

MSC

2.º
CICLO

FCUP
2018

U.PORTO

Seeking species specific compensatory episodes in
Mus musculus a structural and molecular approach

Íris Raquel Teixeira Sereno

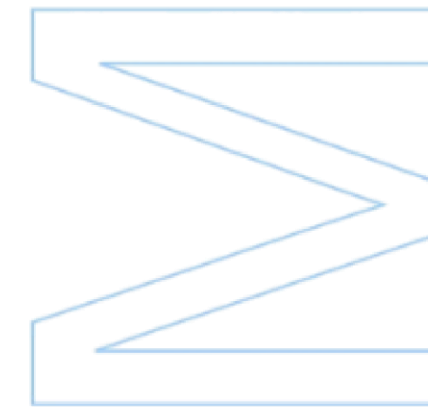
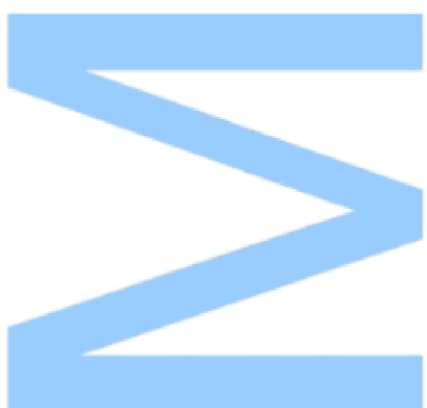
FC

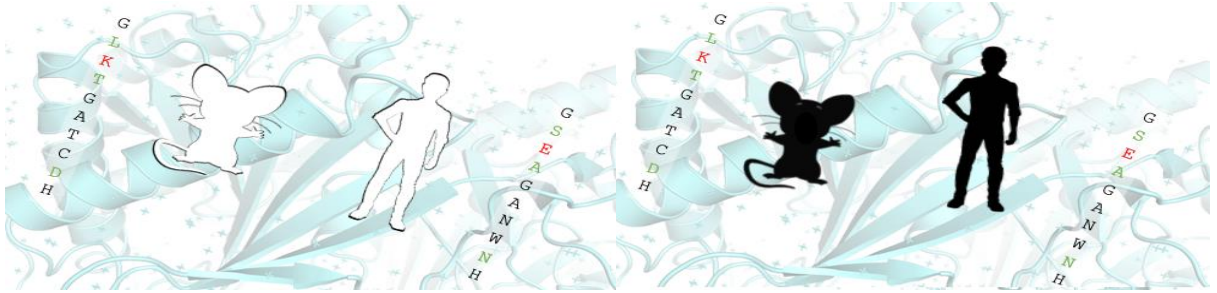


Seeking species-specific compensatory episodes in *Mus musculus*

s structural and molecular approach

Íris Raquel Teixeira Sereno
Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Mestrado em Genética Forense
2018





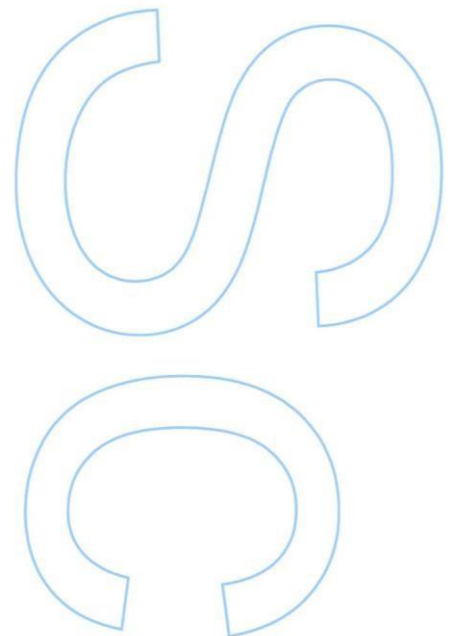
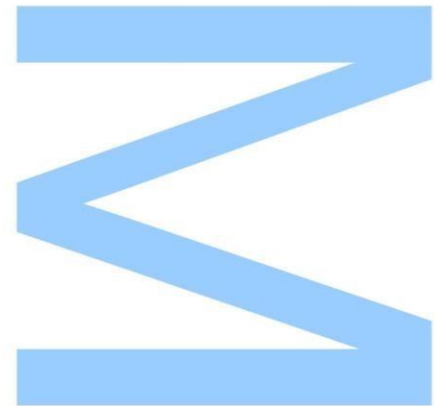
Seeking species-specific compensatory episodes in *Mus musculus*: a structural and molecular approach

Íris Raquel Teixeira Sereno

Dissertação no Âmbito do Mestrado em Genética Forense
Faculdade de Ciências da Universidade do Porto, Departamento de Biologia
2018

Orientador

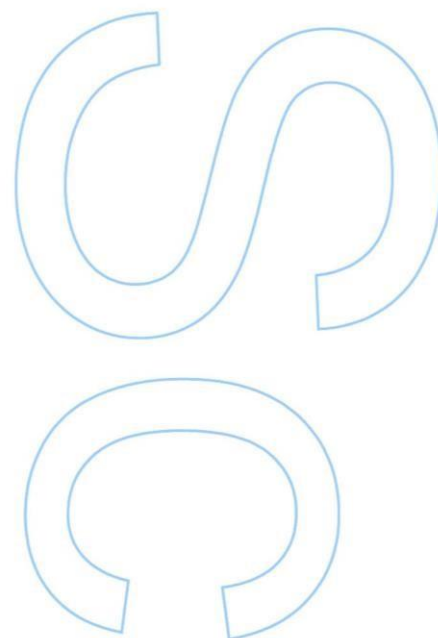
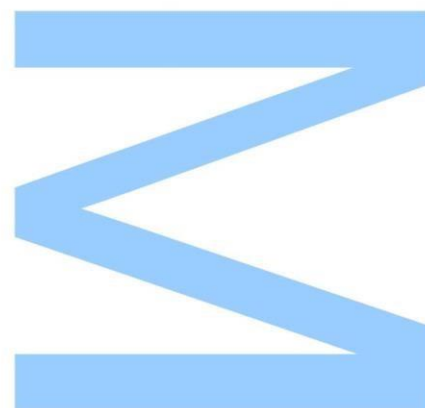
Luísa Cristina da Costa Azevedo



Todas as correções
determinadas pelo
júri, e só essas,
foram efetuadas.

O Presidente do Júri,

Porto,
____/____/____



“Never stop fighting until you arrive at your destined place (...)

Have an aim in life, continuously acquire knowledge,

work hard and have perseverance.”

A. P. J. Abdul Kalam

Agradecimentos

Neste momento que marca o culminar de mais um patamar atingido numa longa escadaria, não poderia deixar de expressar a minha enorme gratidão a todos os que tornaram possível atingir esta etapa.

Em primeiro lugar, gostaria de me dirigir à Professora Doutora Luísa Azevedo, minha orientadora, pela sua constante disponibilidade, pela transmissão de conhecimentos e pelo acompanhamento, assim como pela incessante boa disposição e entusiasmo, mas também por todas as sugestões e críticas que foram, definitivamente, essenciais para me permitir melhorar como pessoa e profissional. Obrigada pela confiança e pela liberdade de ação, sendo estas decisivas para permitir o crescimento e desenvolvimento pessoal. Muito, muito obrigada.

Ao Professor Doutor António Amorim, na qualidade de diretor do Mestrado em Genética Forense, quero deixar um agradecimento especial por se preocupar com a nossa evolução e progressão e por procurar sempre ajudar nas mais diversas situações.

A toda a equipa do grupo Population Genetics and Evolution, pela forma como nos acolheram e se disponibilizaram para ajudar sempre que necessário, permitindo a existência das melhores condições de trabalho e um ambiente espetacular.

Aos colegas de curso e laboratório pelo espírito de entreajuda e pelos momentos de descontração. Garantidamente os obstáculos que foram surgindo seriam mais difíceis de ultrapassar sem a colaboração de todos. Que estes ideais nos acompanhem sempre e nos permitam aprender constantemente com o auxílio do próximo.

Às BioPrincesas – Ana, Ashley, Lia, Rita, Margarida e Maria João, pelo apoio incondicional, companheirismo e mais importante que tudo, pela amizade e por todas as características que vos tornam únicas. Obrigada meninas pela vossa presença, por todas as partilhas e por terem feito destes anos melhores do que alguma vez poderiam ser.

Aos amigos de sempre – Duarte, Frias, Loureiro e Bessa, por me fazerem ver sempre o lado positivo, mesmo quando ele teima em estar camuflado. Por me ajudarem a atravessar todas as adversidades e por nunca deixarem o meu lado. Obrigada por todos os momentos, por suportarem o “lado *nerd*” e pelo apoio incondicional. Sem a vossa boa disposição o mundo não teria tanta cor. O meu sincero obrigada.

Aos meus pais, por acreditarem em mim, por me incentivarem e por suportarem os momentos “*scientific fun fact*” a que foram sendo sujeitos ao longo dos anos. Obrigada por todos os ensinamentos inculcados, nomeadamente persistência, dedicação, vontade de trabalhar e constante procura por ser e dar o melhor de mim. Espero que de alguma forma consiga compensar todo o esforço e o apoio que me deram desde sempre.

À minha família, nomeadamente aos meus avós, por me acompanharem, por ajudarem a ver todas as situações de vários ângulos e por apoiarem as minhas decisões. Ao Gaspar, pelos momentos de animação constante e também por me tirar do sério.

A todos os que ficaram por referenciar e que contribuíram para a pessoa que sou hoje, não me esqueço nenhum de vós. Um enorme obrigada por poder contar com vocês na minha vida e por contribuírem para os mais diversos momentos.

Resumo

As doenças genéticas são predominantemente causadas por alterações na sequência de aminoácidos das proteínas e afetam locais conservados das mesmas. Nos últimos anos têm sido reportados diversos casos de mutações associadas a doenças humanas que surgem em algumas espécies como sendo o alelo nativo. Exemplos incluem doenças como Parkinson e fibrose cística. Assume-se que interações epistáticas entre diferentes aminoácidos contribuam para os diferentes fenótipos observados entre humanos e outras espécies.

Tendo como objetivo inferir as bases dos mecanismos de compensação em *Mus musculus*, este estudo utilizou modulação de proteínas de forma compreender as alterações estruturais provocadas por diferentes combinações de mutações, tendo como objetivo principal encontrar os aminoácidos que têm um papel de compensação para as mutações que, de outra forma, seriam deletérias.

Os resultados obtidos neste trabalho mostram que: (a) os aminoácidos compensatórios estão maioritariamente localizados na proximidade do local da mutação deletéria, (b) as mutações deletérias são mais comuns na superfície da proteína, (c) o impacto estrutural, em alguns casos, envolve uma dinâmica estrutural que vai muito além das simples alterações ao nível dos contactos polares, e (d) o mecanismo de compensação pode envolver mais do que um aminoácido. De uma forma geral, as nossas análises mostram que os chamados mecanismos de compensação molecular podem ser de tal complexidade intrínseca que a simples análise das interações físicas da estrutura da proteína não consegue alcançar, tornando necessário o estudo funcional em tais situações.

Abstract

Genetic diseases are mainly caused by mutations on amino acid sequence in proteins and are mostly found in conserved sites. On the past few years, there have been several reports showing that disease-associated mutations in humans can occur as a wild-type in other organisms. Examples of this phenomenon include mutations associated with Parkinson disease and cystic fibrosis. These mutations have been predicted to be the target of epistatic interactions with a second site, which is likely to balance the outcome of the deleterious substitution, softening or even neutralizing the negative effect.

Aiming to infer the basis of compensatory mechanisms, this study used protein homology modeling to understand the structural changes that mutations induce on the protein structure, but also to find possible compensatory episodes that allow the arise of an otherwise deleterious mutation, as wild-type in *Mus musculus*.

Our results show that: (a) compensatory sites are often located in the vicinity of the mutation site, (b) are more common on the protein surface, (c) the structural impact in some cases falls beyond the simplest H-bond perturbation between distinct backgrounds, and (d) the compensatory process may involve several residues. Overall, our data suggests that given the intrinsic complexity, some compensatory mechanisms may not be highlighted by the study of the physical interactions of the protein structure, but will certainly benefit from functional assays.

Key Words

Amino acid, Bioinformatics, Compensatory episodes, Deleterious mutations, Epistasis, Homology modeling, Interactions, Missense mutation, Modeling, *Mus musculus*, Protein crystal structure, Pathogenic deviations, Species-specific.

Table of contents

Agradecimentos	I
Resumo	III
Abstract.....	IV
Key Words	V
Table Index	VIII
Figure Index.....	IX
List of Acronyms	XIII
Introduction	1
1. Genetics overall	1
1.1. Introductory statement	1
1.2. Mutations as evolutionary drivers.....	2
1.2.1. Forensic Genetics.....	2
1.2.2. Population Genetics.....	3
1.2.3. Molecular Genetics	3
1.3. Amino acids as unities of variability.....	4
1.4. Compensatory events involving mutation-mutation interaction.....	6
2. Animal models	8
3. Bioinformatics: a turning point on biological studies	8
3.1. Structural Bioinformatics: unveiling protein function	9
3.2. Comparative Homology Modeling: an extrapolation to different organisms	9

Main Aims	11
Methodology	12
1. Species Specific Compensatory Pathogenic Deviations.....	12
2. Sequence data.....	14
2.1. Phylogenetic approach	14
2.2. Assessing mutational pathogenicity	14
3. Molecular approach	14
3.1. Polymerase Chain Reaction.....	15
3.2. Sanger sequencing	16
4. Structural approach.....	17
4.1. Structural analysis.....	17
4.2. Homology modeling	17
Results and Discussion.....	18
1. Confirmation of CPDs in the mouse genome	18
2. Structural compensatory episode analysis	19
Concluding remarks	39
Future Perspectives	42
References	44

Table Index

Table 1: Amino acid properties.....	5
Table 2: Set of <i>Mus musculus</i> species-specific compensated mutations candidates. Human phenotypes were retrieved from Online Mendelian Inheritance in Man, https://www.omim.org/ [59].....	12
Table 3: Gene-specific primers used in the PCR method.	15
Table 4: Summary of the results obtained in section 1 and 2. Data comprises molecular analysis, the predicted effect of the mutation by prediction software, the existence of protein structure, the putative compensatory episode, placement of the mutation on the protein structure, location of the putative compensation when compared to the deleterious mutation (being neighborhood considered 10 amino acid distant from the PD) and the number of residues involved in the compensatory episode.	40

Figure Index

Figure 1: Schematic representation of the structure of an amino acid. Adapted from [25]. The R group (side chain) is variable between distinct amino acids.....	4
Figure 2: Illustrative scheme on how compensated mutations can reach fixation in populations. Gene sequence is show in blue, wild-type non-disease causing is shown in green, compensatory site in orange, compensatory mutation in yellow and the deleterious mutation in red.	7
Figure 3: Differences in the H-bond interactions when different amino acids (asparagine and leucine) locate at the same position of the <i>CFTR</i>. Representation based on PDB ID 5UAK.	10
Figure 4: Squematic representation of the workflow followed in this work.....	13
Figure 5: Sanger sequencing validation of compensated mutations. Example of electropherograms of <i>Mus musculus</i> : <i>F8</i> , <i>F11</i> , <i>GALT</i> and <i>KCNQ1</i> mutations are shown.....	18
Figure 6: Illustrative scheme of the modeling process. A: Human and mouse homologous sequences. In red, the possible compensated mutation in mouse is shown; wild-type in human is shown in non-colored shapes while wild-type in mouse is shown in colored shapes; B: Modeling process represented graphically. On the outline of the shape models containing only the interest mutation and comprehending both the mutation and one other variable position between human and mouse are shown. On the center the second step of modeling is shown, consisting ion analyzing the mutation along with two different positions. The next steps should comprise the combinations shown in the center, consisting on the mutation and the three variable positions	19

Figure 7: Alignment of <i>ACBD5</i>. The highlighted residue is 43, the mutation-associated site in humans.....	20
Figure 8: Alignment of <i>ASAH1</i>. The highlighted residue is 23, the mutation-associated site in humans.....	21
Figure 9: Alignment of <i>ATP7B</i>. The highlighted residue is 410, the mutation-associated site in humans.....	22
Figure 10: Representation of the polar contacts around position 410. In red are represented lost connections and in green are shown gained polar contacts, when comparing human wild-type with the mutation.....	22
Figure 11: Alignment of <i>CFTR</i>. The highlighted residue is 1422, the mutation-associated site in humans.....	23
Figure 12: Protein structure models based on <i>CFTR</i> PDB, 5UAK. The position of interest is shown in red, the putative compensatory position in orange and all the positions affected by the deleterious mutation are represented in green. A: Human wild-type; B: Human background containing the deleterious allele Trp1422; C: Human background containing the deleterious mutation and the possible compensation, residue Ala1429.....	24
Figure 13: Alignment of <i>COL4A5</i>. The highlighted residue is 1659, the mutation-associated site in humans.	25
Figure 14: Alignment of <i>F8</i>. The highlighted residue is 296, the mutation-associated site in humans.....	26
Figure 15: <i>F8</i> protein structure (PDB entry 2R7E). Relative position of residues 285 and 320, when compared to residue Leu296.	26
Figure 16: Alignment of <i>F11</i>. The highlighted residue is 511, the mutation-associated site in humans.....	27
Figure 17: Homology modeling of <i>F11</i> using 5EOK PDB entry. A: Wild-type. B: Polar contacts (H-bonds) predicted in the human background in the presence of His511mutation.The yellow circle highlights the loss of an H-bond to amino acid Glu612.....	28

Figure 18: Overview of the target residues of GALT protein. A: Alignment of the human and rodents' GALT protein sequences showing the variations between positions 307 and 314. These positions correspond to a hypervariable region inserted on a highly conserved sequence; B: Detailed view of positions 307-314 on the crystal structure of Human GALT where the human disease-associated mutation (E308K) is lightened in red and possible compensatory residues in green. 29

Figure 19: Four models concerning the process of compensation of the Human mutation in residue 308, associated with Galactosemia. A: Human wild-type; B: Human background containing the deleterious allele Lys308; C: Mouse wild-type; D: Human background containing both deleterious mutation (Lys308) and putative compensatory partners (Leu307, Thr309, Asp314). 30

Figure 20: Reconstruction of the evolution of the hypervariable region of GALT, in Rodents. The starting point was the conserved sequence considered to be ancestral sequence. 31

Figure 21: Alignment of IDUA. The highlighted residue is 103, the mutation-associated site in humans. 32

Figure 22: Alignment of KCNQ1. The highlighted residue is 619, the mutation-associated site in humans. 33

Figure 23: Alignment of LDB3Z4. The highlighted residue is 351, the mutation-associated site in humans. 34

Figure 24: Alignment of NF1. The highlighted residue is 491, the mutation-associated site in humans. 35

Figure 25: Protein structure of PTS: (A) Human wild-type is shown. (B) Human background with the deleterious mutation (Val103Ala) inserted. 36

Figure 26: Alignment of RET. The highlighted residue is 480, the mutation-associated site in humans. 37

Figure 27: Alignment of SGCA. The highlighted residue is 30, the mutation-associated site in humans.....**37**

Figure 28: Alignment of TG. The highlighted residue is 2202, the mutation-associated site in humans.....**38**

List of Acronyms

3D	Three dimensional
%	Percent sign
®	Registered trademark
°C	Degree Celsius
∞	Infinite
ACBD5	Acyl-CoA binding domain containing protein 5
APP	Amyloid beta precursor protein
ASAH1	N-acylsphingosine amidohydrolase 1
ATP7B	Copper-transporting P-type adenosine triphosphatase
CFTR	Cystic fibrosis transmembrane conductance regulator
COL4A5	Collagen type IV alpha 5 chain
Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
BLAST	Basic Local Alignment Search Tool
bp	Base pair
CPD	Compensatory pathogenic deviation

Cys	Cysteine
DNA	Deoxyribonucleic acid
et. al.	<i>et alli</i>
ExoSap	Exonuclease I – Shrimp Alkaline Phosphatase
<i>e.g.</i>	<i>Exempli gratia</i>
F8	Coagulation factor VIII
F11	Coagulation factor XI
GALT	Galactose-1-phosphate uridylyltransferase
Gln	Glutamine
Glu	Glutamic acid
Gly	Glycine
HGMD	Human Gene Mutation Database
His	Histidine
HuRI	Human Reference Protein Interactome Mapping Project
IDUA	Alpha-L-iduronidase
i.e.	<i>Id est</i>
Ile	Isoleucine
KCNQ1	Potassium voltage-gated channel subfamily Q member 1
LDB3Z4	LIM domain-binding 3 isoform 4
Leu	Leucine
Lys	Lysine
Met	Methionine

min	Minutes
mL	Milliliter
NF1	Neurofibromin 1
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PD	Pathogenic deviation
PDB	Protein Data Bank
Phe	Phenylalanine
Pro	Proline
PTS	6-pyruvoyl-tetrahydropterin synthase
RET	Ret proto-oncogene
s	Second
Ser	Serine
SGCA	Sarcoglycan alpha
SGTA	Small glutamine rich tetratricopeptide repeat containing alpha
SGTB	Small glutamine rich tetratricopeptide repeat containing beta
SIFT	Sorting intolerant from tolerant
SNP	Single nucleotide polymorphism
TG	Thyroglobulin
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine

Val	Valine
VNTR	Variable number of tandem repeats
WT	Wild-type

Introduction

1. Genetics overall

1.1. Introductory statement

Genetics is the field of research that studies the difference within and between organisms and how this variation is transferred from generation to generation. This concept was intuitively followed during Human evolution, when domestication of wild species and selective breeding was determinant for the success of the human species. Currently, Genetics is based on the study of gene function and their interactions, with a large number of applications, from food industry to medicine [1]. Human Genetics has allowed breakthroughs on the legal field, namely on identification, paternity and kinship tests and potential homicide investigations [2], as well as in clinical investigations. Genetic investigations in non-human DNA (animals [3], plants [4] and microorganisms [5]) have also been of great relevance in comparative biology, medical genetics as well as forensic investigations.

A gene is the unity of hereditability, a DNA segment that codes the information needed to synthesize proteins. These encoded proteins are responsible for the phenotypical expression of distinct traits. Each phenotypic trait can result from a protein gene or from the interactions between distinct proteins and the levels of co-dependence [1].

1.2. Mutations as evolutionary drivers

Proteins are made by a set of amino acids, which are constituted by a nucleotide sequence. A mutation is the replacement of a certain nucleotide sequence during cell division. These spontaneous (or *de novo*) mutations may have a beneficial, neutral or negative effect. Advantageous mutations are able to increase in frequency and eventually cause adaptative evolutionary changes. Deleterious changes can have their frequency reduced due to their outcome. In addition, some variations are neutral as no notorious effect on individual fitness is observed. These types of mutations can eventually accumulate and reach fixation in a certain population, over time. It is assumed that the great majority of mutations does not have an immediate effect on the fitness, mainly because most of mutations fall in non-coding regions of the genome, but also because changes associated to milder phenotypes are more easily accepted than the ones with a severe effect [1]. Deleterious mutations can affect a high number of important factors, such as sexual selection, which are essential to the prevalence and adaptive success of species [6] and could even cause extinction of populations [7].

1.2.1. Forensic Genetics

There are some specific requirements to be fulfilled in terms of markers used in forensic genetics, such as being easy and cheap to characterize, give profiles simple to compare, not be under selective pressure, be highly polymorphic and have a low mutation rate [8].

Some of the most important markers for forensics framework are minisatellites or variable number of tandem repeats (VNTRs), and microsatellites also known as short tandem repeats (STRs). VNTRs have a repeat sequence from 6 to 100 bp and were firstly used in forensics, but due to limitations were lately replaced by short tandem repeats [9]. STRs are repeats of 1 to 6 bp, but mostly tetranucleotide repeats [10]. These markers are located in non-coding regions of the genome, are highly discriminatory, sensitive, require low DNA quantities, a technically easy and cheap method of genotyping, are suitable for multiplex techniques and are frequent in the genome [11].

But more frequent than the STRs, are the Single Nucleotide Polymorphisms (SNPs). Their number compensate their lack of variability (two alleles per *locus*), being the same power of

discrimination achieved through usage of 50-80 SNPs for each 10 STRs [12]. Except for mitochondrial DNA, SNPs are not widely used. Although they have a lower success rate, due to their low mutation rate (when compared to STRs) [13], they require less amounts of DNA and can work when STR analysis fails [14, 15]. Since they show lower mutation rates, they are better for distinguishing lineages, populations or species, rather than individuals [16].

1.2.2. Population Genetics

With the aim of understand the history of different populations, population genetics examines several phenomena among populations such as adaptation, mutation, selection, genetic drift, recombination and migration. This field intends to monitor fluctuations on population level, in terms of evaluating the relationship between genotype and phenotype and can also detect admixture, bottleneck and founder effect events [17].

1.2.3. Molecular Genetics

The study of genetics is of great importance in terms of diagnosis and personalized treatment since different genotypes induce different phenotypical expression. Regarding the fields of applicability, prevention, genetic counseling and genetic therapy are some of the main interventions of genetics in the field of medicine. Medical genetics does not only involve treatment and counseling of the patient itself. It is also of great importance to extend the analysis to their families not only in terms of conduct, but also concerning genetic studies, to understand the pathology and to advise and prevent risk behaviors [18].

Molecular genetics embraces the understanding of phenotype-genotype relationships either in pathological as in non-pathological context [19]. This subject aims to understand how the variability of the genotype is connected to human diseases, since most of them do not have effective treatment. To this purpose, there are projects as The Human Genome Project [20] which allowed the emergence of new techniques of gene manipulation on non-human models, including manipulation and insertion of DNA segments on an organism, such as the CRISPR/cas9 approach [21, 22]. Frequently, silencing techniques are applied, which aim to knock out a certain gene that allow a better knowledge on human diseases and provide better insights on the possibility of inducing a milder expression of some highly deleterious

syndromes [23]. Genetic engineering became valuable in terms of protein research, allowing to induce loss of function, determination of protein function, place of expression and protein interaction identification [24].

1.3. Amino acids as unities of variability

Each diverse set of amino acids determines a different protein structure and, consequently, a specific function. All amino acids have a R group or variable side chain. There are twenty known amino acids in proteins, each with a different R group, as shown on Figure 1, that gives each amino acid their exclusive properties. Amino acids are linked by covalent bonds, also named peptide bonds, which are made from the link between the amino group of one and the carboxyl group of another [25].

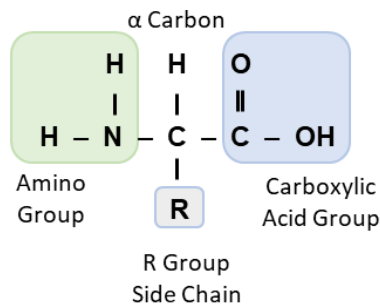


Figure 1: Schematic representation of the structure of an amino acid. Adapted from [25]. The R group (side chain) is variable between distinct amino acids.

Amino acids are usually classified according to their side chain properties. There are several ways in which we can group amino acids, being the most used charge, polarity, hydrophobicity and molecular volume, as seen on Table 1 [26, 27]. These changes can have a milder effect when the substitution consists on chemically similar amino acids or can have a more severe outcome when there is a more abrupt substitution, which can also determinate the frequency in which certain changes can occur, since they are considered more or less disturbing, depending on the substitution event [28-30].

Table 1: Amino acid properties

Amino acid	3 Letter code	1 Letter code	Characteristics
Alanine	Ala	A	Non-polar, neutral
Arginine	Arg	R	Basic polar, positive
Asparagine	Asn	N	Polar, neutral
Aspartic Acid	Asp	D	Acid polar, Negative
Cysteine	Cys	C	Non-polar, neutral
Glutamine	Gln	Q	Polar, neutral
Glutamic Acid	Glu	E	Acid polar, Negative
Glycine	Gly	G	Non-polar, neutral
Histidine	His	H	Basic polar, mostly neutral
Isoleucine	Ile	I	Non-polar, neutral
Leucine	Leu	L	Non-polar, neutral
Lysine	Lys	K	Basic polar, positive
Methionine	Met	M	Non-polar, neutral
Phenylalanine	Phe	F	Non-polar, neutral
Proline	Pro	P	Non-polar, neutral
Serine	Ser	S	Polar, neutral
Threonine	Thr	T	Polar, neutral
Tryptophan	Trp	W	Non-polar, neutral
Tyrosine	Tyr	Y	Polar, neutral
Valine	Val	V	Non-polar, neutral

1.4. Compensatory events involving mutation-mutation interaction

The term “epistatic” was firstly used by Bateson in 1909, to describe interaction between two different loci [31]. There are several ways by which these interactions can occur, one of them is the interaction between nonallelic genes where the expression of one buffers the impact a second one [32]. Results from several studies are increasingly indicative that the effect of a gene on a determined phenotype is caused by the joint interaction between a set of functionally linked genes and their contact with the environment, which is known as genotype-environment interaction [32]. These interplays have been observed in several organisms [33].

Epistasis can also be responsible for the arising of pleiotropy, where a certain gene can affect several phenotypic features, depending on the magnitude of interactions with other genes in the same biological/functional pathway [34].

Epistatic interactions are expected to have two different effects: they can mask the expected outcome or can aggravate their effect, increasing severity. They can also occur in the same molecule (‘intramolecular epistasis’) or between different genes (‘intermolecular epistasis’) [35].

Missense mutations are the main cause of most human pathological diseases, which are associated to a certain phenotypical expression. Some of these mutations appear as the wild-type (WT) allele in non-human genomes. Because most of the human disease-associated mutations locate at conserved positions, one would expect that these mutations would show a similar outcome in related species. These episodes are known as compensated pathogenic deviations (CPDs) [36]. The fact that these changes do not cause visible phenotypes can be related to possible epistatic interactions between the pathogenic mutation and a second position, which is likely to balance the outcome of the deleterious substitution, softening or even neutralizing the negative effect. The interaction between distinct alleles that can act as genetic modifiers results in episodes of structural/functional compensation which are more frequently found in the same gene [37]. There are now numerous cases reported of possible mutation-compensatory pairs, which also include severe human diseases, appearing as wild-type in mouse (e.g. Parkinson); cystic fibrosis, androgen insensitivity syndrome and leukoencephalopathy in rat and mouse; phenylketonuria in rhesus monkey; Niemann-Pick disease in gorilla [38, 39]; human

neonatal hyperammonemia [40] and ornithine transcarbamylase deficiency [41] in chimpanzee; among many others.

The allelic combination between a deleterious mutation and compensatory partners may result from several evolutionary paths. One possible path is that both mutations arise independently on the population, converging afterwards by recombination. An alternative hypothesis suggests simultaneous birth and fixation of both mutations. Finally, one may also consider the birth of an otherwise deleterious mutation in a background where their corresponding compensatory site already exists. This seems to be the case for many compensatory episodes in natural populations [42]. These different pathways are graphically shown in Figure 2.

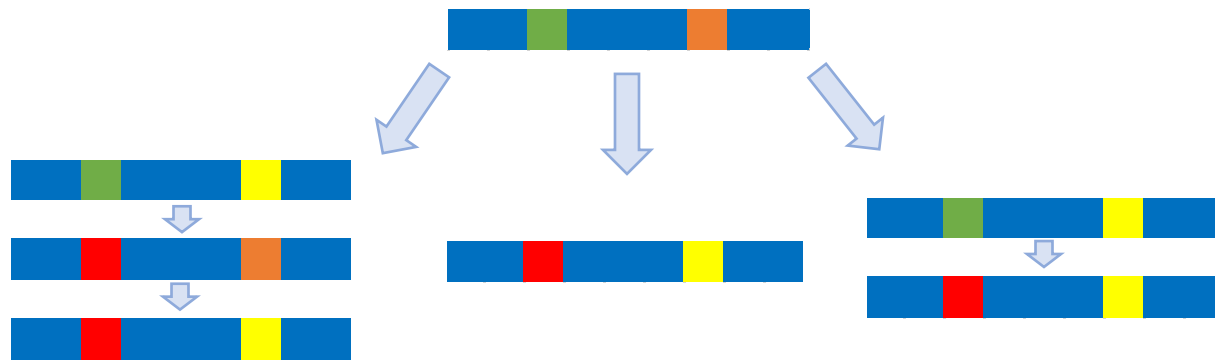


Figure 2: Illustrative scheme on how compensated mutations can reach fixation in populations. Gene sequence is shown in blue, wild-type non-disease causing is shown in green, compensatory site in orange, compensatory mutation in yellow and the deleterious mutation in red.

A wide range of effects are caused by mutations, which include impairment of protein structure, folding, stability and binding, among many others [43, 44]. The interaction between the substitution and structural environment has a crucial role on the development of compensatory mechanisms. These are most likely to arise on the same protein as the pathogenic mutation, although some cases of intermolecular interactions can also occur [45].

Mutations buried deep in the protein structure, with higher number of connections, are less likely to be compensated, while those more accessible are easier to balance, mainly due to the contrast between the simplicity and complexity of the interaction to be compensated. Additionally,

extreme changes might cause large scale protein destabilization, resulting on huge compensatory effort and difficulty [46].

2. Animal models

An animal model is used with the intent of mimicking the human organism in terms of the main aim of the object of study (a certain disease, response to drugs, e.g.). There is a huge variety of animal models available for genetic studies and mouse (*Mus musculus*) is one of the most widely used due to the similarity with the human physiology [47]. In genetics terms, although smaller than the human genome, the mouse genome has a high number of direct human orthologs [48]. In accordance, genetic findings in mice and human have been proven highly advantageous because of concordant data between both species [32].

Genetic study of animal models can help us to understand how compensatory episodes occur and provide us insights on how it would be possible to extrapolate to the human species and other mammalian models, improving quality of prediction, diagnosis, prevention and treatment of diseases, providing better knowledge and directing to develop treatments that might mitigate or even counteract the pathological effect [49].

3. Bioinformatics: a turning point on biological studies

Bioinformatics is a field of study where computational and mathematical components are applied to biological data, consisting in storage, analysis, interpretation and processing of information [33].

The widespread 3rd generation techniques such is the Next-Generation Sequencing (NGS), generate tremendous amounts of information that requires specialized skills and computer power to process the data and to promote the establishment of accurate phenotype-genotype relationships [50]. In this context, Bioinformatics aims to provide huge amount of organized

information, therefore it was of most importance to create two very intuitive categories of platforms, (1) databases, such as libraries, to contain and organize information accordingly to their focus, and (2) tools to aid data processing and analysis, such as gene identification, phylogenies trees, predict three-dimensional (3D) protein conformation, among many others [51].

3.1. Structural Bioinformatics: unveiling protein function

Predicting the three-dimensional protein structure is an important part of bioinformatics. The primary structure, that is the amino acid sequence of the protein, can be obtained through the sequence of the gene that codes that specific protein. Knowledge in terms of protein structure is vital to understand protein function and even predict the effects of possible mutations. In this sense, structural bioinformatics aims to create processes which will allow to manipulate biological data, but also to apply these approaches to solve raised problems [52].

3.2. Comparative Homology Modeling: an extrapolation to different organisms

Comparative homology modeling is a tool used to obtain information on protein structure when experimental data is not available. The process can be described as follows: after selecting the target sequence, a BLAST shall be made [53] to find the homologous sequence, associated with a protein structure. The structure showing highest identity degree should be chosen template. When the most appropriate template is selected, protein modeling can then be performed, considering that a high confidence level is achieved when identity is above 75% [54]. This can be accomplished using modeling programs such as MODELLER [55] or SWISS-MOD [56], that can be visualized in the PyMol software [57], which inclusively has a mutagenesis predictor modeler itself, although it is confined to the neighborhood of the induced change. After obtaining the 3D model it is necessary to evaluate the degree of confidence of the information obtained through

the model. This can be made using quality check software, such is the VERIFY3D [58]. If the model passes these “stress tests”, we can consider that the information obtained can mimic the outcome of a possible conformational change.

Through homology modelling, the visualization of the structural perturbation imposed by replacing a specific amino acid by an alternative one allows better understanding of the phenotypic consequences. An example is shown in Figure 3.

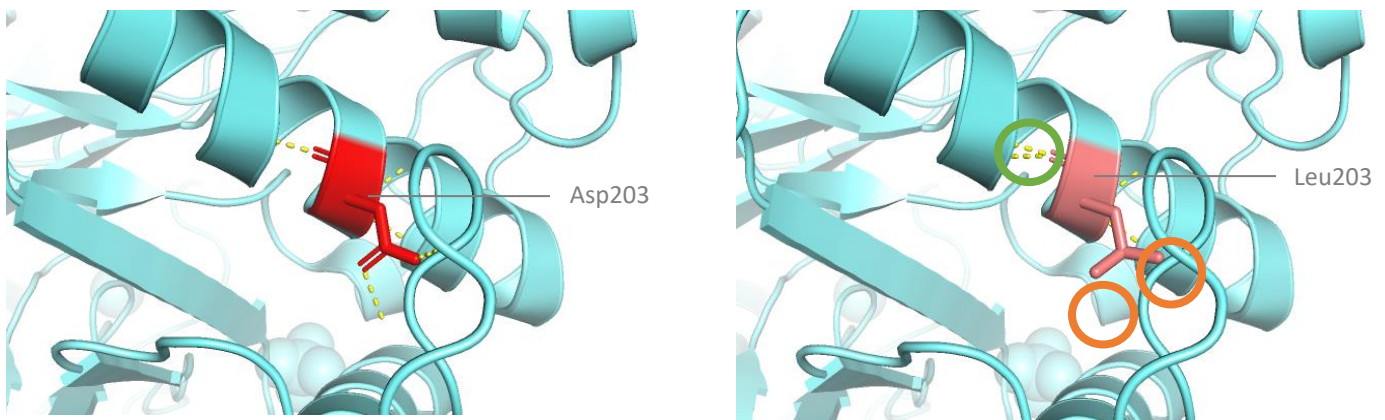


Figure 3: Differences in the H-bond interactions when different amino acids (asparagine and leucine) locate at the same position of the *CFTR*. Representation based on PDB ID 5UAK.

Main Aims

This study was intended to infer the molecular and structural basis of compensatory events found in the mouse (*Mus musculus*) genome.

Specifically, the main aims of this work are:

1. Confirm the existence of the human mutation as wild-type in mouse through molecular methods, involving Polymerase Chain Reaction (PCR) and Sanger sequencing.
2. Explore the interaction between disease-associated mutations and physically interacting residues of compensated mutations previously found by the lab team in the mouse genome, in order to define the most suitable candidate for the compensatory deviation.

This study might allow us to understand the basis of some compensatory events and is expected to contribute to the milestone of personalized and predictive medicine, allowing to improve adapted predictions of risk and even help to direct individual therapeutic strategies. It might also contribute to the development of more suitable mouse models of human diseases, also providing a better insight on how compensatory events occur in non-human lineages.

Methodology

1. Species Specific Compensatory Pathogenic Deviations

The focus of this study was studying seventeen putative *Mus musculus* specific compensatory episodes as is shown in Table 2 [38].

Table 2: Set of *Mus musculus* species-specific compensated mutations candidates. Human phenotypes were retrieved from Online Mendelian Inheritance in Man, <https://www.omim.org/> [59].

Gene	Position	Mutation	Phenotype
ACBD5	43	His>Tyr	Thrombocytopenia
ASAH1	23	His>Asn	Farber lipogranulomatosis
ATP7B	410	Pro>Leu	Wilson disease
CFTR	1422	Arg>Trp	Cystic Fibrosis
COL4A5	1659	Ser>Asp	Alport syndrome
F8	296	Leu>Phe	Hemophilia A
	1511	Val>Ile	Hemophilia A
F11	511	Tyr>His	Factor XI deficiency
			Hemophilia C
GALT	308	Glu>Lys	Galactosemia
IDUA	103	Thr>Pro	Mucopolysaccharidosis
KCNQ1	619	Leu>Met	Jervell and Lange-Nielsen syndrome
LDB3 Z4	351	Thr>Ala	Cardiomyopathy
NF1	491	Tyr>Cys	Neurofibromatosis
PTS	103	Val>Ala	Hyperphenylalaninemia
RET	480	Glu>Lys	Thyroid carcinoma
			Hirschsprung disease
SGCA	30	Pro>Leu	Muscular dystrophy
TG	2202	Pro>Arg	Thyroid dys hormonogenesis

The methodological approach followed in this work was based on the principles documented by several previous studies such as the work of Barešić and Martin [60], where the compensatory episodes may be identified by several stages which are graphically present in Figure 4. This map shows an overall image of the workflow performed in this study and will be the focus of the following sections.

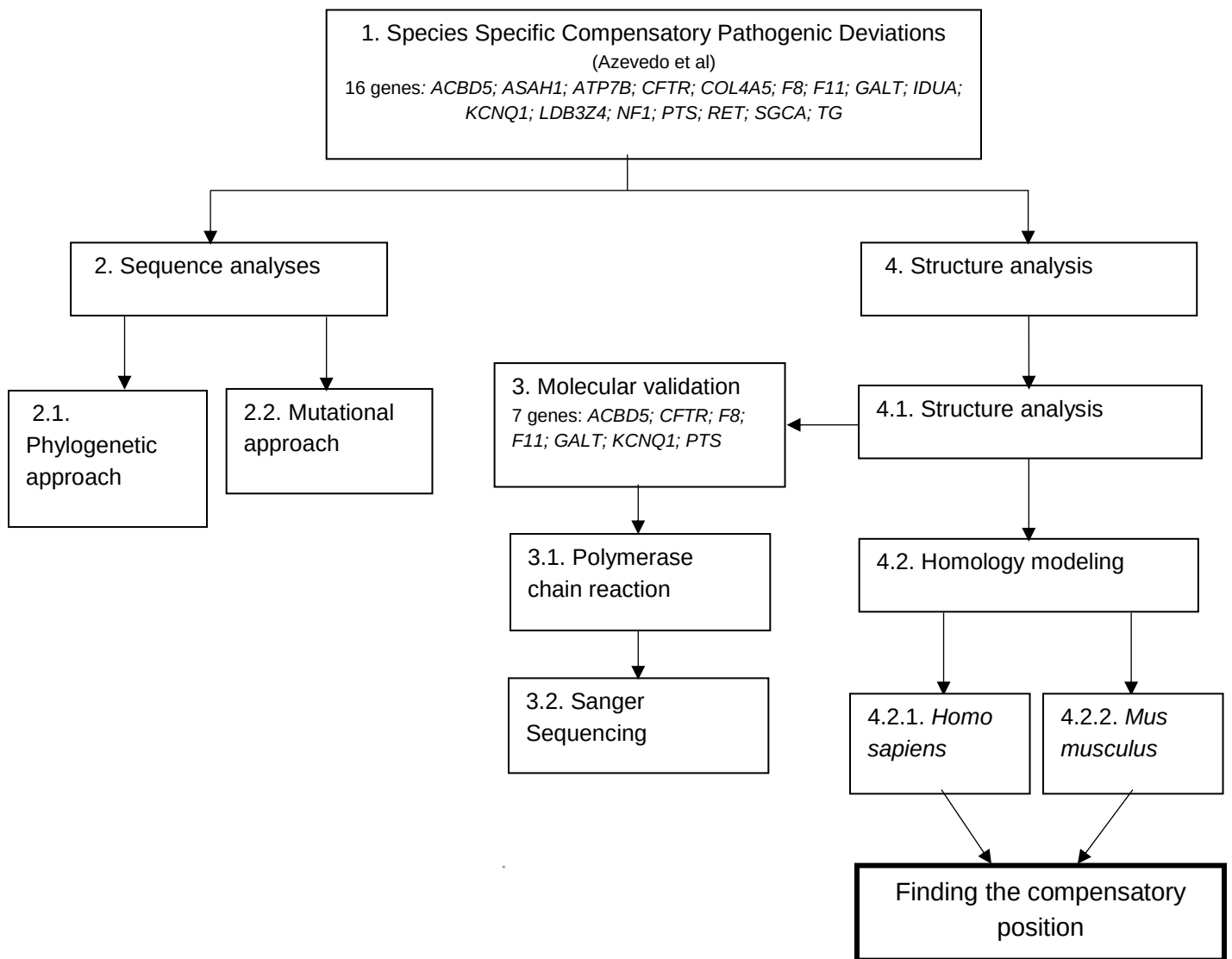


Figure 4: Schematic representation of the workflow followed in this work.

2. Sequence data

Sequences for each of the targeted genes were retrieved from Ensembl database [61]. All available placental mammalian sequences were analyzed. Incomplete and low-quality sequences were excluded from the analysis. Multiple alignment was performed by Clustal Omega [62], inserted on BioEdit program [63], and the resulting outcomes were manually curated.

2.1. Phylogenetic approach

Phylogenetical trees were obtained using MEGA7 [64] software, that allows to infer the molecular relationships between species, regarding each gene of interest. Neighbor-Joining method [65] was used to construct the phylogenetic tree.

2.2. Assessing mutational pathogenicity

All mutations were analyzed in PolyPhen [66] and in SIFT [67] software in order to predict possible damaging effects on the protein.

3. Molecular approach

The practical component of this work was developed at the Population Genetics and Evolution laboratory, of the Instituto de Patologia e Imunologia Molecular da Universidade do Porto, IPATIMUP, at the Instituto Investigação e Inovação em Saúde, i3S.

3.1. Polymerase Chain Reaction

Based on the *Mus musculus* reference sequence retrieved from Ensembl database, sets of primers were designed to amplify the region containing the mutation on the genes of interest, *ACBD5*, *CFTR*, *F8*, *F11*, *GALT*, *KCNQ1* and *PTS*, as shown in Table 3. Primers were designed using Primer3 [68] software and tested on OligoCalc [69]. Extracted DNA from *Mus musculus* was kindly provided by a collaborator.

Table 3: Gene-specific primers used in the PCR method.

Gene	Primer sequence	Amplicon length
<i>ACBD5</i>	F: 5' CGGTGTTTCAGTTTCATGCAGG 3' R: 5' GGCAAACCTCTGGATCACCTTC 3'	171 bp
<i>CFTR</i>	F: 5' CACTCAGAGGTCTCACTAATGC 3' R: 5' GGAGCTAATGGCCTGCTGG 3'	123 bp
<i>F8</i>	F: 5' GTCTACTGGCACGTGATTGG 3' R: 5' GAACTGCCCAAGATCTATCAAG 3'	162 bp
<i>F11</i>	F: 5' GAGACACCTCAGATCCAAACC 3' R: 5' GTTACCTCTTAGTGCTGTGTAC 3'	183 bp
<i>GALT</i>	F: 5' GGGCTAGACTGAAGGCTGAC 3' R: 5' GAACTTCCGGACAGTTGCGG 3'	160 bp
<i>KCNQ1</i>	F: 5' GTCCTTACAGGTGACACAACCTG 3' R: 5' CACAGTCAGTTGTTCGTAGGTG 3'	196 bp
<i>PTS</i>	F: 5' GATCACAAGAACCTGGACCTG 3' R: 5' GAGACAGGTGAGCAGGGTAC 3'	110 bp

Polymerase chain reaction was performed with a total volume of 10 μ L, containing PCR mixture (5 μ L of Qiagen MasterMix, 2 μ L of Qiagen RNAase-free water and 0.5 μ L of each primer, at 10 μ M) and 2 μ L of DNA template.

Since these primers were specifically designed for this experiment, PCR comprised two different stages, concerning the 30 cycles partition. The first 10 cycles consisted in using a lower annealing temperature (55 °C), to decrease the specificity of the reaction, so that every fragment could be amplified. The remaining 20 cycles were performed using a higher annealing temperature (62°C), to increase the specificity of the reaction. PCR program was performed on a GENEAMP® PCR SYSTEM 2700 Applied Biosystem and amplification conditions comprised initial denaturation of 15 min at 95 °C; followed by 10 cycles of 30 s at 95 °C, 30 s at 55 °C, 1 min at 72 °C (stage one); 20 cycles of 30 s at 95 °C, 30 s at 62 °C, 1 min at 72 °C (stage two); and a final extension of 20 min at 72 °C.

PCR products were then visualized after a polyacrylamide gel electrophoresis, performed according to standard procedures.

3.2. Sanger sequencing

PCR amplicons were purified using 1 µL of ExoSap-IT for 1,5 µL of amplified product, in 1 cycle of 15 min at 37°C and 15 min at 85 °C. After purification, 2 µL of ABI Big Dye Terminator Cycle Sequencing Ready Reaction kit v3.1 was added to each sample, followed by 4 min at 96 °C, 35 cycles of 20 s at 96 °C, 20 s at 57 °C, 90 s at 60 °C, and finally 20 min at 60 °C. Final purification was performed using Sephadex ®. Samples were then inserted on ABI 3130, and the amplified products were separated and detected by a capillary electrophoresis process.

4. Structural approach

To analyze protein structure, a search in PDB database [70] was performed in order to retrieve each target Human protein, if available.

4.1. Structural analysis

Visualization and structural analysis of theoretical models was performed using academic version of PyMol software to compare possible interactions between mutations and to analyze the spatial proximity on the original file retrieved from PDB database.

4.2. Homology modeling

To analyze conformational changes, both *Homo sapiens* and *Mus musculus* protein sequences were analyzed in the HHpred tool on MPI Bioinformatics toolkit [71] with the intent of searching, matching and aligning sequences with the correspondent homologous human protein structure, in PIR format, which consists on the alignment of our interest sequence with the one of the protein structure.

Using PIR formats obtained, MODELLER option was used to simulate the effects of several mutations on the human background to understand the structural impact of the pathogenic change on the structure of the human protein. Secondly, *Mus musculus* homologous proteins were modelled using the same strategy. Finally, several theoretical models containing both mutation and possible compensatory site were manually created by inserting the amino acid replacements in the background of interest in order to search for structural conformational compensation.

Results and Discussion

1. Confirmation of CPDs in the mouse genome

From the initial set of proteins, we selected seven (*ACBD5*; *CFTR*; *F8*; *F11*; *GALT*; *KCNQ1*; *PTS*) for which the crystal structures are available, to confirm the presence of the mutation in the mouse genome. A PCR was performed, followed by electrophoretic run to confirm success of the amplification, and finally Sanger sequencing was performed to validate the results. We confirmed the presence of the human disease-associated allele in all cases. Some examples of the sequencing results are shown in Figure 5.

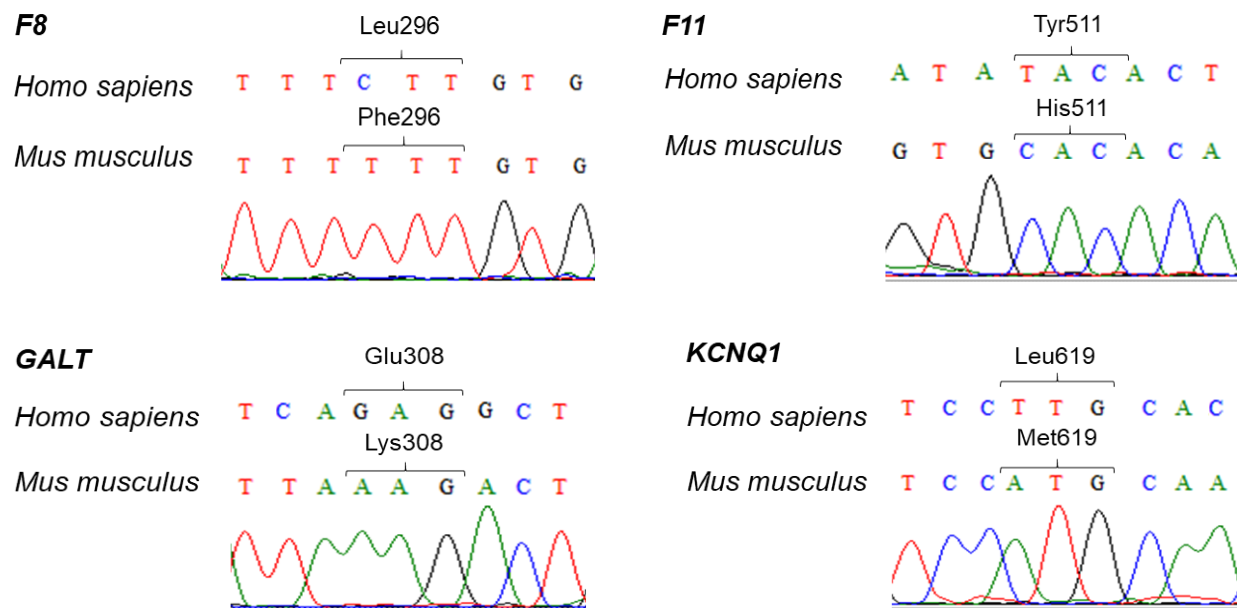


Figure 5: Sanger sequencing validation of compensated mutations. Example of electropherograms of *Mus musculus*: *F8*, *F11*, *GALT* and *KCNQ1* mutations are shown.

2. Structural compensatory episode analysis

In order to evaluate the influence of distinct mutations on protein structure, models concerning differences between human and mouse sequences were elaborated whenever protein structure was available. The effect of the mutation *per se* on a human background was analyzed first. The next round of models contained the specific mutation and a second site, selected by proximity to the original mutation, corresponding to a variable position between human and mouse, as shown in Figure 6. After testing these first combinations, the next step comprised models concerning different combinations of these changes in pairs, then triplets, sequentially, as shown in the center of Figure 6B, until the putative compensatory episodes could be detected.

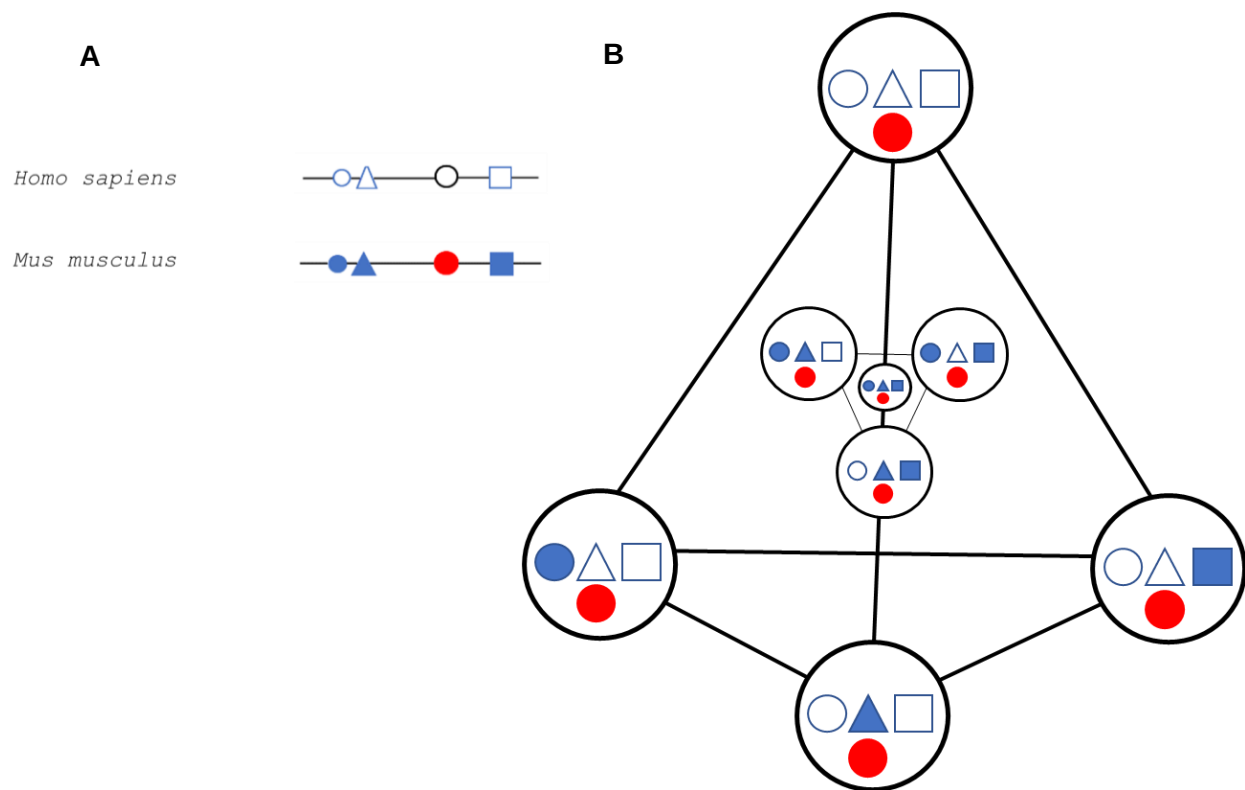


Figure 6: Illustrative scheme of the modeling process. A: Human and mouse homologous sequences. In red, the possible compensated mutation in mouse is shown; wild-type in human is shown in non-colored shapes while wild-type in mouse is shown in colored shapes; B: Modeling process represented graphically. On the outline of the shape models containing only the interest mutation and comprehending both the mutation and one other variable position between human and mouse are shown. On the center the second step of modeling is shown, consisting in analyzing the mutation along with two different positions. The next steps should comprise the combinations shown in the center, consisting on the mutation and the three variable positions

The results obtained by following this strategy will next be described in detail for each gene.

ACBD5

A substitution of a histidine in position 43 to a tyrosine in human Acyl-CoA binding domain containing protein 5 (*ACBD5*) has previously been associated to thrombocytopenia which is an hematologic disorder, characterized by a low number of platelets [72, 73].

The Tyr43 substitution appears as wild type in some rodents, namely in several *Mus* species (Figure 7). The available protein crystal structure did not contain enough residues close to the mutation position to obtain a clear model for the inference of the possible compensatory episode (PDB entry 3FLV). Yet, the surrounding area of the mutation is conserved throughout mammals. In these terms, compensatory episode could include an interaction between residues Arg30 and Arg31, since both co-occur in several organisms with the pathogenic amino acid as the wild type allele.

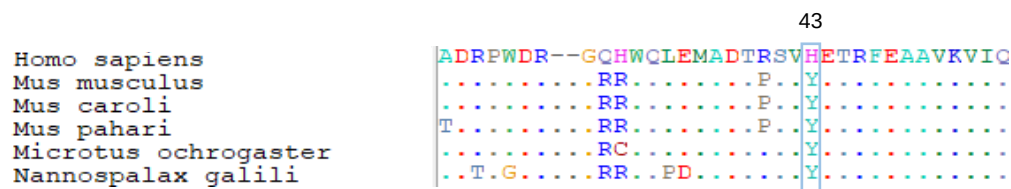


Figure 7: Alignment of *ACBD5*. The highlighted residue is 43, the mutation-associated site in humans.

ASAH1

A mutation at the 23rd residue of human N-acylsphingosine amidohydrolase 1 (*ASAH1*) gene, resulting in a change from histidine to aspartic acid has been previously associated with Farber disease, a metabolic disease, consisting of the malfunction of acid ceramidase [74, 75].

The His23Asp change involves the replacement for differently charged amino acids, which might indicate possible strong impact in protein structure or in terms of substrate binding. Interestingly, the Asp23 residue seems to be *Mus* specific rather than *Mus musculus* specific only (Figure 8). Other mammals also present a different wild-type than human, although different from mouse, which might indicate a tendency for the occurrence of mutations in this position. The most likely compensatory residue is Thr17, since it is present in all mammals with a different WT residue to the one present in the human protein. Notwithstanding, Gln20 and Val21, should not be discarded, since they co-occur in several genomes that also harbors the human deleterious allele as the wild-type.

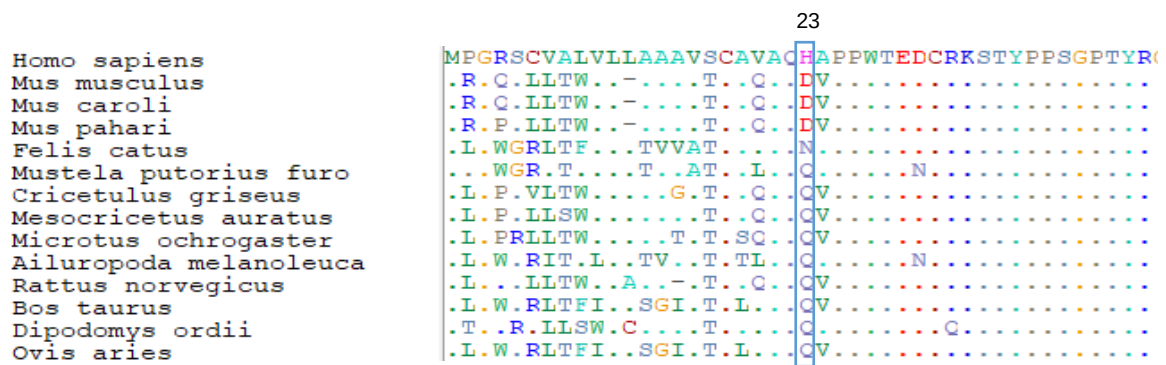


Figure 8: Alignment of *ASA1*. The highlighted residue is 23, the mutation-associated site in humans.

ATP7B

A mutation from proline to leucine in human Copper-transporting P-type adenosine triphosphatase (*ATP7B*) in position 410 has been associated with Wilson disease. Wilson disease is a disorder of copper metabolism, associated with copper accumulation resulting in liver damage and neurologic symptoms [76, 77].

Position 410 belongs to a highly variable region both in *Mus musculus* and *Mus spretus*. Although other *Mus* species show a different WT than human, none is a leucine at this site (Figure 9).

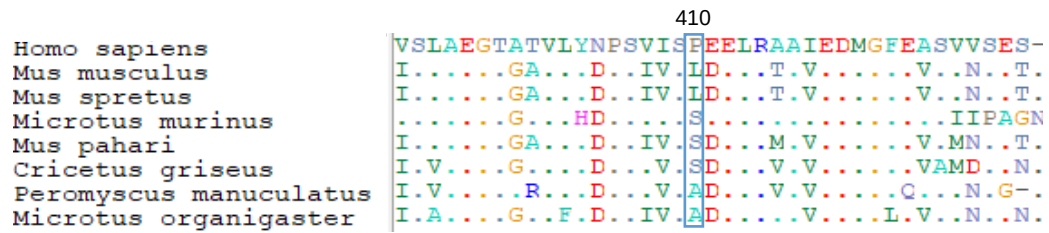


Figure 9: Alignment of ATP7B. The highlighted residue is 410, the mutation-associated site in humans.

Using 2ROP [78] as base to perform model analysis it was possible to obtain the H-bond changes made by the Pro410Leu, when compared to the human wild type, as seen in Figure 10.

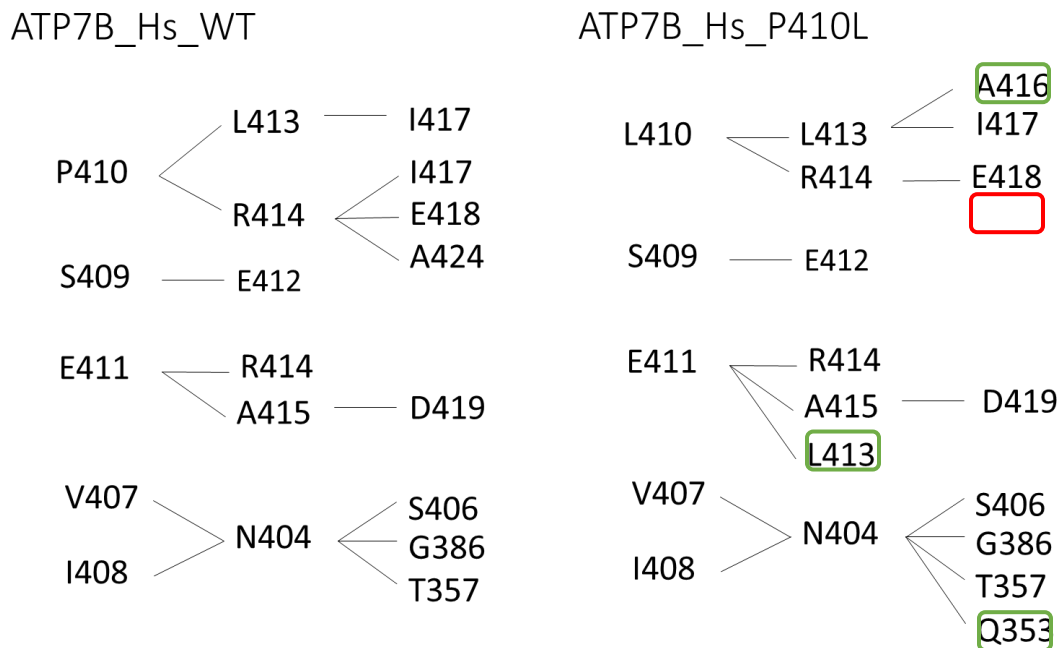


Figure 10: Representation of the polar contacts around position 410. In red are represented lost connections and in green are shown gained polar contacts, when comparing human wild-type with the mutation.

After analyzing several models, it was possible to understand that the interaction between different residues restore partially the H-bonds related to the wild type, but not completely. Therefore, it is not possible to infer a specific residue as being the compensatory episode.

CFTR

Mutations concerning the arginine to tryptophan change in residue 1422, in the human Cystic fibrosis transmembrane conductance regulator (*CFTR*) gene have been associated with cystic fibrosis, which provokes a thickening of the secretions of the body, causing major difficulties on the respiratory tract [79].

The Arg1422Trp substitution could induce great disturbance in protein function since this replacement implies a change in polarity and charge, modifying a positive polar residue to a neutral hydrophobic one. The sequence surrounding the position 1422 revealed to be very conserved among the Rodent order but distinct from the remaining mammals analyzed (Figure 11).

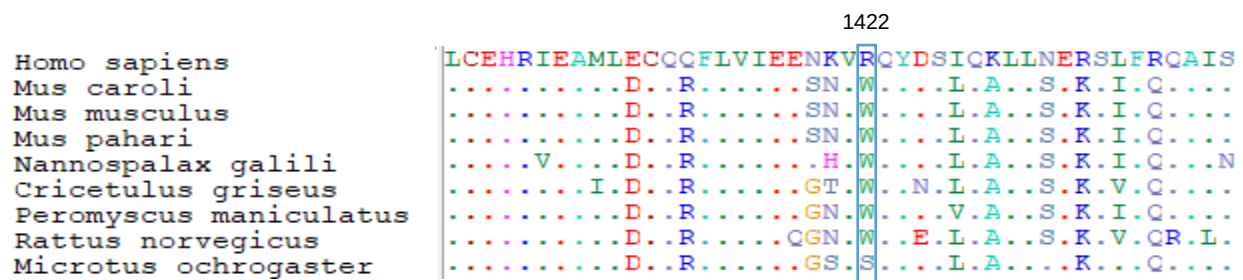


Figure 11: Alignment of *CFTR*. The highlighted residue is 1422, the mutation-associated site in humans.

Through the analysis of *CFTR* protein structure (5UAK [80]), structural models were made and analyzed (Figure 12). Both Asn1420 and Ala1429 residues in *Mus musculus* are possible candidates to be the compensatory partners of the pathogenic mutation, since both restore the lost H-bonds caused by the presence of the otherwise deleterious residue Trp1422.

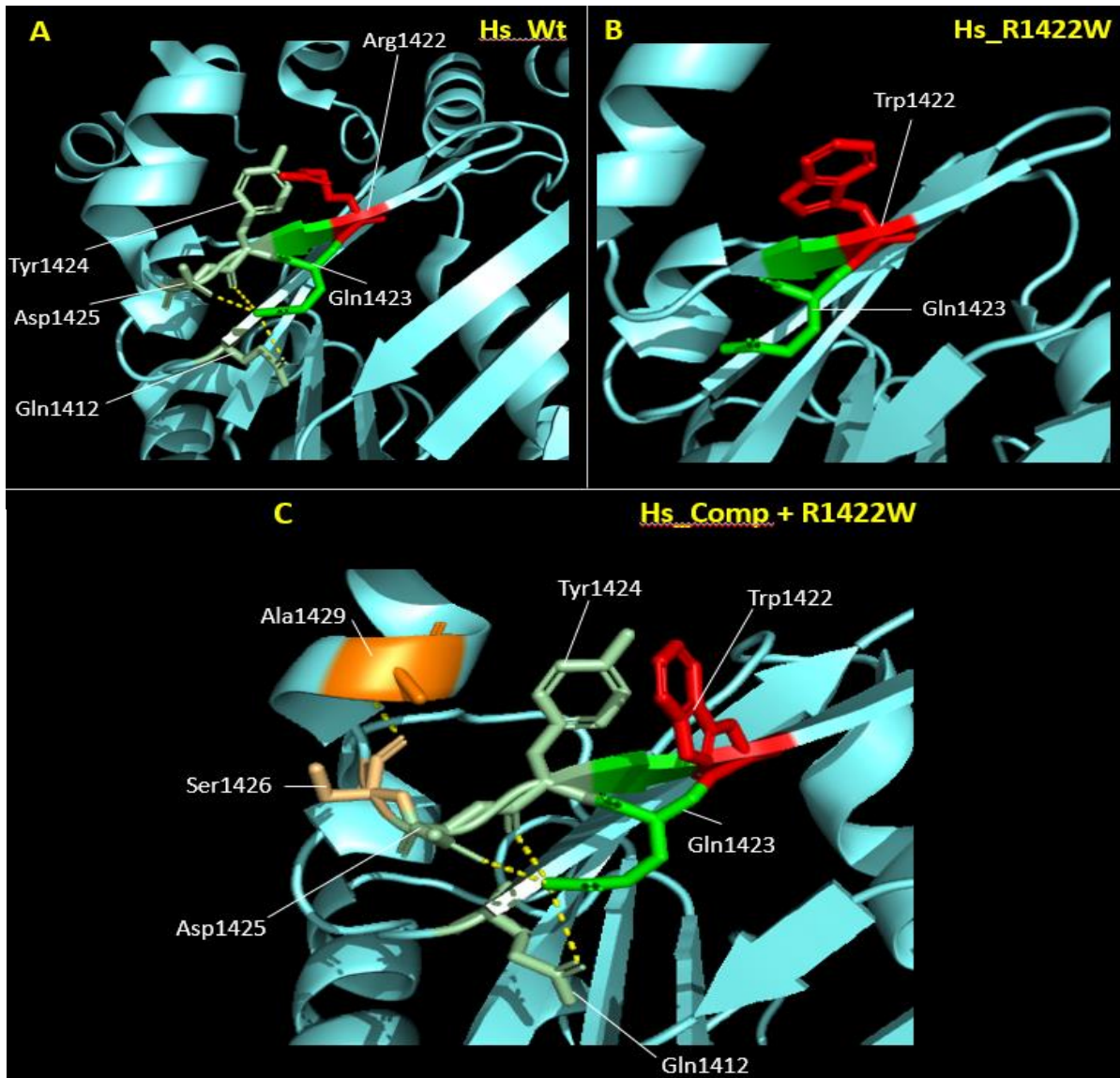


Figure 12: Protein structure models based on CFTR PDB, 5UAK. The position of interest is shown in red, the putative compensatory position in orange and all the positions affected by the deleterious mutation are represented in green. A: Human wild-type; B: Human background containing the deleterious allele Trp1422; C: Human background containing the deleterious mutation and the possible compensation, residue Ala1429.

Although both changes seem to compensate the disruptive effect caused by the deleterious mutation, Ala1429 appears to be the most likely candidate since it is observed in all species and restores the lost H-bonds of the residue 1423, while Asn1420 is a more heterogeneous position and, therefore, not present in all species that have the deleterious mutation as wild-type.

COL4A5

A mutation on human Collagen type IV alpha 5 chain (*COL4A5*), consisting on a change on position 1659 of serine to asparagine has been linked to Alport's Syndrome, which is associated with kidney disease [81, 82].

The neighboring area of residue 1659 is highly conserved within mammals. The single difference between mouse and human is observed at residue 1654, where a valine and a methionine is shown, respectively (Figure 13). Since no experimentally determined 3D structure exists for this protein, it is impossible to infer whether the 1654 is the compensatory site or not. If this is the case, we may infer that the compensatory event could be arising in other organisms, namely Rodents.

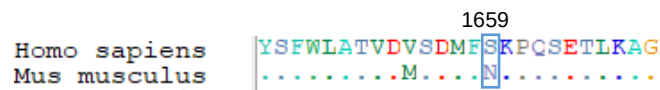


Figure 13: Alignment of COL4A5. The highlighted residue is 1659, the mutation-associated site in humans.

F8

Several changes on the coagulation factor VIII (*F8*) gene have been associated with Hemophilia A, including Leu296Phe [83] and Val1511Ile [84], which is a blood related disease involving coagulation deficiency [35].

The substitution of leucine for phenylalanine on position 296 replaces a highly conserved residue of the protein. Since Phe296 exists in several *Mus* species, the compensatory residue is expected to be present in the three sequences, therefore, the two hypotheses are amino acids Ile285 or Ile320 (Figure 14).

F11

Mutations consisting on a change from a tyrosine to a histidine on residue 511 of coagulation factor XI (*F11*) gene have been linked to the appearance of Factor XI deficiency. This condition, also known as Hemophilia C, is related with clotting problems [86], similar to the previously mentioned *F8*.

The His511 allele appears to be common in several Rodents, especially within *Mus* species (Figure 16).

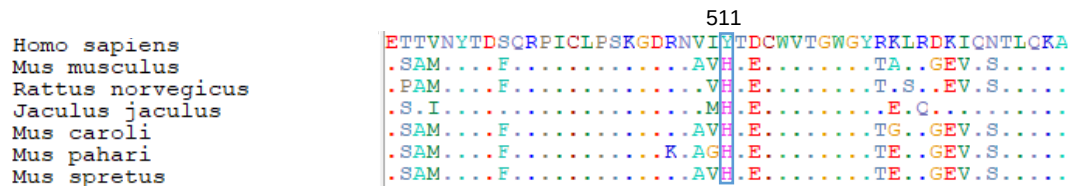


Figure 16: Alignment of *F11*. The highlighted residue is 511, the mutation-associated site in humans.

After modeling analysis, using 5EOK PDB entry [87] as reference, it was possible to observe that the region surrounding residue 511 is not highly affected in terms of the H-bonds when a threonine is replaced by a histidine, where only a connection to residue Glu612 is lost (Figure 17). In this case it is not possible to indicate the most likely compensatory site for this position given the high variability of the amino acids that share the physical space with the residue 511 and due to the minor change in terms of H-bonds.

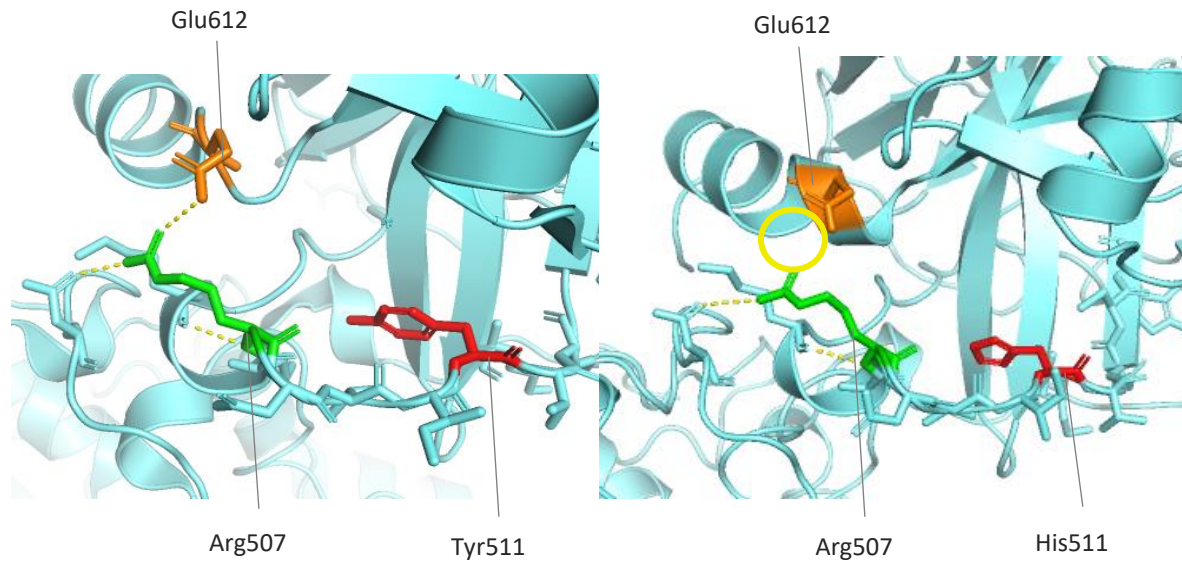


Figure 17: Homology modeling of F11 using 5EOK PDB entry. A: Wild-type. B: Polar contacts (H-bonds) predicted in the human background in the presence of His511mutation. The yellow circle highlights the loss of an H-bond to amino acid Glu612.

GALT

Previous data revealed that a change from a glutamic acid to a lysine in residue 308 of human Galactose-1-phosphate uridylyltransferase (*GALT*) is associated to galactosemia, which is a disorder characterized by a deficiency on the galactose metabolism, that can culminate in liver and renal failure [88, 89].

This position is close to amino acid Asn314, which is associated to Duarte variant (Asn314Asp), also related to galactosemia, as well present as wild-type in mouse, and most likely a polymorphism (because it actually appears more often than classical galactosemia itself) [90]. The analysis of the surroundings of these positions revealed a hypervariable region embedded in a conserved block showing that this human deleterious substitution is not only as the wild-type in mouse, but also in other rodents (Figure 18).

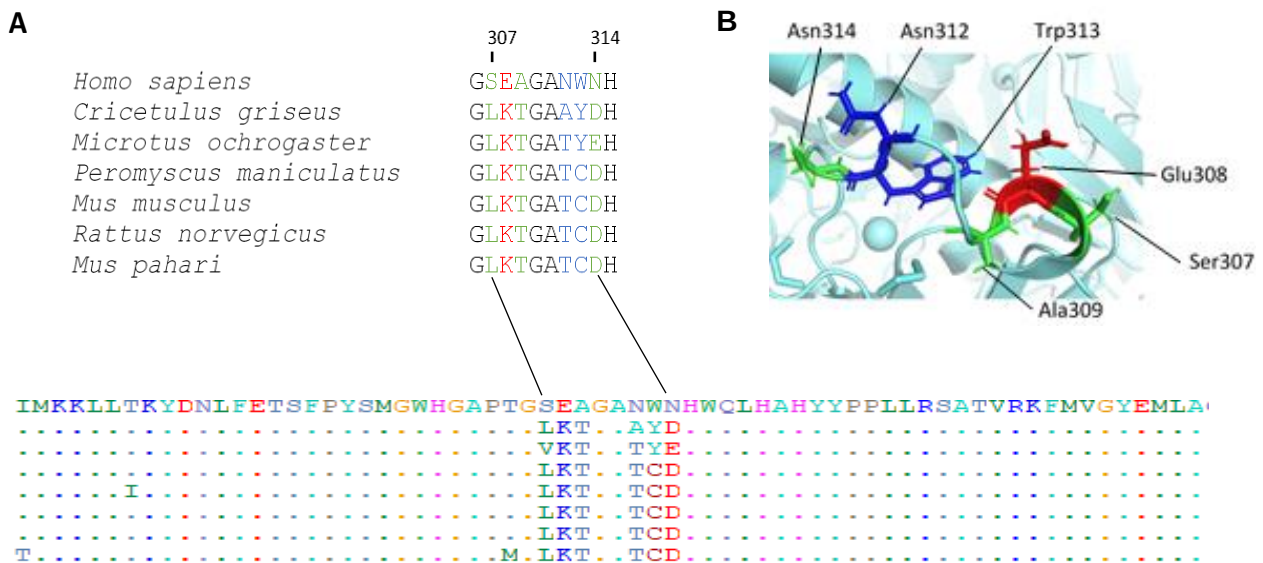


Figure 18: Overview of the target residues of GALT protein. A: Alignment of the human and rodents' GALT protein sequences showing the variations between positions 307 and 314. These positions correspond to a hypervariable region inserted on a highly conserved sequence; B: Detailed view of positions 307-314 on the crystal structure of Human GALT where the human disease-associated mutation (E308K) is lightened in red and possible compensatory residues in green.

The replacement of a negative acid residue to a positive basic is expected to disturb the protein structure, therefore, using PDB entry 5IN3 [91] as template to perform model analysis, it was possible to detect a putative structural compensation for amino acid Lys308 in mouse, that also involves Leu307, Thr309 and Asp314 changes. This complex network of interactions restores the contacts that are lost in the human background carrying a lysine at position 308 (Figure 19).

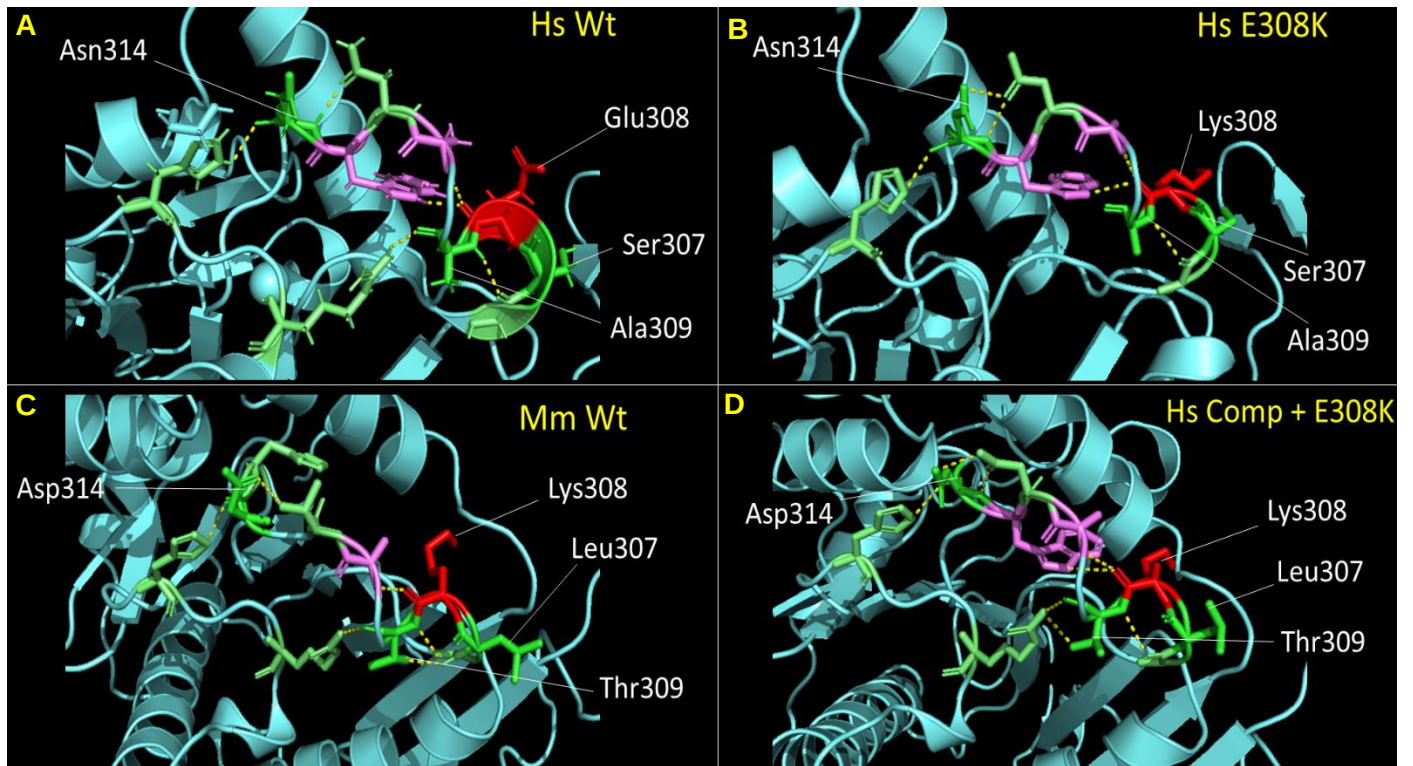


Figure 19: Four models concerning the process of compensation of the Human mutation in residue 308, associated with Galactosemia. A: Human wild-type; B: Human background containing the deleterious allele Lys308; C: Mouse wild-type; D: Human background containing both deleterious mutation (Lys308) and putative compensatory partners (Leu307, Thr309, Asp314).

Since this hipervariable site is present among rodents, but not in any other mammals analysed, it was of great interest to rebuild the evolutionary path of the divergence of Rodents between residues 307 and 314, as shown in Figure 20.

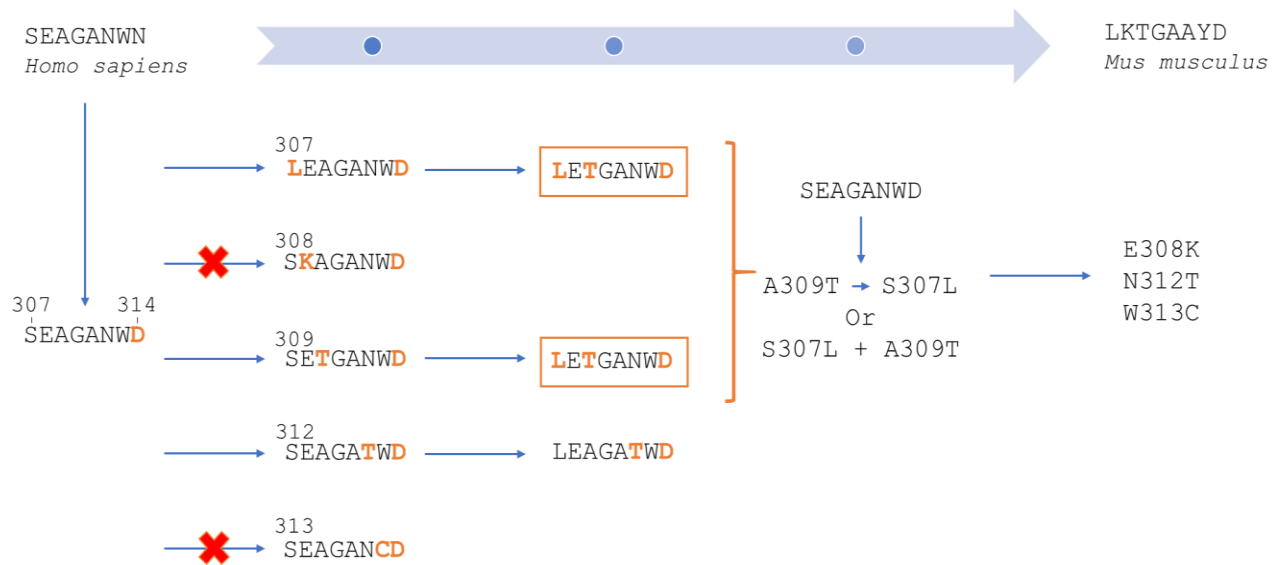


Figure 20: Reconstruction of the evolution of the hypervariable region of *GALT*, in Rodents. The starting point was the conserved sequence considered to be ancestral sequence.

Considering that the Asn314 allele is human-specific, the starting point of this reconstruction contained the same sequence as the human background, but containing the Duarte allele, since it was the ancestral. The first round of models consisted on changing one different variable position in each one. Assuming the principle of parcimony, it was possible to eliminate residues 308 and 313 as being the first changes to occur, since they would induce greater disturbance on protein structure. The second round of models contained the previously modeled variable positions and a new one.

After analyses of every combination, it was possible to understand that the most likely evolutionary paths for this hypervariable motif may have been (a) the appearance of Thr309 allele followed by Leu307, or (b) simultaneous fixation of Leu307 and Thr309, in a background containing the Duarte allele (Asn314), in both cases. After the appearance of these residues, the putative disruptive effect caused by the remaining changes (Lys308, Thr312, Cys313) may have been masked, without inducing severe disruptive disturbance on the protein structure.

IDUA

An exchange from threonine to proline on Alpha-L-iduronidase (*IDUA*) on residue 103 is associated with Mucopolysaccharidosis, which has been associated with a wide range of manifestations such as heart and mental disease [92].

The change from threonine, a polar residue, to proline, a hydrophobic residue could induce great disturbance on the protein structure. This pathogenic mutation appears as wild type in some elements of *Mus* genera and belongs to a neighborhood that is less conserved between mouse and human (Figure 21). Although there is an available structure for this gene, the structure for the surrounding region of the pathological deviation is still not known (PDB entry 4MJ2 [93]), therefore, models derived from this structure, for this site, were not possible to make. On these terms, compensatory residues might comprise one or several of the following spatially close substitutions: Glu95Asp, Val97Ile, Thr99Ala, Gly101Lys and Arg105Gln.

At first, Met108 could also be considered, but since this substitution occurs as Ile108 in two rodent species (*Mus pahari* and *Mus caroli*), it is most likely not related to the compensatory episode, although cannot be discarded without further studies. Changes in residues 95, 99 and 105 are observed across several mammalian species that show different amino acids at position 103, indicating that one, or more, of these positions makes the background permissive for the fixation of different residues, without affecting normal protein function.

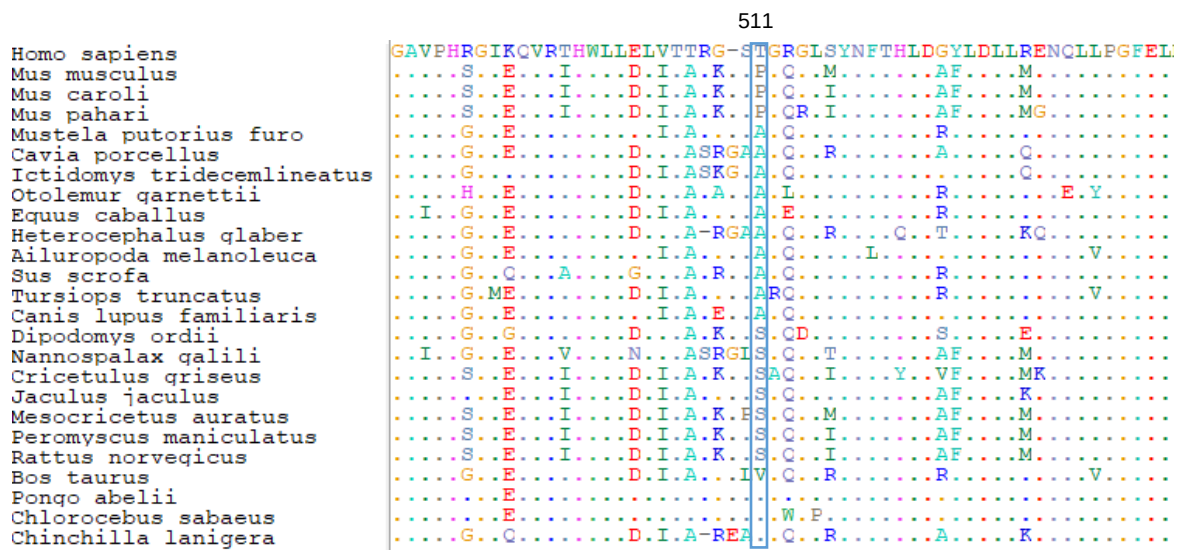


Figure 21: Alignment of *IDUA*. The highlighted residue is 103, the mutation-associated site in humans.

KCNQ1

Previous studies suggested that when a substitution occurs on residue 619 of human potassium voltage-gated channel subfamily Q member 1, where a leucine is replaced by a methionine, Jervell and Lange-Nielsen syndrome is shown. This disease is associated with severe hearing loss and arrhythmia, syncope and sudden death [94, 95].

The available protein models (PDB 3BJ4 and 5VMS) are highly incomplete which turns impossible the structural analysis. Interestingly, the Leu619 position is highly conserved, but it locates in a highly variable region between mammals (Figure 22). Given this high level of polymorphism, it is not possible to suggest a putative compensatory partner although one may expect it to be within this block of sequence.

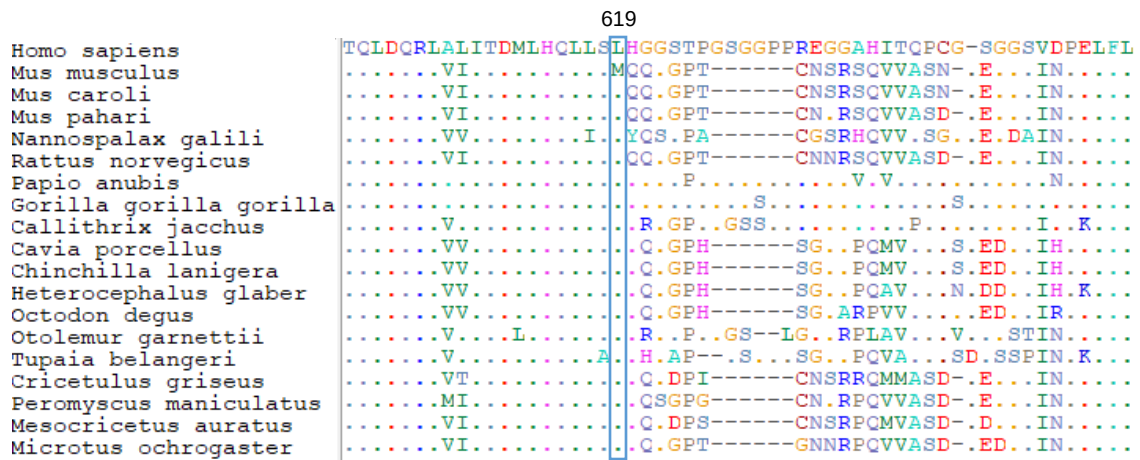


Figure 22: Alignment of *KCNQ1*. The highlighted residue is 619, the mutation-associated site in humans.

LDB3 Z4

Changes from amino acid threonine to alanine on residue 351 of LIM Domain-binding 3 isoform 4 (*LDB3Z4*) is associated with cardiomyopathy, which consists on heart diseases affecting the heart muscle [96].

The surrounding area of residue 351 appears to be highly variable, also the pathogenic mutation appears as a wild type in several organisms, mostly Rodents or members of close orders such as Lagomorphs (Figure 23). In this case the human mutation corresponds to an ancestral allele (Ala351) that is no longer accommodate in the human sequence. The deleterious impact of the mutation in humans is expected to result from the gaining of other specific alleles as is the contiguous Thr350 during human evolution. In other words, the ancestral Ala350-Ala551 background is non-deleterious while the human-specific Thr350-Ala351 revealed to be clinically relevant.

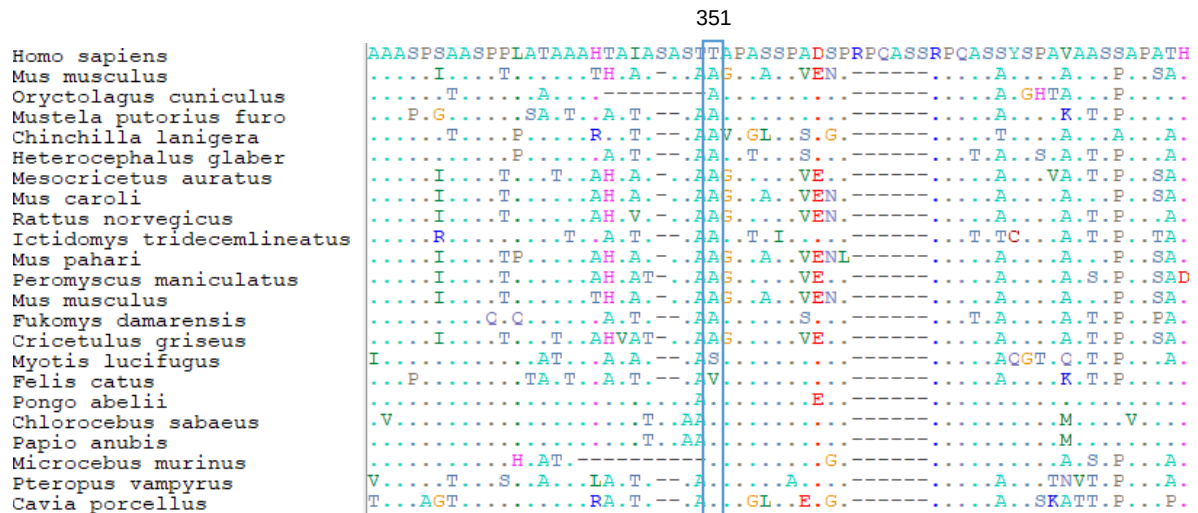


Figure 23: Alignment of *LDB3Z4*. The highlighted residue is 351, the mutation-associated site in humans.

NF1

In residue 491 of Neurofibromin 1 (*NF1*), a change of tyrosine to cysteine induces Neurofibromatosis type 1, where patients show predisposition to develop tumors, and also affects the nervous system, among others [97, 98].

Residues close by to Tyr491Cys are highly conserved within mammals. The substitution of a hydrophobic residue for a polar one could mean high disturbance on the side chain connections. When comparing human and mouse sequences (Figure 24), the closest change in terms of sequential space is the 520 position, however the *Peromyscus maniculatus* does not

carry this putative compensatory site (Ser520). The possibility of a protein-protein interaction instead of a compensatory episode within the same protein could not be discarded. In this context, through the analysis made on The Human Reference Protein Interactome Mapping Project (HuRI) [34], it was detected a possible interaction between *NF1* and Amyloid Beta Precursor Protein (*APP*), which interferes in neural functions [34]. Further studies are suggested in order to evaluate the possibility of this hypothesis.

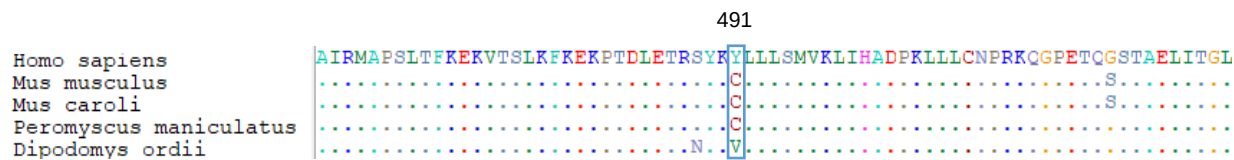


Figure 24: Alignment of *NF1*. The highlighted residue is 491, the mutation-associated site in humans.

PTS

A substitution of valine for alanine, on residue 103 of human 6-pyruvoyl-tetrahydropterin synthase (*PTS*) is associated with hyperphenylalaninemia, which is a condition associated with high concentrations of phenylalanine in the bloodstream [99].

Structural analysis using PDB entry 3I2B, did not reveal any disturbance when a valine is replaced by an alanine in terms of polar contacts (Figure 25). It shall, however, be recalled that the structural impact of any amino acid replacement may involve physical disturbances other than the polar contacts [42]. If this is the case, the structural analysis cannot *per se* be informative on what concerns the most likely compensatory site(s). In this case, the analysis of the multiple sequence alignment couldn't also be of further support since there are multiple positions in which the *Mus* sequence differs from the human.

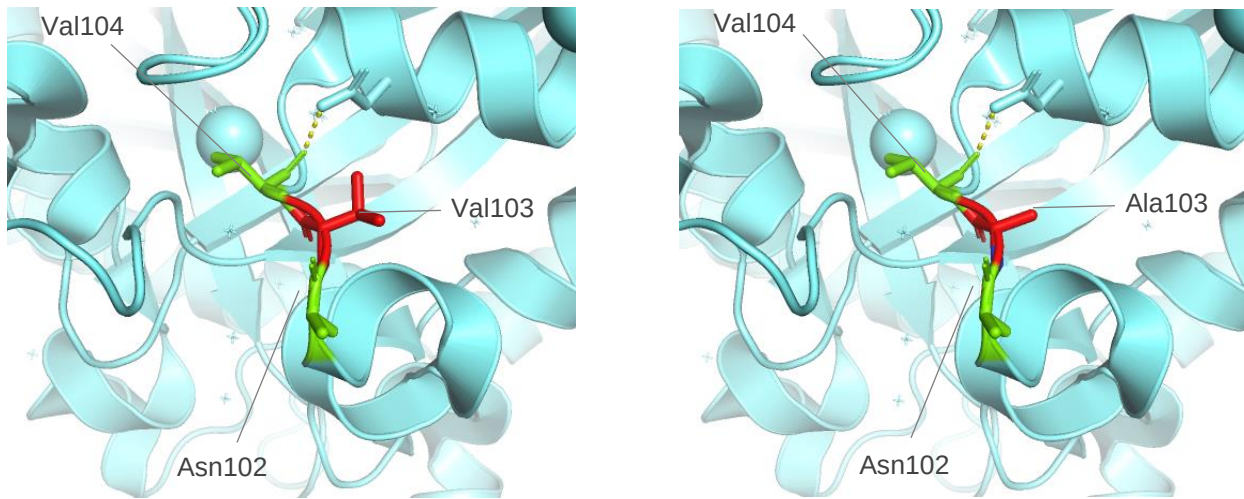


Figure 25: Protein structure of PTS: (A) Human wild-type is shown. (B) Human background with the deleterious mutation (Val103Ala) inserted.

In these terms, it is not possible to suggest a putative compensatory site without further studies to understand the damage caused by the pathogenic mutation, although the Met95Leu change could be the most suitable candidate as it is the closest to the deleterious mutation site.

RET

A mutation on residue 480 of Ret Proto-Oncogene (*RET*) consisting on a change of glutamic acid to lysine induces Hirschsprung's disease, a developmental disorder characterized by the absence of ganglion cells in lower digestive tract, related to the nervous system and connected to thyroid cancer [100, 101].

The substitution of an acid, negative residue to a basic, positive one can affect tremendously protein structure due to the properties of the different amino acids. The surrounding area of residue 480 is highly polymorphic when compared to mouse (Figure 26). Since some other mammals present a Glu480Gln substitution, we expect that the compensatory episode does also exist in these species. Specifically, Glu477Lys, His482Gln and Met484Thr are the closest changes in these sequences. The mutation also occurs as wild-type in *Mus Pahari*, pointing to the possibility of being ancestral to the divergence of *Mus* species.

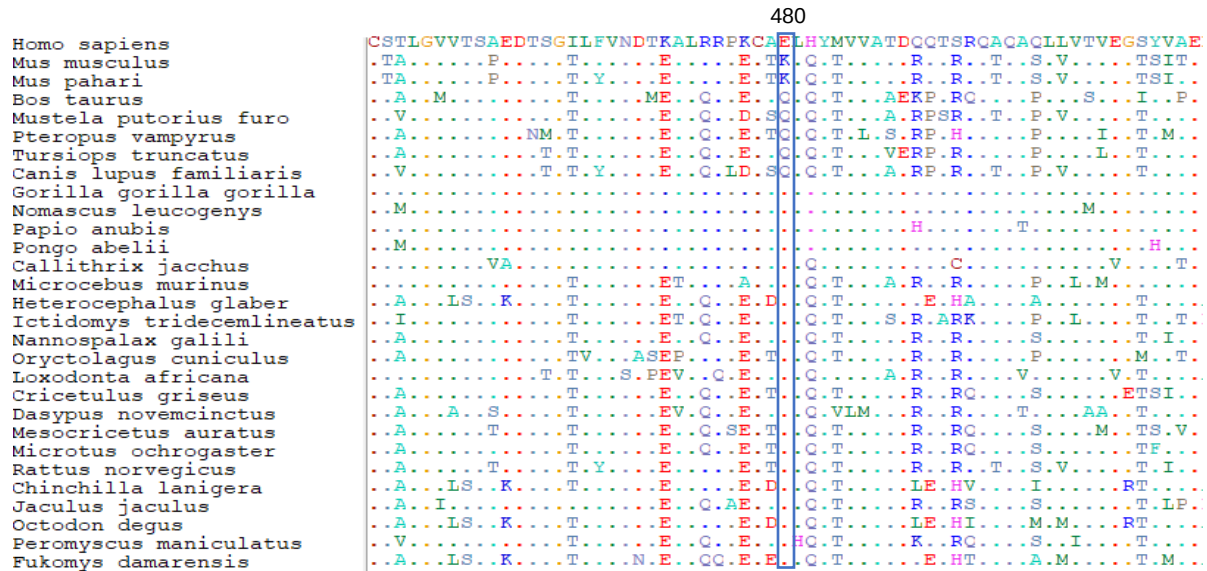


Figure 26: Alignment of RET. The highlighted residue is 480, the mutation-associated site in humans.

SGCA

When a mutation occurs on residue 30 of Sarcoglycan alpha, changing from proline to leucine, a phenotype of muscular dystrophy is observed in humans [102, 103].

The closest variable positions of this possible *Mus musculus* specific compensatory episode are Gly18Arg, Glu22Lys, Thr39Pro, Asp41Glu, Glu43Ala (Figure 27). In addition, protein-protein interactions should not be discarded because according to the HuRI, SGCA can be associated with Small Glutamine Rich Tetratricopeptide Repeat Containing Alpha (SGTA) and Beta (SGTB), that regulate ATPase activity [104].

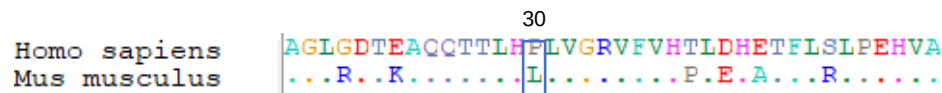


Figure 27: Alignment of SGCA. The highlighted residue is 30, the mutation-associated site in humans.

TG

A mutation from a proline to an arginine on residue 2202 of Thyroglobulin (TG) induces thyroid dyshormonogenesis, which is an endocrine disorder, causing cognitive and motor deficits [105, 106].

The region embracing residue 2202 of TG is a highly variable throughout several mammalian species being, therefore, highly difficult to suggest a compensatory hypothesis without further studies. This substitution appears as wild-type in several rodents, namely on two *Mus* species. Interestingly, a third species, *Mus Pahari*, shows a different substitution, Pro2202Gln, which might indicate a different evolutionary trajectory (Figure 28).

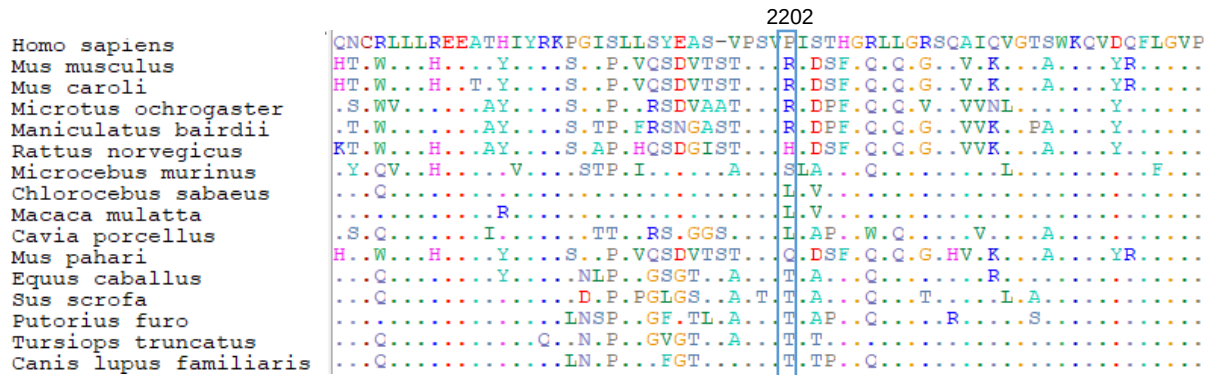


Figure 28: Alignment of TG. The highlighted residue is 2202, the mutation-associated site in humans.

Although highly different between organisms, the surrounding area of the pathological mutation is very similar within the same order. Since this gene is related to hormone regulation and to the immune system [107], different organisms could be exposed to different environmental pressures, which might be driven gene selection on the different orders.

Concluding remarks

In this study, we used multiple sequence alignments and structural data to infer the most likely compensatory sites for human deleterious mutations found as the wild-type allele in the mouse (*Mus musculus*) genome. Although limited in number, we were able to gather important conclusions in what concerns these types of mutations that are clinically relevant, and which are summarized in Table 4.

Most of human mutations could not be detected as pathogenic by either of the *in silico* tools that are widely used in genetics (Polyphen and SIFT). This is a critical point that indicated that the sequence conservation is indeed the most important feature underlying these tools [38]. If the human deleterious allele is found as wild-type in other species the conservation scores will be lower and therefore the final score will also be underestimated.

Table 4: Summary of the results obtained in section 1 and 2. Data comprises molecular analysis, the predicted effect of the mutation by prediction software, the existence of protein structure, the putative compensatory episode, placement of the mutation on the protein structure, location of the putative compensation when compared to the deleterious mutation (being neighborhood considered 10 amino acid distant from the PD) and the number of residues involved in the compensatory episode.

Gene		ACBD5	ASAH1	ATP7B	CFTR	COL4A5	F8 (A)	F8 (B)	F11	GALT	IDUA	KCNQ1	LDB3 Z4	NF1	PTS	RET	SGCA	TG
Location of the mutation on protein structure	Damaging effect on PolyPhen or SIFT			✓		✓	✓											
	Protein Structure	✓		✓	✓		✓		✓	✓	✓	✓		✓	✓	✓		
	Putative compensation				✓					✓								
Location of the putative compensation	Surface (exposed)	✓		✓	✓					✓								
	Buried (harder to access)						✓											
	Neighborhood (of original mutation)				✓					✓								
Number of residues involved	Distant (from original mutation)																	
	One				✓													
	More than one									✓								

The fact that only 3 of the 17 studied mutations are shown to be possibly damaging in SIFT or PolyPhen shows itself as an interesting result. It would be expected that all mutations (or most of them) would appear as damaging, since they are associated to human diseases, some of which with a severe phenotype. The obtained data could be possibly explained by the occurrence of these compensatory episodes.

As expected, in general, obtained data suggests that compensatory mutations seem to be favoring intramolecular epistasis by arising in the same protein and on the proximity of the deleterious mutation. We could also notice that they often locate in the protein surface, in accordance to previous data [36], being, in this case, easier to access and, therefore, easily compensated [108]. Compensatory episodes consist mainly in simpler forms of compensation, involving a reduced number of residues although, in some cases, interactions between two or

more residues are determinant to the success of the epistatic interactions between a conditional deleterious mutation and the physically interacting partners [109]. Finally, intermolecular interactions for some of these compensated mutations cannot be ruled without experimental validation.

The results provided by this thesis can promote the development of better diagnostic techniques, since they help to predict the effect of deleterious mutation in human proteins. It is also expected that, on a different perspective, these results open the doors of the evolutionary history of groups or individuals, which is known as biogeographical ancestry [110], and will be of great usage in criminal investigations, aiding to reduce the pool of suspects and narrowing the possibilities in forensic investigations.

Future Perspectives

It is expected that these results give an overlook on the possible species-specific compensatory episodes in mice, aiming to lead to a better understanding of the evolutionary process of mutations and interacting residues, through the analysis of multiple sequence alignments and 3D structure. The follow-up of this work may be summarized as follow:

1. Protein functional assays

Functional essays are of great importance to understand not only the disturbance induced by a mutation on protein function. It should be performed to evaluate the severity of the mutation on protein function and to confirm the predicted compensatory allele.

2. Bioinformatics tool development

Due to the high number of deleterious mutations, it would be time-consuming and technical demanding to perform functional essays in all of them. Therefore, in the future we plan to develop new bioinformatics tools to help the assessment of the deleterious effect of compensated mutations.

3. Stabilization of protein structure

By understanding how compensatory episodes occur, it will be possible to infer how these compensations occur. Based on this, it will be possible, as a long-term perspective to try to mimic the effect of the compensatory event. This can be made with the help of proteins that stabilize protein structures, known as molecular chaperones.

These next steps might allow to give an insight of possible compensatory mechanisms and could permit to produce, in a near future, manipulated mice strains using CRISP/Cas9 techniques, for example, expressing a disease with minor phenotypical traits, or even neutralize completely the pathological effect and it may, eventually, be extrapolated to humans.

References

* Important

** Very Important

1. Suzuki DT, Griffiths AJ: **An introduction to genetic analysis**: WH Freeman and Company.; 1976.
2. Antonio A, Bruce B: **Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation**, vol. 2: World Scientific; 2016.
3. Parson W, Pegoraro K, Niederstätter H, Föger M, Steinlechner M: **Species identification by means of the cytochrome b gene**. *International Journal of Legal Medicine* 2000, **114**(1-2):23-28.
4. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH: **Use of DNA barcodes to identify flowering plants**. *Proceedings of the National Academy of Sciences* 2005, **102**(23):8369-8374.
5. Hoffmaster AR, Fitzgerald CC, Ribot E, Mayer LW, Popovic T: **Molecular subtyping of Bacillus anthracis and the 2001 bioterrorism-associated anthrax outbreak, United States**. *Emerging infectious diseases* 2002, **8**(10):1111.
6. Cowperthwaite MC, Bull J, Meyers LA: **From bad to good: Fitness reversals and the ascent of deleterious mutations**. *PLoS computational biology* 2006, **2**(10):e141.
7. Charlesworth B, Charlesworth D: **Some evolutionary consequences of deleterious mutations**. *Genetica* 1998, **102**:3.
8. Goodwin W, Linacre A, Hadi S: **An introduction to forensic genetics**, vol. 2: John Wiley & Sons; 2011.

9. Jeffreys AJ, Allen MJ, Armour JA, Collick A, Dubrova Y, Fretwell N, Guram T, Jobling M, May CA, Neil DL: **Mutation processes at human minisatellites**. *Electrophoresis* 1995, **16**(1):1577-1585.
10. Ruitberg CM, Reeder DJ, Butler JM: **STRBase: a short tandem repeat DNA database for the human identity testing community**. *Nucleic acids research* 2001, **29**(1):320-322.
11. Gill P, Sparkes R, Fereday L, Werrett DJ: **Report of the European Network of Forensic Science Institutes (ENSFI): formulation and testing of principles to evaluate STR multiplexes**. *Forensic Science International* 2000, **108**(1):1-29.
12. Fan J-B, Chen X, Halushka MK, Berno A, Huang X, Ryder T, Lipshutz RJ, Lockhart DJ, Chakravarti A: **Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays**. *Genome Research* 2000, **10**(6):853-860.
13. Senge T, Madea B, Junge A, Rothschild MA, Schneider PM: **STRs, mini STRs and SNPs; A comparative study for typing degraded DNA**. *Legal Medicine* 2011, **13**(2):68-74.
14. Dixon L, Murray C, Archer E, Dobbins A, Koumi P, Gill P: **Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes**. *Forensic science international* 2005, **154**(1):62-77.
15. Sobrino B, Brión M, Carracedo A: **SNPs in forensic genetics: a review on SNP typing methodologies**. *Forensic science international* 2005, **154**(2-3):181-194.
16. Frudakis T, Venkateswarlu K, Thomas M, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK: **A classifier for the SNP-based inference of ancestry**. *Journal of Forensic Sciences* 2003, **48**(4):771-782.
17. Keats BJ, Sherman SL: **Population genetics**. In: *Emery and Rimoin's Principles and Practice of Medical Genetics*. Elsevier; 2013: 1-12.
18. Hampel H, Bennett RL, Buchanan A, Pearlman R, Wiesner GL: **A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: referral indications for cancer predisposition assessment**. *Genetics in Medicine* 2015, **17**(1):70.

19. Lehner B: **Molecular mechanisms of epistasis within and between genes.** *Trends in Genetics* 2011, **27**(8):323-331.
20. Consortium IHGS: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860.
21. Green MR, Sambrook J: **Isolation and Quantification of DNA.** *Cold Spring Harbor Protocols* 2018, **2018**(6):pdb. top093336.
22. Ming S, Tian-Rui X, Ce-Shi C: **The big bang of genome editing technology: development and application of the CRISPR/Cas9 system in disease animal models.** *Zoological Research* 2016, **37**(4):191.
23. Meister G, Tuschl T: **Mechanisms of gene silencing by double-stranded RNA.** *Nature* 2004, **431**(7006):343.
24. Smith M: **DNA Sequence Analysis in Clinical Medicine, Proceeding Cautiously.** *Frontiers in molecular biosciences* 2017, **4**:24.
25. * Gu J, Bourne PE: **Structural bioinformatics**, vol. 44: John Wiley & Sons; 2009.
26. Taylor WR: **The classification of amino acid conservation.** *Journal of theoretical Biology* 1986, **119**(2):205-218.
27. Bondi A: **van der Waals volumes and radii.** *The Journal of physical chemistry* 1964, **68**(3):441-451.
28. De Juan D, Pazos F, Valencia A: **Emerging methods in protein co-evolution.** *Nature Reviews Genetics* 2013, **14**(4):nrg3414.
29. * Ferrer-Costa C, Orozco M, de la Cruz X: **Characterization of compensated mutations in terms of structural and physico-chemical properties.** *Journal of molecular biology* 2007, **365**(1):249-256.
30. Wang X, Minasov G, Shoichet BK: **Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs.** *Journal of molecular biology* 2002, **320**(1):85-95.

31. ** Cordell HJ: **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.** *Human Molecular Genetics* 2002, **11**(20):2463-2468.
32. Adams DJ, Ackert-Bicknell CL: **Genetic regulation of bone strength: a review of animal model studies.** *BoneKEy reports* 2015, **4**.
33. Bayat A: **Bioinformatics.** *Bmj* 2002, **324**(7344):1018-1022.
34. Aydin D, Weyer SW, Müller UC: **Functions of the APP gene family in the nervous system: insights from mouse models.** *Experimental Brain Research* 2012, **217**(3):423-434.
35. Antonarakis S, Rossiter J, Young M, Horst J, de Moerloose P, Sommer S, Ketterling R, Kazazian HJ, Negrier C, Vinciguerra C: **Factor VIII gene inversions in severe hemophilia A: results of an international consortium study.** *Blood* 1995, **86**(6):2206-2212.
36. * Barešić A, Hopcroft LE, Rogers HH, Hurst JM, Martin AC: **Compensated pathogenic deviations: analysis of structural effects.** *Journal of molecular biology* 2010, **396**(1):19-30.
37. * Azevedo L, Suriano G, van Asch B, Harding RM, Amorim A: **Epistatic interactions: how strong in disease and evolution?** *Trends in Genetics* 2006, **22**(11):581-585.
38. ** Azevedo L, Mort M, Costa AC, Silva RM, Quelhas D, Amorim A, Cooper DN: **Improving the in silico assessment of pathogenicity for compensated variants.** *European Journal of Human Genetics* 2016, **25**(1):2-7.
39. * Xu J, Zhang J: **Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis.** *Molecular biology and evolution* 2014, **31**(7):1787-1792.
40. * Zhang G, Pei Z, Krawczak M, Ball EV, Mort M, Kehrer-Sawatzki H, Cooper DN: **Triangulation of the human, chimpanzee, and Neanderthal genome sequences identifies potentially compensated mutations.** *Human mutation* 2010, **31**(12):1286-1293.

41. * Suriano G, Azevedo L, Novais M, Boscolo B, Seruca R, Amorim A, Ghibaudi EM: **In vitro demonstration of intra-locus compensation using the ornithine transcarbamylase protein as model.** *Human molecular genetics* 2007, **16**(18):2209-2214.
42. ** Ivankov DN, Finkelstein AV, Kondrashov FA: **A structural perspective of compensatory evolution.** *Current opinion in structural biology* 2014, **26**:104-112.
43. * Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, De Koning A, Dokholyan NV, Echave J: **The interface of protein structure, protein biophysics, and molecular evolution.** *Protein Science* 2012, **21**(6):769-785.
44. Lorch M, Mason JM, Sessions RB, Clarke AR: **Effects of mutations on the thermodynamics of a protein folding reaction: implications for the mechanism of formation of the intermediate and transition states.** *Biochemistry* 2000, **39**(12):3480-3485.
45. DePristo MA, Weinreich DM, Hartl DL: **Missense meanderings in sequence space: a biophysical view of protein evolution.** *Nature Reviews Genetics* 2005, **6**(9):678-687.
46. ** Azevedo L: **Compensatory Evolution in Disease-Associated Genes.** *eLS*.
47. Giacomotto J, Ségalat L: **High-throughput screening and small animal models, where are we?** *British journal of pharmacology* 2010, **160**(2):204-216.
48. Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**(6915):520-562.
49. Moore JH: **A global view of epistasis.** *Nature genetics* 2005, **37**(1):13-14.
50. Lu JT, Campeau PM, Lee BH: **Genotype-phenotype correlation: promiscuity in the era of next-generation sequencing.** *Obstetrical & Gynecological Survey* 2014, **69**(12):728-730.
51. Roy AK: **Bioinformatics and its application.**

52. Altman RB, Dugan JM: **Defining bioinformatics and structural bioinformatics.** *Structural Bioinformatics* 2003, **44**:1-14.
53. Kent WJ: **BLAT—the BLAST-like alignment tool.** *Genome research* 2002, **12**(4):656-664.
54. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins: Structure, Function, and Bioinformatics* 1991, **9**(1):56-68.
55. Eswar N, Webb B, Marti-Renom MA, Madhusudhan M, Eramian D, Shen My, Pieper U, Sali A: **Comparative protein structure modeling using Modeller.** *Current protocols in bioinformatics* 2006, **15**(1):5.6. 1-5.6. 30.
56. Schwede T, Kopp J, Guex N, Peitsch MC: **SWISS-MODEL: an automated protein homology-modeling server.** *Nucleic acids research* 2003, **31**(13):3381-3385.
57. DeLano WL: **The PyMOL molecular graphics system.** <http://www.pymol.org> 2002.
58. Eisenberg D, Lüthy R, Bowie JU: **VERIFY3D: assessment of protein models with three-dimensional profiles.** In: *Methods in enzymology*. vol. 277: Elsevier; 1997: 396-404.
59. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic acids research* 2005, **33**(suppl_1):D514-D517.
60. ** Barešić A, Martin AC: **Compensated pathogenic deviations.** *Biomolecular concepts* 2011, **2**(4):281-292.
61. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG: **Ensembl 2018.** *Nucleic acids research* 2017, **46**(D1):D754-D761.
62. Sievers F, Higgins DG: **Clustal Omega, accurate alignment of very large numbers of sequences.** In: *Multiple sequence alignment methods*. Springer; 2014: 105-116.
63. Hall T, Biosciences I, Carlsbad C: **BioEdit: an important software for molecular biology.** *GERF Bull Biosci* 2011, **2**(1):60-61.

64. Kumar S, Stecher G, Tamura K: **MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets**. *Molecular biology and evolution* 2016, **33**(7):1870-1874.
65. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees**. *Molecular biology and evolution* 1987, **4**(4):406-425.
66. Adzhubei I, Jordan DM, Sunyaev SR: **Predicting functional effect of human missense mutations using PolyPhen-2**. *Current protocols in human genetics* 2013, **76**(1):7.20. 21-27.20. 41.
67. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic acids research* 2003, **31**(13):3812-3814.
68. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG: **Primer3—new capabilities and interfaces**. *Nucleic acids research* 2012, **40**(15):e115-e115.
69. Kibbe WA: **OligoCalc: an online oligonucleotide properties calculator**. *Nucleic acids research* 2007, **35**(suppl_2):W43-W46.
70. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Research* 2000, **28**(1):235-242.
71. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V: **A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core**. *Journal of molecular biology* 2018, **430**(15):2237-2243.
72. Punzo F, Mientjes E, Rohe C, Scianguetta S, Amendola G, Oostra B, Bertoli-Avella A, Perrotta S: **A mutation in the acyl-coenzyme A binding domain-containing protein 5 gene (ACBD5) identified in autosomal dominant thrombocytopenia**. *Journal of Thrombosis and Haemostasis* 2010, **8**(9):2085-2087.
73. Izak M, Bussel JB: **Management of thrombocytopenia**. *F1000prime reports* 2014, **6**.
74. Zhang Z, Mandal AK, Mital A, Popescu N, Zimonjic D, Moser A, Moser H, Mukherjee AB: **Human acid ceramidase gene: novel mutations in Farber disease**. *Molecular genetics and metabolism* 2000, **70**(4):301-309.

75. Nasreen A, Sunil AK, Pulak MR, Sujit BK: **Clinical diagnosis of Farber's disease-a rare case report.** *Int J Health Sci Res* 2017, **7**(7):388-390.
76. Lee BH, Kim JH, Lee SY, Jin HY, Kim KJ, Lee JJ, Park JY, Kim GH, Choi JH, Kim KM: **Distinct clinical courses according to presenting phenotypes and their correlations to ATP7B mutations in a large Wilson's disease cohort.** *Liver International* 2011, **31**(6):831-839.
77. Huster D, Kühne A, Bhattacharjee A, Raines L, Jantsch V, Noe J, Schirrmeister W, Sommerer I, Sabri O, Berr F: **Diverse functional properties of Wilson disease ATP7B variants.** *Gastroenterology* 2012, **142**(4):947-956. e945.
78. Banci L, Bertini I, Cantini F, Rosenzweig AC, Yatsunyk LA: **Metal Binding Domains 3 and 4 of the Wilson Disease Protein: Solution Structure and Interaction with the Copper(I) Chaperone HAH1.** *Biochemistry* 2008, **47**(28):7423-7429.
79. Riordan JR: **CFTR function and prospects for therapy.** *Annu Rev Biochem* 2008, **77**:701-726.
80. Liu F, Zhang Z, Csanády L, Gadsby DC, Chen J: **Molecular Structure of the Human CFTR Ion Channel.** *Cell* 2017, **169**(1):85-95.e88.
81. Wang F, Zhao D, Ding J, Zhang H, Zhang Y, Yu L, Xiao H, Yao Y, Zhong X, Wang S: **Skin biopsy is a practical approach for the clinical diagnosis and molecular genetic analysis of X-linked Alport's syndrome.** *The Journal of Molecular Diagnostics* 2012, **14**(6):586-593.
82. Kruegel J, Rubel D, Gross O: **Alport syndrome—insights from basic and clinical research.** *Nature Reviews Nephrology* 2013, **9**(3):170.
83. Salviato R, Belvini D, Radossi P, Sartori R, Pierobon F, Zanotto D, Zanon E, Castaman G, Gandini G, Tagariello G: **F8 gene mutation profile and ITT response in a cohort of Italian haemophilia A patients with inhibitors.** *Haemophilia* 2007, **13**(4):361-372.

84. Lin S-Y, Su Y-N, Hung C-C, Tsay W, Chiou S-S, Chang C-T, Ho H-N, Lee C-N: **Mutation spectrum of 122 hemophilia A families from Taiwanese population by LD-PCR, DHPLC, multiplex PCR and evaluating the clinical application of HRM.** *BMC medical genetics* 2008, **9**(1):53.
85. Shen BW, Spiegel PC, Chang C-H, Huh J-W, Lee J-S, Kim J, Kim Y-H, Stoddard BL: **The tertiary structure and domain organization of coagulation factor VIII.** *Blood* 2008, **111**(3):1240-1247.
86. Zivelin A, Bauduer F, Ducout L, Peretz H, Rosenberg N, Yatuv R, Seligsohn U: **Factor XI deficiency in French Basques is caused predominantly by an ancestral Cys38Arg mutation in the factor XI gene.** *Blood* 2002, **99**(7):2448-2454.
87. Wong SS, Østergaard S, Hall G, Li C, Williams PM, Stennicke H, Emsley J: **A novel DFP tripeptide motif interacts with the coagulation factor XI apple 2 domain.** *Blood* 2016, **127**(23):2915-2923.
88. Elsas L, Langley S, Steele E, Evinger J, Fridovich-Keil J, Brown A, Singh R, Fernhoff P, Hjelm L, Dembure P: **Galactosemia: a strategy to identify new biochemical phenotypes and molecular genotypes.** *American journal of human genetics* 1995, **56**(3):630.
89. Bosch AM: **Classical galactosaemia revisited.** *Journal of inherited metabolic disease* 2006, **29**(4):516-525.
90. McCorvie TJ, Timson DJ: **Structural and molecular biology of type I galactosemia: Disease-associated mutations.** *IUBMB life* 2011, **63**(11):949-954.
91. McCorvie TJ, Kopec J, Pey AL, Fitzpatrick F, Patel D, Chalk R, Shrestha L, Yue WW: **Molecular basis of classic galactosemia from the structure of human galactose 1-phosphate uridylyltransferase.** *Human Molecular Genetics* 2016, **25**(11):2234-2244.
92. Li P, Wood T, Thompson JN: **Diversity of mutations and distribution of single nucleotide polymorphic alleles in the human α -L-iduronidase (IDUA) gene.** *Genetics in Medicine* 2002, **4**(6):420.

93. Bie H, Yin J, He X, Kermode AR, Goddard-Borger ED, Withers SG, James MNG: **Insights into Mucopolysaccharidosis I from the structure and action of α -L-Iduronidase.** *Nature chemical biology* 2013, **9**(11):739-745.
94. Tester DJ, Will ML, Haglund CM, Ackerman MJ: **Compendium of cardiac channel mutations in 541 consecutive unrelated patients referred for long QT syndrome genetic testing.** *Heart Rhythm* 2005, **2**(5):507-517.
95. Coto E, García-Fernández FJ, Calvo D, Salgado-Aranda R, Martín-González J, Alonso B, Iglesias S, Gómez J: **An elderly Jervell and Lange-Nielsen patient heterozygous compound for two new KCNQ1 mutations.** *American Journal of Medical Genetics Part A* 2017, **173**(3):749-752.
96. Marziliano N, Mannarino S, Nespoli L, Diegoli M, Pasotti M, Malattia C, Grasso M, Pilotto A, Porcu E, Raisaro A: **Barth syndrome associated with compound hemizyosity and heterozygosity of the TAZ and LDB3 genes.** *American Journal of Medical Genetics Part A* 2007, **143**(9):907-915.
97. Fahsold R, Hoffmeyer S, Mischung C, Gille C, Ehlers C, Kucukceylan N, Abdel-Nour M, Gewies A, Peters H, Kaufmann D: **Minor lesion mutational spectrum of the entire NF1 gene does not explain its high mutability but points to a functional domain upstream of the GAP-related domain.** *The American Journal of Human Genetics* 2000, **66**(3):790-818.
98. Hirbe AC, Gutmann DH: **Neurofibromatosis type 1: a multidisciplinary approach to care.** *The Lancet Neurology* 2014, **13**(8):834-843.
99. Leuzzi V, Carducci C, Carducci C, Pozzessere S, Burlina A, Cerone R, Concolino D, Donati M, Fiori L, Meli C: **Phenotypic variability, neurological outcome and genetics background of 6-pyruvoyl-tetrahydropterin synthase deficiency.** *Clinical genetics* 2010, **77**(3):249-257.
100. Julies MG, Moore SW, Kotze MJ, du Plessis L: **Novel RET mutations in Hirschsprung's disease patients from the diverse South African population.** *European Journal of Human Genetics* 2001, **9**(6):419.

101. So M-T, Leon TY-Y, Cheng G, Tang CS-M, Miao X-P, Cornes BK, Ngo DN, Cui L, Ngan ES-W, Lui VC-H: **RET mutational spectrum in Hirschsprung disease: evaluation of 601 Chinese patients.** *PLoS One* 2011, **6**(12):e28986.
102. Carrie A, Piccolo F, Leturcq F, De Toma C, Azibi K, Beldjord C, Vallat J, Merlini L, Voit T, Sewry C: **Mutational diversity and hot spots in the alpha-sarcoglycan gene in autosomal recessive muscular dystrophy (LGMD2D).** *Journal of medical genetics* 1997, **34**(6):470-475.
103. Reddy HM, Hamed SA, Lek M, Mitsuhashi S, Estrella E, Jones MD, Mahoney LJ, Duncan AR, Cho Ka, Macarthur DG: **Homozygous nonsense mutation in SGCA is a common cause of limb-girdle muscular dystrophy in Assiut, Egypt.** *Muscle & nerve* 2016, **54**(4):690-695.
104. Philp LK, Butler MS, Hickey TE, Butler LM, Tilley WD, Day TK: **SGTA: A New Player in the Molecular Co-Chaperone Game.** *Hormones and Cancer* 2013, **4**(6):343-357.
105. Citterio CE, Machiavelli GA, Miras MB, Gruñeiro-Papendieck L, Lachlan K, Sobrero G, Chiesa A, Walker J, Muñoz L, Testa G: **New insights into thyroglobulin gene: molecular analysis of seven novel mutations associated with goiter and hypothyroidism.** *Molecular and cellular endocrinology* 2013, **365**(2):277-291.
106. Siffo S, Adrover E, Citterio CE, Miras MB, Balbi VA, Chiesa A, Weill J, Sobrero G, González VG, Papendieck P: **Molecular analysis of thyroglobulin mutations found in patients with goiter and hypothyroidism.** *Molecular and cellular endocrinology* 2017.
107. Webster Marketon JI, Glaser R: **Stress hormones and immune function.** *Cellular Immunology* 2008, **252**(1):16-26.
108. * Storz JF: **Compensatory mutations and epistasis for protein function.** *Current Opinion in Structural Biology* 2018, **50**:18-25.
109. ** Kondrashov AS, Sunyaev S, Kondrashov FA: **Dobzhansky–Muller incompatibilities in protein evolution.** *Proceedings of the National Academy of Sciences* 2002, **99**(23):14878-14883.

110. Kayser M, De Knijff P: **Improving human forensics through advances in genetics, genomics and molecular biology.** *Nature Reviews Genetics* 2011, **12**(3), 179.