efficiency and higher quality results in several ways: First, nucleotides contain more informational resolution than the amino acids they encode. DNA or mRNA sequences may have been collected instead of protein sequences, as the VNTR patterns or potential distortions in those patterns could have been more easily recognised at the nucleotide level. Secondly; when available, multiple CEL sequences for each species could have been included to show isoforms and intraspecies variation. Thirdly, algorithms might have been established to recognise and annotate VNTRs, as well as automatically creating datasets and phylogenetic figures. This would have greatly reduced the time spent on this type of analysis. Finally, more accurate phylogenetic trees could probably have been constructed. We used the NCBI Taxonomy Database, but this taxonomy is only approximately phylogenetic. For example, humans and chimpanzees are closer evolutionary relatives to each other than to the Western gorilla. However, in the trees constructed in this thesis, all three species branched off from each other at the same time.

8.3 Physical properties of the simulated VNTR region

8.3.1 The disorder of the CEL-3R-DAN VNTR

The main outcome of the MD simulations was that the CEL-3R-DAN VNTR displayed intrinsic disorder. Despite the formation of some temporary secondary structures and the overall compaction of the protein, it never displayed a fixed tertiary structure (Figure 7.22). This propensity for disorder was true for both the unmodified h3R-DAN-U and *O*-glycosylated h3R-DAN-G structures. The prediction is in agreement with every article through the years which has ever proposed or mentioned that the CEL VNTR is intrinsically disordered (*Reue et al.*, 1991; *Chen et al.*, 1995; *Terzyan et al.*, 2000; *Johansson et al.*, 2018).

Although the VNTR is disordered, it can still form secondary structures and be restricted in the types of conformations it exhibits. Indeed, we observed that the typical 'GDSE' sequence seen on the ends of the VNTR repeats often formed turn structures (Figures 7.24 and 7.25). The flexible Gly residue likely helps in achieving the bond angles needed for turns. The hydrophilic Asp, Ser and Glu residues were extending away from the turn and into the solvent, which may also have reinforced the bond angles necessary to form the turn. These turns usually consisted of 4 residues, similar to β-turns. However, the turns did not display the typical backbone-backbone H-bonds. Rather, the turns sporadically displayed sidechain-sidechain and sidechain-backbone H-bonds, often involving the Asp sidechain. Thus, the observed turns were more similar to the reported 'Asx turn' (*D'mello et al.*, 2022).

Further, the hydrophobic residues in the VNTR, like Val or Ala, tended to contribute to β-bridges or β-sheets. The strands of these β-bridges and β-sheets usually formed in proximity to turns, meaning they likely reinforced each other. The hydrophobic Ala and Val residues are not typically seen in other PEST sequences (*Rechsteiner and Rogers*, 1996), which might indicate that the formation of these small secondary structures are particularly important for the function of CEL.

Lastly, the parts of the VNTR sequence consisting of Pro residues rarely took part in the formation of secondary structures. This is in agreement with the rigidity of the φ- and ψ-angles of the Pro residue and its resistance to conform to secondary structures such as α-helices and β-sheets. However, Pro residues are known to participate in turns, which was rarely observed in the simulated trajectories.

Moreover, the secondary structures and other intramolecular forces in the CEL-3R-DAN VNTR made the simulated structure become less linear and more bunched together (Figure 7.22). The turns made the protein strands change direction, back towards the rest of the protein. Meanwhile, the β-bridges or β-sheets reinforced the interactions between the strands. The secondary structures would sporadically form and disappear, leading to the protein structure shrinking and expanding. This property is similar to the more large-scale phenomenon of condensates (see Section 4.5 'Intrinsically disordered protein regions'). It is possible that longer VNTRs or multiple VNTRs could form these "condensing" interactions at larger scales.

8.3.2 The conformational restrictions by *O*-glycosylations

There was not observed any profound differences in the conformations of the unmodified h3R-DAN-U and the *O*-glycosylated h3R-DAN-G. The *O*-glycosylations did not prevent the formation of secondary structures (Figure 7.24) nor the compaction of the protein (Figure 7.23). Still, the *O*-glycosylations did seem to restrict the conformations of the structure. The h3R-DAN-G structure displayed a more narrow range of compaction prevalence and the secondary structures that formed were more stable. The less volatile properties of h3R-DAN-G indicate that the *O*-glycosylations contribute to stabilising the conformations of the protein.

The effect of the *O*-glycosylations may have been more apparent if we had simulated larger VNTRs or multiple VNTRs. In this hypothetical scenario, the *O*-glycans could have served as sterical hindrance against the formation of compact, protein-protein interactions. The structure used in this thesis tended to have the *O*-glycans extend away from the protein and towards the solvent. In contrast, the *O*-glycans in a larger system of interacting VNTR proteins may have clashed with other proteins and glycans. Such collisions could possibly have prevented the formation of these large-scale protein-protein interactions and condensates.

8.3.3 Reflections on the MD simulations

The aim of applying MD simulation to predict the conformations of the CEL-3R-DAN VNTR was mainly successful. We sampled some of the conformational space of the structures by running replicate simulations. Analysis and comparison of the conformations revealed some differences, but mostly similarities, between the unmodified h3R-DAN-U and the *O*-glycosylated h3R-DAN-G. However, the simulations are just predictions and may be incorrect. We need experimental proof in some form to confirm or reject the predicted findings.

Moreover, the physical properties of the CEL-3R-DAN VNTR could be further characterised by more simulations with different setups. Firstly, the *O*-glycosylations in the h3R-DAN-G structure were somewhat arbitrarily designed. The glycan mass was estimated from the Western blots of CEL-3R-DAN expressed in HEK293 cells, which does not necessarily represent the *O*-glycosylations of CEL-3R-DAN in human acinar cells. Further research on the CEL VNTR *O*-glycosylations should be conducted in order to design models with more accurate *O*-glycan structure. Secondly, the simulated structures were protonated at pH 7. The ER and Golgi in the CEL secretion pathway work at pH 8 and 6, respectively. The secondary structures and conformations of the CEL VNTR could change within such a pH range. Lastly, the application of a different force field than CHARMM36m for the MD simulation may yield different results. The force field Amberff99SB-disp is a potential candidate, as it has actually been reported to be more accurate than CHARMM36m (*Wang*, 2021).

8.4 Final thoughts on the properties and role of the VNTR

From the combined results of the phylogenetic analysis, cellular experiment and MD analysis, we can reflect on the general properties and functions of the CEL VNTR.

The VNTR residues are well conserved within eutherian mammals, with a typical repeat sequence of 'EATPVPPTGDS'. As seen in the MD simulations, these residues and their order are probably important for the overall conformation of the VNTR. The residues at the ends of the repeat ('GDSE') tended to form turns. The hydrophobic residues Ala and Val tended to form β-bridges or β-sheets. Meanwhile, the Pro residues formed unstructured coils. The alternation of these occasional turns, β-structures and unstructured coils seems to stabilise the compaction of the VNTR, regardless of being *O*-glycosylated or not.

The post-VNTR sequence seems to be able to participate in the occasional secondary structures that arise in the VNTR, but the post-VNTR may also have some other, unknown function. This is because this sequence exhibits conserved residues that are different from the VNTR. The best example is the C-terminal Phe residue which was conserved in virtually all eutherian mammals (Figure 7.5). The aromatic Phe sidechain could be important for pi-pi interactions. Thus, there may exist an undiscovered, hydrophobic binding partner of the post-VNTR (other than the CEL active site).

As mentioned above, the CEL VNTR is unlikely to play any important role in the enzymatic activity of the globular domain. Indeed, a collaborating group in St. Louis assessed the enzymatic activity of various CEL variants (*Cassidy et al.*, 2020; *Gravdal et al.*, 2021). They also tested the activities of CEL-3R-USA and CEL-3R-DAN (Xunjun Xiao, personal correspondence). In all analyses, there was not identified any differences in enzymatic activity that were considered of biological importance. However, if the VNTR does actually have any enzymatic function in CEL, then it may be to work as a tether or a flexible linker between the active site and the post-VNTR sequence, as proposed by *Chen et al.* (1995). This tether will direct the post-VNTR to block the catalytic site when CEL is still present in the pancreas and has not been stimulated by mixing with bile salts.

A potential property of the CEL VNTR is the formation of condensates (see Section 4.5 'Intrinsically disordered protein regions'). As mentioned before, a condensate is a diffuse mass of cellular compounds, such as proteins, which are held together by a vast amount of weak forces. We predicted that the short CEL-3R VNTR becomes compact and dense due to intramolecular forces (Figure 7.24). Perhaps the CEL VNTR regions employ these forces to participate in large-scale condensates in the ER, Golgi or somewhere else within the cell. Furthermore, the *O*-glycosylations could serve to break up such forces by steric hindrance and increased solvation.

The pathology attributed to some CEL variants could perhaps be due to different propensities to participate in condensates. For example, too strong interactions could mean that the CEL proteins would be unable to depart from the condensate, resulting in aggregations. Or, too weak interactions could mean that CEL proteins will not participate in the condensates, which could somehow impede the secretion pathway.

Indeed, mutations in IDRs may lead to a change in residue composition, and subsequently, a pathogenic shift in condensation propensity. One example is presented in a recent paper published by *Mensah et al.* (2023), where frameshift mutations in the IDP-encoding *HMGB1* gene may have altered the proteins ability to condensate. The frameshift mutations led to an increased number of positively charged Arg residues and caused bodily malformations in humans. The authors compiled and published lists of proteins with C-terminal IDRs and frameshifts that increase the number of Arg residues. The CEL protein was not included in these lists, although it contains a C-terminal IDR and has pathogenic frameshift variants with increased residue numbers for Arg (CEL-MODY, see Section 4.4.1 'CEL-MODY'). The reason CEL was not included in the lists, was because it was annotated with a C-terminal "mucin-like domain" in the Pfam database (Denes Hnisz, personal correspondence). Regardless, these findings highlight the potential function of the CEL VNTR as a participant in condensates, and not just a vessel for *O*-glycosylation.

8.5 Conclusion

The VNTR region of CEL was investigated using cellular experiments and *in silico* analyses. We found that the CEL-3R variant, discovered in a Danish diabetes family, exhibited levels of aggregation comparable to the confirmed pathogenic variant CEL-HYB. We also found that the CEL VNTR likely evolved sometime after the origin of Synapsida, but before the formation of Eutheria and Metatheria. Further, we found that the typical 'EATPVPPTGDS' residues in the VNTR repeats are well conserved within eutherian mammals and that they are likely important for forming alternating turns, β-bridges, β-sheets and unstructured coils. These semi-stable secondary structures induce the VNTR to become compacted, but still disordered. We found no obvious link between the number of VNTR repeats and the morphology or physiology of mammalian species.

Still, our analyses make it possible to propose some hypotheses about the role of the VNTR in the CEL protein. It is striking that this region evolved along with mammals when we know that the protein is present in the mother's milk used to feed the offspring. It is tempting to speculate that the VNTR stabilizes the CEL protein, aids in lipid digestion in the gut, or that it has some beneficiary effect on the microbiota in the intestines.

9 Future prospects

There are many possibilities for continuing the study of the CEL VNTR and the CEL-3R-DAN variant.

- **Wet lab experiments**
  - Repeat the optimised Western blotting protocol to obtain statistically significant results
  - Perform confocal microscopy to observe cellular localisation of CEL variants by immunofluorescence
  - Do Western blotting and confocal microscopy using antibodies for ER stress markers
  - Use the methods CD, NMR, SAXS, etc. to study the IDP properties of the CEL VNTR. Compare with the predicted properties by MD simulation
- **Phylogeny**
  - Expand the sequencing of *CEL* in Mammalia, especially Prototheria
  - Perform a DNA sequence search on the *CEL* gene. Nucleotide sequences confer more ”resolution” than amino acid sequences
  - Automate the sequence search and subsequent filtering and VNTR annotation
- **Molecular modelling**
  - Simulate a longer VNTR (like in CEL-WT) to see if the condensation properties of the CEL-3R-DAN simulation will scale, or if new properties are found.
  - Simulate the VNTR with protonation assuming pH 8 or 6 to mimic the environments of ER and Golgi, respectively. See if compaction properties change
  - Simulate multiple CEL VNTR molecules to see if they show the same intermolecular interactions as the intramolecular forces seen in this thesis. Do the same with *O*-glycosylations to see if this prevents aggregation.
  - Simulate the entire CEL protein to see if the globular domain and VNTR form any interactions.
  - Search for potential interaction partners between the CEL post-VNTR and other proteins.

10 References

A-Kader, H. H., and F. K. Ghishan, The Pancreas, *Textbook of Clinical Pediatrics*, p. 1925, doi:10.1007/978-3-642-02202-9_198, 2012.

Abouakil, N., E. Mas, N. Bruneau, A. Benajiba, and D. Lombardo, Bile salt-dependent lipase biosynthesis in rat pancreatic AR 4-2 J cells: Essential requirement of N-linked oligosaccharide for secretion and expression of a fully active enzyme, *Journal of Biological Chemistry*, *268*(34), 25,755–25,763, doi:10.1016/s0021-9258(19)74454-8, 1993.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool, *Journal of molecular biology*, *215*(3), 403–410, doi:10.1016/S0022-2836(05)80360-2, 1990.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic acids research*, *25*(17), 3389–3402, doi:10.1093/NAR/25.17.3389, 1997.

Altschul, S. F., J. C. Wootton, E. M. Gertz, R. Agarwala, A. Morgulis, A. A. Schäffer, and Y. K. Yu, Protein database searches using compositionally adjusted substitution matrices, *The FEBS journal*, *272*(20), 5101–5109, doi:10.1111/J.1742-4658.2005.04945.X, 2005.

Angielczyk, K. D., and C. F. Kammerer, Mammalian Evolution, Diversity and Systematics, in *Mammalian Evolution, Diversity and Systematics*, chap. 5, pp. 117–198, De Gruyter, doi:10.1515/9783110341553-005, 2018.

Ashraf, H., J. P. Colombo, V. Marcucci, J. Rhoton, and O. Olowoyo, A Clinical Overview of Acute and Chronic Pancreatitis: The Medical and Surgical Management, *Cureus*, doi:10.7759/CUREUS.19764, 2021.

Baba, T., D. Downs, K. W. Jackson, J. Tang, and C. S. Wang, Structure of human milk bile salt activated lipase, *Biochemistry*, *30*(2), 500–510, doi:10.1021/BI00216A028, 1991.

Baxevanis, A. D., G. D. Bader, and D. S. Wishart, *Bioinformatics, 4th ed.*, 625 pp., Wiley, 2020.

Bengtsson-Ellmark, S. H., J. Nilsson, M. Orho-Melander, K. Dahlenborg, L. Groop, and G. Bjursell, Association between a polymorphism in the carboxyl ester lipase gene and serum cholesterol profile, *European Journal of Human Genetics 2004 12:8*, *12*(8), 627–632, doi:10.1038/sj.ejhg.5201204, 2004.

Bläckberg, L., and O. Hernell, Further characterization of the bile salt-stimulated lipase in human milk, *FEBS Letters*, *157*(2), 337–341, doi:10.1016/0014-5793(83)80571-7, 1983.

Blackberg, L., O. Hernell, T. Olivecrona, L. Dommelöf, and M. R. Malinov, The bile salt-stimulated lipase in human milk is an evolutionary newcomer derived from a non-milk protein, *FEBS Letters*, *112*(1), 51–54, doi:https://doi.org/10.1016/0014-5793(80)80125-6, 1980.

Bläckberg, L., K. A. Ängquist, and O. Hemell, Bile salt-stimulated lipase in human milk: Evidence for its synthesis in the lactating mammary gland, *FEBS Letters*, *217*(1), 37–41, doi:10.1016/0014-5793(87)81237-1, 1987.

Blackberg, L., M. Stromqvist, M. Edlund, K. Juneblad, L. Lundberg, L. Hansson, O. Hernell, S. Astra Hassle, and P. Research, Recombinant Human-Milk Bile-Salt-Stimulated Lipase, *European Journal of Biochemistry*, *228*(3), 817–821, 1995.

Bonnefond, A., R. Unnikrishnan, A. Doria, M. Vaxillaire, R. N. Kulkarni, V. Mohan, V. Trischitta, and P. Froguel, Monogenic diabetes, *Nature Reviews Disease Primers 2023 9:1*, *9*(1), 1–16, doi:10.1038/s41572-023-00421-w, 2023.

Brucale, M., B. Schuler, and B. Samorì, Single-molecule studies of intrinsically disordered proteins, *Chemical Reviews*, *114*(6), 3281–3317, doi:10.1021/CR400297G, 2014.

Bruneau, N., A. Nganga, E. A. Fisher, and D. Lombardo, O-Glycosylation of C-terminal tandem-repeated sequences regulates the secretion of rat pancreatic bile salt-dependent lipase, *The Journal of biological chemistry*, *272*(43), 27,353–27,361, doi:10.1074/JBC.272.43.27353, 1997.

Bulleid, N. J., Disulfide Bond Formation in the Mammalian Endoplasmic Reticulum, *Cold Spring Harbor Perspectives in Biology*, *4*(11), doi:10.1101/CSHPERSPECT.A013219, 2012.

Cassidy, B. M., S. Zino, K. Fjeld, A. Molven, M. E. Lowe, and X. Xiao, Single nucleotide polymorphisms in CEL-HYB1 increase risk for chronic pancreatitis through proteotoxic misfolding, *Human Mutation*, *41*(11), 1967–1978, doi:10.1002/HUMU.24105, 2020.

Chen, J. C., et al., Structure of bovine pancreatic cholesterol esterase at 1.6 Å: Novel structural features involved in lipase activation, *Biochemistry*, *37*(15), 5107–5117, doi:10.1021/BI972989G, 1998.

Chen, W., J. Helenius, I. Braakman, and A. Helenius, Cotranslational folding and calnexin binding during glycoprotein synthesis., *Proceedings of the National Academy of Sciences of the United States of America*, *92*(14), 6229, doi:10.1073/PNAS.92.14.6229, 1995.

Cole, J. B., and J. C. Florez, Genetics of diabetes mellitus and diabetes complications, *Nature Reviews Nephrology*, *16*(7), 377–390, doi:10.1038/s41581-020-0278-5, 2020.

Dalva, M., et al., Copy number variants and VNTR length polymorphisms of the carboxyl-ester lipase (CEL) gene as risk factors in pancreatic cancer, *Pancreatology*, *17*(1), 83–88, doi:10.1016/J.PAN.2016.10.006, 2017.

DiPersio, L. P., and D. Y. Hui, Aspartic acid 320 is required for optimal activity of rat pancreatic cholesterol esterase., *Journal of Biological Chemistry*, *268*(1), 300–304, doi:10.1016/S0021-9258(18)54149-1, 1993.

DiPersio, L. P., R. N. Fontaine, and D. Y. Hui, Identification of the active site serine in pancreatic cholesterol esterase by chemical modification and site-specific mutagenesis., *The Journal of Biological Chemistry*, *265*(28), 16,801–16,806, doi:10.1016/S0021-9258(17)44832-0, 1990.

DiPersio, L. P., R. N. Fontaine, and D. Y. Hui, Site-specific mutagenesis of an essential histidine residue in pancreatic cholesterol esterase, *Journal of Biological Chemistry*, *266*(7), 4033–4036, doi:10.1016/S0021-9258(20)64279-X, 1991.

DiPersio, L. P., C. P. Carter, and D. Y. Hui, Exon 11 of the rat cholesterol esterase gene encodes domains important for intracellular processing and bile salt-modulated activity of the protein., *Biochemistry*, *33*(11), 3442–3448, doi:10.1021/BI00177A038, 1994.

D'mello, V. C., G. Goldsztejn, V. Rao Mundlapati, V. Brenner, E. Gloaguen, F. Charnay-Pouget, D. J. Aitken, and M. Mons, Characterization of Asx Turn Types and Their Connate Relationship with β-Turns, *Chemistry – A European Journal*, *28*(25), e202104,328, doi:10.1002/CHEM.202104328, 2022.

Dunker, A. K., C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic´, Intrinsic Disorder and Protein Function, *Biochemistry*, doi:10.1021/BI012159, 2002.

Erdős, G., and Z. Dosztányi, Analyzing Protein Disorder with IUPred2A, *Current Protocols in Bioinformatics*, *70*(1), doi:10.1002/CPBI.99, 2020.

Erlanson, C., p-Nitrophenylacetate as a Substrate for a Carboxyl-ester Hydrolase in Pancreatic Juice and Intestinal Content., *Scandinavian Journal of Gastroenterology*, *5*(5), 333–336, doi:10.1080/00365521.1970.12096599, 1970.

Fjeld, K., et al., A recombined allele of the lipase gene CEL and its pseudogene CELP confers susceptibility to chronic pancreatitis, *Nature Genetics*, *47*(5), 518–522, doi:10.1038/ng.3249, 2015.

Fjeld, K., et al., Length of Variable Numbers of Tandem Repeats in the Carboxyl Ester Lipase (CEL) Gene May Confer Susceptibility to Alcoholic Liver Cirrhosis but Not Alcoholic Chronic Pancreatitis, *PLOS ONE*, *11*(11), e0165,567, doi:10.1371/JOURNAL.PONE.0165567, 2016.

Fjeld, K., et al., The genetic risk factor CEL-HYB1 causes proteotoxicity and chronic pancreatitis in mice, *Pancreatology : official journal of the International Association of Pancreatology (IAP) ... [et al.]*, *22*(8), 1099–1111, doi:10.1016/J.PAN.2022.11.003, 2022.

Freed, L. M., C. M. York, M. Hamosh, J. A. Sturman, and P. Hamosh, Bile salt-stimulated lipase in non-primate milk: longitudinal variation and lipase characteristics in cat and dog milk, *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism*, *878*(2), 209–215, doi:10.1016/0005-2760(86)90148-7, 1986.

Gopal, S. M., S. Wingbermühle, J. Schnatwinkel, S. Juber, C. Herrmann, and L. V. Schäfer, Conformational Preferences of an Intrinsically Disordered Protein Domain: A Case Study for Modern Force Fields, *The journal of physical chemistry. B*, *125*(1), 24–35, doi:10.1021/ACS.JPCB.0C08702, 2021.

Gravdal, A., Functional characterization of Carboxyl Ester Lipase variants causing diabetes and exocrine dysfunction, *Bergen Open Research Archive*, pp. 1–79, 2016.

Gravdal, A., et al., The position of single-base deletions in the VNTR sequence of the carboxyl ester lipase (CEL) gene determines proteotoxicity, *Journal of Biological Chemistry*, *296*, 100,661, doi:10.1016/j.jbc.2021.100661, 2021.

Hagen, G., S. Müller, M. Beato, and G. Suske, Cloning by recognition site screening of two novel GT box binding proteins: a family of Sp1 related genes., *Nucleic Acids Research*, *20*(21), 5519–5525, doi:10.1093/NAR/20.21.5519, 1992.

Halbrook, C. J., C. A. Lyssiotis, M. Pasca, D. Magliano, and A. Maitra, Pancreatic cancer: Advances and challenges, *Cell*, *186*(8), 1729–1754, doi:10.1016/j.cell.2023.02.014, 2023.

Hansson, L., L. Blackberg, M. Edlund, L. Lundberg, M. Stromqvist, and O. Hernell, Recombinant human milk bile salt-stimulated lipase. Catalytic activity is retained in the absence of glycosylation and the unique proline-rich repeats., *Journal of Biological Chemistry*, *268*(35), 26,692–26,698, doi:10.1016/S0021-9258(19)74368-3, 1993.

Hart, P. A., et al., Type 3c (pancreatogenic) diabetes mellitus secondary to chronic pancreatitis and pancreatic cancer, *Lancet Gastroenterol Hepatol*, *1*(3), 226, doi:10.1016/S2468-1253(16)30106-6, 2016.

Hernell, O., and T. Olivecrona, Human milk lipases. I. Serum stimulated lipase, *Journal of Lipid Research*, *15*(4), 367–374, doi:10.1016/s0022-2275(20)36784-5, 1974.

Hess, J., P. Angel, and M. Schorpp-Kistner, AP-1 subunits: quarrel and harmony among siblings, *Journal of Cell Science*, *117*(25), 5965–5973, doi:10.1242/JCS.01589, 2004.

Hidalgo, M., S. Cascinu, J. Kleeff, R. Labianca, J. M. Löhr, J. Neoptolemos, F. X. Real, J. L. Van Laethem, and V. Heinemann, Addressing the challenges of pancreatic cancer: Future directions for improving outcomes, *Pancreatology*, *15*(1), 8–18, doi:10.1016/J.PAN.2014.10.001, 2015.

Higuchi, S., Y. Nakamura, and S. Saito, Characterization of a VNTR polymorphism in the coding region of the CEL gene, *Journal of Human Genetics 2002 47:4*, *47*(4), 213–215, doi:10.1007/s100380200027, 2002.

Hilger-Eversheim, K., M. Moser, H. Schorle, and R. Buettner, Regulatory roles of AP-2 transcription factors in vertebrate development, apoptosis and cell-cycle control, *Gene*, *260*(1-2), 1–12, doi:10.1016/S0378-1119(00)00454-6, 2000.

Holmes, R. S., and L. A. Cox, Comparative Structures and Evolution of Vertebrate Carboxyl Ester Lipase (CEL) Genes and Proteins with a Major Role in Reverse Cholesterol Transport, *Cholesterol*, *2011*, 15, doi:10.1155/2011/781643, 2011.

Huang, J., S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. De Groot, H. Grubmüller, and A. D. Mackerell, CHARMM36m: an improved force field for folded and intrinsically disordered proteins, *Nature Publishing Group*, *14*(1), doi:10.1038/nMeth.4067, 2016.

Hui, D. Y., K. Hayakawa, and J. Oizumi, Lipoamidase activity in normal and mutagenized pancreatic cholesterol esterase (bile salt-stimulated lipase)., *Biochemical Journal*, *291*(Pt 1), 65, doi:10.1042/BJ2910065, 1993.

Hunter, J. D., Matplotlib: A 2D graphics environment, *Computing in Science & Engineering*, *9*(3), 90–95, doi:10.1109/MCSE.2007.55, 2007.

Iakoucheva, L. M., A. L. Kimzey, C. D. Masselon, J. E. Bruce, E. C. Garner, C. J. Brown, A. K. Dunker, R. D. Smith, and E. J. Ackerman, Identification of Intrinsic Order and Disorder in the DNA Repair Protein XPA, *Protein Science, 10(3):560-571*, *10*(3), 560–571, doi:10.1110/PS.29401, 2001.

Ilic, M., and I. Ilic, Epidemiology of pancreatic cancer, *World Journal of Gastroenterology*, *22*(44), 9694–9705, doi:10.3748/WJG.V22.I44.9694, 2016.

Jellas, K. E., et al., The mucinous domain of pancreatic carboxyl-ester lipase (CEL) contains core 1/core 2 O-glycans that can be modified by ABO blood group determinants, *Journal of Biological Chemistry*, *293*(50), 19,476–19,491, doi:10.1074/JBC.RA118.001934, 2018.

Jellas, K. E., et al., Two New Mutations in the CEL Gene Causing Diabetes and Hereditary Pancreatitis: How to Correctly Identify MODY8 Cases, *The Journal of clinical endocrinology and metabolism*, *107*(4), E1455–E1466, doi:10.1210/CLINEM/DGAB864, 2022.

Jo, S., T. Kim, V. G. Iyer, and W. Im, CHARMM-GUI: A web-based graphical user interface for CHARMM, *Journal of Computational Chemistry*, *29*(11), 1859–1865, doi:10.1002/JCC.20945, 2008.

Jo, S., K. C. Song, H. Desaire, A. D. MacKerell, and W. Im, Glycan reader: Automated sugar identification and simulation preparation for carbohydrates and glycoproteins, *Journal of Computational Chemistry*, *32*(14), 3135–3141, doi:10.1002/JCC.21886, 2011.

Johansson, B. B., et al., Diabetes and pancreatic exocrine dysfunction due to mutations in the carboxyl ester lipase gene-maturity onset diabetes of the young (CEL-MODY): A protein misfolding disease, *Journal of Biological Chemistry*, *286*(40), 34,593–34,605, doi:10.1074/JBC.M111.222679, 2011.

Johansson, B. B., et al., The role of the carboxyl ester lipase (CEL) gene in pancreatic disease, *Pancreatology*, *18*(1), 12–19, doi:10.1016/J.PAN.2017.12.001, 2018.

Kabsch, W., and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, *22*(12), 2577–2637, doi:10.1002/BIP.360221211, 1983.

Kadonaga, J. T., K. R. Carner, F. R. Masiarz, and R. Tjian, Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain, *Cell*, *51*(6), 1079–1090, doi:10.1016/0092-8674(87)90594-0, 1987.

Kahraman, S., et al., Abnormal exocrine–endocrine cell cross-talk promotes β-cell dysfunction and loss in MODY8, *Nature Metabolism 2022 4:1*, *4*(1), 76–89, doi:10.1038/s42255-021-00516-2, 2022.

Kannius-Janson, M., U. Lidberg, K. Hultén, A. Gritli-Linde, G. Bjursell, and J. Nilsson, Studies of the regulation of the mouse carboxyl ester lipase gene in mammary gland, *Biochemical Journal*, *336*(3), 577–585, doi:10.1042/BJ3360577, 1998.

Kannius-Janson, M., U. Lidberg, G. Bjursell, and J. Nilsson, The tissue-specific regulation of the carboxyl ester lipase gene in exocrine pancreas differs significantly between mouse and human., *Biochemical Journal*, *351*(Pt 2), 367, doi:10.1042/0264-6021:3510367, 2000.

Kumar, V. B., T. Sasser, J. B. Mandava, A. Sadi, and C. Spilburg, Identification of 5′ flanking sequences that affect human pancreatic cholesterol esterase gene expression, *Biochem Cell Biol.*, *75*(3), 247–254, doi:https://doi.org/10.1139/bcb-75-3-247, 1997.

Kyte, J., and R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, *Journal of molecular biology*, *157*(1), 105–132, doi:10.1016/0022-2836(82)90515-0, 1982.

Leach, A. R., *Molecular Modelling Principles and Applications*, 2nd ed., 354 pp., Pearson Education Limited, 2001.

Lee, J., M. Hitzenberger, M. Rieger, N. R. Kern, M. Zacharias, and W. Im, CHARMM-GUI supports the Amber force fields, *Journal of Chemical Physics*, *153*(3), 35,103, doi:10.1063/5.0012280, 2020.

Lee, J., et al., CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field, *Journal of Chemical Theory and Computation*, *12*(1), 405–413, doi:10.1021/ACS.JCTC.5B00935, 2016.

Letunic, I., and P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation, *Nucleic acids research*, *49*(W1), W293–W296, doi:10.1093/NAR/GKAB301, 2021.

Lidberg, U., J. Nilsson, K. Strömberg, G. Stenman, P. Sahlin, S. Enerbäck, and G. Bjursell, Genomic organization, sequence analysis, and chromosomal localization of the human car□yl ester lipase (CEL) gene and a CEL-like (CELL) gene, *Genomics*, *13*(3), 630–640, doi:10.1016/0888-7543(92)90134-E, 1992.

Livingstone, C. D., and G. J. Barton, Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation, *Computer applications in the biosciences : CABIOS*, *9*(6), 745–756, doi:10.1093/BIOINFORMATICS/9.6.745, 1993.

Lombardo, D., and O. Guy, Studies on the substrate specificity of a carboxyl ester hydrolase from human pancreatic juice. II. Action on cholesterol esters and lipid-soluble vitamin esters, *Biochimica et Biophysica Acta (BBA) - Enzymology*, *611*(1), 147–155, doi:10.1016/0005-2744(80)90050-9, 1980.

Lombardo, D., O. Guy, and C. Figarella, Purification and characterization of a carboxyl ester hydrolase from human pancreatic juice, *Biochimica et Biophysica Acta (BBA) - Enzymology*, *527*(1), 142–149, doi:10.1016/0005-2744(78)90263-2, 1978.

Lombardo, D., F. Silvy, I. Crenon, E. Martinez, A. Collignon, E. Beraud, and E. Mas, Pancreatic adenocarcinoma, chronic pancreatitis, and MODY-8 diabetes: is bile salt-dependent lipase (or carboxyl ester lipase) at the crossroads of pancreatic pathologies?, *Oncotarget*, *9*(15), 12,513, doi:10.18632/ONCOTARGET.23619, 2018.

Loomes, K. M., Structural Organisation of Human Bile-salt-activated Lipase Probed by Limited Proteolysis and Expression of a Recombinant Truncated Variant, *Article in European Journal of Biochemistry*, 1995.

Loomes, K. M., and H. E. Senior, Bile salt activation of human cholesterol esterase does not require protein dimerisation, *FEBS Letters*, *405*(3), 369–372, doi:10.1016/S0014-5793(97)00215-9, 1997.

López-Maury, L., S. Marguerat, and J. Bähler, Tuning gene expression to changing environments:

from rapid responses to evolutionary adaptation, *Nature Reviews Genetics 2008 9:8*, *9*(8), 583–593, doi:10.1038/nrg2398, 2008.

Luchini, C., et al., Inflammatory and tumor-like lesions of the pancreas:, *Pathologica - Journal of the Italian Society of Anatomic Pathology and Diagnostic Cytopathology*, *112*(3), 197–209, doi:10.32074/1591-951X-168, 2020.

Mackay, K., and R. M. Lawn, Characterization of the mouse pancreatic/mammary gland cholesterol esterase-encoding cDNA and gene, *Gene*, *165*(2), 255–259, doi:10.1016/0378-1119(95)00564-M, 1995.

Madeyski, K., U. Lidberg, G. Bjursell, and J. Nilsson, Characterization of the gorilla carboxyl ester lipase locus, and the appearance of the carboxyl ester lipase pseudogene during primate evolution, *Gene*, *239*(2), 273–282, doi:10.1016/S0378-1119(99)00410-2, 1999.

Maldonado-Valderrama, J., P. Wilde, A. Maclerzanka, and A. MacKie, The role of bile salts in digestion, *Advances in Colloid and Interface Science*, *165*(1), 36–46, doi:10.1016/J.CIS.2010.12.002, 2011.

Mao, X. T., S. J. Deng, R. L. Kang, Y. C. Wang, Z. S. Li, W. B. Zou, and Z. Liao, Homozygosity of short VNTR lengths in the CEL gene may confer susceptibility to idiopathic chronic pancreatitis, *Pancreatology*, *21*(7), 1311–1316, doi:10.1016/J.PAN.2021.09.001, 2021.

Mao, X. T., et al., The CEL-HYB1 Hybrid Allele Promotes Digestive Enzyme Misfolding and Pancreatitis in Mice, *Cellular and Molecular Gastroenterology and Hepatology*, *14*(1), 55–74, doi:10.1016/J.JCMGH.2022.03.013, 2022.

McNaught, A. D., Nomenclature of carbohydrates (IUPAC recommendations 1996), *Pure and Applied Chemistry*, *68*(10), 1919–2008, doi:10.1351/PAC199668101919, 1996.

Mensah, M. A., et al., Aberrant phase separation and nucleolar dysfunction in rare genetic diseases, *Nature*, *614*(7948), 564–571, doi:10.1038/s41586-022-05682-1, 2023.

Mészáros, B., G. Erdös, and Z. Dosztányi, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, *Nucleic Acids Research*, *46*(W1), W329–W337, doi:10.1093/NAR/GKY384, 2018.

Mirdita, M., K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, ColabFold: making protein folding accessible to all, *Nature Methods 2022 19:6*, *19*(6), 679–682, doi:10.1038/s41592-022-01488-1, 2022.

Morgan, R. G., J. Barrowman, H. Filipek-Wender, and B. Borgström, The lipolytic enzymes of rat pancreatic juice, *Biochimica et Biophysica Acta (BBA) - Enzymology*, *167*(2), 355–366, doi:10.1016/0005-2744(68)90214-3, 1968.

NCBI, CEL carboxyl ester lipase [Homo sapiens (human)] - Gene, [Accessed: 2023-02-12], 2023a.

NCBI, CELP carboxyl ester lipase pseudogene [Homo sapiens (human)] - Gene, [Accessed: 2023-02-17], 2023b.

Nilsson, J., L. Bläckberg, P. Carlsson, S. Enerbäck, O. Hernell, and G. Bjursell, cDNA cloning of human-milk bile-salt-stimulated lipase and evidence for its identity to pancreatic carboxylic ester hydrolase, *European Journal of Biochemistry*, *192*(2), 543–550, doi:10.1111/J.1432-1033.1990.TB19259.X, 1990.

Nyberg, L., A. Farooqi, L. Bläckberg, R. D. Duan, Å. Nilsson, and O. Hernell, Digestion of ceramide by human milk bile salt-stimulated lipase, *Journal of Pediatric Gastroenterology and Nutrition - Jpgn*, *27*(5), 560–567, doi:10.1097/00005176-199811000-00013, 1998.

Ollis, D. L., et al., The α/β hydrolase fold, *Protein Engineering*, *5*(3), 197–211, doi:10.1093/protein/5.3.197, 1992.

Olsson, M. H., C. R. Søndergaard, M. Rostkowski, and J. H. Jensen, PROPKA3: Consistent treatment of internal and surface residues in empirical p K a predictions, *Journal of Chemical Theory and Computation*, *7*(2), 525–537, doi:10.1021/CT100578Z, 2011.

Paetzel, M., A. Karla, N. C. Strynadka, and R. E. Dalbey, Signal peptidases, *Chemical reviews*, *102*(12), 4549–4579, doi:10.1021/CR010166Y, 2002.

Park, S. J., J. Lee, D. S. Patel, H. Ma, H. S. Lee, S. Jo, and W. Im, Glycan Reader is improved to recognize most sugar types and chemical modifications in the Protein Data Bank, *Bioinformatics*, *33*(19), 3051–3057, doi:10.1093/BIOINFORMATICS/BTX358, 2017.

Park, S. J., et al., CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates, *Glycobiology*, *29*(4), 320–331, doi:10.1093/GLYCOB/CWZ003, 2019.

Pasqualini, E., N. Caillol, E. Mas, N. Bruneau, D. Lexa, and D. Lombardo, Association of bile-salt-dependent lipase with membranes of human pancreatic microsomes is under the control of ATP and phosphorylation., *Biochemical Journal*, *327*(Pt 2), 527, doi:10.1042/BJ3270527, 1997.

Pedretti, A., A. Mazzolari, S. Gervasoni, L. Fumagalli, and G. Vistoli, The VEGA suite of programs: an versatile platform for cheminformatics and drug design projects, *Bioinformatics*, *37*(8), 1174–1175, doi:10.1093/BIOINFORMATICS/BTAA774, 2021.

Pellegrini, S., et al., Generation of β Cells from iPSC of a MODY8 Patient with a Novel Mutation in the Carboxyl Ester Lipase (CEL) Gene, *The Journal of Clinical Endocrinology & Metabolism*, *106*(5), e2322–e2333, doi:10.1210/CLINEM/DGAA986, 2021.

Ræder, H., et al., Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction, *Nature Genetics 2006 38:1*, *38*(1), 54–62, doi:10.1038/ng1708, 2006.

Ramji, D. P., and P. Foka, CCAAT/enhancer-binding proteins: structure, function and regulation. Biochem J 365:561-575 CCAAT/enhancer-binding proteins : structure, function and regulation, *Biochemical Journal*, *365*, 561–575, doi:10.1042/BJ20020508, 2002.

Rechsteiner, M., and S. W. Rogers, PEST sequences and regulation by proteolysis, *Trends in Biochemical Sciences*, *21*(7), 267–271, doi:10.1016/S0968-0004(96)10031-1, 1996.

Reue, K., et al., cDNA cloning of carboxyl ester lipase from human pancreas reveals a unique proline-rich repeat unit, *Journal of Lipid Research*, *32*(2), 267–276, doi:10.1016/s0022-2275(20)42088-7, 1991.

Rice, P., L. Longden, and A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite, *Trends in Genetics*, *16*(6), 276–277, doi:10.1016/S0168-9525(00)02024-2, 2000.

Robustelli, P., S. Piana, and D. E. Shaw, Developing a molecular dynamics force field for both folded and disordered protein states, *Proceedings of the National Academy of Sciences of the United States of America*, *115*(21), E4758–E4766, doi:10.1073/pnas.1800690115, 2018.

Rueden, C. T., J. Schindelin, M. C. Hiner, B. E. DeZonia, A. E. Walter, E. T. Arena, and K. W. Eliceiri, ImageJ2: ImageJ for the next generation of scientific image data, *BMC Bioinformatics*, *18*(1), 1–26, doi:10.1186/S12859-017-1934-Z, 2017.

Schindelin, J., et al., Fiji: an open-source platform for biological-image analysis, *Nature Methods 2012 9:7*, *9*(7), 676–682, doi:10.1038/nmeth.2019, 2012.

Schoch, C. L., et al., NCBI Taxonomy: a comprehensive update on curation, resources and tools, *Database*, *2020*, doi:10.1093/DATABASE/BAAA062, 2020.

Schramm, A., C. Bignon, S. Brocca, R. Grandori, C. Santambrogio, and S. Longhi, An arsenal of methods for the experimental characterization of intrinsically disordered proteins–how to choose and combine them?, *Archives of biochemistry and biophysics*, *676*, 108,055, doi:10.1016/j.abb.2019.07.020, 2019.

Schwitzgebel, V. M., Many faces of monogenic diabetes, *J Diabetes Invest*, *5*, 121–133, doi:10.1111/jdi.12197, 2014.

Shin, Y., and C. P. Brangwynne, Liquid phase condensation in cell physiology and disease, *Science*, *357*(6357), doi:10.1126/SCIENCE.AAF4382, 2017.

Sievers, F., and D. G. Higgins, Clustal Omega for making accurate alignments of many protein sequences, *Protein Science*, *27*(1), 135–145, doi:10.1002/PRO.3290, 2018.

Sievers, F., et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Molecular systems biology*, *7*, doi:10.1038/MSB.2011.75, 2011.

Søndergaard, C. R., M. H. Olsson, M. Rostkowski, and J. H. Jensen, Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values, *Journal of Chemical Theory and Computation*, *7*(7), 2284–2295, doi:10.1021/CT200133Y, 2011.

Strous, G. J., and J. Dekker, Mucin-Type Glycoproteins, *Critical Reviews in Biochemistry and Molecular Biology*, *27*(1-2), 57–92, doi:10.3109/10409239209082559, 1992.

Taylor, A. K., J. L. Zambaux, I. Klisak, T. Mohandas, R. S. Sparkes, M. C. Schotz, and A. J. Lusis, Carboxyl ester lipase: A highly polymorphic locus on human chromosome 9qter, *Genomics*, *10*(2), 425–431, doi:10.1016/0888-7543(91)90328-C, 1991.

Terzyan, S., C.-S. Wand, D. Downs, B. Hunter, and X. C. Zhang, Crystal structure of the catalytic domain of human bile salt activated lipase, *Protein Science*, *9*(9), 1783–1790, doi:10.1110/PS.9.9.1783, 2000.

Tjora, E., A. Gravdal, M. Cnop, T. Engjom, B. B. Johansson, G. G. Dimcevski, A. Molven, and K. Fjeld, Protein misfolding in combination with other risk factors in CEL-HYB1-mediated chronic pancreatitis, *European Journal of Gastroenterology & Hepatology*, *41*(11), 1–5, doi:10.1097/MEG.0000000000001963, 2020.

Tompa, P., Intrinsically unstructured proteins evolve by repeat expansion, *BioEssays*, *25*(9), 847–855, doi:10.1002/BIES.10324, 2003.

Torsvik, J., et al., Mutations in the VNTR of the carboxyl-ester lipase gene (CEL) are a rare cause of monogenic diabetes, *Human Genetics*, *127*(1), 55–64, doi:10.1007/S00439-009-0740-8, 2010.

Torsvik, J., et al., Endocytosis of Secreted Carboxyl Ester Lipase in a Syndrome of Diabetes and Pancreatic Exocrine Dysfunction, *Journal of Biological Chemistry*, *289*(42), 29,097–29,111, doi:10.1074/jbc.M114.574244, 2014.

Touvrey, C., C. Courageux, V. Guillon, R. Terreux, F. Nachon, and X. Brazzolotto, X-ray structures of human bile-salt activated lipase conjugated to nerve agents surrogates, *Toxicology*, *411*, 15–23, doi:10.1016/j.tox.2018.10.015, 2019.

Touw, W. G., C. Baakman, J. Black, T. A. Te Beek, E. Krieger, R. P. Joosten, and G. Vriend, A series of PDB-related databanks for everyday needs, *Nucleic Acids Research*, *43*(D1), D364–D368, doi:10.1093/NAR/GKU1028, 2015.

Varki, A., et al., Symbol nomenclature for graphical representations of glycans, *Glycobiology*, *25*(12), 1323–1324, doi:10.1093/GLYCOB/CWV091, 2015.

Verine, A., J. Le Petit-Thevenin, L. Panicot-Dubois, A. Valette, and D. Lombardo, Phosphorylation of the Oncofetal Variant of the Human Bile Salt-dependent Lipase, *Journal of Biological Chemistry*, *276*(15), 12,356–12,361, doi:10.1074/jbc.m008658200, 2001.

Walkowska, J., N. Zielinska, P. Karauda, R. S. Tubbs, K. Kurtys, and Ł. Olewnik, The Pancreas and Known Factors of Acute Pancreatitis, *Journal of Clinical Medicine 2022, Vol. 11, Page 5565*, *11*(19), 5565, doi:10.3390/JCM11195565, 2022.

Wang, C. S., and J. A. Hartsuck, Bile salt-activated lipase. A multiple function lipolytic enzyme, *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism*, *1166*(1), 1–19, doi:10.1016/0005-2760(93)90277-G, 1993.

Wang, C. S., M. E. Martindale, M. M. King, and J. Tang, Bile-salt-activated lipase: effect on kitten growth rate, *The American Journal of Clinical Nutrition*, *49*(3), 457–463, doi:10.1093/AJCN/49.3.457, 1989.

Wang, W., Recent advances in atomic molecular dynamics simulation of intrinsically disordered proteins, *Physical Chemistry Chemical Physics*, *23*(2), 777–784, doi:10.1039/D0CP05818A, 2021.

Wang, X., C. S. Wang, J. Tang, F. Dyda, and X. C. Zhang, The crystal structure of bovine bile salt activated lipase: insights into the bile salt activation mechanism, *Structure*, *5*(9), 1209–1218, doi:10.1016/S0969-2126(97)00271-2, 1997.

Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life, *Journal of Molecular Biology*, *337*(3), 635–645, doi:10.1016/J.JMB.2004.02.002, 2004.

Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton, Jalview Version 2—a multiple sequence alignment editor and analysis workbench, *Bioinformatics*, *25*(9), 1189–1191, doi:10.1093/BIOINFORMATICS/BTP033, 2009.

Williams, J. A., Regulation of pancreatic acinar cell function, *Current Opinion in Gastroenterology*, *22*, 498–504, doi:10.1097/01.mog.0000239863.96833.c0, 2006.

Wright, P. E., and H. J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *Journal of Molecular Biology*, *293*(2), 321–331, doi:10.1006/JMBI.1999.3110, 1999.

Wynne, K., B. Devereaux, A. Dornhorst, and C. Katie Wynne, Diabetes of the exocrine pancreas, *Diabetes of the Exocrine pancreas Article in Journal of Gastroenterology and Hepatology*, doi:10.1111/jgh.14451, 2018.

Xiao, X., G. Jones, W. A. Sevilla, D. B. Stolz, K. E. Magee, M. Haughney, A. Mukherjee, Y. Wang, and M. E. Lowe, A Carboxyl Ester Lipase (CEL) Mutant Causes Chronic Pancreatitis by Forming Intracellular Aggregates That Activate Apoptosis, *The Journal of biological chemistry*, *291*(44), 23,224–23,236, doi:10.1074/JBC.M116.734384, 2016.

Yadav, D., and A. B. Lowenfels, The Epidemiology of Pancreatitis and Pancreatic Cancer, *Gastroenterology*, *144*(6), 1252–1261, doi:10.1053/J.GASTRO.2013.01.068, 2013.

Zachos, F. E., *Mammalian Phylogenetics: A Short Overview of Recent Advances*, 1–18 pp., Springer, Cham, doi:10.1007/978-3-319-65038-8_6-1, 2020.

Zou, W. B., et al., No Association Between CEL–HYB Hybrid Allele and Chronic Pancreatitis in Asian Populations, *Gastroenterology*, *150*(7), 1558–1560.e5, doi:10.1053/J.GASTRO.2016.02.071, 2016.

11 Appendix

The following supplemental files will be available along with the thesis:

- **S1**: The vertebrate, non-mammalian dataset
- **S2**: The mammalian dataset with annotation data
- **S3**: The full VNTR annotations of the 156 mammalian sequences
- **S4**: The full secondary structure plots for all the simulations

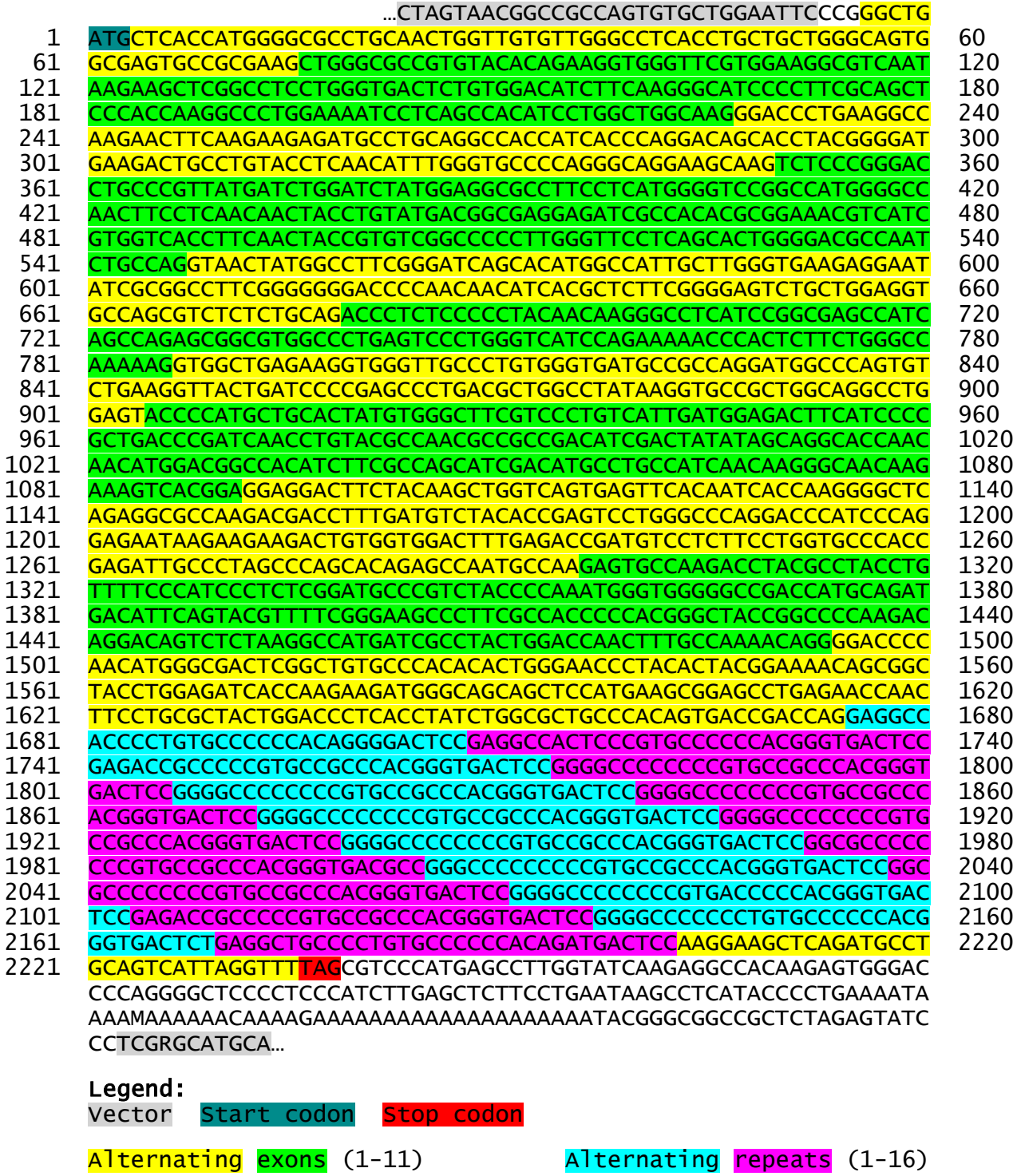


Figure 11.1: **Sequencing of CEL-WT inserted in vector pcDNA3.** Nucleotides in open reading-frame are numerated. Preceding and succeeding non-coding nucleotides are not numerated. Sanger sequencing performed with instrument ABI 3500xL (Applied Biosystems).

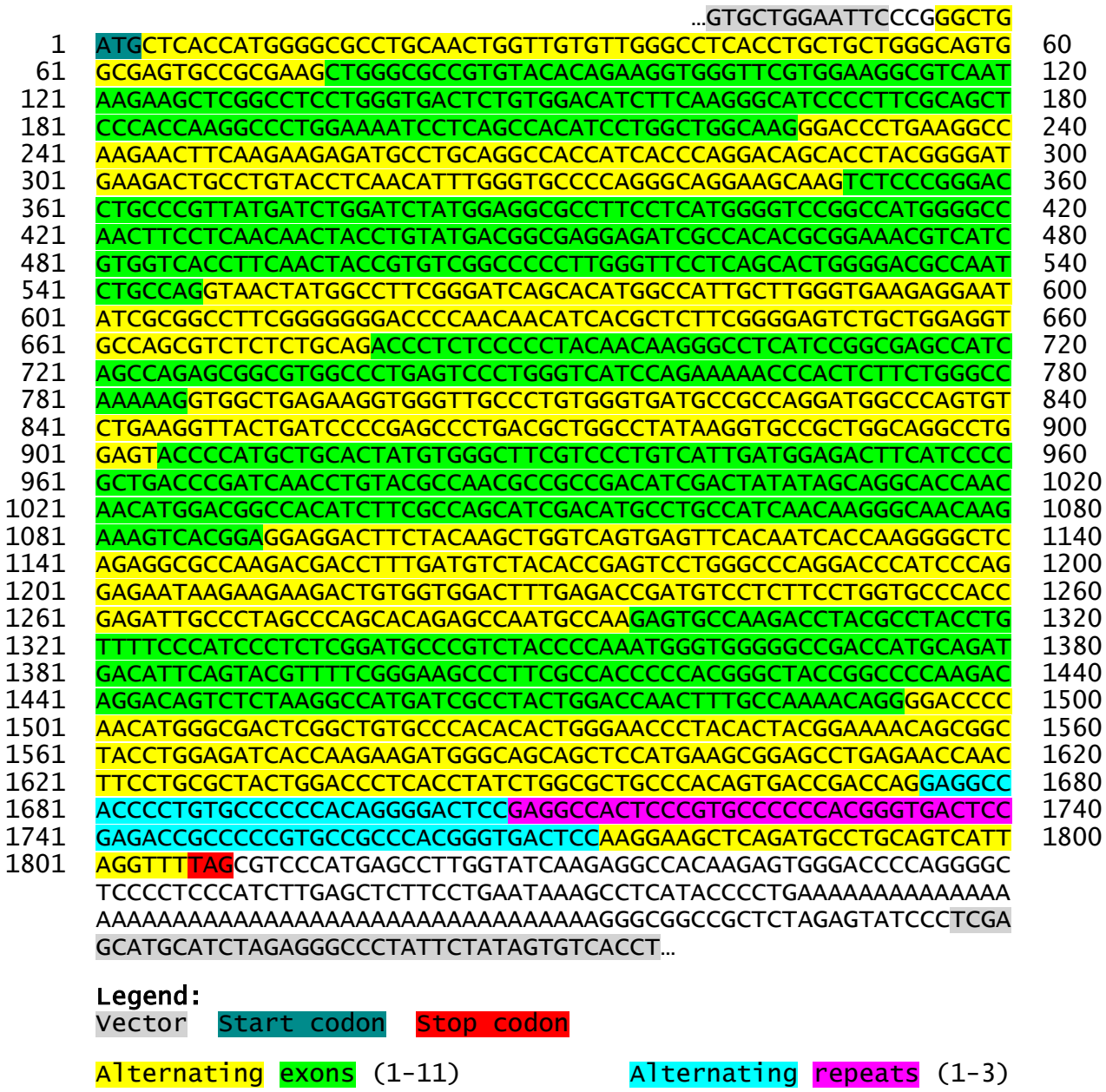


Figure 11.2: **Sequencing of CEL-3R-USA inserted in vector pcDNA3.** Nucleotides in open reading-frame are numerated. Preceding and succeeding non-coding nucleotides are not numerated. Sanger sequencing performed with instrument ABI 3500xL (Applied Biosystems).

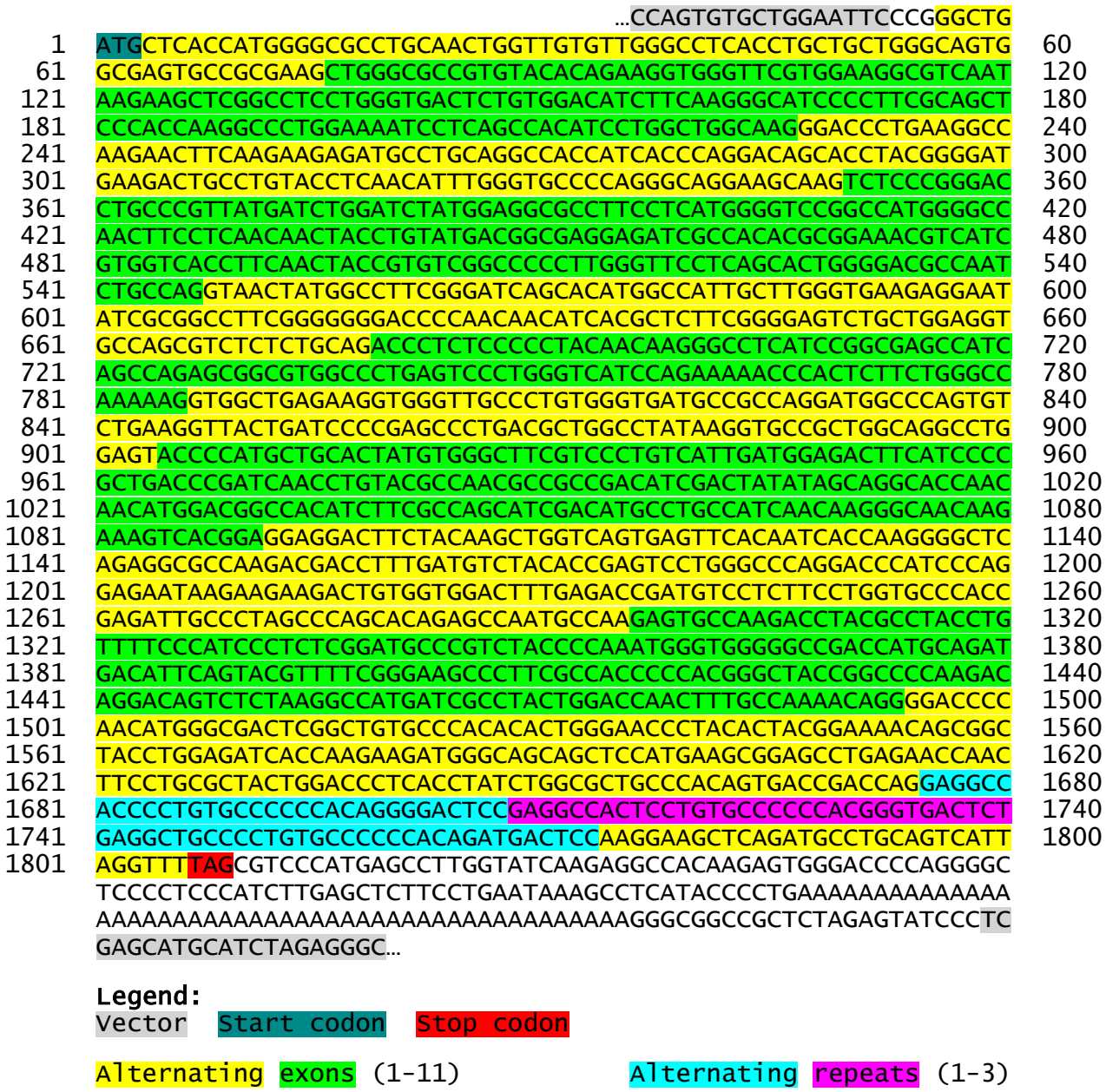


Figure 11.3: **Sequencing of CEL-3R-DAN inserted in vector pcDNA3.** Nucleotides in open reading-frame are numerated. Preceding and succeeding non-coding nucleotides are not numerated. Sanger sequencing performed with instrument ABI 3500xL (Applied Biosystems).

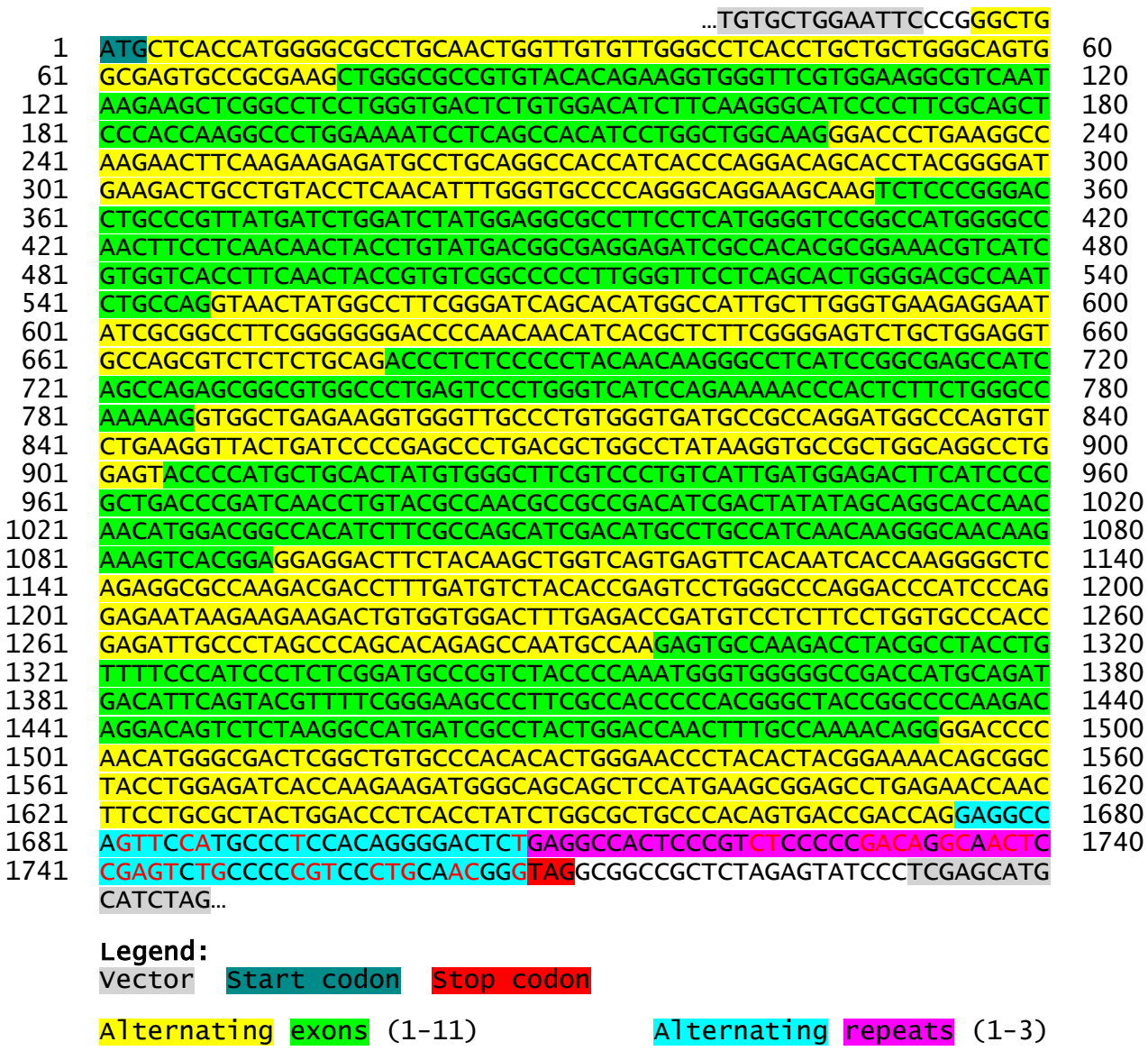


Figure 11.4: **Sequencing of CEL-HYB inserted in vector pcDNA3.** Red coloured text: VNTR nucleotides that are different from CEL-WT. Nucleotides in open reading-frame are numerated. Preceding and succeeding non-coding nucleotides are not numerated. Sanger sequencing performed with instrument ABI 3500xL (Applied Biosystems).

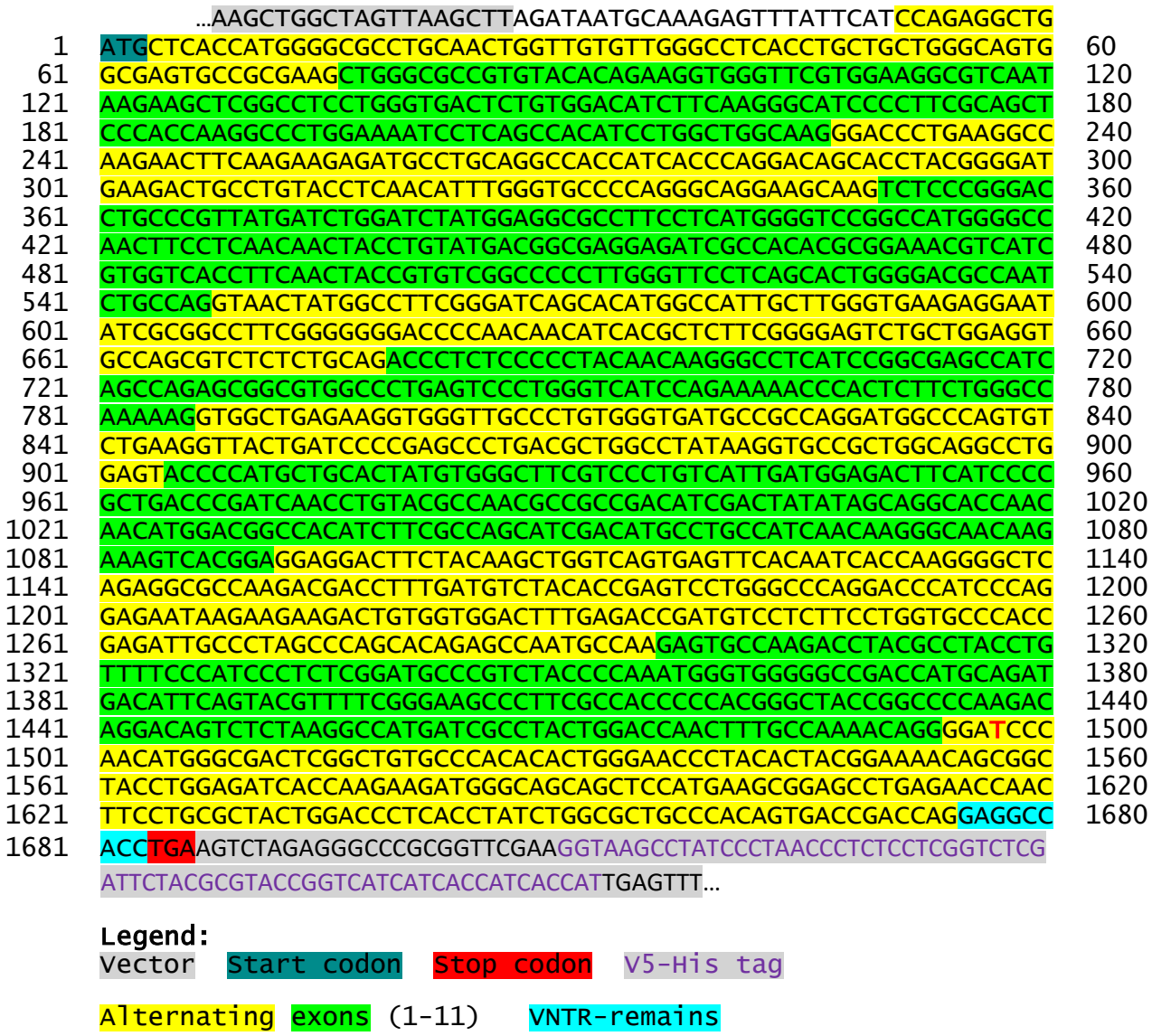


Figure 11.5: **Sequencing of CEL-TRUNC inserted in vector pcDNA3.1/V5-His B.** Red coloured text; T1497 was the only pre-VNTR nucleotide not matching with the other CEL variants. Nucleotides in open reading-frame are numerated. Preceding and succeeding non-coding nucleotides are not numerated. Sanger sequencing performed with instrument ABI 3500xL (Applied Biosystems).

…TATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCC
CCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTA
TGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATG
CGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGT
CTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCA
AAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAG
GTCTATATAAGCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAA
ATTAATACGACTCACTATAGGGAGACCCAAGCTTGGTACCGAGCTCGGATCCACTAGTAA
CGGCCGCCAGTGTGCTGGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATG
CATCTAGAGGGCCCTATTCTATAGTGTCACCTAAATGCTAGAGCTCGCTGATCAGCCTCG
ACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACC
CTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGT
CTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGAT
TGGGAAGACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGGCTTCTGAGGCGGAA
AGAACCAGCTGGGGCTCTAGGGGGTATCCCCACGCGCCCTGTAGCGGCGCATTAAGCGCG
GCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCT
CCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTA
AATC…

Legend:

Vector

T7 forward primer 5'- TAATACGACTCACTATAGGG

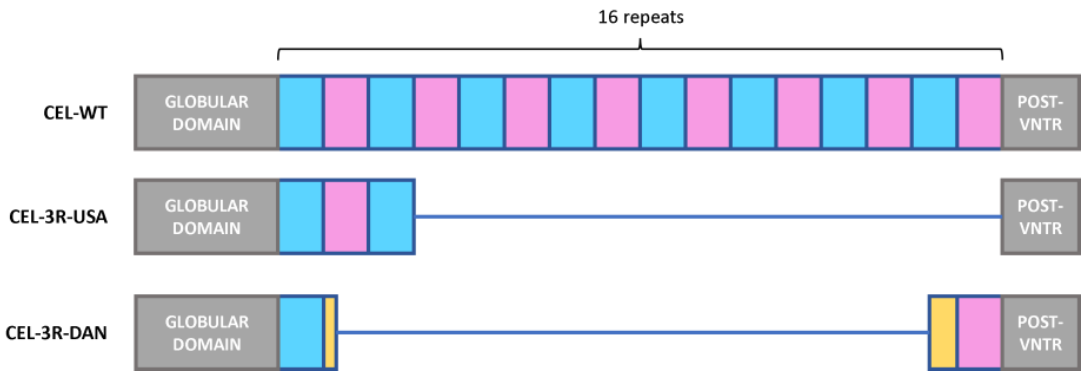
BGH reverse primer 5'- TAGAAGGCACAGTCGAGG

EcoR1 restriction site G|AATTC

XhoI1 restriction site C|TCGAG

Figure 11.6: **Sequencing of empty vector pcDNA3 insertion site.** Sequence is annotated for primer binding sites and restriction sites for restriction-based insertions. Sanger sequencing performed with instrument ABI 3500xL (Applied Biosystems).

A



B

```
CEL-WT         1 GAGGCCACCCCTGTGCCCCCCACAGGGGACTCCGAGGCCACTCCCGTGCCCCCCACGGGT 60
CEL-3R-USA     1 GAGGCCACCCCTGTGCCCCCCACAGGGGACTCCGAGGCCACTCCCGTGCCCCCCACGGGT 60
CEL-3R-DAN     1 GAGGCCACCCCTGTGCCCCCCACAGGGGACTCCGAGGCCACT---------------- 42

CEL-WT        61 GACTCCGAGACCGCCCCCGTGCCGCCCACGGGTGACTCCGGGGCCCCCCCCCGTGCCGCCC 120
CEL-3R-USA    61 GACTCCGAGACCGCCCCCGTGCCGCCCACGGGTGACTCC---------------------- 99
CEL-3R-DAN        ------------------------------------------------------------

CEL-WT       121 ACGGGTGACTCCGGGGCCCCCCCCCGTGCCGCCCACGGGTGACTCCGGGGCCCCCCCCGTG 180
CEL-3R-USA        ------------------------------------------------------------
CEL-3R-DAN        ------------------------------------------------------------

CEL-WT       181 CCGCCCACGGGTGACTCCGGGGCCCCCCCCGTGCCGCCCACGGGTGACTCCGGGGCCCCC 240
CEL-3R-USA        ------------------------------------------------------------
CEL-3R-DAN        ------------------------------------------------------------

CEL-WT       241 CCCGTGCCGCCCACGGGTGACTCCGGGGCCCCCCCCGTGCCGCCCACGGGTGACTCCGGC 300
CEL-3R-USA        ------------------------------------------------------------
CEL-3R-DAN        ------------------------------------------------------------

CEL-WT       301 GCCCCCCCCGTGCCGCCCACGGGTGACGCCGGGCCCCCCCCCGTGCCGCCCACGGGTGAC 360
CEL-3R-USA        ------------------------------------------------------------
CEL-3R-DAN        ------------------------------------------------------------

CEL-WT       361 TCCGGCGCCCCCCCCGTGCCGCCCACGGGTGACTCCGGGGCCCCCCCCGTGACCCCCACG 420
CEL-3R-USA        ------------------------------------------------------------
CEL-3R-DAN        ------------------------------------------------------------

CEL-WT       421 GGTGACTCCGAGACCGCCCCCGTGCCGCCCACGGGTGACTCCGGGGCCCCCCCTGTGCCC 480
CEL-3R-USA        ------------------------------------------------------------
CEL-3R-DAN    43 ----------------------------------------------------CCTGTGCCC  51

CEL-WT       481 CCCACGGGTGACTCTGAGGCTGCCCCTGTGCCCCCCACAGATGACTCC 528
CEL-3R-USA       ------------------------------------------------
CEL-3R-DAN    52 CCCACGGGTGACTCTGAGGCTGCCCCTGTGCCCCCCACAGATGACTCC  99

                 Legend:
                 CEL-WT:       Alternating repeats (1-16)

                 CEL-3R-USA:   Repeat 1  Repeat 2  Repeat 3

                 CEL-3R-DAN:   Repeat 1  Repeat 2  Repeat 3
```

Figure 11.7: **Alignments of *CEL-3R-USA* and *CEL-3R-DAN* against reference *CEL-WT*.** Repeats of *CEL-3R-USA* and *CEL-3R-DAN* are aligned with matching repeats in *CEL-WT*. **A.** Illustration of coloured boxes which represent the individual repeats. Globular domain and post-VNTR are not to scale. **B.** Nucleotide alignment. Sanger sequencing performed with instrument ABI 3500xL. Pairwise alignment was performed manually.

Table 11.1: **Strings for filtering BLASTP hits.** BLASTP hits with description entries containing any of these strings were removed from the dataset. Filtrations were case-insensitive.

Filtered strings
[Bos indicus x Bos taurus]
bche
carboxylester
cholin
crystal
fatty acyl
hcg
hypothetic
like
lipase [Larimichthys
loc4
lysopho
mkiaa
neuro
ngln
n-myc
partial
polym
predict
pyreth
tandem
truncated
uncharac
unnamed

Table 11.2: **Runtimes for the performed MD simulations of h3R-DAN-U and h3R-DAN-G.** Column 'Runtime (min)' displays the total number of minutes it took to compute each simulation. Column 'Sim. time (ns)' shows the total simulated time. The last column exhibits the ratio of runtime and simulated time. In other words, the ratio tells how many minutes the program had to run in real time to produce 1 ns of simulated time.

	Runtime (min)	Sim. time (ns)	Runtime/Sim. time (min/ns)
h3R-DAN-U			
Large	27469.4	955	28.8
Replica 1	4205.2	500	8.4
Replica 2	4059.8	500	8.1
Replica 3	4079.6	500	8.2
h3R-DAN-G			
Large	10585.3	500	21.2
Replica 1	4033.8	500	8.1
Replica 2	4212.7	500	8.4
Replica 3	4217.8	500	8.4

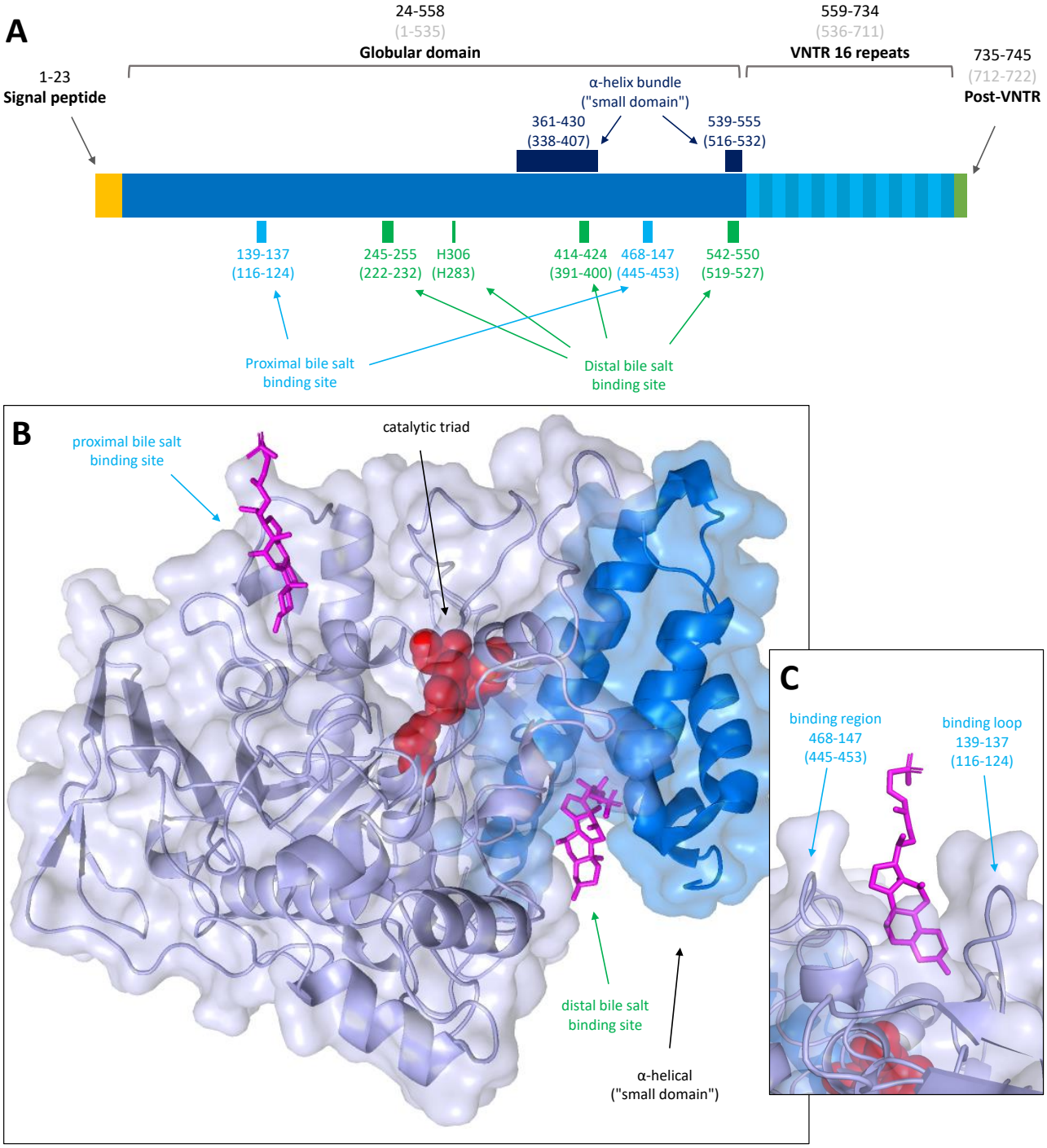
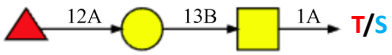


Figure 11.8: **Bile salt binding sites in the CEL protein.** **A.** Diagram showing approximate bile salt binding sites for a human* CEL-variant with 16 repeats. Numbers denote the amino acid positions for the fully intact CEL protein (numbers in parentheses denote positions for without signal peptide). Coloured boxes indicate approximate regions which make up the α-helical bundle (or ”small domain”) and bile salt binding sites. **B.** X-ray crystallography of bovine CEL globular domain bound to taurocholate (PDB accession: 1AQL). Annotated by colour: taurocholate is purple, atoms of the catalytic triad are red, α-helical bundle is blue. Visualised in PyMOL. **C.** Close-up of taurocholate in proximal bile salt binding site. The taurocholate is nestled up against residues 468-147 and 139-137. (Own illustration).

* Bile salt-binding annotations were based on bovine CEL. See X. Wang, C. S. Wang, J. Tang, F. Dyda, and X. C. Zhang. The crystal structure of bovine bile salt activated lipase: insights into the bile salt activation mechanism. *Structure*, 5(9):1209–1218, sep 1997. ISSN 0969-2126. doi:10.1016/S0969-2126(97)00271-2.

glycan A (0.512 kDa)



IUPAC:

aLFuc(1→2)bDGal(1→3)aDGalNAc(1→)PROA

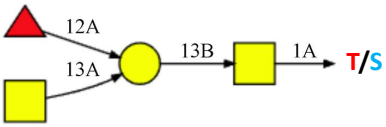
GRS:

1 AGALNA

2 - 13B: BGAL

3 - - 12A: AFUC

glycan B (0.716 kDa)



IUPAC:

aLFuc(1→2)[aDGalNAc(1→3)]bDGal(1→3)aDGalNAc(1→)PRO

GRS:

1 AGALNA

2 - 13B: BGAL

3 - - 12A: AFUC

4 - - 13A: AGALNA

Figure 11.9: **Determining the *O*-glycosylations for simulation of the CEL-3R-DAN VNTR. A** The two glycans prepared used for the *O*-glycosylations of h3R-DAN-G. Visualised in SNFG format (*Varki et al.*, 2015). Glycan structures are also represented in the text formats IUPAC extended 2-Carb (*McNaught*, 1996) and GRS (*Park et al.*, 2017). See 'References' for full reference information.