U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

PhD

PhD

PhD

PhD

PhD

3.º
CICLO

FCUP
2024

U.PORTO

Mutational change of X-linked disease genes in both the pathological and the evolutionary context

Catarina Serrano

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Mutational change of X-linked disease genes in both the pathological and the evolutionary context

Catarina Serrano
Doutoramento em Biologia
Departamento de Biologia
Faculdade de Ciências da Universidade do Porto
2024

PhD

PhD

PhD

PhD

PhD

Mutational change of X-linked disease genes in both the pathological and the evolutionary context

Catarina Tavares Serrano
Doutoramento em Biologia
Departamento de Biologia
2024

Orientador
Doutora Luísa Azevedo, Investigadora Auxiliar, Instituto de Ciências Biomédicas Abel Salazar (ICBAS)

Coorientador
Professora Doutora Maria João Prata, Professora Associada com agregação da Faculdade de Ciências da Universidade do Porto, Investigadora do i3S/IPATIMUP

Acknowledgements

I am sincerely grateful to the Fundação para a Ciência e Tecnologia (FCT) for awarding me the PhD scholarship (SFRH/BD/137925/2018). This scholarship was instrumental in supporting my academic journey and research pursuits. Additionally, I appreciate the funding provided by project the POCI-01-0145-FEDER-29723, which significantly contributed to the successful execution of my research project.

I owe a tremendous debt of gratitude to my supervisor, Luisa Azevedo, PhD, whose guidance, expertise, and support have been indispensable throughout my Ph.D. journey. I am equally thankful to my co-supervisor, Maria João Prata, PhD, for the valuable insights and collaboration. I would like to express my gratitude to António Amorim, the group leader, for his leadership of the Population Genetics and Evolution team at i3S.

I extend my deepest gratitude to my family, whose unwavering support and understanding have been my pillars of strength throughout this challenging journey. Their encouragement provided the motivation to persevere during moments of doubt and fatigue.

A special and heartfelt thank you goes to my husband, Levi, for his patience and encouragement throughout the demanding hours and intellectual challenges of my doctoral studies. Your unconditional support has fueled my determination, and I am grateful for every moment shared with you.

A thank you to baby Camila for being a source of joy and inspiration. Your innocent laughter and presence have been a constant source of comfort throughout this journey. Your presence has made this academic pursuit more meaningful.

Abstract

Mutations play an important role in both pathological and evolutionary contexts, since they influence genetic diversity in general, including the fraction that accounts to species adaptation. Natural selection acts on genetic diversity, favoring mutations that improve the overall fitness while acting against those that are detrimental. Exploring the complex interactions between mutational changes in both pathological and evolutionary scenarios allows not only a better understanding of the factors modelling diversity with pathogenic potential, but also underscores the dual nature of mutations as contributors to diversity in both health and disease.

Predictably, if an amino acid residue is evolutionarily conserved over a substantial period of time, its substitution may be associated with some degree of functional impairment. This reasoning can be broken when compensated mutations come into play, a scenario that commonly arises when a mutation appears as the wild-type allele in some species but is known to cause disease in closed related lineages, as the result of epistatic interactions between the amino acid residues. Taking advantage of the peculiar transmission properties of X-linked genes, from which arises the hemizygous state of males in most mammals, this study aimed to increase the knowledge about deleterious mutations in humans that supposedly are compensated mutations in non-human mammals. It was addressed how this likely possibility could challenge the assessment of pathogenicity of human mutations and its weight as players underlying human genetic diseases. The analysis of 10 X-linked genes (*ABCD1, AR, BMP15, COL4A5, EDA, F9, GLA, G6PD, MECP2* and *OTC*) revealed a total of 72 compensated mutations, and only 7 were concurrently and simultaneously predicted as deleterious by all tools (Polyphen-2, PON-P2 and FATHMM). The current tools, which as a rule overlook the genetic background, revealed to be quite imprecise in predicting the pathogenicity of human mutations, particularly when dealing with potential compensated genetic variants. Furthermore, two amino acid positions assumed to be compensated mutations besides showing signs of positive selection, demonstrated high evolutionary tolerance, supporting the idea that both residues are compensated.

Still regarding compensated mutations, a structural and molecular dynamics approach enabled the identification of a compensated mutation and its corresponding compensatory partner. In the *F9* gene, the human pathogenic mutation Glu270Lys causes a conformational alteration in 39-loop, leading to functional impairment. However, another mutation in a neighboring residue (Thr271Pro) was found to

counteract the conformational change promoted by Glu270Lys, mitigating its deleterious effect.

Human mutations that cause diseases may also be associated with genomic segments that underwent duplication in the past, events that were crucial during evolution. The analysis of a partial duplication within an X-linked segment in the *Macaca* genus, revealed no signs for selective forces acting on that specific genomic segment, suggesting to be a likely duplicate without functional consequences.

The objective of this thesis was not only to enhance the comprehension of the impact of genetic diversity on disease but also to establish a foundation for future investigations into evolutionary facets of human diversity.

Keywords: X-chromosome, Epistasis, Mutation, Duplication, Evolution, Selection, Compensated Mutation, Disease.

Resumo

As mutações desempenham um papel importante tanto em contextos patológicos como evolutivos, uma vez que influenciam a diversidade genética no seu todo incluindo a fração envolvida em processos de adaptação de espécies. A seleção natural atua sobre a diversidade genética favorecendo ou desfavorecendo mutações que melhoram ou prejudicam a sobrevivência, respetivamente. Explorar as complexas interações entre mutações em cenários patológicos e evolutivos permite não apenas uma melhor compreensão dos fatores que modelam a diversidade genética humana com potencial patogénico, mas também destacar a dualidade das mutações como contribuintes para a diversidade na saúde e na doença.

É comum assumir que se um aminoácido se encontra conservado durante um período substancial de tempo numa dada linhagem evolutiva, a sua substituição deverá implicar algum grau de comprometimento funcional. Porém, tal assunção pode não ser aplicável quando estão em causa mutações compensadas, cenário que frequentemente se coloca quando uma mutação associada a doença numa espécie surge como alelo de referência noutras espécies filogeneticamente relacionadas, indiciando, assim, o envolvimento de interações epistáticas entre resíduos de aminoácidos. Tirando partido das propriedades de transmissão de genes ligados ao cromossoma X, das quais decorrem a hemizigotia nos machos na maioria dos mamíferos, esta tese visou aumentar o conhecimento sobre mutações deletérias em humanos que aparentemente podem ser variantes compensadas. Numa primeira etapa foi analisada como mutações humanas ligadas ao X previamente descritas como possivelmente compensadas poderiam desafiar a previsão de patogenicidade usando as ferramentas atualmente disponíveis de avaliação de mutações humanas quanto ao seu peso como fatores subjacentes a doenças genéticas. A análise de 10 genes ligados ao cromossoma X (*ABCD1*, *AR*, *BMP15*, *COL4A5*, *EDA*, *F9*, *GLA*, *G6PD*, *MECP2* e *OTC*) revelou que num total de 72 mutações assumidamente compensadas, apenas 7 foram consistentemente avaliadas como patogénicas por todas as ferramentas usadas (Polyphen-2, PON-P2 e FATHMM). Tais ferramentas, ignoram, no entanto, o contexto genético, e daí revelaram ser bastante imprecisas na predição de patogenicidade de mutações humanas, particularmente quando estão em causa variantes possivelmente compensadas. Dois dos aminoácidos classificados como sendo mutações compensadas, exibiram não só sinais de seleção positiva, como demonstraram grande tolerância à mudança numa escala evolutiva, suportando a ideia de ambos os resíduos serem compensados.

Ainda quanto a mutações compensadas, através de abordagens de análise estrutural conjugada com dinâmica molecular, foi possível identificar no gene *F9* uma mutação compensada e o respetivo parceiro compensatório. Em humanos, a mutação patogénica Glu270Lys é responsável pela alteração conformacional num segmento contendo um 39-loop que compromete a função proteica. No entanto, uma outra mutação localizada num resíduo vizinho, Thr271Pro, reverteu a alteração conformacional provocada por Glu270Lys, mitigando o seu impacto deletério.

Mutações humanas responsáveis por doenças podem estar também associadas a segmentos genómicos que foram sujeitos a eventos de duplicação, processos que desempenharam papéis evolutivos cruciais. Neste estudo, identificámos uma duplicação parcial duma região ligada ao cromossoma X compreendendo os genes codificadores da OTC (ornithine transcarbamylase) e RPGR (retinitis pigmentosa GTPase regulator), abrangendo 0.1 Mb no cromossoma 9 no género Macaca. Nesse segmento genómico, não foram detetados sinais de atuação de seleção, indicando que tal duplicação aparentemente não teve consequências funcionais.

Esta tese teve como objetivo enriquecer a compreensão do impacto da diversidade genética em doenças, mas também abrir caminhos para futuras investigações nas várias vertentes, incluindo evolutiva, em que pode ser encarada a diversidade genética humana.

Palavras-chaves: Cromossoma X, Epistasia, Mutação, Duplicação, Evolução, Seleção, Mutações Compensadas, Doença.

Table of Contents

List of Figures

List of Tables

Abbreviations

Ala: Alanine

Arg: Arginine

Asn: Asparagine

Asp: Aspartic acid

BEB: Bayes empirical Bayes

CCR5: Coreceptor C-C chemokine receptor type 5

CDS: Coding sequence

CYP: Cytochrome P450

CGB: β subunit of chorionic gonadotropin

CPD: Compensated pathogenic deviations

CRISPR: Clustered regularly interspaced short palindromic repeats

DNA: Deoxyribonucleic acid

FATHMM: Functional analysis through Hidden Markov Model

Gln: glutamine

GTR: General time reversible

Gly: Glycine

Glu: Glutamic acid

His: Histidine

HGMD: Human Gene Mutation Database

Hox: Homeobox

Ile: Isoleucine

Leu: Leucine

LHB: β subunit of the luteinizing hormone

LRT: Likelihood ratio test

Lys: Lysine

MAS: Multiple sequence alignment

Met: Methionine

PAR: Pseudo autosomal region

PKU: Phenylpyruvic oligophrenia phenylketonuria

Pro: Proline

Provean: Protein variation effect analyzer

HIV: Human immunodeficiency virus

HPA: Hyperphenylalaninaemia

SIFT: Sorting intolerant from tolerant

Thr: Tryptophan

Val: Valine

Chapter 1: General Introduction

Genetic diversity

Genetic diversity, in a broader sense, is the measure of the variability found within genomes. Traditional notions of genetic diversity were formulated within the population framework, to assess the disparities among individuals in a population or between populations of identical species. Nevertheless, genetic diversity can be envisioned at other levels, including individuals, species and even genes or any genomic segment [1]. Irrespective of the level, it is dependent on genetic variation, which is a fundamental resource for evolutionary change.

As the number of sequenced genomes increased, the possibility to compare specific genomic segments has also augmented, allowing robust statistical inferences. The main sources of genetic diversity are mutation and recombination [2]. Mutation might arise by replacements at the DNA level that create distinct allelic forms between individuals, populations, or species. Recombination is the mechanism by which novel genetic combinations (backgrounds) are produced. If the reshuffled genetic material is transmitted to the offspring, then it can be spread through a population. Finally, random genetic drift is the process by which fluctuations in the frequencies of alleles occur due to random processes.

The coexistence of multiple allelic forms at a specific gene position within a population is a common event. Conventionally, when the minor allelic form reaches a frequency ≥1% in a population, it is known as a polymorphism. Polymorphisms are crucial to determine the level of diversity in populations, and although often neutral, they can represent the raw material for adaptation [3]. Genetic diversity is crucial to allowing responses to environmental changes, thus contributing to the conservation of species. Human genetic diversity modulates several aspects of the human health, including susceptibility to many diseases [4], [5], [6], [7].

Genetic diversity differs markedly among species. For example, the intraspecific genomic variation in *Drosophila simulans* is 3%, while in *Homo sapiens* is only 0,1% [8], [9]. Such variation in genetic diversity might be partially explained by differences in the divergence times or the germline mutation rates across species [10], [11]. However, other factors also contribute to the disparity of genetic diversity between species, in

particular the rate of allele loss and fixation [2]. For neutral *loci* this rate is highly influenced by genetic drift, whose strength is inversely proportional to the population effective size. Since the effective population is closely linked with the demographic history, both play an important role in modulating genetic diversity within a population or species [12], [13].

Unraveling the complexities of genome diversity in humans and the intricate interplay between gene variants and population-specific adaptations, holds profound implications for the advancement of precision medicine and a better understanding of human health. In this perspective, a detailed assessment of genetic diversity, achieved by combining the gene evolutionary history with molecular, structural, and functional data, seems critical to improve the knowledge on genotype-phenotype relationships, and as a consequence, the accuracy and reliability of predictions in health and disease.

Mendelian diseases and phenotype genotype correlation

In the mid-19th century, Gregor Mendel, based on experiments conducted with pea plants, was pioneer in uncovering the fundamental principles of inheritance. His discoveries would lay the foundation of a field of science that became known as Mendelian genetics [14]. Importantly, Mendel's principles explained well the inheritance of some human genetic disorders, most of which were soon realized to be caused by mutations in a single gene. These disorders, known as Mendelian, result from functional impairments in proteins, including many enzymes. Mendelian diseases can show different patterns of inheritance: autosomal dominant or recessive, and X-linked also dominant or recessive. In autosomal dominant diseases, the development of the disorder depends on the presence of just a mutated allele, which is inherited from one the parent, although it can also arise by a *de novo* mutation. Some examples include Huntington's disease or familial hypercholesterolemia [15]. In autosomal recessive diseases, the presence of two copies of disabled alleles is required, which again are inherited from each of the parents, but cases involving at least one *de novo* mutation were also documented. Carriers, i.e. subjects with one normal and one mutated allele, do not manifest symptoms. Cystic fibrosis and phenylketonuria are among the diseases with this inheritance mode [16], [17], [18]. X-linked disorders are caused by variants in genes located in the X chromosome. When the causative allele is recessive, the disease is more common in males because the presence of a pathogenic variant in their unique copy of the X chromosome is enough to induce disease, whereas in females the disease only appears when both their X chromosomes bear a disabled allele. Hemophilia and Duchenne Muscular Dystrophy are two of the many instances of X-linked diseases [19], [20].

There are disorders that result from the combination of multiple genetic, environmental, and behavioral factors. These are called complex diseases [21].Type I diabetes, obesity, and non-familial forms of cancer, are examples of complex diseases that result from the contribution of numerous genetic and environmental factors that together account to disease susceptibility [22], [23], [24].

Phenotype and genotype are two fundamental concepts in genetics, that contemplate the observable traits in an organism and their genetic composition, respectively. They are considered two distinct levels of biological abstraction, under the paradigm that the phenotype is the manifestation of the genotype [25]. However, the relationship between phenotype and genotype is exceedingly complicated. In fact, a single gene may have

consequences on different traits, or multiple genes might influence a given single trait[26].

The complexity of the interplay between genotype and phenotype can be illustrated by the *PAH* gene and its associated genetic disorder, Phenylpyruvic oligophrenia phenylketonuria (PKU). *PAH* encodes the enzyme phenylalanine hydroxylase, whose deficiency is the result of different loss of function mutations in the gene. Enzyme deficiency leads to the accumulation of phenylalanine in the body, a phenotype known as Hyperphenylalaninaemia (HPA). The classical symptoms of HPA include, besides those consequents to the neurological impairments, lighter skin and hair compared to unaffected family members. As a matter of fact, patients tend to have lower levels of melanin, the pigment responsible for skin and hair color, because deficiency in phenylalanine hydroxylase also results in lower melanin production [27]. In this case, a single genetic mutation (in *PAH*) has a pleiotropic effect on both neurological and pigmentation traits. Still, while the enzyme deficiency promoted by PKU is always correlated with an accumulation of phenylalanine in toxic levels (HPA), not all abnormal levels of phenylalanine are associated to a PKU phenotype [28]. Other forms of HPA can result from different genetic mutations affecting enzymes or proteins involved in the phenylalanine metabolism pathway [29]. In respect to melanin, it is well established that multiple genes are involved in its synthesis and distribution in the skin. That's why skin color is often presented as a classical trait with polygenic inheritance. Many alleles in different genes contribute to the overall variation in skin color, accounting thus to the wide range of skin tones that humans can exhibit. Such phenotypic continuum is the result of an outstanding number of multiallelic combinations involving different loci [30].

There is no doubt that a better comprehension of genotype-phenotype relationship is especially relevant in the context of genetic disorders. Significant steps have been taken regarding Mendelian diseases, concerning which recent findings have contributed to design more personalized approaches to prevent and treat diseases. However, there is still a long way to go, not only for simple Mendelian diseases but particularly for complex ones. Strong efforts are being made to identify individuals at risk of developing complex diseases based on their genetic composition. The challenge is how and when will be possible to translate the knowledge in effective heath care improvements and advances towards personalized medicine [31].

Sex chromosome evolution and the uniqueness of the X-chromosome

In many animals the primary sex determination relies on specific chromosomes that define different chromosomal sex-determination systems, such as XY, ZW or ZZ [32]. Most mammals present the XY system, where the females are the homogametic sex (XX) and males the heterogametic one (XY). Since males have two different sex chromosomes, they can produce two types of gametes- either with X or with Y, whereas females only produce gametes with X. The two sex chromosomes have evolved from an ancestral pair of autosomes [33]. Around 160 million years ago, a proto-Y chromosome acquired a gene, *SRY*, that would become the male-determining gene in the Theria (marsupial and eutherian mammals) [34], [35], [36]. This proto-Y chromosome has undergone substantial structural changes, including multiple inversions. The accumulated alterations progressively blocked recombination with the X chromosome, except in a short region where pairing was necessary to guarantee correct chromosome segregation during meiosis. This is the pseudo autosomal region (PAR), where high homology between the X and the Y chromosomes was maintained [37]. Contrarily to other mammals, humans contain two PARs in the sex chromosomes, PAR1 and PAR2. PAR2 is a human-specific genomic segment, having been created by duplication of material from X to Y since the divergence of human and chimpanzee lineages [38]. As the non-recombination region of the Y increased, deleterious mutations were exposed and to maintain male fitness large deletions encompassing deleterious and non-functional DNA regions occurred, prompting the degeneration of Y chromosome [39], [40].

The unequal dose expression between males (XY) and females (XX) was overcome by the emergence of a dosage compensation mechanism that assured the almost complete inactivation of one of the two X chromosomes in mammalian females [41]. Since early in the evolutionary process, the Y-chromosome underwent a rapid decay in size and gene content, in such a way that currently it only retains ~3% of the ancestral genes, compared with ~98% in the X-chromosome. Among the few Y-chromosome genes that persisted, some are known to be crucial to male gonadal sex determination and spermatogenesis [37]. Comparatively, the X-chromosome has a much more dense and diverse gene content, being enriched in brain-expressed genes and genes related to muscle function, sex and reproduction [42], [43], [44].

The X-chromosome has some hallmarks that makes it unique. Males are hemizygous, which implies that the presence of a single copy of a disease-associated allele is enough to the expression of a given phenotype, which in females only manifests when they are homozygous [45]. Population and comparative genomic data revealed that the rate of divergence between species is higher for X-linked genes than for autosomes, a phenomenon called “faster X-evolution effect” [46]. The accelerated evolutionary changes that the X experienced in comparison to autosomes has deeply shaped the patterns of diversity across the chromosome. As so, and compared to autosomes, the X-chromosome represents a unique resource to study mutations associated with Mendelian traits, all the more that the hemizygous state of males turns possible more uncomplicated analyses.

General notions on compensatory epistasis

The phenotypic effect of a genetic variant can vary widely in different individuals depending on the genetic background in which it occurs. Epistasis is the name given to this context-dependency of mutation effects, that might involve complex relationships between genes, alleles or other genomic segments (reviewed in [47]). In other words, epistasis is the interaction between two or more genes, alleles or genomic regions [48]. Collectively, these genetic interactions significantly account to the phenotypic diversity within individuals and, on an evolutionary timescale, to the genomic architecture of different species [49], [50], [51].

The term epistasis was coined by Bateson in the early twentieth century ([52]) to describe the discrepancy between the predicted phenotypic outcomes based on the effects of each of two individual genes and the observed phenotypic outcome under the combined effects of both *loci*. Soon after, epistasis began to be applied with other related meanings, to accommodate not only the wide variety of genetic interactions but also the different levels at which they can be assessed [50]. As the sense of epistasis was broadening, the terminology used to distinguish types of epistasis also proliferated. In the specialized literature it is common to find categorizations such as the ones presented below, although these definitions may define categories that are not mutually exclusive and often complement each other in terms of scope and practical applications ([53]):

- Positive and negative epistasis: positive or negative when the effect of the presence of two variations is better or worse, respectively, than expected;
- Synergistic and antagonistic epistasis: synergistic or antagonistic epistasis when the absolute value of the effect of presence of two variants (either deleterious or beneficial) is larger or smaller, respectively, than the absolute value of the sum of the effects of the two individual variations;
- Pairwise and higher-order epistasis: depending on the interaction involving two or more genetic variants;
- Statistical epistasis: referring to the average epistasis between two genetic variants when measured on genetic backgrounds (at other loci) that are randomly sampled from a population;
- Functional epistasis: used to describe protein–protein interactions;
- Intragenic (or intramolecular) and intergenic (or intermolecular epistasis): if interactions occur between sequence variants in the same gene or between variants in different genes, respectively;

- Sign and magnitude epistasis: when the variant changes the sign/direction of the effect, or when it only changes the strength of the effect, respectively;
- Reciprocal sign epistasis: a complex type of epistasis that occurs when the effects of the mutations are dependent on each other, and the direction of the effect is reversed in the reciprocal genetic backgrounds. In simpler terms, deleterious mutations assessed individually become beneficial in combination with other mutation(s), or for the contrary, individual beneficial mutations become deleterious in combination with others.

Many examples of epistatic interactions have already been identified [54]. However, knowledge of the molecular mechanisms underlying such interactions is still very scarce. What is well established is that mechanisms are multiple, might involve very high levels of complexity and that much work needs yet to be done to dissect their understanding. Among the molecular mechanisms of epistasis are protein stability thresholds, functional redundancy, and metabolic pathway structure [54].

In the context of sign epistasis, that encompassing interactions in which a mutation's effect is conditional on the genetic background, there are many reports on pairs of mutations individually neutral, beneficial or deleterious that in combination with others change the direction of the outcome [55]. For instance, the negative/deleterious effect of a given mutation might be fully reversed by interaction with a nearby amino acid position. This mutation is commonly referred to as a compensated mutation (or as in its original form-compensated pathogenic deviation, CPD) [56], [57].

Considering that a protein's function is determined by the interactions between residues in the three-dimensional space, and that many amino acid alterations confer different physical and chemical properties to the protein, few doubts remain that epistasis might contribute to the phenotypic diversity observed between individuals. Acknowledging that compensated mutations are important players in molecular evolution, relevant theoretical work on compensatory evolution was developed allowing the prediction of different evolutionary frameworks [58].

Detecting compensated mutations is not easy. It requires approaches that combining genomic, experimental and computational methods might inform on more or less complex genetic interactions involved in functional recovery. One strategy to detect compensated mutations is by the cross-comparison of orthologous sequences from humans and non-human species [56], [59], [60]. When a human deleterious mutation appears as the wild-type allele in a nonhuman species, a reasonable explanation is that

the disease-associated residue reached high frequency or even became fixed in the nonhuman species due to the existence of one or more compensatory residues in the same protein or another protein that interacts with the first [56], [61], [62], [63], [64], [65]. This kind of analysis relies on accurate genome sequencing data, as well on robust bioinformatics tools able to identify changes that potentially restore functional integrity. Experimental methods also play a role in identifying compensatory epistasis. Reverse genetics experiments, such as site-directed mutagenesis, can be conducted to introduce additional mutations into the genetic makeup of an organism that already carries a primary mutation ([66], [67]). The observation of subsequent changes in traits or overall fitness can indicate whether the introduced mutations rescued or not a given phenotype [66], [67], [68], [69]. Empirical evidence of interactions between the compensated mutations and their compensatory partners are gathered through a panoply of tests, growth experiments, among other ways, conducted under different conditions. More recently, computational models have emerged as valuable means for predicting compensatory interactions. They take advantage of new statistical approaches and machine learning techniques that can analyze large datasets aiming at identify genetic variants that show patterns of co-occurrence with primary mutations. Furthermore, these models integrate variables such as protein-protein interactions, gene expression profiles, and predicted functional consequences of mutations [70], [71].

The impact of epistasis on the relationship between genotype and phenotype is a topic that has captured great interest in genetic research [72], [73]. The realization that some harmful mutations can be compensated by the presence of certain residues in other positions is enriching our understanding on the road between genetic variants and the associated phenotypic consequences. However, a critical question that compensated mutations raise is to understand how is overcome the burden of a novel deleterious mutation during the evolutionary time. A previous study demonstrates that in mammals the deleterious mutations are expected to occur in backgrounds already carrying the compensatory partner(s) or originate from the same mutational event that originate the compensatory partner [74]. Therefore, the dimension time must be integrated when addressing the nature of epistatic interactions and regulatory networks that govern the translation of genetic information into a wide range of observable characteristics. Deciphering the consequences associated with the interaction between multiple variants will contribute to a broader comprehension of how proteins maintains a delicate balance between genetic stability and phenotypic diversity.

Evolution through gene duplication

Molecular evolution is an inexorable process that shapes the diversity of all living beings. It operates through many mechanisms, including gene duplication, which is the gene replication of an existing gene, leading to the emergence of additional elements in the genome. Duplicated genes can undergo different fates, conventionally ascribed as neofunctionalization, subfunctionalization and non-functionalization [75], [76], [77], [78]. The fate of newly arisen genes depends on various factors, including selective pressures acting on organisms carrying duplicated genes, the functional significance of the copied genes and the mutational processes occurring after the replication event.

The acquisition of novel functions by one of the members of the duplicates is referred to as neofunctionalization. Functional divergence over time increases the operational diversity of a gene family, potentially contributing to the evolution of novel traits. Duplications or multi-replications by themselves can be neofunctionally advantageous, as shows the gene *CCL3L1*, which encodes the C-C Motif Chemokine Ligand 3 Like 1, a potent HIV-1-suppressive chemokine and a ligand for HIV coreceptor C-C chemokine receptor type 5, known as CCR5. It was demonstrated that there are significant interindividual and interpopulation differences in gene copy number of *CCL3L1*, and also that high copy number of the gene confers lower susceptibility to HIV infection [79].

The evolutionary process of gene duplication(s) followed by neofunctionalization can be illustrated by the six human genes encoding the β subunit of chorionic gonadotropin-*CGB* genes. Gonadotropin is a hormone present in some primate's placenta, thought to be responsible for early pregnancy maintenance. The human *CGB* genes that arose through gene duplications were likely associated with the acquisition of new functionalities turning the placenta more efficient [80]. All six CGB genes are duplicated copies of the β subunit of the luteinizing hormone gene (LHB), crucial to trigger ovulation [81].

After gene duplication the evolution of the duplicates can lead to subfunctionalization, which happens when both copies retain the original functions, but each copy becomes specialized in performing a subset of the ancestral functions. Change in the expression pattern of both copies is mainly due to mutations in regulatory regions. There are various known cases of subfunctionalization, such as in the Hox-genes, in the MADS-box genes or in the immunoglobulin gene superfamily [82]. Taking an example from the Hox-genes, the paralogous *Hoxa1* and *Hoxb1*, which play crucial roles in patterning the hindbrain during the mouse embryonic development, evolved by subfunctionalization. Hoxa1 is

highly sensitive to retinoic acid, and plays a role for segment identity in rhombomere 5 in the hindbrain. Hoxa1 activates its paralogue Hoxb1, which is important for rhombomere 4 identity. The expression of Hoxb1 is, however, maintained by autoregulation [83]. Studies have indicated that after the event of duplication, the paralogues diverged in such way that the autoregulation degenerated in Hoxa1 while high retinoic acid sensitivity degenerated in Hoxb1, which in this case fits a scenario of spatial/temporal subfunctionalization (complementary expression in rhombomeres 4 and 5) [84]. Each gene acquired unique regulatory elements leading to the fine-tuning of their roles in hindbrain development.

Non-functionalization is another possible outcome following gene duplication. One of the copies retains the ancestral function, whereas the other diverge and accumulates disruptive mutations over time, rendering it nonfunctional. This is one of the processes that originates pseudogenes. When the presence of one of the copies of the duplicated gene is enough, mutations in the extra copy are not under stringent selective forces, steering the conversion into a pseudogene prone to become fixed in populations [85], [86]. During the process of pseudogenization, mutations accumulate rapidly and the gene degenerates. It is well documented that numerous gene duplication and loss events have underlie the diversity of genes encoding enzymes in cytochrome P450 (CYP) family, concerning which the human genome contains dozens of pseudogenes [87]. Whilst CYP genes are present in all metazoan, its number varies widely among species as a consequence of the complex evolutionary processes of the entire gene family. Many CYP genes have undergone lineage-specific duplication and other rearrangement events, whereby some duplicated genes have retained the ancestral functions or similar while others became pseudogenes. Gains and losses of CYP genes occurred relatively frequently in different lineages, supposedly driven by their substrate and metabolic specificities [87].

Gene duplication was indeed a major contributor to the dynamic change of genomes and thus to the renewal of the patterns of genetic diversity over evolutionary time scales.

Thesis outline and main objectives

The initial chapter of this thesis has provided a comprehensive introduction to fundamental topics that are essential for understanding the impact of mutations on the evolution of species integrating the pathogenic context. Since the research undertaken had focus on missense mutations associated to Mendelian diseases, that urged an exploration of subjects such as genetic diversity and factors hampering to achieve reliable genotype-phenotype correlations. A brief overview of compensatory epistasis was also given, summarizing the importance of compensatory mutations and the approaches used to investigate this complex issue. An introduction to gene evolution through duplication was also provided as one goal of the study was to investigate a case related to a X-linked gene.

The second chapter presents data based on the exploration of amino acid positions that were previously associated with diseases in 10 X-linked genes. Levels of conservation, alongside the selective constraints the sites were upon, were major of interest. The uniqueness of X chromosome, its degree of conservation along mammals and the common male hemizygousity, make it an ideal candidate to understand the conservation of mutations associated to human diseases, particularly compensated mutations.

The third chapter, which deals with genetic variants and their role as phenotype modifiers, includes one published work (DOI: 10.1007/S00439-021-02307-X) that demonstrates the compensatory behavior of a mutation through molecular dynamic.

The fourth chapter, titled “Partial gene duplication”, contains a published article (DOI: 10.1016/j.gene.2022.146997) reporting an event of partial duplication detected on the genomic region where the *OTC* gene locates.

The 5th chapter is dedicated to concluding remarks and major findings of the work.

The main objective of this doctoral thesis is to understand X-linked genetic variation in terms of disease and evolution. In more detail, the aims of the present work were to:

- Determine the conservation degree of X-linked disease-associated missense mutations.
- Study events of mutational epistasis in mammals.
- Search for gene duplication events in mammalian X-linked genes and assess their functional status.

Chapter 2: Conservation degree of X-linked disease-associated missense mutations

Introduction

The relevance of an amino acid has been positively correlated with its conservation along a phylogeny, explaining why highly conserved amino acid positions are assumed to have relevant functional roles. Hence, mutations in those positions are expected to be deleterious [88], [89]. Many *in silico* methods have been developed to predict the outcome of missense mutations, most of which rely on protein conservation to score pathogenicity. Since the available prediction tools are based on different databases and algorithms, none of which deprived of advantages and disadvantages, the use of multiple software is strongly recommended to strength evidence on the potential pathogenicity of a given variant [90]. Prediction tools such SIFT (Sorting Intolerant From Tolerant) and Provean (Protein Variation Effect Analyzer) apply evolutionary conservation-based algorithms centered in a score given to the conservation of a residue through the analyses of Multiple Sequence Alignment [91], [92], [93]. Others, such Polyphen2, combine an evolutionary approach with the consequences in the protein structure [94]. There are also machine learning-based approaches based on random forest probability scores, as for instance PON-P2, which classifies a mutation in pathogenic, neutral or unknown significance [95]. All these tools are widely used in clinical genetics [90], [96]. However, these tools cannot handle the effect of genetic background and how it can influence, positively or negatively, the overall outcome of a given variant [65]. That might be a very limiting problem when are at skate possible compensated mutations, in particular when they were signalized due to be deleterious in humans while wild-type in closely related species, raising the hypothesis that the wild-type background also contains another variant that neutralizes the effect of the pathogenic mutation in humans [56], [61], [74]. The numerous reports on compensated mutations that emerged through comparative genomics ([59], [97], [98], [99]), led to estimate that they constitute approximately 4% of all the human pathogenic missense variants [71]. However, this kind of epistatic interactions is still understudied and rarely considered in the evolutionary analyses aiming at predicting the pathogenicity of a variant.

To further explore this subject, this study concentrated in human mutations in X-linked disease-associated genes with the aim to assess potential inaccuracies provided by

commonly used predictors of pathogenicity when dealing with putative compensated mutations.

Material and Methods

Gene data

A previous study used a machine learning approach to detect compensated mutations in the whole genome of mammal species [71]. A total of 57 genes located in the X-chromosome revealed signs of having associated compensated mutations. From the initial set of 57 genes, we selected the 10 predicted to have at least three compensated mutations for further analysis. These genes are listed in **Table 1**.

Table 1 Genes selected to study X-linked disease-associated missense mutations

Gene	Description
ABCD1	ATP Binding Cassette Subfamily D Member 1
AR	Androgen Receptor
BMP15	Bone Morphogenetic Protein 15
COL4A5	Collagen Type IV Alpha 5 Chain
EDA	Ectodysplasin A
F9	Coagulation Factor IX
GLA	Galactosidase Alpha
G6PD	Glucose-6-Phosphate Dehydrogenase
MECP2	Methyl-CpG Binding Protein 2
OTC	Ornithine Transcarbamylase

Multiple Sequence alignment

For each gene, the human Coding Sequence (CDS) and the orthologous CDSs were retrieved from Ensembl Genome Browser v97 [100]. A python script was created to filter the non-human mammal orthologues. Sequences with more than 1% of low complexity regions (“N”) and/or with a size at least 25% different from the human reference sequence were removed from the dataset considered in further analyses. Multiple sequence alignment (MSA) containing different mammals was performed. The MSAs were built using PRANK[101]. All alignments were done using a codon model with default parameters. Additionally, a manual inspection was implemented for the alignments.

Amino acid position classification

X-linked disease-associated missense mutations retrieved from the Human Gene Mutation Database (HGMD) [102] were annotated throughout the corresponding human genes. Amino acid positions with disease-associated missense mutations were primarily classified as 'invariant', 'inter-species polymorphism' or 'compensated mutation'. Residues divergent from the human wild-type were assigned into two distinct categories: 'inter-species polymorphisms' or 'compensated mutations'. Inter-species polymorphisms were assumed to be variants not associated with disease. A deleterious human variant that appeared as the wild-type in other mammalian species will be herein mentioned as a compensated mutation. The amino acid positions containing an invariant residue across all mammals were classified as invariants.

In silico prediction of functional impact

Three *in silico* prediction tools were used to evaluate the functional impact of compensated mutations: FATHMM, Polyphen2 and PON-P2. FATHMM considers various features, such as sequence conservation, protein domain annotations, and functional annotations, to make predictions about the pathogenicity or functional impact of a given variant [98]. Polyphen2 uses an evolutionary approach combined with the analyses of the protein structure [86]. PON-P2 is a machine learning-based tool, which classifies the mutation in pathogenetic, neutral or unknown using a random forest probability score [95].

Selection analysis

In order to detect amino acids under positive selection, the site model from PAMLv4 was used [103]. It requires a phylogenetic tree, which was pruned in RAxML Version 8 [104]. The tree was calculated through the GTR gamma model, with the remaining parameters as default.

The ratio between nonsynonymous and synonymous mutations (w=dN/dS) is conventionally classified into 3 categories: $w<1$, $w=1$, $w>1$, indicating purifying selection, neutral selection, and positive selection, respectively. To identify amino acids under positive selection, 6 tests were used: M0 against all other models, M1a vs. M2a and M7 vs. M8. The null hypothesis for all models was M0, which considers a unique w ratio in all codons. M1a classifies codon sites into 2 groups ($w<1$ and $w=1$), while M2a does consider an extra class ($w>1$), allowing the inference of positive selection. M7 uses a beta distribution to describe the w variation among sites, accepting a w ratio from 0 to 1.

M8 also uses a beta distribution to compute the w ratio along the sites, but it considers a range for w from 0 to a w>1.

The package Codeml from PAML that computes the log Likelihood values (lnL) for each model, was used to perform a likelihood ratio test (LRT=2x(lnL1-lnL0). In order to access the statistical significance, the P value was obtained by comparing the LRT values in each pair against the X^2 distribution, where the degree of freedom was the difference of number of parameters between the null and alternative hypothesis. A significant P value (P<0,05) commonly means that the model assuming sites under positive selection fits better the data rather than models providing w>1.

Results and Discussion

Classification of amino acid positions harboring disease-associated mutation

We start by retrieving human disease-associated missense mutations from HGMD corresponding to the ten human genes listed in **Table 1**. A total of 1598 amino acid positions with one or more disease-associated missense mutations were classified into invariant, inter-species polymorphism or compensated mutations. The MSA analysis revealed that 81,9% of amino acid positions persisted invariant among all mammals. Amino acid positions with divergent amino acids were distributed in polymorphisms, accounting to 13,7% of total amino acid positions, and likely compensated mutations, which represented 4,4% of total amino acid positions where deleterious alleles were uniquely reported in humans. This data is graphically illustrated in **Figure 1**. As expected, this data is aligned with previous results from [71], where it was demonstrated that compensated mutations represent “at least 4% of all known missense variants causing human-inherited disease”.

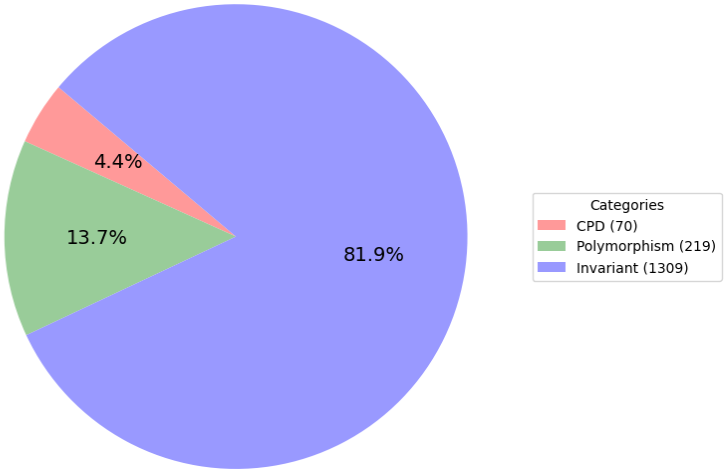


Figure 1 Classification of amino acid positions harboring disease-associated mutation in invariant, compensated mutation or polymorphism.

Assessment of functional impact of missense mutations located in non-conserved amino acid positions

From the total of 289 amino acid positions classified as likely polymorphism and CPD (219 and 70, respectively), 79 had more than one mutation annotated, which resulted in a total of 383 mutations (72 CPDs and 311 polymorphisms) that were subsequently subjected to pathogenicity prediction (**Figure 2-A and Appendix I**). PolyPhen-2 and especially FATHMM offer the most conservative predictions regarding the pathogenicity of variants. According to Polyphen-2, 23 out of the 72 CPDs and 175 out of the 311 polymorphisms were predicted as being deleterious. The PON-P2 estimated a total of 91 polymorphisms and 10 CPDs to be not tolerated in humans. In its turn, FATHMM was more permissive in detecting the damaging effect of CPDs and polymorphisms, only excluding from being deleterious 5 and 3 variants, respectively. The good discriminative power of FATHMM compared to other tools was previous observed in [105], [106], [107]. Also, the high number of unclassified missense mutations was very high in the PON-P2 tool, a situation previously reported [105].

Focusing in the 72 putative compensated mutations that were analyzed by PolyPhen-2, FATHMM and PON-P2 (**Figure 2-A**), it is notorious a low overlap in the deleteriousness prediction for this kind of variants (**Figure 2-B**). Among these mutations, 7 were consistently classified as deleterious (5 in *ABCD1*, 1 in *COL4A5* and 1 in *GLA*). This 7 CPDs have one common feature: they are located in extremely conserved positions. In more detail, from the total range of species analyzed only one (Ala616Thr, Pro543Leu,

Arg518Trp and Met403Lys in ABCD1; Gly45Arg in COL4A5 and Pro40Leu in GLA) or two (Asp555Asn in *ABCD1*) presented an alternative amino acid to the human reference.

As observed, deleterious alleles in humans found as wild-type alleles in other species (putative CPDs) can easily be predicted as tolerated. The need to choose the best prediction tools in the classification of variants was an issue discussed before [108], and our analyses further reinforce the need for prioritizing prediction tools that can deal with such variants.

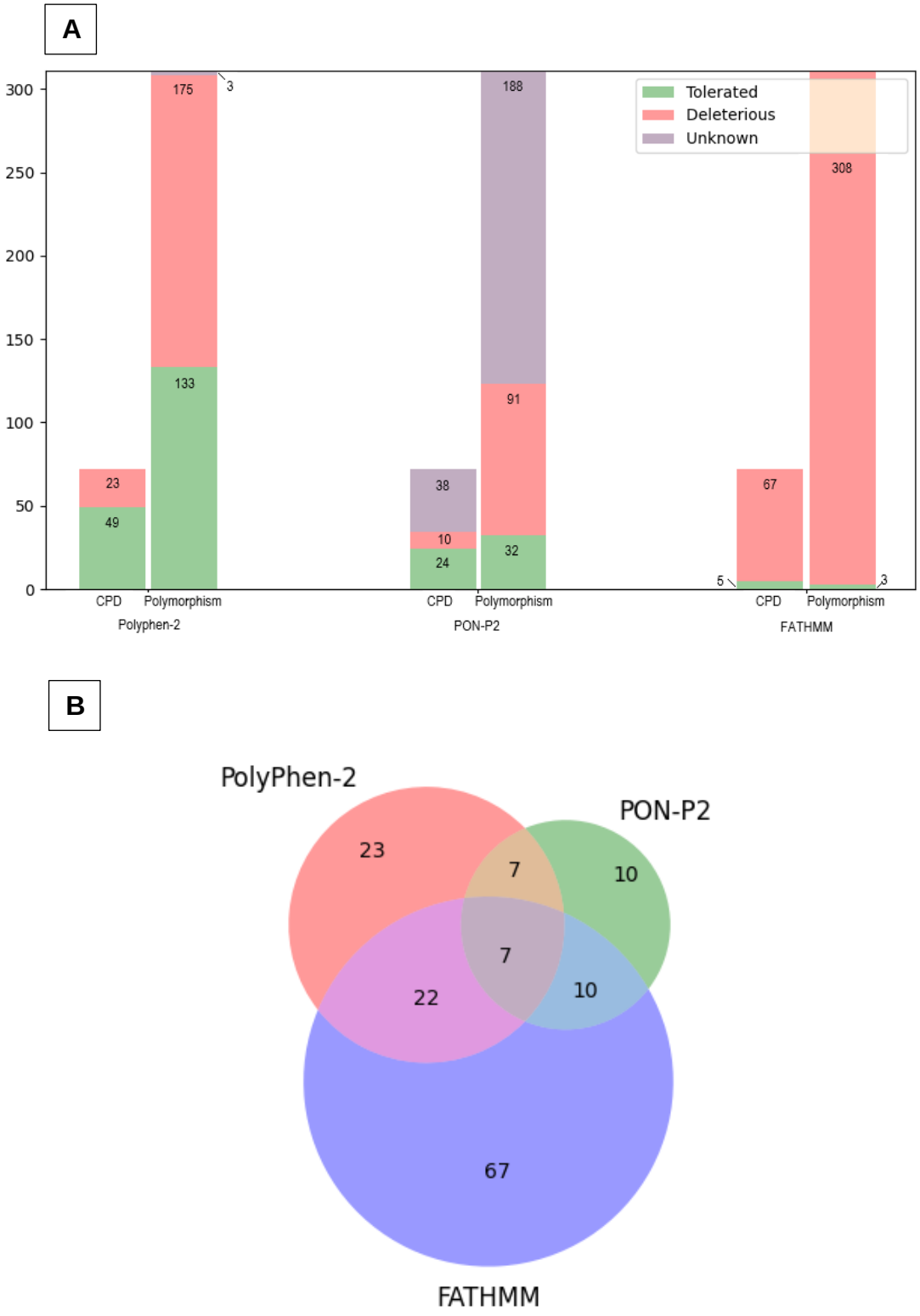


Figure 2 Results from pathogenicity prediction tools. A: Classification of compensated mutations and polymorphisms present in mammalian orthologous sites associated with human disease**. B:** Venn diagram with the number of missense mutations located in CPD amino acid positions predicted as deleterious by Polyphen-2, PON-P2 and FATHMM (graph construed using library matplotlib from Python language).

Evaluation of selective pressures

To address whether natural selection contributed to shape diversity in the 10 genes listed in **Table 1**, an evaluation of selection pressure was pursued using Codeml. The Codeml embedded models M0, M1a, M2a, M7 and M8 were used to test if some codons were under positive selection. The ln and parameter number for each model is present in **Table 2**. According to the statistical test of significance present in **Table 3**, the null hypothesis (M0), where all sites share a common w ratio, was rejected in all genes (**Table 3**). So, with a P value lower than 0,05 it was accepted that the w ratio varied between codons. To evaluate whether or not some amino acids were under positive selection, the tests M1a vs. M2a and M7 vs. M8 were done. The first test (M1a vs. M2a) was non-significant for most of the genes, except for *F9* and *GLA*, indicating that the model M1a, which does not allow w >1, was the best fitting model for the majority of genes. In contrast, *AR* and *EDA* were the only genes demonstrating no significant differences between the M7 and M8 models, supporting no signals of positive selection along their codons. The test M7 vs. M8 exhibited significant differences in *ABCD1*, *BMP15*, *COL4A5*, *F9*, *GLA*, *MECP2* and *OTC*.

Both tests, M1a vs. M2a and M7 vs. M8, revealed different results in detecting signals of positive selection in the gene codons, it is important to consider their differences. The test between models (M1a and M2a) is more conservative and for that reason less powerful than M7 vs M8. For instance, M7 vs. M8, due to its algorithm being less stringent, is more powerful and less accurate [109]. The existence of codons with a selective force >1 is corroborated by both tests (M1a vs. M2a and M7 vs. M8) only in *F9* and *GLA.* So, according to a more prudent interpretation only *F9* and *GLA* seem to have codons under positive selection, whereas assuming a less cautious view *ABCD1*, *BMP15*, *COL4A5*, *F9*, *GLA*, *MECP2* and *OTC* all contain positively selected codons.

These 7 genes that appeared to have codons subject to positive selection were further tested with Bayes empirical Bayes (BEB) to identify the amino acids positions under positive selection [110]. This method assigns the posterior probability of the w value estimated for a given site; when such probability is higher than 95% for sites with w>1, it allows to infer that they are positively selected [110]. The BEB test for M7 vs. M8 predicted the existence of 3,1,8,7,1 and 2 sites under positive selection (P>95%) for *ABCD1*, *COL4A5*, *F9*, *GLA, MECP2* and *OTC* genes, respectively (**Appendix II**). Two of these variants are Arg26 in *OTC* and Arg3 in *F9*. The 26th amino acid site from the OTC protein harbors two mutations that result in ornithine transcarbamylase deficiency: Arg26Gln and Arg26Pro. The amino acid glutamine (Gln) is found as a wild-type allele

in 2 species (*Heterocephalus glaber* and *Octodon degus*) and 12 other species also have a different amino acid (tryptophan or glycine) compared to the human reference. In turn, concerning Arg3His, which is responsible for Hemophilia B, His3 was found as wild-type allele in 9 species (*Nannospalax galili*, *Cricetulus griseus*, *Peromyscus maniculatus bairdii*, *Microtus ochrogaster*, *Mus musculus*, *Mus spretus*, *Mus caroli*, *Dipodomys ordii* and *Meriones unguiculatus*). In 32 other mammals, different amino acids were found at the same position, namely: cysteine, leucine, asparagine, phenylalanine, glutamine and methionine.

Through the lens of evolution, both sites show considerable tolerance for different variants, in line with the result of the selection analyses. Collectively, these findings reinforce the possibility of the sites implicated in compensatory epistatic mechanisms having a role in shaping the phenotypic manifestation of specific variants.

Table 2 Lnl and np for each model

Gene	Model	lnL	np
ABCD1	M0	-25178.93	131
	M1a	-24788.44	132
	M2a	-24788.44	134
	M7	-24526.41	132
	M8	-24515.38	134
AR	M0	-21913.63	111
	M1a	-21601.09	112
	M2a	-21601.09	114
	M7	-21459.17	112
	M8	-21457.34	114
BMP15	M0	-14648.39	119
	M1a	-14405.25	120
	M2a	-14405.25	122
	M7	-14373.08	120
	M8	-14367.19	122
COL4A5	M0	-36265.98	103
	M1a	-35489.73	104
	M2a	-35487.50	106
	M7	-35463.30	104
	M8	-35434.04	106
EDA	M0	-6499.189	85
	M1a	-6386.464	86
	M2a	-6386.464	88

	M7	-6377.481	86
	M8	-6377.477	88
F9	M0	-15644.58	141
	M1a	-15008.84	142
	M2a	-14969.99	144
	M7	-14947.99	142
	M8	-14908.49	144
GLA	M0	-15616.44	139
	M1a	-14906.90	140
	M2a	-14870.83	142
	M7	-14866.98	140
	M8	-14825.61	142
G6PD	M0	-13510.87	117
	M1a	-13179.71	118
	M2a	-13178.28	120
	M7	-13067.64	118
	M8	-13060.19	120
MECP2	M0	-11815.60	135
	M1a	-11693.69	136
	M2a	-11693.69	138
	M7	-11641.04	136
	M8	-11635.17	138
OTC	M0	-8828.16	149
	M1a	-8631.83	150
	M2a	-8629.23	152
	M7	-8610.52	150
	M8	-8603.09	152

Table 3 Likelihood ratio test for model comparisons

Gene	Test	df	LRT
ABCD1	M0 vs. M1a	1	780,98**
	M0 vs. M2a	3	780,98**
	M0 vs. M7	1	1305,03**
	M0 vs. M8	3	1327,10**
	M1a vs. M2a	2	0,00
	M7 vs. M8	2	22,07**
AR	M0 vs. M1a	1	625,08**
	M0 vs. M2a	3	625,08**
	M0 vs. M7	1	908,90**
	M0 vs. M8	3	912,56**

	M1a vs. M2a	2	0,0
	M7 vs. M8	2	3,68
BMP15	M0 vs. M1a	1	486,28**
	M0 vs. M2a	3	486,28**
	M0 vs. M7	1	550.62**
	M0 vs. M8	3	562,40**
	M1a vs. M2a	2	0
	M7 vs. M8	2	11,78**
COL4A5	M0 vs. M1a	1	1552,49**
	M0 vs. M2a	3	1556,95**
	M0 vs. M7	1	1605,35**
	M0 vs. M8	3	1663,87**
	M1a vs. M2a	2	4,45
	M7 vs. M8	2	58,51**
EDA	M0 vs. M1a	1	225,45**
	M0 vs. M2a	3	225,45**
	M0 vs. M7	1	243,42**
	M0 vs. M8	3	243,42**
	M1a vs. M2a	2	0
	M7 vs. M8	2	0,008
F9	M0 vs. M1a	1	1271,48**
	M0 vs. M2a	3	1349,18**
	M0 vs. M7	1	1393,18**
	M0 vs. M8	3	1472,18**
	M1a vs. M2a	2	77,70**
	M7 vs. M8	2	79,00**
GLA	M0 vs. M1a	1	1419,08**
	M0 vs. M2a	3	1491,22**
	M0 vs. M7	1	1498,93**
	M0 vs. M8	3	1581,67**
	M1a vs. M2a	2	72,14**
	M7 vs. M8	2	82,74**
G6PD	M0 vs. M1a	1	662,33**
	M0 vs. M2a	3	665,18**
	M0 vs. M7	1	886,46**
	M0 vs. M8	3	901,36**
	M1a vs. M2a	2	0,24
	M7 vs. M8	2	14,90*
MECP2	M0 vs. M1a	1	243,83**
	M0 vs. M2a	3	243,83**
	M0 vs. M7	1	349,12**

	M0 vs. M8	3	360,85**
	M1a vs. M2a	2	0
	M7 vs. M8	2	11,73**
OTC	M0 vs. M1a	1	392,66**
	M0 vs. M2a	3	397,86**
	M0 vs. M7	1	435,28**
	M0 vs. M8	3	450,14**
	M1a vs. M2a	2	5,2
	M7 vs. M8	2	14,86**

Conclusion

Sequence comparisons between species showed that most amino acid positions where disease-associated missense mutations were previously detected in humans remained invariant among other mammals. The high degree of conservation indicated that those positions were subject to strong evolutionary constraints.

The pathogenicity assessment of CPDs mutations and inter-species polymorphisms using prediction tools provided discrepant results. Our analysis revealed that while the tools used are effective in scoring the pathogenic effect of novel mutations, their performance is lower with CPDs. This highlights the need for multi-faceted approaches in mutation pathogenicity prediction, integrating evolutionary, structural, and functional data and to account the existence of compensated mutations when choosing the prediction tool to use. Perhaps, with the democratization of artificial intelligence technologies and more robust machine learning methodologies, it will be possible to create new tools to account such unique variation as compensated mutations, which will contribute to early disease management and diagnosis.

Positive selection signals were detected in a number of codons of several genes, suggesting a scenario of adaptive evolution. The identification of amino acids under positive selection, particularly in amino acid positions associated with compensated mutations like Arg26 in *OTC* and Arg3 in *F9*, demonstrates the dynamic nature of evolutionary processes and the interplay of compensatory effects. The evolutionary tolerance at both sites, coupled with signs of positive selection, support the hypothesis of compensatory evolutionary mechanisms.

Only deciphering the intricate relationship between evolutionary conservation, mutation pathogenicity and selective pressures, will offer the key insights into the molecular mechanisms underlying human diseases.

Chapter 3: Genetic variants interaction as phenotypic modifiers

Compensatory epistasis explored by molecular dynamics simulations

Catarina Serrano; Carla S.S. Teixeira; David N. Cooper; João Carneiro; Monica Lopes-Marques; Peter D. Stenson; António Amorim; Maria J. Prata; Sérgio F. Sousa; Luísa Azevedo

Human Genetics, 2021, vol. 140, no. 9, pp. 1329–1342

(doi: 10.1007/s00439-021-02307-x)

ABSTRACT

A non-negligible proportion of human pathogenic variants are known to be present as wild-type in at least some non-human mammalian species. The standard explanation for this finding is that molecular mechanisms of compensatory epistasis can alleviate the mutations' otherwise pathogenic effects. Examples of compensated variants have been described in the literature but the interacting residue(s) postulated to play a compensatory role have rarely been ascertained. In this study, the examination of five human X-chromosomally-encoded proteins (FIX, GLA, HPRT1, NDP and OTC) allowed us to identify several candidate compensated variants. Strong evidence for a compensated/compensatory pair of amino acids in the coagulation FIXa protein (involving residues 270 and 271) was found in a variety of mammalian species. Both amino acid residues are located within the 60-loop, spatially close to the 39-loop that performs a key role in coagulation serine proteases. To understand the nature of the underlying interactions, molecular dynamics simulations were performed. The predicted conformational change in the 39-loop consequent to the Glu270Lys substitution (associated with hemophilia B) appears to impair the protein's interaction with its substrate but, importantly, such steric hindrance is largely mitigated in those proteins that carry the compensatory residue (Pro271) at the neighboring amino acid position.

Human Genetics (2021) 140:1329–1342
https://doi.org/10.1007/s00439-021-02307-x

ORIGINAL INVESTIGATION

Compensatory epistasis explored by molecular dynamics simulations

Catarina Serrano[1,2,3] · Carla S. S. Teixeira[4] · David N. Cooper[5] · João Carneiro[6] · Mónica Lopes-Marques[1,2,3] · Peter D. Stenson[5] · António Amorim[1,2,3] · Maria J. Prata[1,2,3] · Sérgio F. Sousa[4] · Luísa Azevedo[1,2,3]

Received: 16 April 2021 / Accepted: 20 June 2021 / Published online: 26 June 2021

Abstract
A non-negligible proportion of human pathogenic variants are known to be present as wild type in at least some non-human mammalian species. The standard explanation for this finding is that molecular mechanisms of compensatory epistasis can alleviate the mutations' otherwise pathogenic effects. Examples of compensated variants have been described in the literature but the interacting residue(s) postulated to play a compensatory role have rarely been ascertained. In this study, the examination of five human X-chromosomally encoded proteins (FIX, GLA, HPRT1, NDP and OTC) allowed us to identify several candidate compensated variants. Strong evidence for a compensated/compensatory pair of amino acids in the coagulation FIXa protein (involving residues 270 and 271) was found in a variety of mammalian species. Both amino acid residues are located within the 60-loop, spatially close to the 39-loop that performs a key role in coagulation serine proteases. To understand the nature of the underlying interactions, molecular dynamics simulations were performed. The predicted conformational change in the 39-loop consequent to the Glu270Lys substitution (associated with hemophilia B) appears to impair the protein's interaction with its substrate but, importantly, such steric hindrance is largely mitigated in those proteins that carry the compensatory residue (Pro271) at the neighboring amino acid position.

Introduction

It is almost a truism that the evolutionary conservation of any given amino acid residue in a particular protein reflects its relevance to that protein's function. Thus, if a specific amino acid position is evolutionarily highly conserved, this is indicative of enduring functionality for the position and any substitution in humans is predicted to be associated with pathogenicity (de Beer et al. 2013; Kumar et al. 2009; Subramanian and Kumar 2006). The corollary to this is that as the evolutionary conservation of the site in which the amino acid substitution occurs decreases, the lower is the severity of the resulting clinical phenotype (Miller and Kumar 2001; Wacey et al. 1994).

This notwithstanding, the genetic context in which any given pathogenic variant occurs may be responsible for exacerbating its deleterious effect or, conversely, for ameliorating its pathogenicity through an epistatic mechanism of compensation between variants (Genin et al. 2008; Jordan

Sérgio F. Sousa and Luísa Azevedo contributed equally to this manuscript.

✉ Sérgio F. Sousa
sergiofsousa@med.up.pt

✉ Luísa Azevedo
lazevedo@ipatimup.pt

[1] i3S, Instituto de Investigação e Inovação em Saúde, Population Genetics and Evolution Group, Universidade do Porto, Rua Alfredo Allen 208, 4200-135 Porto, Portugal

[2] IPATIMUP-Institute of Molecular Pathology and Immunology, University of Porto, Rua Júlio Amaral de Carvalho 45, 4200-135 Porto, Portugal

[3] Department of Biology, Faculty of Sciences, University of Porto, Rua Do Campo Alegre, s/n, 4169-007 Porto, Portugal

[4] UCIBIO/REQUIMTE, BioSIM, Departamento de Biomedicina, Faculdade de Medicina da Universidade do Porto, Porto, Portugal

[5] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK

[6] CIIMAR, Interdisciplinary Centre of Marine and Environmental Research, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, 4450-208 Matosinhos, Portugal

et al. 2015; van Leeuwen et al. 2016). Such a pairwise compensatory interaction is exemplified by the occurrence of deleterious variants in humans that in non-human species constitute wild-type residues (Azevedo et al. 2009, 2016; Mouse Genome Sequencing Consortium 2002; Ferrer-Costa et al. 2007; Kondrashov et al. 2002; Marin et al. 2019; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). If a deleterious variant in a human disease-associated protein does not have the same detrimental effect on an orthologous protein in other evolutionarily related species, the toleration of the variant in non-human genomes can be explained, at least in some cases, by antagonistic epistatic interactions, in which the effect of a deleterious allele is ameliorated by the co-occurrence of another variant that acts to introduce a compensatory amino acid residue into the same protein or an interacting partner protein (Jordan et al. 2015; Kondrashov et al. 2002; Suriano et al. 2007). Such sequence variants have been termed 'compensated pathogenic deviations' (CPDs) (Kondrashov et al. 2002). Numerous descriptions of CPDs have emerged through comparative analyses of genomes from multiple mammalian species (Chimpanzee Sequencing and Analysis Consortium 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; Xue et al. 2015; Zhang et al. 2010, 2011). Further examples of CPDs were identified among human inherited pathogenic variants, and it has been estimated that at least 3.7% of human pathogenic missense variants correspond to CPDs (Azevedo et al. 2016). Despite the increasing number of CPDs reported in the literature, only rarely has the compensatory site been identified (Azevedo et al. 2009; Jordan et al. 2015; Kondrashov et al. 2002) and even more rarely has experimental validation been obtained (e.g., Jordan et al. 2015; Suriano et al. 2007).

Since in almost all mammals the X-chromosome is present as a single copy in males, the process of identifying and validating CPDs can be simplified by examining X-linked genes. Therefore, taking advantage of this natural model system, in which the action of purifying selection is expected to be fully exposed in hemizygous individuals, we opted to study X-chromosome-encoded disease proteins. Five X-chromosome-encoded proteins were specifically targeted viz. coagulation factor IX (FIX), alpha-galactosidase A (GLA), hypoxanthine–guanine phosphoribosyltransferase (HPRT1), norrin (NDP) and ornithine transcarbamylase (OTC).

With few exceptions, eutherian X-chromosomes are evolutionarily highly conserved both in terms of gene content and function, with only a few genes known to escape dosage compensation and monoallelic expression. We, therefore, operated under the assumption that if an allele that was deleterious in humans appeared to be fixed in a non-human species, it was very likely to co-occur in the latter with compensatory partner amino acid residue(s); otherwise, it would have been long since removed from the population under the action of purifying selection. This study represents a concerted attempt to explore the molecular basis of the epistatic interactions between two residues in mammalian genomes. To address the issue, we employed protein modeling and molecular dynamics simulations, key tools for the study of intra-molecular amino acid interactions (Castellana et al. 2021; O'Rourke et al. 2016).

Methods

Proteins and disease-causing variants

Missense mutation density (i.e. the number of pathogenic missense variants per protein sequence length in amino acids) was determined from data retrieved from the Human Gene Mutation Database (HGMD) (Stenson et al. 2020). The five proteins encoded by X-linked genes with the highest mutation density and with high-resolution crystal structures available from the Protein Data Bank (Berman et al. 2000) were selected for this study: FIX, GLA, HPRT1, NDP and OTC. These proteins are associated, respectively, with the following Mendelian conditions: hemophilia B (OMIM 300746), Fabry disease (OMIM 300644), Lesch-Nyhan syndrome (OMIM 308000), Norrie disease (OMIM 300658) and ornithine transcarbamylase deficiency (OMIM 300461). A graphical representation of the analyses performed is shown in Fig. 1 and is detailed below.

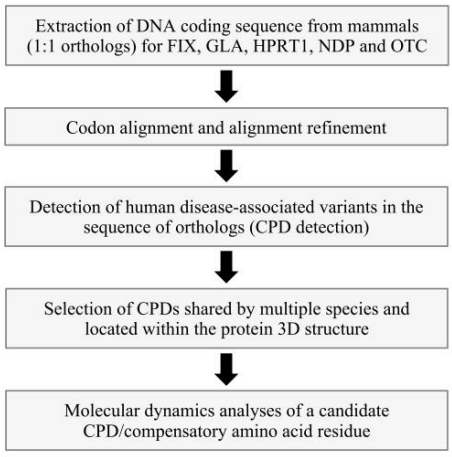


Fig. 1 Workflow of the strategy used in this work

Multiple sequence alignment and evolutionary analysis

DNA coding sequences (CDS) from placental mammals were retrieved from Ensembl Genome Browser v.97 and v.100 (Cunningham et al. 2019; Yates et al. 2020). The corresponding protein accession numbers are given in Supplementary Table S1. Using an in-house Python script, sequences with more than 1% unknown nucleotides in a DNA sequence and/or with a size difference of at least 25% compared to the human reference sequence were removed from the dataset to ensure valid comparisons between bona fide orthologous sequences and to avoid the analysis of less well-characterized sequences. A codon-based multiple sequence alignment (MSA) was built using PRANK (Loytynoja and Goldman 2010) and the resulting alignments were manually inspected.

CPD variant dataset and their putative compensatory partners

A total of 628, 428, 259, 170 and 76 disease-causing missense variants for FIX, GLA, OTC, HPRT1 and NDP, respectively, were retrieved from the HGMD (Stenson et al. 2020). Protein residues in the mammalian orthologues that harbored the human deleterious variants as wild-type amino acids were classified as potential CPDs. To maximize the probability of correctly identifying the corresponding compensatory partners, only cases in which the CPD was shared by at least six mammalian species were considered, under the assumption that the larger the number of species harboring a given CPD, the higher the probability that they would share a recent common ancestor and, therefore, the higher the probability that all the species in which the CPD occurs will share the same compensatory partner (Supplementary Table S2).

Three-dimensional (3D) analysis was performed on the crystal structure of the serine protease catalytic domain of coagulation FIXa (the active enzymatic form of FIX) (6MV4 (Vadivel et al. 2019)) and a dimeric structure of α-galactosidase (3s5z (Guce et al. 2011)) available from the Protein Data Bank (Berman et al. 2000). For each CPD shared by at least six species and located within the crystal structure of the corresponding protein (6MV4 and 3S5Z for FIXa and GLA, respectively), putative compensatory variants were identified by protein sequence comparisons. A variant was considered to be a putative compensatory partner for a CPD when it co-occurred in all species in which the CPD was present and the amino acid residue differed from the wild-type found in humans. The molecular distance between the CPD and its putative compensatory variant was determined with PyMOL software (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC) (Supplementary Table S3).

Human coagulation FIXa model setup and parameterization

A molecular model for the serine protease domain of wild-type human coagulation FIXa was prepared from 6MV4 (Vadivel et al. 2019), an X-ray crystallographic structure with a resolution of 1.37 Å. The protonation states of all the amino acid residues were predicted using PropKA version 3.0 at pH 7.0 (Olsson et al. 2011). The system was prepared using the AMBER18 software package and LEAP, using the ff14SB force field (Maier et al. 2015). Charges on the system were neutralized through the addition of counterions (Na^+) and the system was placed in a TIP3P water box with a minimum distance of 12 Å between the protein surface and the side of the box, using the LEAP module of AMBER. For the mutants, the initial wild-type molecule was modeled with the mutagenesis feature in PyMOL 1.7.2.1 using the Dunbrack rotamer libraries available (Shapovalov and Dunbrack 2011). Selection of the initial amino-acid conformation of each mutant was based on the dominant conformer predicted, excluding clashes with other amino acid residues in the protein. This computational mutagenesis protocol has been previously used successfully in the study of other enzymes (Ferreira et al. 2017). The mutant systems were subject to the same solvation and neutralization protocol described for the wild-type protein. The modeled FIXa systems are listed in Table 1.

Molecular dynamics simulations

All systems were subjected to four consecutive energy minimization stages to remove clashes prior to the molecular dynamics (MD) simulation. In these four stages, the minimization procedure was applied to the following atoms of the system: 1—water molecules (2500 steps); 2—hydrogen atoms (2500 steps); 3—side chains of the amino acid residues (2500 steps); 4—full system (10,000 steps). The energy minimized systems were then subjected to a two-stage MD equilibration procedure: in the first stage (50 ps), the systems

Table 1 Protein models with FIXa variants simulated through molecular dynamics

Protein	Position 270	Position 271
Wild-type	Glutamate	Threonine
Glu270Lys Mutant	Lysine	Threonine
Glu270Lys + Thr271Pro Mutant	Lysine	Proline
Thr271Pro Mutant	Glutamate	Proline

were gradually heated to 310.15 K using a Langevin thermostat at constant volume (NVT ensemble); in the second stage (50 ps), the density of the systems was further equilibrated at 310.15 K.

Finally, MD production simulation runs were performed for 100 ns for the wild-type FIXa protein and for the three mutants listed in Table 1. These were performed with an NPT ensemble at constant temperature (310.15 K, Langevin thermostat) and pressure (1 bar, Berendsen barostat), with periodic boundary conditions, with an integration time of 2.0 fs using the SHAKE algorithm to constrain all covalent bonds involving hydrogen atoms. A 10 Å cutoff for non-bonded interactions was used during the entire molecular simulation procedure. Coordinates were saved at each 10 ps. This procedure (energy minimization to MD production simulation) was repeated twice to obtain two MD replicates for each FIXa variant listed in Table 1. Final trajectories were analyzed in terms of backbone Root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), cluster analysis, hydrogen bonds formed, solvent accessible surface areas (SASAs) and dynamic cross-correlation maps (DCCM). The DCCMs were obtained with the Bio3D R package (Grant et al. 2006).

Results and discussion

Identification and evolutionary conservation of CPDs in non-human placental mammals

For each protein (FIX, GLA, HPRT1, NDP and OTC), an MSA alignment derived from placental mammalian sequences was analyzed and manually annotated with disease-causing variants. Initially, homologous mammalian amino acid positions harboring a human pathogenic allele as the wild-type in any of the species studied were considered to be potential CPDs. The complete set of the potential CPDs identified is given in Fig. 2a and Supplementary Table S2. The proportion of pathogenic missense variants assumed to be CPDs varied between 0.5% (HPRT1) and 4.2% (GLA) in the set of proteins examined (Fig. 2b), which is in accord with previous findings (Azevedo et al. 2016). In some cases, the CPDs were detected in only one or very few species, whereas in other cases the CPDs were spread over a large number of species. For example, the FIX pathogenic variants Glu142Lys, Val257Ile and Glu323Lys were present in about one half of

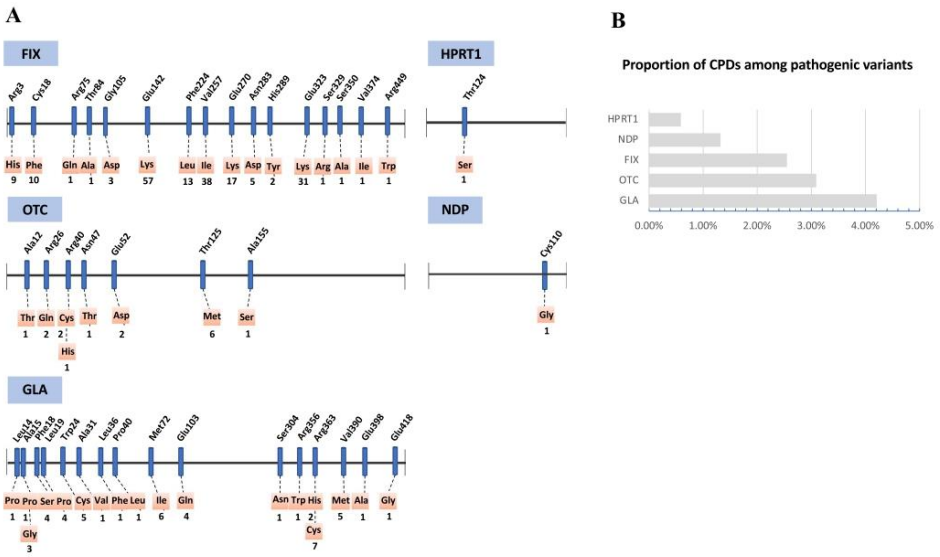


Fig. 2 **A** CPDs detected through the comparison of FIX, GLA, HPRT1, NDP and OTC in different mammalian species. Orange spheres represent the human deleterious missense variants that correspond to the wild-type amino acid residues in non-human mammalian species. Numbers below the orange spheres indicate the number of species in which the CPD was found. **B** Proportion of CPDs found in each protein

the mammalian species examined whereas Arg75Gln and Thr84Ala were each found in only one species (Fig. 2a, Supplementary Table S2 and Supplementary Fig. S1).

Identification of putative compensatory amino acid partners

Under the hypothesis that our candidate CPDs are tolerated in the genomes of non-human mammalian species due to compensatory interactions with other amino acids, we next attempted to identify the putative compensatory partners so that the CPD/compensatory pairs could be studied by MD. The strategy adopted was to seek amino acid differences between the human protein and its orthologues harboring the CPDs, under the assumption that any difference between the human and non-human sequences would represent a potential compensatory amino acid position. If the analysis was based on comparisons made with one or a very few species, the amino acid differences with respect to the human protein would be too numerous to reach meaningful conclusions. Therefore, to increase the prospect of finding bona fide compensatory sites among those common to a group of species while being distinct from the human sequence, we adopted a conservative approach whereby the results were filtered so that we considered only CPDs present in at least six mammalian species. A total of seven candidate CPDs (Arg3His, Cys18Phe, Glu142Lys, Phe224Leu, Val257Ile, Glu270Lys and Glu323Lys) in FIX, two (Met72Ile and Arg363Cys) in GLA and one (Thr125Met) in OTC, matched this criterion. Because the OTC Thr125Met has been previously studied by enzyme activity assays (Suriano et al. 2007), it was not explored further by MD simulations.

Since the available 3D structures for FIXa and GLA do not encompass the entirety of these proteins, only three candidate CPDs in FIXa (Val257Ile, Glu270Lys and Glu323Lys) and two in GLA (Met72Ile and Arg363Cys) were contained within the retrieved 3D structures. The list of all shared amino acid differences between the human and other mammalian FIXa and GLA proteins for these five CPDs are shown in Supplementary Table S3. In four out of the five putative CPDs, the number of differences between the human and non-human proteins was too high to identify the compensatory site(s) with any degree of certainty, according to the aforementioned criteria. However, in the case of Glu270Lys in FIXa, two putative compensatory residues were disclosed: Pro271 and Ile368. As far as Ile368 is concerned, the 3D distance to the mutated amino acid at residue 270 (24 Å, Supplementary Table S3) renders it unlikely to be the compensatory partner. By contrast, the immediate proximity of Pro271 reinforces our argument that Pro271 could be the sought-after compensatory site. We, therefore, retained the FIXa 270/271 pair to be further investigated by MD simulations. Thus, from our initial list of 44 possible CPDs, we ended up with a single high-confidence CPD/compensatory pair suitable for MD simulations. This protracted process demonstrates how challenging the identification of a bona fide compensatory partner for a given CPD can be, a process which can be hampered by the extensive background genetic variability that exists between orthologous protein sequences.

Despite the juxtaposition of Lys270 and Pro271 in the FIXa 3D structure, the question nevertheless remained as to whether they constituted a genuine CPD/compensatory pair. We next recruited MD simulations as a tool to establish whether the putative compensatory variant might serve to stabilize the protein that would otherwise be destabilized by the pathogenic variant. To this end, MD simulations were performed and compared i) between sequences containing Glu270 or Lys270, to assess whether the differences between both simulations could provide clues as to the changes in function and likely pathogenicity associated with Lys270, and ii) between sequences containing different combinations of Glu270Lys and Thr271Pro to establish if the variant Thr-271Pro could act as a suppressor of the pathogenic effect of Glu270Lys.

Analyses of the impact of amino acid variants at positions 270 and 271 in the catalytic domain of FIXa

To investigate the possible interactions between Glu270Lys and Thr271Pro, four FIXa structures were prepared containing the four variants described in Table 1 (wild-type, Glu270Lys, Glu270Lys + Thr271Pro and Thr271Pro); and the RMSD of the alpha carbon atoms (Cα) of all residues from the FIXa catalytic domain were calculated along the 100 ns of the MD simulations.

The analysis of the graphical representation of the RMSD of all structures (Fig. 3) shows that the protein containing the Glu270Lys deleterious variant has a distinctive profile when compared with the other three FIXa structures. In line with the data from both the graphical RMSD representation and the 2D RMSD plots, the calculated RMSD mean and respective standard deviation of the RMSD values reveal that the FIXa Glu270Lys protein has a higher mean RMSD and a higher standard deviation when compared with the other three protein variants (Supplementary Table S4). The addition of the compensatory variant (Glu270Lys + Thr271Pro) reduced the RMSD of the protein to a value closer to that of the wild-type protein. The RMSD profiles also indicated that FIXa proteins stabilize at 40 ns of simulation, with the exception of the FIXa Glu270Lys protein which shows an irregular pattern with a higher standard deviation. Analysis of equilibrium properties was based on the last 60 ns of the MD simulations. These results are in agreement with the analysis of the RMSD of the Cα atoms from the residues 265

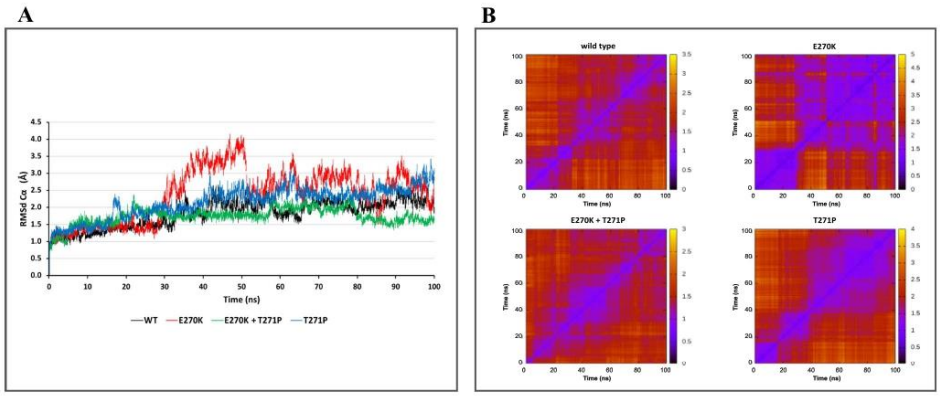


Fig. 3 **A** Graphical representation of the RMSd of the alpha carbon atoms from the catalytic domain residues from the wild-type, Glu270Lys, Glu270Lys + Thr271Pro and Thr270Pro FIXa proteins along the 100 ns of MD simulation. **B** 2D RMSd plots for the MD simulations of the different FIXa variants. The simulation time is plotted on both the *X* and *Y* axes with the RMSd values represented as a gradation in color

to 274 of the loop containing the mutated residues (60-loop) (Supplementary Table S4). The analysis of the overall solvent accessible surface area (SASA) of the different proteins (Supplementary Table S4) did not reveal any meaningful differences.

To evaluate the influence of the Glu270Lys, Glu270Lys + Thr271Pro and Thr271Pro mutations in the dynamic behavior of the FIXa protein, DCCMs of Cα atoms of all residues from the catalytic domain were calculated for the four variants. Comparison of the four resulting DCCMs (Fig. 4) clearly shows a distinct pattern for the FIXa with the Glu270Lys variant, with an increased level of correlated and anti-correlated motions, particularly in the ranges between residues 257 and 277 (which includes the 60-loop) and between residues 377 and 477.

At this stage, both RMSD and DCCM analyses clearly indicated that the structural impact of the human deleterious variant Glu270Lys is particularly evident in the 60-loop but that such an impact is ameliorated by the inclusion of the putative compensatory partner Thr271Pro.

Analyses of the flexibility of the FIXa catalytic domain

To identify the amino acid residues associated with the differences in behavior observed between the four FIXa protein sequences in the previous analyses, a Root-Mean-Square Fluctuation (RMSF) calculation of the Cα atoms of all residues from the catalytic domain was performed. RMSF measures the relative positional variability (i.e. flexibility) of the different amino acid positions along the protein backbone, in other words, how much a particular residue fluctuates during a simulation, thereby allowing a distinction to be made between highly and poorly mobile regions of a protein. The analysis of the four different profiles demonstrated that the residues which stand out in the overall profile are Lys247 (from the 39-loop), plus residues Arg384 and Lys387. The flexibility of these three residues is clearly increased in the Glu270Lys protein and reduced in the Glu270 + Thr271 protein, becoming closer to that of the wild-type protein (Fig. 5a).

The analysis of the RMSF values from the residues that comprise the active site (Fig. 5b; active site) further demonstrates that the flexibility of residue Ser411 (which contributes to the catalytic triad His221-Asp269-Ser365) is markedly higher in the wild type FIXa. All three mutant FIXa proteins present similar RMSF values, even though a small difference is evident in the double mutant FIXa Glu270Lys + Thr271Pro, whose RMSF is slightly higher and closer to the wild-type. The flexibility of the 39-loop (Fig. 5b: 39-loop) is significantly affected by the introduction of the Glu270Lys variant. The RMSF values of this loop are clearly higher for the FIXa Glu270Lys protein than for the remaining three proteins. The introduction of a putative compensatory Thr271Pro variant serves to revert this increase in flexibility to values closer to the wild-type, whereas the single mutant Thr271Pro does not appear to affect the 39-loop.

Analysis of the 60-loop (Fig. 5b; 60-loop), where the Glu270Lys and Thr271Pro variants are located (yellow bars

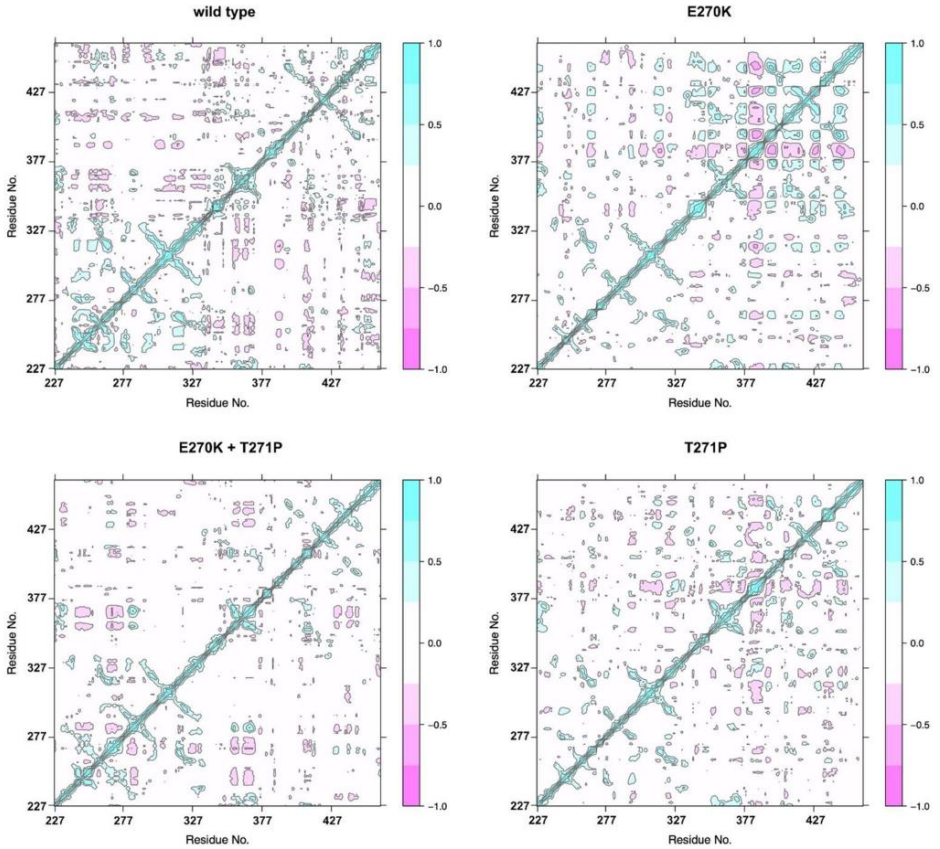


Fig. 4 DCCM analyses of Cα atoms of all amino acid residues from the FIXa catalytic domain

in Fig. 5b), reveals some small differences between the four RMSF profiles, although the overall behavior of the loop in terms of RMSF is similar between the four FIXa proteins.

The flexibility of the region between residues 375–393 (Fig. 5b, last group) is profoundly disturbed by the introduction of the Glu270Lys variant. Residues 384 and 387 from the FIXa Glu270Lys protein clearly stand out from the RMSF profile. The introduction of the compensatory Thr-271Pro variant, however, reverses this effect and lowers the flexibility of residues 385–390 to values significantly lower than those obtained for the wild-type FIXa protein. The Thr271Pro replacement on its own does not appear to have such a marked effect in this region, although it increases the flexibility of residues 385 and 386 to RMSF values very close to those obtained for the FIXa Glu270Lys mutant.

Taken together, the results obtained from the RMSF analyses are in agreement with data from both the RMSD and DCCM analyses in which the structural disturbance caused by the FIXa disease-associated variant Glu270Lys was less dramatic in relation to the wild type when the Thr271Pro was added to the background, thereby reinforcing the compensatory effect that a proline at position 271 has when it co-occurs in *cis* with a lysine at position 270. To explain the structural differences between the Glu270Lys mutant and the wild type a cluster analysis was used as described in the next section.

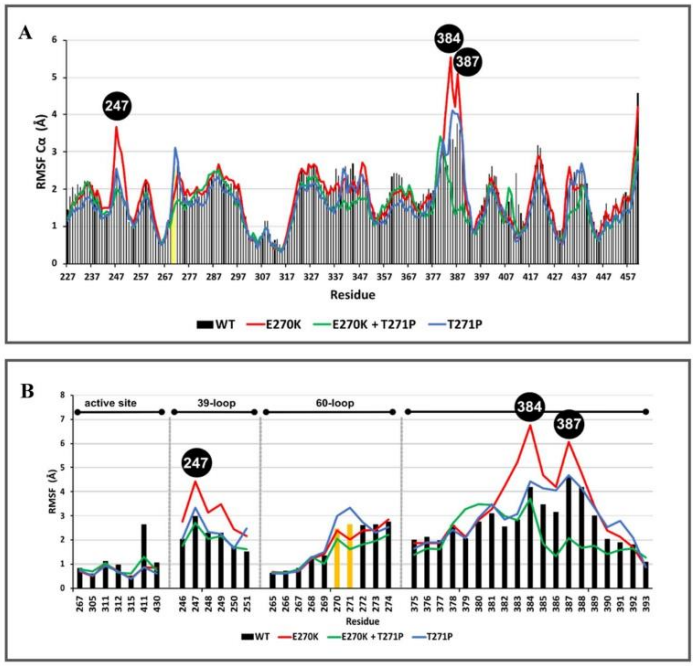


Fig. 5 **A** Graphical representation of the RMSF of the Cα atom of each amino acid residue in the catalytic domain of wild-type, Glu270Lys, Glu270Lys + Thr271Pro and Thr270Pro proteins along the last 60 ns of MD simulation. **B** Graphical representation of the RMSF of all atoms from the residues of the active site, 39-loop, 60-loop and 375–393 of wild-type, Glu270Lys, Glu270Lys + Thr271Pro and Thr270Pro proteins along the last 60 ns of MD simulations

Cluster analysis of the 39-loop and residues 375–393

To obtain the representative structures of the FIXa variants containing different conformations for the 39-loop and residues 375–393, a cluster analysis was performed of the protein conformations adopted along the MD for the four variants, based on the variability at these regions. The MD structures of each FIXa protein were grouped into two clusters according to the RMSD for these regions. Table 2 summarizes the proportion of structures along each MD simulation that comprise each cluster. The wild-type, Glu270Lys + Thr271Pro and Thr271Pro proteins adopt a set of conformations distributed between two main clusters: a dominant cluster, that represents between 72.6% (WT) and 74.7% (Thr271Pro) of the conformations, and an alternative cluster, representing between 27.4% (WT) and 25.3% (Thr271Pro) of the conformations observed in the MD simulations. By contrast, Glu270Lys remains almost locked in a single dominant conformation during 98.5% of the simulation.

To obtain further insights into the major structural differences between the wild-type and mutant FIXa proteins, representative structures of the dominant cluster of each FIXa protein were structurally aligned. This alignment showed that Lys247 (39-loop) from the Glu270Lys + Thr271Pro structure adopts a rather distinctive conformation when compared with the other three versions of the protein. In addition, residues Arg384 and Lys387 also assume distinctive conformations in

Table 2 Summary of the cluster analysis with the estimated proportion of time spent in each cluster during the last 60 ns of MD simulation

	Wild-type	Glu270Lys	Glu270Lys + Thr271Pro	Thr271Pro
Dominant cluster (%)	72.6	98.5	71.4	74.7
Alternative cluster (%)	27.4	1.5	28.6	25.3

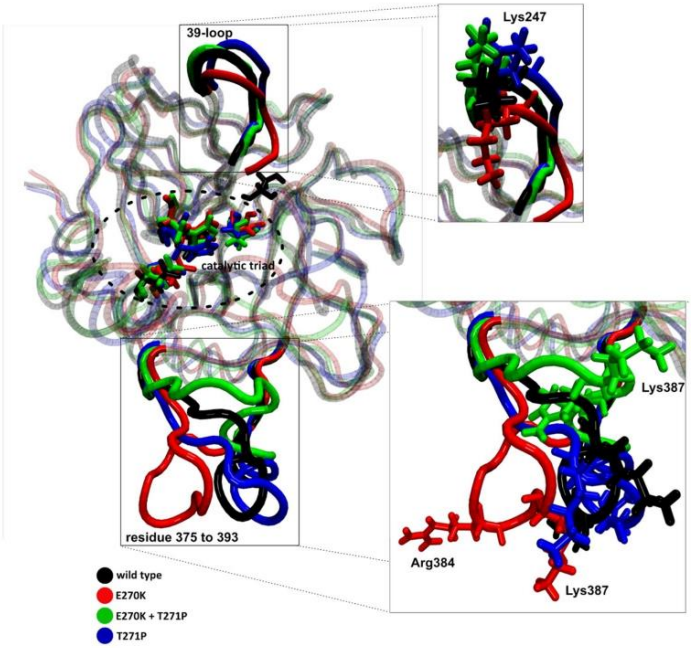


Fig. 6 Visual molecular dynamics (VMD) representation of the alignment of the representative structure from the dominant cluster from each FIXa protein. The residues with the higher RMSF variation are represented in licorice

the Glu270Lys and Glu270Lys + Thr271Pro proteins when compared both with each other and with the other two proteins (Fig. 6). Finally, the analysis of the 2D RMSD plots for the 100 ns of MD simulations of the 39-loop and residues 375 to 393 from the different FIXa variants shows that the Glu270Lys protein has the highest RMSD variation (1–10 Å) during the last 70 ns of the MD simulation whereas the Glu270Lys + Thr271Pro protein has the lowest range of RMSD variation (1–5 Å) during the 100 ns of simulation (Supplementary Fig. S2).

Both the mean RMSD and SASA from the 39-loop and residues 375–393 are consistent with the RMSF values that suggested relevant conformational changes in these particular locations between the different FIXa proteins (Supplementary Table S5).

In-depth analysis of the structural changes resulting from the Glu270Lys replacement

Analysis of the tertiary structure of the different proteins (Fig. 6) showed that residues Arg384 and Lys387 are too distant from the 60-loop to be able to interact directly with the loop and therefore with amino acid positions 270 and 271. Although the results from the RMSF, RMSD and SASA analyses clearly show that these residues undergo a conformational change in the Glu270Lys mutant protein that is not observed in the other proteins, it is not possible to directly correlate these changes with the Glu270Lys replacement. Nevertheless, we cannot rule out that the observed alterations in this region could have resulted from the influence of Lys270 via long-range electrostatic interactions or from the rearrangement of the network of hydrogen bonds.

On the other hand, analysis of Lys247 (part of the 39-loop, Fig. 6) revealed that it is spatially close to Glu270 and Thr271 (60-loop) and may therefore be influenced by the mutated residues. More specifically, inspection of the 3D structure of FIXa harboring Glu270Lys shows that the conformational change observed at Lys247 results from the formation of a stable hydrogen bond between the positively charged ε-amino group of Lys247 and the oxygen from the peptide bond of Cys268. Analysis of the distance between the nitrogen atom of the ε-amino group of Lys247 and the oxygen atom of the Cys268 confirms that these residues are closer together in the FIXa Glu270Lys protein than in the other three 3D structures during the last 60 ns of the simulation. Given this result, we then attempted to clarify how

Lys247 (from the 39-loop) is affected by residues 270 and 271 of the FIXa protein.

Closer inspection of the spatial arrangement of the moiety surrounding residue Cys268 clearly indicates that a bulky and charged amino acid such as a lysine requires free access to bend and become hydrogen bonded to the oxygen of Cys268. In the Glu270Lys + Thr271Pro and Thr-271Pro mutant proteins, the protruding side chain of Pro271 appears to prevent access of the lysine to the Cys268 oxygen (Fig. 7a). In the case of the wild-type and Glu270Lys proteins, the 271 residue is the same but, the proximity of a negatively charged Glu or a positively charged Lys affects its RMSF and hence its conformation. The RMSF of Thr271 is 2.65 Å for the wild-type and 2.02 Å for the Glu270Lys protein. The higher RMSF for the wild-type protein is consequent to the hydrogen bond interactions that the hydroxyl group of Thr271 may establish with Glu270 and which are absent in the Glu270Lys mutant.

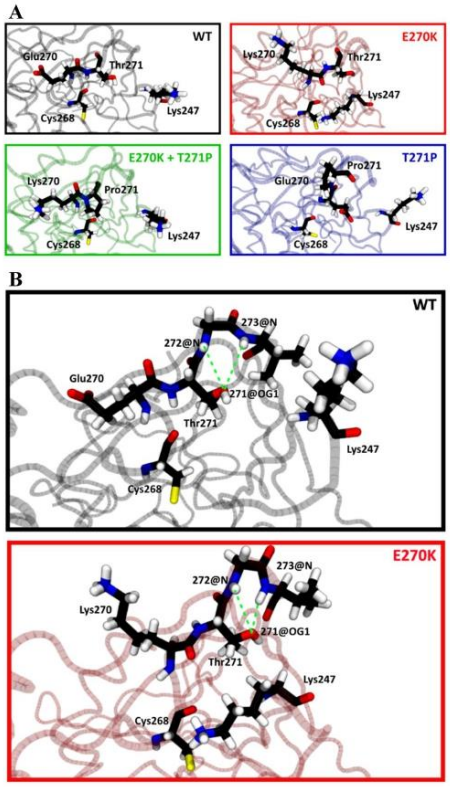


Fig. 7 VMD representation of **A** the spatial arrangement of residues Glu/Lys270, Thr/Pro271, Cys268 and Lys247 in the four FIXa proteins. **B** Hydrogen bonds established by the 271@OG1 atom from Thr271 in the wild-type and Glu270Lys mutant proteins

To evaluate how the absence of a nearby glutamate residue influences the conformational freedom of Thr271 in the wild-type and Glu270Lys mutant, the potential hydrogen bonds established between the Thr271 hydroxyl oxygen (271@OG1) and the neighboring amino acid residues were identified and characterized. Two hydrogen bonds were identified (represented by green dotted lines in Fig. 7b), one of them between the Thr271@OG1 and the nitrogen from the peptide bond of Gly272 (272@N); the other between the Thr271@OG1 and the nitrogen from the peptide bond of Val273 (273@N). According to the data in Supplementary Table S6, both Thr271@OG1–272@N and Thr271@OG1–273@N distances and standard deviations are greater for the wild-type protein suggesting that in the FIXa Glu270Lys protein the side chain of Thr271 is stabilized in a conformation that allows Lys247 to interact with Cys268. The Thr271 from the wild-type protein has greater conformational freedom which precludes access of Lys247 to the Cys268 oxygen.

Functional implications of replacements at residues 270 and 271 of the FIX protein

At a functional level, it is well established that the 39-loop has a vital role in all serine proteases of coagulation (Yang and Rezaie 2013), accounting for the restriction of substrate and inhibitor specificity. We, therefore, examined whether the human Glu270Lys deleterious variant might interfere with these functions. The alignment of the representative structure from the dominant cluster in the Glu270Lys simulation with the high-resolution structure of the Michaelis complex between pentasaccharide-activated antithrombin (AT) and human FIXa (PDB ID 3KCG) (Johnson et al. 2010) revealed that the conformational change of the 39-loop in the Glu270Lys mutant protein is likely to impair its interaction with the reactive center loop (RCL) from AT (Fig. 8). Since AT is a suicide-substrate inhibitor of FIXa (Law et al. 2006), its binding mode is expected to be similar to the binding mode of the natural substrate of FIXa—coagulation factor X. The alignment of catalytic domains from the compensated protein and from the 3KCG crystallographic structure showed that the presence of a proline at position 271 serves to reverse the conformational change observed in the Glu270Lys in the 39-loop, thereby strengthening

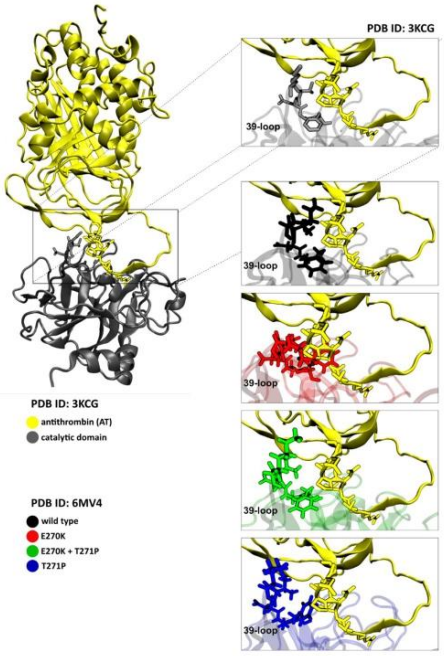


Fig. 8 VMD representation of the alignment between the catalytic domains of the representative structures of the dominant clusters taken from the MD simulations studies on the different FIXa protein variants and the catalytic domain from the crystallographic structure

the evidence supporting a structural compensatory effect for Pro271 when in cis with Lys270. The FIX protein with compensatory amino acid Pro271 on a wild-type background behaved similarly to the wild-type protein with Thr271, and it may therefore be assumed that Pro271 per se should not significantly affect FIXa activity.

Evolutionary history of the FIX Glu270Lys CPD and its compensatory partner

Having obtained strong evidence through MD simulations that positions 270 and 271 harbor, respectively, a CPD and its *cis*-compensatory partner, we sought to complete our analysis by integrating an evolutionary perspective into the alternative combinations involving the Glu270Lys-Thr271Pro pair during mammalian evolution. To this end, we constructed a phylogenetic tree of FIXa with sequences encompassing our amino acid residues of interest at positions 270 and 271.

The ancestral Pro271 appears to have mutated to a threonine in the primate lineage that preceded the divergence of Old World and New World monkeys approximately 43 MYA (Fig. 9). Mammalian FIX homologs harboring Pro271 have, over evolutionary time, also accepted residues with very different biochemical properties at position 270 such as Lys and His (basic), Glu (acidic), Val (nonpolar), and Gln and Asn (polar). However, once threonine became fixed at position 271, only two different residues with similar biochemical properties (both acidic) were subsequently found at position 270, namely aspartate and glutamate. This concurs with the considerable flexibility conferred by Pro271 in terms of its toleration of an evolutionary replacement at position 270. It should also be noted that residue Lys270 emerged recurrently (on at least four occasions) during mammalian evolution, but always on a background where residue Pro271 is present (Supplementary Table S7), suggesting that the two residues have indeed been subject to compensatory interaction.

In summary, we have provided evidence showing that is possible to identify CPDs and their compensatory partners by means of a combination of fine-scale computational and physicochemical methods. In the case of the CPD/compensatory pair identified, characterized and validated in this study, the lack of toleration of a FIX Glu270Lys substitution in humans (as evidenced by the hemophilia B phenotype) appears to be due to the replacement of the ancestral Pro271 by a threonine prior to the radiation of the primate lineage. Such epistatic interactions between amino acid residues are likely to have shaped the acceptability of certain amino acid residues on an evolutionary timescale. One consequence of this is that novel variants identified in human disease-associated proteins will sometimes be located at less evolutionarily conserved sites and may, therefore, escape detection by conventional methods of pathogenicity assessment that rely heavily upon measures of evolutionary conservation (Azevedo et al. 2016). In other words, a low level of evolutionary conservation at a particular amino acid residue in a given protein should not automatically be interpreted as a predictor of negligible pathogenicity when that residue is substituted. To circumvent this issue, variant prioritization should take into account the epistatic interactions between amino acid residues (e.g. Kim et al. 2019), to potentiate the identification of pathogenic mutations that occur at less well-conserved amino acid positions.

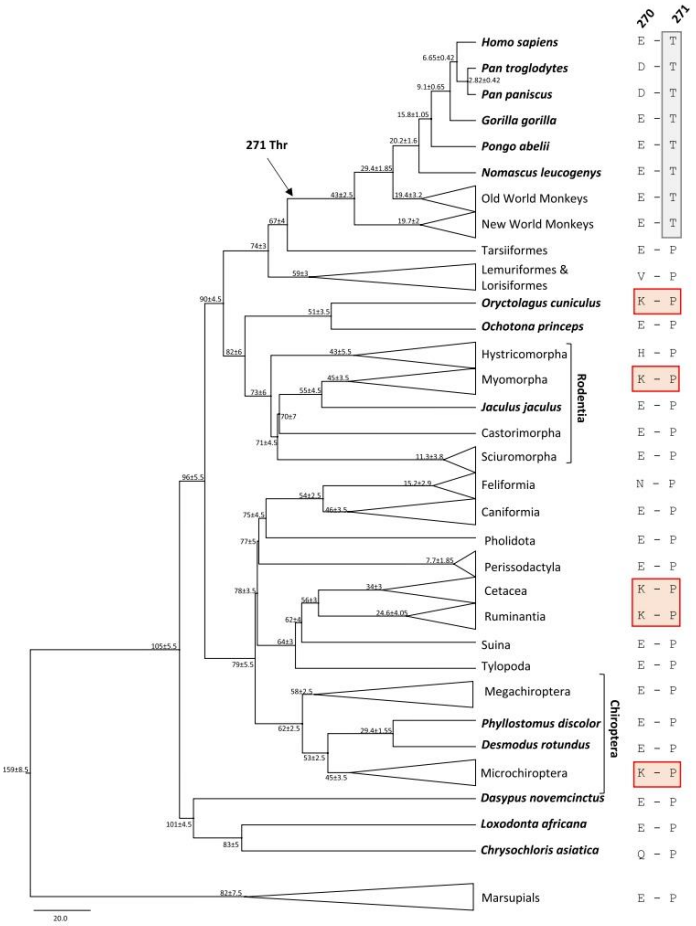


Fig. 9 Species tree of major mammalian lineages demonstrating the evolution of the residues homologous to the amino acid positions 270 and 271 of human FIX, constructed in TimeTree (Kumar et al. 2017). Values at tree nodes indicate the approximate time (in MYA) of divergence of each lineage

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-021-02307-x.

Funding This work was supported by Fundo Europeu de Desenvolvimento Regional (FEDER) through the COMPETE 2020—Operational Programme for Competitiveness and Internationalization, Portugal 2020 and by Foundation for Science and Technology (FCT) [POCI-01-0145-FEDER-007274, POCI-01-0145-FEDER-29723]; Foundation for Science and Technology [UIDB/04423/2020, UIDP/04423/2020, UIDP/04378/2020, UIDB/04378/2020, SFRH/BD/137925/2018 to C.S.]; Infraestrutura Nacional de Computação Distribuída (INCD) [01/SAICT/2016 number 022153, CPCA/A00/7140/2020, CPCA/A00/7145/2020] through funds from Foundation for Science and Technology and Fundo Europeu de Desenvolvimento Regional; Foundation for Science and Technology [under the transitional rule of Decree Law 57/2016, amended by Law 57/2017 to J.C.]; Foundation for Science and Technology [2020.01423.CEECIND to S.S.].

Code availability Not applicable.

Declarations

Conflict of interest The authors are unaware of any conflict of interest.

Availability of data Sequences used in this study were retrieved from the Ensembl Genome Browser (https://www.ensembl.org) and are shown in Supplementary Table S1. Human disease-causing missense variants were retrieved from HGMD (http://www.hgmd.cf.ac.uk). The crystal structure of FIXa (6MV4) and GLA (3S5Z) were retrieved from the Protein Data Bank (PDB) (https://www.rcsb.org).

References

Azevedo L, Carneiro J, van Asch B, Moleirinho A, Pereira F, Amorim A (2009) Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components. BMC Genom 10:266. https://doi.org/10.1186/1471-2164-10-266

Azevedo L, Mort M, Costa AC, Silva RM, Quelhas D, Amorim A, Cooper DN (2016) Improving the in silico assessment of pathogenicity for compensated variants. Eur J Hum Genet 25:2–7. https://doi.org/10.1038/ejhg.2016.129

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235

Castellana S, Biagini T, Petrizzelli F, Parca L, Panzironi N, Caputo V, Vescovi AL, Carella M, Mazza T (2021) MitImpact 3: modeling the residue interaction network of the respiratory chain subunits. Nucleic Acids Res 49:D1282–D1288. https://doi.org/10.1093/nar/gkaa1032

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87. https://doi.org/10.1038/nature04072

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, Dodiya KJ, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Laird MR, Lavidas I, Liu Z, Loveland JE, Marugan JC, Maurel T, McMahon AC, Moore B, Morales J, Mudge JM, Nuhn M, Ogeh D, Parker A, Parton A, Patricio M, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sparrow H, Stapleton E, Szuba M, Taylor K, Threadgold G, Thormann A, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Yates AD, Zerbino DR, Flicek P (2019) Ensembl 2019. Nucleic Acids Res 47:D745–D751. https://doi.org/10.1093/nar/gky1113

de Beer TAP, Laskowski RA, Parks SL, Sipos B, Goldman N, Thornton JM (2013) Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 Genomes Project dataset. PLoS Comput Biol 9:e1003382. https://doi.org/10.1371/journal.pcbi.1003382

Ferreira P, Sousa SF, Fernandes PA, Ramos MJ (2017) Improving the catalytic power of the DszD enzyme for the biodesulfurization of crude oil and derivatives. Chemistry 23:17231–17241. https://doi.org/10.1002/chem.201704057

Ferrer-Costa C, Orozco M, de la Cruz X (2007) Characterization of compensated mutations in terms of structural and physico-chemical properties. J Mol Biol 365:249–256. https://doi.org/10.1016/j.jmb.2006.09.053

Genin E, Feingold J, Clerget-Darpoux F (2008) Identifying modifier genes of monogenic disease: strategies and difficulties. Hum Genet 124:357–368. https://doi.org/10.1007/s00439-008-0560-2

Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csuros M, Glasscock J, Harris RA, Havlak P et al (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234. https://doi.org/10.1126/science.1139247

Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22:2695–2696. https://doi.org/10.1093/bioinformatics/btl461

Guce AI, Clark NE, Rogich JJ, Garman SC (2011) The molecular basis of pharmacological chaperoning in human alpha-galactosidase. Chem Biol 18:1521–1526. https://doi.org/10.1016/j.chembiol.2011.10.012

Johnson DJ, Langdown J, Huntington JA (2010) Molecular basis of factor IXa recognition by heparin-activated antithrombin revealed by a 1.7-Å structure of the ternary complex. Proc Natl Acad Sci USA 107:645–650. https://doi.org/10.1073/pnas.0910144107

Jordan DM, Frangakis SG, Golzio C, Cassa CA, Kurtzberg J, Task Force for Neonatal G, Davis EE, Sunyaev SR, Katsanis N (2015) Identification of *cis*-suppression of human disease mutations by comparative genomics. Nature 524:225–229. https://doi.org/10.1038/nature14497

Kim D, Han SK, Lee K, Kim I, Kong J, Kim S (2019) Evolutionary coupling analysis identifies the impact of disease-associated variants at less-conserved sites. Nucleic Acids Res 47:e94–e94. https://doi.org/10.1093/nar/gkz536

Kondrashov AS, Sunyaev S, Kondrashov FA (2002) Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci USA 99:14878–14883. https://doi.org/10.1073/pnas.232565499

Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ (2009) Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. Genome Res 19:1562–1569. https://doi.org/10.1101/gr.091991.109

Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol 34:1812–1819. https://doi.org/10.1093/molbev/msx116

Law RH, Zhang Q, McGowan S, Buckle AM, Silverman GA, Wong W, Rosado CJ, Langendorf CG, Pike RN, Bird PI, Whisstock JC (2006) An overview of the serpin superfamily. Genome Biol 7:216. https://doi.org/10.1186/gb-2006-7-5-216

Loytynoja A, Goldman N (2010) webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinform 11:579. https://doi.org/10.1186/1471-2105-11-579

Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) ff14SB: improving the accuracy of protein side

chain and backbone parameters from ff99SB. J Chem Theory Comput 11:3696–3713. https://doi.org/10.1021/acs.jctc.5b00255
Marin O, Aguirre J, de la Cruz X (2019) Compensated pathogenic variants in coagulation factors VIII and IX present complex mapping between molecular impact and hemophilia severity. Sci Rep 9:9538. https://doi.org/10.1038/s41598-019-45916-3
Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10:2319–2328. https://doi.org/10.1093/hmg/10.21.2319
Olsson MH, Sondergaard CR, Rostkowski M, Jensen JH (2011) PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. J Chem Theory Comput 7:525–537. https://doi.org/10.1021/ct100578z
O'Rourke KF, Gorman SD, Boehr DD (2016) Biophysical and computational methods to analyze amino acid interaction networks in proteins. Comput Struct Biotechnol J 14:245–251. https://doi.org/10.1016/j.csbj.2016.06.002
Shapovalov MV, Dunbrack RL Jr (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive Kernel density estimates and regressions. Structure 19:844–858. https://doi.org/10.1016/j.str.2011.03.019
Stenson PD, Mort M, Ball EV, Chapman M, Evans K, Azevedo L, Hayden M, Heywood S, Millar DS, Phillips AD, Cooper DN (2020) The Human Gene Mutation Database (HGMD((R))): optimizing its use in a clinical diagnostic or research setting. Hum Genet 139:1197–1207. https://doi.org/10.1007/s00439-020-02199-3
Subramanian S, Kumar S (2006) Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. BMC Genom 7:306. https://doi.org/10.1186/1471-2164-7-306
Suriano G, Azevedo L, Novais M, Boscolo B, Seruca R, Amorim A, Ghibaudi EM (2007) In vitro demonstration of intra-locus compensation using the ornithine transcarbamylase protein as model. Hum Mol Genet 16:2209–2214. https://doi.org/10.1093/hmg/ddm172
Vadivel K, Schreuder HA, Liesum A, Schmidt AE, Goldsmith G, Bajaj SP (2019) Sodium-site in serine protease domain of human coagulation factor IXa: evidence from the crystal structure and molecular dynamics simulations study. J Thromb Haemost 17:574–584. https://doi.org/10.1111/jth.14401
van Leeuwen J, Pons C, Mellor JC, Yamaguchi TN, Friesen H, Koschwanez J, Usaj MM, Pechlaner M, Takar M, Usaj M, VanderSluis B, Andrusiak K, Bansal P, Baryshnikova A, Boone CE, Cao J, Cote A, Gebbia M, Horecka G, Horecka I, Kuzmin E, Legro N, Liang W, van Lieshout N, McNee M, San Luis BJ, Shaeri F, Shuteriqi E, Sun S, Yang L, Youn JY, Yuen M, Costanzo M, Gingras AC, Aloy P, Oostenbrink C, Murray A, Graham TR, Myers CL, Andrews BJ, Roth FP, Boone C (2016) Exploring genetic suppression interactions on a global scale. Science 354:aag0839. https://doi.org/10.1126/science.aag0839
Wacey AI, Krawczak M, Kakkar VV, Cooper DN (1994) Determinants of the factor IX mutational spectrum in haemophilia B: an analysis of missense mutations using a multi-domain molecular model of the activated protein. Hum Genet 94:594–608. https://doi.org/10.1007/BF00206951
Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562. https://doi.org/10.1038/nature01262
Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN, de Manuel M, Hernandez-Rodriguez J, Lobon I, Siegismund HR, Pagani L, Quail MA, Hvilsom C, Mudakikwa A, Eichler EE, Cranfield MR, Marques-Bonet T, Tyler-Smith C, Scally A (2015) Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science 348:242–245. https://doi.org/10.1126/science.aaa3952
Yang L, Rezaie AR (2013) Residues of the 39-loop restrict the plasma inhibitor specificity of factor IXa. J Biol Chem 288:12692–12698. https://doi.org/10.1074/jbc.M113.459347
Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Marugan JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker A, Parton A, Patricio M, Sakthivel MP, Salam AA, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M, Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Flint B, Frankish A, Hunt SE, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR, Flicek P (2020) Ensembl 2020. Nucleic Acids Res 48:D682–D688. https://doi.org/10.1093/nar/gkz966
Zhang G, Pei Z, Krawczak M, Ball EV, Mort M, Kehrer-Sawatzki H, Cooper DN (2010) Triangulation of the human, chimpanzee, and Neanderthal genome sequences identifies potentially compensated mutations. Hum Mutat 31:1286–1293. https://doi.org/10.1002/humu.21389
Zhang G, Pei Z, Ball EV, Mort M, Kehrer-Sawatzki H, Cooper DN (2011) Cross-comparison of the genome sequences from human, chimpanzee, Neanderthal and a Denisovan hominin identifies novel potentially compensated mutations. Hum Genom 5:453–484. https://doi.org/10.1186/1479-7364-5-5-453

Chapter 4: Partial gene duplication

A partial duplication of an X-linked gene exclusive of a primate lineage (Macaca)

Catarina Serrano; Monica Lopes-Marques; António Amorim; Maria J. Prata; Luísa Azevedo

Gene, 2023, vol. 851

(doi: 10.1016/j.gene.2022.146997)

Abstract

Gene duplication plays a significant role in evolution. Paralogous gene copies may be lost due to the successive accumulation of deleterious mutations or remain active in the genome. In this work, a partial duplication of an X-linked region in the Macaca genus is identified and explored. Genomic comparisons reveal that the duplication encompasses the genes encoding ornithine transcarbamylase (OTC) and retinitis pigmentosa GTPase regulator (RPGR), spanning over 0.1 Mb on the chromosome 9 of Macaca. According to our analyses, the duplicated region of chromosome 9 involves partial coding sequences of both OTC and RPGR genes. Analyses of the selective pressures did not reveal significant differences in the ratio between nonsynonymous and synonymous mutations ($w<1$), suggesting that no selective pressures were acting in the evolutionary process. Reports for a biological role regarding some partial duplications exist in the literature, therefore, although being rare events, partial duplications of functionally important genes are worthy of study so that their impact can be explored.

Gene 851 (2023) 146997

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

A partial duplication of an X-linked gene exclusive of a primate lineage (*Macaca*)

Catarina Serrano [a], Mónica Lopes-Marques [a,*], António Amorim [b,c], Maria João Prata [a], Luísa Azevedo [a,*]

[a] i3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal
[b] IPATIMUP-Institute of Molecular Pathology and Immunology, University of Porto, Portugal
[c] Department of Biology, Faculty of Sciences, University of Porto, Portugal

ARTICLE INFO

Edited by Andre van Wijnen

Keywords:
Gene duplication
RPGR
OTC
Primates
Macaca

ABSTRACT

Gene duplication plays a significant role in evolution. Paralogous gene copies may be lost due to the successive accumulation of deleterious mutations or remain active in the genome. In this work, a partial duplication of an X-linked region in the *Macaca* genus is identified and explored. Genomic comparisons reveal that the duplication encompasses the genes encoding ornithine transcarbamylase (OTC) and retinitis pigmentosa GTPase regulator (RPGR), spanning over 0.1 Mb on the chromosome 9 of *Macaca*. According to our analyses, the duplicated region of chromosome 9 involves partial coding sequences of both *OTC* and *RPGR* genes. Analyses of the selective pressures did not reveal significant differences in the ratio between nonsynonymous and synonymous mutations ($\omega<1$), suggesting that no selective pressures were acting in the evolutionary process. Reports for a biological role regarding some partial duplications exist in the literature, therefore, although being rare events, partial duplications of functionally important genes are worthy of study so that their impact can be explored.

1. Introduction

Gene duplication plays a significant role in phenotypic variation, as it can lead to the emergence of new traits increasing morphological and physiological diversity (Zhang, 2003; Katju and Bergthorsson, 2013). The most common fate of duplicated genes is gene loss due to random accumulation of mutations that impair function (Lynch and Conery, 2000). Some duplicated genes can remain active, maintaining the global function while accumulating complementary loss-of-function mutations, in such a way that the two encoded products complement each

Abbreviations: ANKRD30A, ankyrin repeat domain 30A; ATP8A2P1, ATPase phospholipid transporting 8A2 pseudogene 1; BAG1, BAG cochaperone 1; Blastn, basic local alignment search tool nucleotide; CaMKII, calcium/calmodulin dependent protein kinase II gamma; camk2g1, calcium/calmodulin dependent protein kinase II gamma 1; camk2g2, calcium/calmodulin dependent protein kinase II gamma 2; ch, chromosome; CYBB, cytochrome *B*-245 beta chain; DEPP1, decidual protein induced by progesterone 1; DMD, dystrophin; dN, non-synonymous substitution rate; DUXA, double homeobox A; dS, synonymous substitution rate; DYNTL3, dynein light chain Tctex-type 3; EIF3L, eukaryotic translation initiation factor 3 subunit L; ERVK-8, endogenous retrovirus group K member 8; EXOV, exonuclease V; EXOVL, exonuclease V-like; FTL, ferritin light chain; FXYD6, FXYD domain containing ion transport regulator 6; FXYD6P2, fxyd domain containing ion transport regulator 6 pseudogene 2; GTR, general time reversible; HGMD, human gene mutation database; H2AL3, H2A.L variant histone 3; HvARM1, hordeum vulgare armadillo 1; HYPM, huntingtin interacting protein M; Kb, kilo base pairs; lnL, log likelihood; LRT, likelihood ratio test; Mb, mega base pairs; MTND1P18, mitochondrially encoded nadh:ubiquinone oxidoreductase core subunit 1 pseudogene 18; NAHR, nonallelic homologous recombination; NBPF, neuroblastoma breakpoint family; NCBI, national center for biotechnology information; NCR3LG1, natural killer cell cytotoxicity receptor 3 ligand 1; NHEJ, nonhomologous end joining; OMIM, online mendelian inheritance in man; OTC, ornithine transcarbamylase; PMM1, phosphomannomutase 1; PMM2, phosphomannomutase 2; RASSF4, ras association domain family member 4; RN7SL314P, RNA 7SL, cytoplasmic 314, pseudogene; RPGR, retinitis pigmentosa gtpase regulator; RPS2, ribosomal protein s2; RPS19, ribosomal protein s19; SMS, smart model selection; SRA, sequence read archive; SRPX, sushi repeat containing protein x-linked; SYTL5, synaptotagmin like 5; TACC1P1, transforming acidic coiled-coil containing protein 1 pseudogene 1; TLK2, tousled like kinase 2; TMEM161B, transmembrane protein 161b; TMEM161BP1, transmembrane protein 161b pseudogene 1; TSPAN7, tetraspanin 7; UBTF, upstream binding transcription factor; VCX, variable charge x-linked; VN1R1, vomeronasal 1 receptor 1; VN1R53P, vomeronasal 1 receptor 53 pseudogene; w, strength of selection; XK, X-linked kx blood group; ZNF22, zinc finger protein 22; ZNF37A, Zinc finger protein 37a.

* Corresponding authors at: i3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal.
E-mail addresses: mmarques@i3s.up.pt (M. Lopes-Marques), lazevedo@ipatimup.pt (L. Azevedo).

https://doi.org/10.1016/j.gene.2022.146997
Received 19 July 2022; Received in revised form 12 October 2022; Accepted 18 October 2022
Available online 22 October 2022

other, assuring the original function – subfunctionalization (Lynch and Force, 2000; Gout and Lynch, 2015; Veitia, 2017). An example of subfuncionalization is found in the vertebrate hemoglobin cluster of genes, as duplicates have diverged in functional properties but altogether transport the oxygen in the organism (Storz, 2016). Alternatively, one copy can retain the original function while the other copy acquires a novel function through the accumulation of mutations – neofunctionalization (Birchler and Yang, 2022). Retinoic acid receptors evolved by neofunctionalization of duplicated copies that diverged in both ligand-binding capacity and expression patterns (Escriva et al., 2006). Both subfuncionalization and neofunctionalization are associated to the emergence of biological novelties during the evolution of species.

In mammals, about 10% of all lineage-specific duplicates show signatures of positive selection (Han et al., 2009). Popesco M. et al. (Popesco et al., 2006) by studying the *NBPF* genes demonstrated that a surprising number of human-specific duplications underwent positive selection. Although the relationship was not totally uncovered, a high proportion of positively selected duplicated genes was showed to be related to neural or cognitive functions and response to social stress among primates (Han et al., 2009; Sapolsky, 2005).

An interesting question concerning some duplications is whether one of the paralogues can act as a shield against disease. On that regard, Hsiao and Vitkup (Vitkup, 2008) reported that human genes with highly identical paralogues (90% sequence identity) are more likely to harbor deleterious mutations than those with less identical copies, likely due to functional compensation by the paralogue counterpart. This is the case of two paralogues of the γ *CaMKII* gene (*camk2g1* and *camk2g2*) known to be involved a compensatory mechanism, where *camk2g2* was able to restore the phenotype and reproductive viability in a homozygous mutant of *camk2g1* (Rothschild et al., 2020).

Duplications may not embrace the full length of the gene. Although partial duplications may lack the parental function, they may evolve to fulfil novel functions (Simioni et al., 2021; Siggs et al., 2017; Rajaraman et al., 2018; Hu et al., 1988; Huang et al., 2022). For example, a partial copy of *HvARM1* in grass genomes was shown to be responsible for the increased resistance against powdery mildew infection (Rajaraman et al., 2018). Despite their putative importance, partial gene duplicates have been sparsely studied. Herein, we examine in detail a previously unidentified partial duplication of the gene encoding the ornithine transcarbamylase (OTC) protein (OMIM 300461, (Suriano et al., 2007; Lopes-Marques et al., 2021) in the chromosome 9 of macaques (OTC-like). (Suriano et al., 2007; Lopes-Marques et al., 2021). The mechanism of origin, the genetic structure of the duplicated region and the evolutionary forces acting on the partially duplicated *OTC* were investigated. While the biological importance, if any, of this partial duplication remains to be proven, this study is contributing to the advance of the knowledge on the mechanisms underlying partial duplicates in mammals.

2. Material and methods

2.1. Sequence retrieval and multiple sequence alignment

DNA coding sequences for the human X-linked *OTC* gene and mammalian orthologues were retrieved from Ensembl Genome Browser release number 103 (Supplementary Table 1). To ensure accurate comparisons between orthologous, sequences with more than 1% of unknown nucleotides and a size difference of at least 5% to the human reference sequence were removed. Since *Macaca nemestrina* sequence was incomplete in Ensembl Genome Browser release number 103 (ENSMNET00000014858), the first three exons were inferred by SRA projects (SRX6972979, SRX6972980, SRX6972982). A codon-based multiple sequence alignment was built using PRANK algorithm (Löytynoja and Goldman, 2010) and the subsequent alignments were manually checked. Final sequence alignment contained 63 sequences and 1062 positions.

2.2. Sequence read archive (SRA) confirmation

Publicly available SRA projects of 18 primates were investigated via blastn searches using *Macaca mulata* OTC coding sequence (ENSMMUT00000106596.1) as query (Supplementary Table 2). Top hits were downloaded and aligned to *Macaca mulata OTC* sequence. Sequence alignments were manually inspected to identify heterozygote individuals.

2.3. Comparative genomics

Using data collected from NCBI, *OTC* and neighboring genes were mapped onto the respective chromosomic region regarding *Homo sapiens* (genome assembly GRCh38.p14), *Macaca mulatta* (genome assembly Mmul_10), *Macaca fascicularis* (genome assembly MFA1912RKSv2) and *Macaca nemestrina* (genome assembly Mnem_1.0). The same strategy was pursued with *OTC-like* (LOC698302) in *Macaca mulatta* genome. *ANKRD30A* and *TMEM161B*, two genes that were in the vicinity of *OTC-like* in *Macaca mulatta*, were used as anchor genes to inspect the corresponding genomic region in Homo sapiens, *Macaca fascicularis* and *Macaca nemestrina*. The synteny maps were constructed by inspecting at least four neighboring genes up and downstream of the target gene in each species.

2.4. Phylogenetic and selection analyses

The best fitting model of evolution was inferred using a smart model selection (SMS), where GTR + R was inferred as the best model (Lefort et al., 2017). The phylogenetic trees were reconstructed using the maximum likelihood method implemented in PhyML 3.0 (Guindon et al., 2010). Branch support values were assessed with aBayes (Anisimova et al., 2011). Tree visualization and edition were performed in iTOL 5.0 (Letunic and Bork, 2021). Additionally, the multiple sequence alignment and the phylogenetic gene tree were used to analyze the possibility of the macaque clade (*Macaca fascicularis*, *Macaca mulatta* and *Macaca nemestrina*) experiencing different selective pressures. The selective pressures were estimated using the branch model from CODEML, implemented in PAML (Yang, 2007). The log Likelihood values (lnL) estimated by CODEML were used to perform a likelihood ratio test (LRT = 2x(lnL0-lnL1). To access the statistical significance, the P value was obtained by comparing the LRT value against Chi^2 distribution, where the degree of freedom was the difference of number of parameters between the null and alternative hypotheses. A significant P value ($p < 0.05$) means that the macaque clade is more likely to have a different selective force, w, from the background branches.

3. Results and discussion

3.1. Partial duplication of the OTC gene in the genome of Macaques

In mammals, *OTC* is a X-linked gene, for which no paralogous had been described. A blastn search against the *Macaca mulatta* genome using the human *OTC* coding sequence (ENST00000039007.5) as query, revealed the existence of a partial duplication of the *OTC* gene, annotated at the Ensembl database (release number 103) as ornithine carbamoyltransferase, mitochondrial-like (*OTC-like*) (LOC698302). The partial duplication here found in the *Macaca* lineage is localized in the chromosome 9 (9: 38,405,731–38,423,758) and comprises 298 nucleotides that match the first three exons of the X-linked *OTC* gene (ENSMMUT00000106596) with a pairwise identity of 98.75%. The nucleotide sequence corresponding to the first exon is 100% identical between the X-linked and the duplicated copy, the second exon differs in three nucleotides (97.8% identity), and the third exon has a single nucleotide difference between the two sequences (98.8% pairwise

identity) (Fig. 1A and 1B). The genomic region containing both intronic and exonic sequences from *OTC* and *OTC-like* of *Macaca mulatta* revealed a global pairwise identity of 91.3% (16585 sites identical out of 18162).

Remarkably, the *Macaca mulatta OTC-like* sequence harbored a nucleotide difference that results in an amino acid substitution that matches a deleterious variant (NM_000531.5:c.119G>A [p.Arg40His]) (Fig. 1.B), associated with OTC deficiency in humans (Lopes-Marques et al., 2021; Augustin et al., 2000; Mavinakere et al., 2001). Prior sequencing of *Macaca mulatta* genome allowed to identify the wild-type residue of the X-linked OTC position 40 as histidine, which corresponds to the human deleterious variant (Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007; Yan et al., 2011). This finding may have resulted from the presence of the unknown and highly similar duplicate of *OTC*, the *OTC-like* sequence.

To clarify the issue, we analyzed the SRA projects available at NCBI for the *Macaca* genus. The SRX5279654, SRX5014286 and SRX6972979 sequencing projects of males of the *Macaca mulatta*, *Macaca fascicularis* and *Macaca nemestrina* species, respectively, revealed three interspecific polymorphic positions: T/A at position 96, A/G at nucleotide 119 and T/C at position 144 (Fig. 2). The variants at positions 96 and 119 imply amino acids changes (histidine/glutamine and histidine/arginine), whereas the T/C at position 144 results in a synonymous substitution (Fig. 2). This apparent heterozygosity at the X-linked *OTC* in male samples of *Macaca mulatta*, *Macaca fascicularis* and *Macaca nemestrina*, can only be explained by the presence of a (partial) *OTC* duplicate in these species. SRA projects were additionally accessed for a total of 18 primates (Supplementary Table 2) but no further evidence for duplication was found in any other of these primates.

3.2. Analysis of the OTC-like genomic region

We next analyzed the region flanking *OTC* gene in *Homo sapiens*, *Macaca mulatta*, *Macaca fascicularis* and *Macaca nemestrina* (Fig. 3A). In *Macaca mulatta*, *OTC-like*, localized in chromosome 9, is flanked by *VN1R1* and *EIF3L*, both pseudogenes, and other genes such *TLK2-like* and *RPGR-like*. A pairwise alignment between the DNA coding sequence of *RPGR-like* (LOC106996533) and DNA coding sequence of the *RPGR* transcript variant X2 (XM_028842536.1:412–2859) revealed that *RPGR-like* includes a partial copy of exon 18 and an entire copy of exon 19 of *RPGR*, with an identity of 98.6% (290 identical nucleotides of a total of 294). *RPGR* is a neighboring gene of *OTC* in the X-chromosome. The identification of a partial duplication of *RPGR-like* gene together with the partial duplication of the *OTC-like* suggests that both sequences duplicated in a single event involving this genomic region (Fig. 3B). The duplicated region spans approximately 0.1 Mb, including the *RPGR* and *OTC* genes. The analysis of this region in *Macaca fascicularis* does not show the presence of an *OTC-like* copy. However, the presence of a *RPGR-like* copy, in combination with the detection of a heterozygous male of *Macaca fascicularis* in the SRX5014286 for the X-linked gene, *OTC*, suggests that the duplication of this region is also shared by *Macaca fascicularis*. So, the apparent absence of the *OTC-like* in this region in *Macaca fascicularis* is likely to be due to poor genome coverage or assembly. Similarly, in *Macaca nemestrina* poor coverage of this genomic region, may also explain the lack of this duplicate (Fig. 3A and Fig. 3B).

3.3. Possible mechanisms for the origin of the OTC-like in the Macaca genome

Gene duplication is the major source involved in the birth of novel genes, playing thus important role in genome evolution (Cazzanelli et al., 2018; Moriyama and Koshiba-Takeuchi, 2018; Lallemand et al., 2020; Samonte and Eichler, 2002). A duplicated segment of DNA can result from whole genome duplication, unequal crossing (which promotes tandem gene duplications), retroposition (when a messenger RNA is retrotranscribed onto complementary DNA and reinserted into the genome), or by segmental duplication (Bailey et al., 2001). The analysis of the synteny maps of X-chromosome and chromosome 9 of *Macaca mulatta* revealed that the *OTC-like* gene retains part of the intronic sequence, and the duplication is not confined to *OTC*, but instead also includes *RPGR* (Fig. 3A and 3B), therefore it is likely that the duplication of *RPGR* and *OTC* represent a lineage-specific segmental duplication (Vollger et al., 2022).

Segmental duplications are in most of the cases, highly identical long sequences (1–200 Kb), that may contain genic structures as introns and exons (Abdullaev et al., 2021; Nurk et al., 2022) originated from non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ) or replication slippage, although the discernment between these processes would require the fine-scale analysis of the

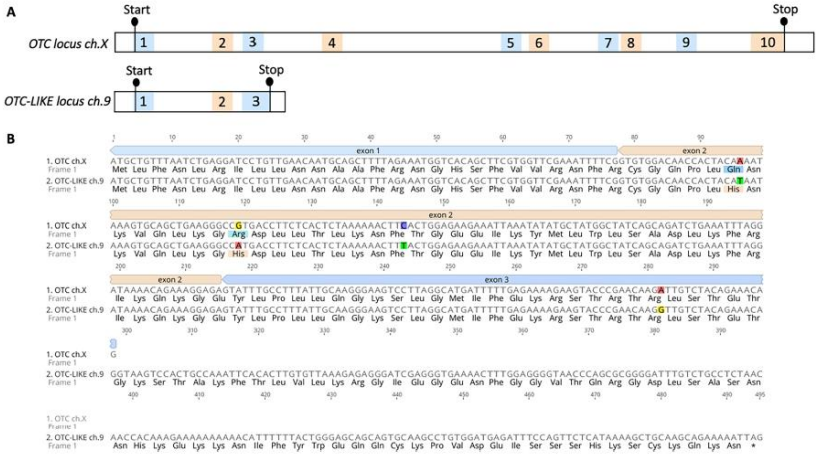


Fig. 1. OTC and OTC-like comparison in *Macaca mulatta*. A. Locus structure of *OTC* and *OTC-like*: *OTC* gene contains 10 exons and *OTC-like* is a partial duplication of the first three exons of *OTC*. B. Pairwise alignment of *OTC* and *OTC-like*, highlighting the divergence between the two sequences.

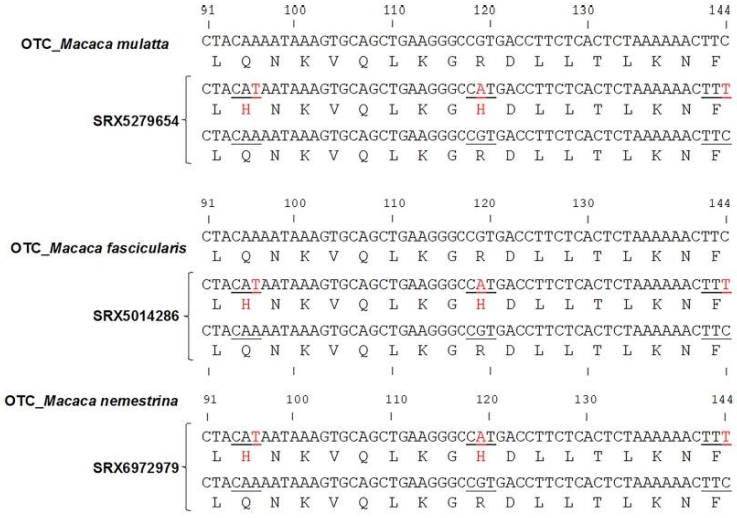


Fig. 2. SRA Project SRX5279654, SRX5014286 and SRX6972979 of three males of *Macaca mulatta*, *Macaca fascicularis* and a *Macaca nemestrina* species, respectively. Red color stands for nucleotide and amino acids specific to *OTC-like* and codons harboring interspecific variable positions are underlined. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

duplicated sequence (Zhang et al., 2004). Segmental duplications have been found in humans (Mao et al., 2021; She et al., 2008) and other mammals (e.g. (Bailey and Eichler, 2006; Marques-Bonet et al., 2009). Whilst duplicated DNA blocks are significant contributors to genomic diversity (Li et al., 2009), they also can have important roles in disease by creating genetic instability (Roos et al., 2019). To date, no segmental duplicates involving either *OTC* or *RPGR* genes have been reported, therefore, the putative biological significance of our finding remains to be determined.

3.4. Evolutionary history of the duplication

The genus *Macaca* is one of the anthropoid primates that successfully dispersed into different continents (Song et al., 2021). The oldest known fossils of *Macaca* in Europe and North Africa date to around 5.3–5.9 million years ago (Cardoso-Moreira et al., 2016). The emergence of *Macaca* remains in Asia is slightly more recent going back to around 3.6–4.9 million years ago (Cardoso-Moreira et al., 2016). There are some uncertainties in the species classification of the genus, but the existence of at least four species groups is accepted: the Silenus-Sylvanus group, including *Macaca nemestrina*; the Fascicularis group, including *Macaca fascicularis* and *Macaca mulatta*; the Sinica group (e.g., *Macaca sinica* and *Macaca radiata*) and the Arctoides group (e.g., *Macaca arctoides*) (Song et al., 2021). The *Macaca* lineage encompassing the Fascicularis group diverged at about 10.3–11.1 million years ago (Kondrashov and Kondrashov, 2006). Accordingly, *OTC* duplication may have occurred still in the common ancestor of *Macaca mulatta* and *Macaca fascicularis*, since it is shared by these two species. However, since gene flow has been reported between *Macaca mulatta* and *Macaca fascicularis* (Shiu et al., 2006), another hypothesis that could explain the presence of the duplicated copy in both species, is the emergence of the duplicate in one of the *Macaca* species and its presence in the other result from hybridization.

To investigate the possibility of the partial duplication conferring some buffering capacity to the X-linked gene *OTC*, and thus promoting a relaxation of the selective forces acting on it, a selection analysis was followed. Since the partial duplicate of *OTC* seems to be specific to the macaque clade, it was inspected if this clade was experiencing a significant difference in the evolutionary forces acting on it (w1), when compared to other mammalian species. Two evolutionary models were tested: a model where the entire phylogeny has the same selective force (model 0) and another one, where the macaque clade experiences a different selective force (Fig. 4A). The comparison of both models predicted the single ratio model, model 0, as the most accurate, the value of the Likelihood Ratio Test (LRT) being 0.8197 for a p value of 0.3653 (Fig. 4B). Hence, it is assumed a non-significant divergent selective force along the macaque lineage and a single selective force for the phylogeny ($w = 0.08976$).

Some previous works suggested the beneficial effect of positive selection in maintaining and fixing duplicated genes (Vallender et al., 2008; Stenson et al., 2020; Peng et al., 2007). In primates there is evidence that correlates duplication occurrences and adaptive evolution with the brain expansion in primates (She et al., 2008). However, taking in consideration the selection analysis, the duplicate retention in the case of Macaques does not seem to have been driven positive selection in X-linked *OTC* gene. On the contrary, that lineage appears to have been evolutionarily constrained, that is under negative selection forces ($w = 0.08976$), consistent with the functional impairment that a missense mutation may cause. According to HGMD database (Stenson et al., 2020), more than 300 missense deleterious variants have been found in the human *OTC* which demonstrates its low tolerability to non-synonymous mutations. The same rational can be applied to non-human primates as well. Therefore, the lack of a differential selective force in macaque lineage may indicates that the partial duplication is not playing a role in compensating OTC.

Paralogs can provide additional protein copies with the same function or create new genes, making the study of gene duplication of foremost importance. Prior works focused on the locus of *OTC* and its neighbor genes (*DMD, XK, CYBB, RPGR*) could not detect duplicates for *OTC* and *RPGR* in vertebrates (Peng et al., 2007). Our study shows for

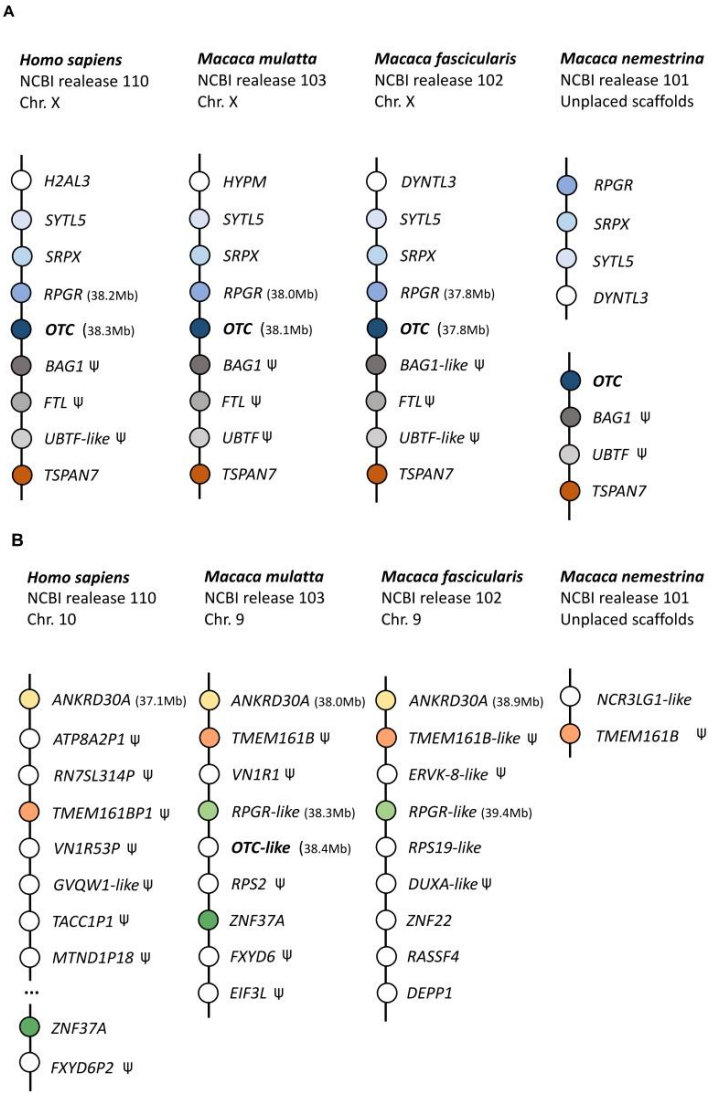
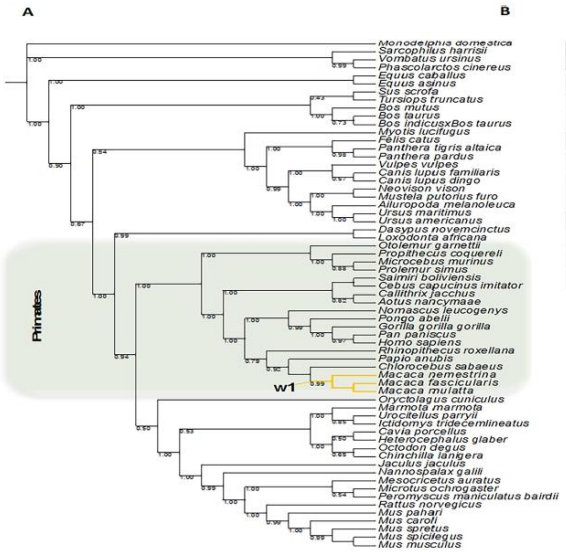


Fig. 3. Synteny maps for *Homo sapiens*, *Macaca mulatta* and *Macaca fascicularis*. A. Synteny map for the X-linked *OTC gene*. B. Synteny map for the *OTC-like* locus. Chromosome (Chr.) and location in Mb is given to *OTC*, *RPGR*, *OTC-like*, *RPGR-like*, *ANKRD30A* and *TMEM161B*. Orthologues genes can be inferred by color assignment.

the first time the presence of a partial duplication of *OTC* and a partial duplication of *RPGR* in *Macaca* genus. In general, partial duplications have been rather rarely described, probably because they are simply overlooked assuming that such small elements are useless genomic regions. Despite the few studies focusing on the characterization of partial duplications, they may have an importance biological role, as was recently reported (Huang et al., 2022), in respect to a small duplicate of *EXOVL* (*EXOV*) rapidly evolved critical developmental functions in *Arabidopsis thaliana*. Signals of positive selection were observed in *EXOV*, allowing the evolution of the regulatory domain and thus the emergence of new functional roles, even without the catalytic region of EXOVL.



One ratio model vs. Two ratio model	
Model 0	
w0	0.08976
lnL	-5315.528409
Model Alternative	
w0	0.08999
w1	0.00010
lnL	-5315.118575
Model Comparison	
LRT	0.8197
df	1
p value	0.3653

Fig. 4. Selection analysis. A. OTC phylogenetic tree, where w1 is the selective force in the macaque branch. Branch support values are provided for each node. The primates are highlighted in green and the *Macaca's* branch is colored in yellow. B. Lnl for model 0 and model alternative, comparison and significance.

In this study, a segmental duplication of approximately 0.1 Mb, including partial copies *OTC* and *RPGR*, was found to have occurred in some species of the genus *Macaca*. Since *Macaca mulatta* is commonly used in biomedical research future studies should dissect better this and other duplications, as well their putative effects on the genome. This work aims to bring attention to the need for new and finer perspectives in understanding the worth of segmental duplication as an evolutionary process.

Ethical statement

There is no ethical statement to declare.

CRediT authorship contribution statement

Catarina Serrano: Investigation, Methodology, Writing – original draft. **Mónica Lopes-Marques:** Investigation, Methodology, Writing – review & editing. **António Amorim:** Writing – review & editing. **Maria João Prata:** Investigation, Writing – review & editing. **Luísa Azevedo:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work was supported by Fundo Europeu de Desenvolvimento Regional (FEDER) through the COMPETE 2020 - Operational Programme for Competitiveness and Internationalization, Portugal 2020 and by Foundation for Science and Technology (FCT) [POCI-01-0145-FEDER-007274, POCI-01-0145-FEDER-29723]. CS holds a Foundation for Science and Technology PhD grant (SFRH/BD/137925/2018).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2022.146997.

References

Zhang, J., 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18, 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8.

Katju, V., Bergthorsson, U., 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. Front. Genet. 4, 273. https://doi.org/10.3389/FGENE.2013.00273/BIBTEX.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155. https://doi.org/10.1126/SCIENCE.290.5494.1151.

Lynch, M., Force, A., 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154, 459–473. https://doi.org/10.1093/GENETICS/154.1.459.

Gout, J.F., Lynch, M., 2015. Maintenance and loss of duplicated genes by dosage subfunctionalization. Mol. Biol. Evol. 32, 2141–2148. https://doi.org/10.1093/MOLBEV/MSV095.

Veitia, R.A., 2017. Gene duplicates: agents of robustness or fragility? Trends Genet. 33, 377–379. https://doi.org/10.1016/J.TIG.2017.03.006.

Storz, J.F., 2016. Gene duplication and evolutionary innovations in hemoglobin-oxygen transport. Physiology 31 (3), 223–232. https://doi.org/10.1152/PHYSIOL.00060.2015.

Birchler, J.A., Yang, H., 2022. The multiple fates of gene duplications: deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. Plant Cell 34, 2466–2474. https://doi.org/10.1093/PLCELL/KOAC076.

Escriva, H., Bertrand, S., Germain, P., Robinson-Rechavi, M., Umbhauer, M., Cartry, J., Duffraisse, M., Holland, L., Gronemeyer, H., Laudet, V., Stern, D., 2006. Neofunctionalization in vertebrates: the example of Retinoic Acid Receptors. PLoS Genet. 2 (7), e102. https://doi.org/10.1371/JOURNAL.PGEN.0020102.

Han, M.V., Demuth, J.P., McGrath, C.L., Casola, C., Hahn, M.W., 2009. Adaptive evolution of young gene duplicates in mammals. Genome Res. 19 (5), 859–867. https://doi.org/10.1101/GR.085951.108.

Vitkup, T.-L., 2008. Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet 4, 1000014. https://doi.org/10.1371/journal.pgen.1000014.

Rothschild, S.C., Ingram, S.R., Lu, F.-I., Thisse, B., Thisse, C., Parkerson, J.A., Tombes, R. M., 2020. Genetic compensation of γ CaMKII, an evolutionarily conserved gene. Gene 742. https://doi.org/10.1016/J.GENE.2020.144567.
Simioni, P., Cagnin, S., Sartorello, F., Sales, G., Pagani, L., Bulato, C., Gavasso, S., Nuzzo, F., Chemello, F., Radu, C.M., Tormene, D., Spiezia, L., Hackeng, T.M., Campello, E., Castoldi, E., 2021. Partial F8 gene duplication (factor VIII Padua) associated with high factor VIII levels and familial thrombophilia. Blood 137 (17), 2383–2393. https://doi.org/10.1182/BLOOD.2020008168.
Siggs, O.M., Javadiyan, S., Sharma, S., Souzeau, E., Lower, K.M., Taranath, D.A., Black, J. o., Pater, J., Willoughby, J.G., Burdon, K.P., Craig, J.E., 2017. Partial duplication of the CRYBB1-CRYBA4 locus is associated with autosomal dominant congenital cataract. Eur J Hum Genet 25 (6), 711–718. https://doi.org/10.1038/EJHG.2017.33.
Popesco, M.C., MacLaren, E.J., Hopkins, J., et al., 2006. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. Science 313, 1304–1307. https://doi.org/10.1126/SCIENCE.1127980.
Rajaraman, J., Douchkov, D., Lück, S., Hensel, G., Nowara, D., Pogoda, M., Rutten, T., Meitzel, T., Brassac, J., Höfle, C., Hückelhoven, R., Klinkenberg, J., Trujillo, M., Bauer, E., Schmutzer, T., Himmelbach, A., Mascher, M., Lazzari, B., Stein, N., Kumlehn, J., Schweizer, P., 2018. Evolutionarily conserved partial gene duplication in the Triticeae tribe of grasses confers pathogen resistance. Genome Biol. 19 (1) https://doi.org/10.1186/S13059-018-1472-7/FIGURES/3.
Hu, X.Y., Burghes, A.H., Ray, P.N., Thompson, M.W., Murphy, E.G., Worton, R.G., 1988. Partial gene duplication in Duchenne and Becker muscular dystrophies. J. Med. Genet. 25 (6), 369–376. https://doi.org/10.1136/JMG.25.6.369.
Huang, Y., Chen, J., Dong, C., Sosa, D., Xia, S., Ouyang, Y., Fan, C., Li, D., Mortola, E., Long, M., Bergelson, J., 2022. Species-specific partial gene duplication in Arabidopsis thaliana evolved novel phenotypic effects on morphological traits under strong positive selection. Plant Cell 34 (2), 802–817. https://doi.org/10.1093/PLCELL/KOAB291.
Suriano, G., Azevedo, L., Novais, M., Boscolo, B., Seruca, R., Amorim, A., Ghibaudi, E.M., 2007. In vitro demonstration of intra-locus compensation using the ornithine transcarbamylase protein as model. Hum. Mol. Genet. 16 (18), 2209–2214. https://doi.org/10.1093/HMG/DDM172.
Lopes-Marques, M., Pacheco, A.R., Peixoto, M.J., Cardoso, A.R., Serrano, C., Amorim, A., Prata, M.J., Cooper, D.N., Azevedo, L., 2021. Common polymorphic OTC variants can act as genetic modifiers of enzymatic activity. Hum Mutat 42 (8), 978–989. https://doi.org/10.1002/HUMU.24221.
Löytynoja, A., Goldman, N., 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinf. 11, 579. https://doi.org/10.1186/1471-2105-11-579.
Lefort, V., Longueville, J.E., Gascuel, O., 2017. SMS: Smart Model Selection in PhyML. Mol Biol Evol 34, 2422–2424. https://doi.org/10.1093/MOLBEV/MSX149.
Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate Maximum-Likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59 (3), 307–321. https://doi.org/10.1093/SYSBIO/SYQ010.
Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., Gascuel, O., 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst. Biol. 60 (5), 685–699. https://doi.org/10.1093/SYSBIO/SYR041.
Letunic, I., Bork, P., 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research 49:W293–W296. https://doi.org/10.1093/NAR/GKAB301.
Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24, 1586–1591. https://doi.org/10.1093/molbev/msm088.
Augustin, L., Mavinakere, M., Morizono, H., Tuchman, M., 2000. Expression of wild-type and mutant human ornithine transcarbamylase genes in Chinese hamster ovary cells and lack of dominant negative effect of R141Q and R40H mutants. Pediatr Res 48, 842–846. https://doi.org/10.1203/00006450-200012000-00023.
Mavinakere, M., Morizono, H., Shi, D., Allewell, N.M., Tuchman, M., 2001. The clinically variable R40H mutant ornithine carbamoyltransferase shows cytosolic degradation of the precursor protein in CHO cells. J. Inherit. Metab. Dis. 24 (6), 614–622. https://doi.org/10.1023/A:1012726207870.
Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316, 222–234. https://doi.org/10.1126/science.1139247.
Yan, G., Zhang, G., Fang, X., Zhang, Y., Li, C., Ling, F., Cooper, D.N., Li, Q., Li, Y., van Gool, A.J., Du, H., Chen, J., Chen, R., Zhang, P., Huang, Z., Thompson, J.R., Meng, Y., Bai, Y., Wang, J., Zhuo, M., Wang, T., Huang, Y., Wei, L., Li, J., Wang, Z., Hu, H., Yang, P., Le, L., Stenson, P.D., Li, B.o., Liu, X., Ball, E.V., An, N.a., Huang, Q., Zhang, Y., Fan, W., Zhang, X., Li, Y., Wang, W., Katze, M.G., Su, B., Nielsen, R., Yang, H., Wang, J., Wang, X., Wang, J., 2011. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat. Biotechnol. 29 (11), 1019–1023. https://doi.org/10.1038/nbt.1992.
Cazzanelli, G., Pereira, F., Alves, S., Francisco, R., Azevedo, L., Carvalho, P.D., Almeida, A., Côrte-Real, M., Oliveira, M.J., Lucas, C., Sousa, M.J., Preto, A., 2018. The yeast Saccharomyces cerevisiae as a model for understanding RAS Proteins and their role in human tumorigenesis. Cells 7, 14. https://doi.org/10.3390/cells702001433.
Moriyama, Y., Koshiba-Takeuchi, K., 2018. Significance of whole-genome duplications on the emergence of evolutionary novelties. Briefings in Functional Genomics 17, 329–338. https://doi.org/10.1093/bfgp/ely007.
Lallemand, T., Leduc, M., Landès, C., Rizzon, C., Lerat, E., 2020. An overview of duplicated gene detection methods: why the duplication mechanism has to be accounted for in their choice. Genes (Basel) 11 (9), 1046. https://doi.org/10.3390/genes11091046.
Samonte, R., Eichler, E., 2002. Segmental duplications and the evolution of the primate genome. Nat Rev Genet 3, 65–72. https://doi.org/10.1038/nrg705.
Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., Eichler, E.E., 2001. Segmental duplications: organization and impact within the current Human Genome Project assembly. Genome Res. 11 (6), 1005–1017. https://doi.org/10.1101/gr.gr-1871r.
Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., Hoekzema, K., Porubsky, D., Li, R., Nurk, S., Koren, S., Miga, K.H., Phillippy, A.M., Timp, W., Ventura, M., Eichler, E.E., 2022. Segmental duplications and their variation in a complete human genome. Science 376 (6588). https://doi.org/10.1126/science.abj6965.
Abdullaev, E.T., Umarova, I.R., Arndt, P.F., 2021. Modelling segmental duplications in the human genome. BMC Genomics 22, 496. https://doi.org/10.1186/s12864-021-07789-7.
Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S.J., Diekhans, M., Logsdon, G.A., Alonge, M., Antonarakis, S.E., Borchers, M., Bouffard, G.G., Brooks, S. Y., Caldas, G.V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L.G., Dishuck, P.C., Durbin, R., Dvorkina, T., Fiddes, I.T., Formenti, G., Fulton, R.S., Fungtammasan, A., Garrison, E., Grady, P.G.S., Graves-Lindsay, T.A., Hall, I.M., Hansen, N.F., Hartley, G.A., Haukness, M., Howe, K., Hunkapiller, M.W., Jain, C., Jain, M., Jarvis, E.D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V.V., Marschall, T., McCartney, A.M., McDaniel, J., Miller, D.E., Mullikin, J.C., Myers, E.W., Olson, N.D., Paten, B., Peluso, P., Pevzner, P.A., Porubsky, D., Potapova, T., Rogaev, E.I., Rosenfeld, J.A., Salzberg, S. L., Schneider, V.A., Sedlazeck, F.J., Shafin, K., Shew, C.J., Shumate, A., Sims, Y., Smit, A.F.A., Soto, D.C., Sović, I., Storer, J.M., Streets, A., Sullivan, B.A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B.P., Wenger, A., Wood, J.M.D., Xiao, C., Yan, S.M., Young, A.C., Zarate, S., Surti, U., McCoy, R.C., Dennis, M.Y., Alexandrov, I.A., Gerton, J.L., O'Neill, R.J., Timp, W., Zook, J.M., Schatz, M.C., Eichler, E.E., Miga, K.H., Phillippy, A.M., 2022. The complete sequence of a human genome. Science 376 (6588), 44–53. https://doi.org/10.1126/science.abj6987.
Zhang, L., Lu, H.H.S., Chung, W.-y., et al., 2004. Patterns of segmental duplication in the human genome. Mol. Biol. Evol. 22, 135–141. https://doi.org/10.1093/molbev/msh262.
Mao, Y., Catacchio, C.R., Hillier, L.W., Porubsky, D., Li, R., Sulovari, A., Fernandes, J.D., Montinaro, F., Gordon, D.S., Storer, J.M., Haukness, M., Fiddes, I.T., Murali, S.C., Dishuck, P.C., Hsieh, PingHsun, Harvey, W.T., Audano, P.A., Mercuri, L., Piccolo, I., Antonacci, F., Munson, K.M., Lewis, A.P., Baker, C., Underwood, J.G., Hoekzema, K., Huang, T.-H., Sorensen, M., Walker, J.A., Hoffman, J., Thibaud-Nissen, F., Salama, S. R., Pang, A.W.C., Lee, J., Hastie, A.R., Paten, B., Batzer, M.A., Diekhans, M., Ventura, M., Eichler, E.E., 2021. A high-quality bonobo genome refines the analysis of hominid evolution. Nature 594 (7861), 77–81. https://doi.org/10.1038/s41586-021-03519-x.
Sapolsky, R.M., 2005. The influence of social hierarchy on primate health. Science 308, 648–652. https://doi.org/10.1126/SCIENCE.1106477.
She, X., Cheng, Z.e., Zöllner, S., Church, D.M., Eichler, E.E., 2008. Mouse segmental duplication and copy number variation. Nat Genet 40 (7), 909–914. https://doi.org/10.1038/ng.172.
Bailey, J., Eichler, E., 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7, 552–564. https://doi.org/10.1038/nrg1895.
Marques-Bonet, T., Girirajan, S., Eichler, E.E., 2009. The origins and impact of primate segmental duplications. Trends Genet 25, 443–454. https://doi.org/10.1016/j.tig.2009.08.002.
Li, J., Han, K., Xing, J., Kim, H.-S., Rogers, J., Ryder, O.A., Disotell, T., Yue, B., Batzer, M. A., 2009. Phylogeny of the macaques (Cercopithecidae: Macaca) based on Alu elements. Gene 448 (2), 242–249. https://doi.org/10.1016/J.GENE.2009.05.013.
Roos, C., Kothe, M., Alba, D.M., Delson, E., Zinner, D., 2019. The radiation of macaques out of Africa: Evidence from mitogenome divergence times and the fossil record. J. Hum. Evol. 133, 114–132. https://doi.org/10.1016/J.JHEVOL.2019.05.017.
Song, Y., Jiang, C., Li, K.H., Li, J., Qiu, H., Price, M., Fan, Z.-X., Li, J., 2021. Genome-wide analysis reveals signatures of complex introgressive gene flow in macaques (genus Macaca). Zool. Res. 42, 433. https://doi.org/10.24272/J.ISSN.2095-8137.2021.038.
Cardoso-Moreira, M., Arguello, J.R., Gottipati, S., Harshman, L.G., Grenier, J.K., Clark, A. G., 2016. Evidence for the fixation of gene duplications by positive selection in Drosophila. Genome. Res. 26 (6), 787–798. https://doi.org/10.1101/GR.199323.115.
Kondrashov, F.A., Kondrashov, A.S., 2006. Role of selection in fixation of gene duplications. J. Theor. Biol. 239, 141–151. https://doi.org/10.1016/J.JTBI.2005.08.033.
Shiu, S.-H., Byrnes, J.K., Pan, R., Zhang, P., Li, W.-H., 2006. Role of positive selection in the retention of duplicate genes in mammalian genomes. Proc. Natl. Acad. Sci. 103, 2232–2236. https://doi.org/10.1073/PNAS.0510388103.
Vallender, E.J., Mekel-Bobrov, N., Lahn, B.T., 2008. Genetic basis of human brain evolution. Trends. Neurosci. 31 (12), 637–644. https://doi.org/10.1016/J.TINS.2008.08.010.
Stenson, P.D., Mort, M., Ball, E.V., Chapman, M., Evans, K., Azevedo, L., Hayden, M., Heywood, S., Millar, D.S., Phillips, A.D., Cooper, D.N., 2020. The Human Gene Mutation Database (HGMD ®): optimizing its use in a clinical diagnostic or research setting. Hum. Genet. 139 (10), 1197–1207. https://doi.org/10.1007/S00439-020-02199-3.
Peng, J., Redman, C.M., Wu, X.u., Song, X., Walker, R.H., Westhoff, C.M., Lee, S., 2007. Insights into extensive deletions around the XK locus associated with McLeod

phenotype and characterization of two novel cases. Gene 392 (1-2), 142–150. https://doi.org/10.1016/J.GENE.2006.11.023.

Further reading

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., Ravikesavan, R., 2013. Gene duplication as a major force in evolution. J. Genet. 92, 155–161. https://doi.org/10.1007/S12041-013-0212-8.

Chapter 5: Main Conclusions

Mutational changes in X-linked disease genes carry profound implications both in the pathological and evolutionary contexts. From a pathological perspective, mutations in genes located on the X chromosome can lead to a range of genetic recessive disorders that are more frequent in males due to their hemizygous state. Evolutionarily, they can illuminate the dynamic nature of the genome, showcasing how genetic variations influence population diversity and fitness. While deleterious mutations have immediate negative effects on individuals from a given species, over evolutionary timescales they can contribute to the overall genetic diversity that in certain environmental circumstances drives adaptation and speciation. Thus, examining X-linked genes through both pathological and evolutionary lenses offers critical insights into the multifaceted roles these genes play in health or disease, and in the broader mechanisms of evolution.

In this work we first addressed missense mutations in 10 X-linked genes that previously were identified as likely compensated mutations in some mammalian species. This investigation revealed that current pathogenicity prediction tools often classifies those putative compensated mutations as deleterious due to their apparent non-conservation across phylogenetic lineages, raising strong doubts on the accuracy of the obtained predictions.

Through analyzing a range of genetic mutations, it was possible to identify a pathogenic amino acid substitution that was actually neutralized by two contiguous residues in the F9 protein found in a variety of non-human mammalian species. The two residues were located closely to a region crucial to the protein function, seemingly illuminating why a human deleterious mutation was rescued in non-humans by a compensatory partner. This research emphasizes the need for a deeper analysis of genetic mutations, moving beyond the simplistic pathogenic versus non-pathogenic classification, towards considering the dynamic and context-dependent nature of genetic variations.

Also, this thesis enriched the catalogue of known gene duplication events, based on the discovery of a partial gene duplication of *OTC* and *RGPR* within species of Macaca genus that are commonly used as model in biomedical research. Our findings underscore the need to further dissect the functional effects of segmental duplications in human model organisms, besides encouraging a more comprehensive explorations of the significance of such events in evolutionary biology.

We hope to have open novel hints for future research engaging to deepen our understanding on genetic variation and their evolutionary and pathological implications. Firstly, because expanding the analysis of compensated mutations across a broader range of genes and organisms represents a step to elucidate their prevalence and impact. The challenge is now to investigate the molecular mechanisms underlying compensation, particularly through experimental approaches such as CRISPR gene editing that ultimately can reveal how specific compensatory changes mitigate the effects of deleterious mutations. Additionally, the discovery of partial gene duplications in the Macaca genus raises new questions regarding the functional evolution of genomic segments. Future studies should explore the biological roles of these partial duplications, assessing their contribution to phenotypic diversity and adaptability.

Much still needs to be done towards the development of more sophisticated models to accurately distinguish between truly human potential pathogenic and benign mutations, answer that cannot be discerned without integrating compensatory epistasis. This endeavor is crucial for advancing genetic diagnostics and personalized medicine, believing that only more accurate previsions of the effects of mutations can guide strategies for better therapeutic patient care.

References

[1] Z. Ma, L. Li, and Y. P. Zhang, “Defining Individual-Level Genetic Diversity and Similarity Profiles,” *Sci Rep*, vol. 10, no. 1, 2020, doi: 10.1038/s41598-020-62362-8.

[2] H. Ellegren and N. Galtier, “Determinants of genetic diversity,” *Nature Reviews Genetics*, vol. 17, no. 7. 2016. doi: 10.1038/nrg.2016.58.

[3] R. C. Lewontin and J. L. Hubby, “A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura.,” *Genetics*, vol. 54, no. 2, 1966, doi: 10.1093/genetics/54.2.595.

[4] L. Quintana-Murci and A. G. Clark, “Population genetic tools for dissecting innate immunity in humans,” *Nature Reviews Immunology*, vol. 13, no. 4. 2013. doi: 10.1038/nri3421.

[5] W. Bodmer, “Genetic characterization of human populations: From abo to a genetic map of the british people,” *Genetics*, vol. 199, no. 2, 2014, doi: 10.1534/genetics.114.173062.

[6] S. Hake and J. Ross-Ibarra, “Genetic, evolutionary and plant breeding insights from the domestication of maize,” *eLife*, vol. 4. 2015. doi: 10.7554/eLife.05861.

[7] M. P. Soares and G. Weiss, “The Iron age of host–microbe interactions,” *EMBO Rep*, vol. 16, no. 11, 2015, doi: 10.15252/embr.201540558.

[8] J. B. Lack *et al.*, “The drosophila genome nexus: A population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population,” *Genetics*, vol. 199, no. 4, 2015, doi: 10.1534/genetics.115.174664.

[9] 1000 Genomes Project Consortium, “1000 Genomes Project Consortium, A global reference for human genetic variation,” *Nature*, vol. 526, 2015.

[10] M. Lynch, “Evolution of the mutation rate,” *Trends in Genetics*, vol. 26, no. 8, 2010, doi: 10.1016/j.tig.2010.05.003.

[11] A. Hodgkinson and A. Eyre-Walker, “Variation in the mutation rate across mammalian genomes,” *Nature Reviews Genetics*, vol. 12, no. 11. 2011. doi: 10.1038/nrg3098.

[12] E. M. Leffler *et al.*, “Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species?,” *PLoS Biol*, vol. 10, no. 9, 2012, doi: 10.1371/journal.pbio.1001388.

[13] N. Alcala and S. Vuilleumier, “Turnover and accumulation of genetic diversity across large time-scale cycles of isolation and connection of populations,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1794, 2014, doi: 10.1098/rspb.2014.1369.

[14] G. Mendel, “Experiments in Plant Hybridization,” *Verhandlungen des naturforschenden Vereines in Brünn*, vol. 4, 1866.

[15] P. McColgan and S. J. Tabrizi, “Huntington’s disease: a clinical review,” *European Journal of Neurology*, vol. 25, no. 1. 2018. doi: 10.1111/ene.13413.

[16] K. V. Chemello, J. García-Nafría, A. Gallo, C. Martín, G. Lambert, and D. Blom, “Lipoprotein metabolism in familial hypercholesterolemia,” *J Lipid Res*, vol. 62, 2021, doi: 10.1016/J.JLR.2021.100062.

[17] Q. Chen, Y. Shen, and J. Zheng, “A review of cystic fibrosis: Basic and clinical aspects,” *Animal Models and Experimental Medicine*, vol. 4, no. 3. 2021. doi: 10.1002/ame2.12180.

[18] F. J. van Spronsen, N. Blau, C. Harding, A. Burlina, N. Longo, and A. M. Bosch, “Phenylketonuria,” *Nature Reviews Disease Primers*, vol. 7, no. 1. 2021. doi: 10.1038/s41572-021-00267-0.

[19] J. García-Chávez and A. Majluf-Cruz, “Hemofilia,” *Gac Med Mex*, vol. 149, no. 3, 2013, doi: 10.29303/jku.v11i3.666.

[20] H. Moser, “Duchenne muscular dystrophy: Pathogenetic aspects and genetic prevention,” *Human Genetics*, vol. 66, no. 1. 1984. doi: 10.1007/BF00275183.

[21] J. Craig, “Complex Diseases: Research and Applications,” Nature Education.

[22] S. Asslih, O. Damri, and G. Agam, “Neuroinflammation as a common denominator of complex diseases (Cancer, diabetes type 2, and neuropsychiatric disorders),” *International Journal of Molecular Sciences*, vol. 22, no. 11. 2021. doi: 10.3390/ijms22116138.

[23] R. Bergholdt *et al.*, “Integrative analysis for finding genes and networks involved in diabetes and other complex diseases,” *Genome Biol*, vol. 8, no. 11, 2007, doi: 10.1186/gb-2007-8-11-r253.

[24] C. M. Friedenreich, C. Ryder-Burbidge, and J. McNeil, “Physical activity, obesity and sedentary behavior in cancer etiology: epidemiologic evidence and biologic mechanisms,” *Molecular Oncology*, vol. 15, no. 3. 2021. doi: 10.1002/1878-0261.12772.

[25] W. Johannsen, “The Genotype Conception of Heredity,” *Am Nat*, vol. 45, no. 531, 1911, doi: 10.1086/279202.

[26] G. S. Fisch, “Whither the genotype-phenotype relationship? An historical and methodological appraisal,” *American Journal of Medical Genetics, Part C: Seminars in Medical Genetics*, vol. 175, no. 3. 2017. doi: 10.1002/ajmg.c.31571.

[27] J. J. Mitchell, Y. J. Trakadis, and C. R. Scriver, “Phenylalanine hydroxylase deficiency,” *Genetics in Medicine*, vol. 13, no. 8. 2011. doi: 10.1097/GIM.0b013e3182141b48.

[28] N. Blau, “Genetics of Phenylketonuria : Then and Now Then : A Brief History of Phenylketonuria,” *Human mutation* , vol. 37, no. 6, 2016.

[29] N. Blau, B. Thöny, R. G. H. Cotton, and K. Hyland, “Disorders of tetrahydrobiopterin and related biogenic amines,” *The metabolic and molecular bases of inherited disease*, vol. 2, 2001.

[30] G. S. Barsh, “What controls variation in human skin color?,” *PLoS Biology*, vol. 1, no. 1. 2003. doi: 10.1371/journal.pbio.0000027.

[31] P. M. Visscher, “Challenges in understanding common disease,” *Genome Medicine*, vol. 9, no. 1. 2017. doi: 10.1186/s13073-017-0506-1.

[32] D. Bachtrog *et al.*, “Sex Determination: Why So Many Ways of Doing It?,” *PLoS Biol*, vol. 12, no. 7, 2014, doi: 10.1371/journal.pbio.1001899.

[33] S. Ohno, “Sex Chromosomes and Sex-Linked Genes,” vol. 1, 1967, doi: 10.1007/978-3-642-88178-7.

[34] J. K. Abbott, A. K. Nordén, and B. Hansson, “Sex chromosome evolution: historical insights and future perspectives,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 284, no. 1854, May 2017, doi: 10.1098/RSPB.2016.2806.

[35] J. W. Foster and J. A. Marshall Graves, “An SRY-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene,” *Proc Natl Acad Sci U S A*, vol. 91, no. 5, pp. 1927–1931, Mar. 1994, doi: 10.1073/PNAS.91.5.1927.

[36] L. Potrzebowski, N. Vinckenbosch, A. C. Marques, F. Chalmel, B. Jégou, and H. Kaessmann, “Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes,” *PLoS Biol*, vol. 6, no. 4, pp. 709–716, Apr. 2008, doi: 10.1371/JOURNAL.PBIO.0060080.

[37] J. Subrini and J. Turner, “Y chromosome functions in mammalian spermatogenesis,” *Elife*, vol. 10, p. 0, Oct. 2021, doi: 10.7554/ELIFE.67345.

[38] M. T. Ross *et al.*, “The DNA sequence of the human X chromosome.,” *Nature*, vol. 434, no. 7031, pp. 325–37, Mar. 2005, doi: 10.1038/nature03440.

[39] C. Stern, “The Problem of Complete Y-Linkage in Man,” *Am J Hum Genet*, vol. 9, no. 3, p. 147, Sep. 1957.

[40] J. A. Marshall Graves, “The degenerate Y chromosome--can conversion save it?,” *Reprod Fertil Dev*, vol. 16, no. 5, pp. 527–534, 2004, doi: 10.10371/RD03096.

[41] J. A. M. Graves, “Evolution of vertebrate sex chromosomes and dosage compensation,” *Nat Rev Genet*, vol. 17, no. 1, pp. 33–46, Jan. 2016, doi: 10.1038/NRG.2015.2.

[42] L. D. Hurst and J. P. Randerson, “An eXceptional chromosome,” *Trends in Genetics*, vol. 15, no. 10, pp. 383–385, Oct. 1999, doi: 10.1016/S0168-9525(99)01809-0.

[43] G. M. Saifi and H. S. Chandra, “An apparent excess of sex- and reproduction-related genes on the human X chromosome.,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 266, no. 1415, p. 203, Jan. 1999, doi: 10.1098/RSPB.1999.0623.

[44] S. Bortoluzzi *et al.*, “A Comprehensive, High-Resolution Genomic Transcript Map of Human Skeletal Muscle,” *Genome Res*, vol. 8, no. 8, p. 817, 1998, doi: 10.1101/GR.8.8.817.

[45] T. A. Gurbich and D. Bachtrog, “Gene content evolution on the X chromosome.,” *Curr Opin Genet Dev*, vol. 18, no. 6, pp. 493–8, Dec. 2008, doi: 10.1016/j.gde.2008.09.006.

[46] D. Bachtrog, J. D. Jensen, and Z. Zhang, “Accelerated Adaptive Evolution on a Newly Formed X Chromosome,” *PLoS Biol*, vol. 7, no. 4, p. e1000082, 2009, doi: 10.1371/JOURNAL.PBIO.1000082.

[47] J. Domingo, P. Baeza-Centurion, and B. Lehner, “The Causes and Consequences of Genetic Interactions (Epistasis),” *Annual Review of Genomics and Human Genetics*, vol. 20. 2019. doi: 10.1146/annurev-genom-083118-014857.

[48] T. N. Starr and J. W. Thornton, “Epistasis in protein evolution,” *Protein Science*. 2016. doi: 10.1002/pro.2897.

[49] Ö. Carlborg and C. S. Haley, “Epistasis: Too often neglected in complex trait studies?,” *Nature Reviews Genetics*, vol. 5, no. 8. 2004. doi: 10.1038/nrg1407.

[50] H. J. Cordell, “Epistasis: What it means, what it doesn’t mean, and statistical methods to detect it in humans,” *Human Molecular Genetics*, vol. 11, no. 20. 2002. doi: 10.1093/hmg/11.20.2463.

[51] T. F. C. Mackay, “Epistasis and quantitative traits: Using model organisms to study gene-gene interactions,” *Nature Reviews Genetics*, vol. 15, no. 1. 2014. doi: 10.1038/nrg3627.

[52] Bateson W., “Mendel’s Principles of Heredity,” *Cambridge University Press*, vol. 86, no. 2169, 1909, doi: 10.1038/086407a0.

[53] M. S. Johnson, G. Reddy, and M. M. Desai, “Epistasis and evolution: recent advances and an outlook for prediction,” *BMC Biology*, vol. 21, no. 1. 2023. doi: 10.1186/s12915-023-01585-3.

[54] B. Lehner, “Molecular mechanisms of epistasis within and between genes,” *Trends in Genetics*, vol. 27, no. 8. 2011. doi: 10.1016/j.tig.2011.05.007.

[55] D. J. Kvitek and G. Sherlock, “Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape,” *PLoS Genet*, vol. 7, no. 4, 2011, doi: 10.1371/journal.pgen.1002056.

[56] A. S. Kondrashov, S. Sunyaev, and F. A. Kondrashov, “Dobzhansky-Muller incompatibilities in protein evolution,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 23, pp. 14878–14883, Nov. 2002, doi: 10.1073/pnas.232565499.

[57] A. D. Kern and F. A. Kondrashov, “Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs,” Nat Genet, vol. 36, no. 11, pp. 1207–1212, Nov. 2004, doi: 10.1038/ng1451.

[58] M. Kimura, “The role of compensatory neutral mutations in molecular evolution,” J Genet, vol. 64, no. 1, 1985, doi: 10.1007/BF02923549.

[59] “Evolutionary and Biomedical Insights from the Rhesus Macaque Genome Rhesus Macaque Genome Sequencing and Analysis Consortium,” Science (1979), vol. 316, no. 5822, pp. 222–34, 2007, doi: 10.1126/science.1139247.

[60] G. Yan et al., “Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques,” Nat Biotechnol, vol. 29, no. 11, pp. 1019–1023, Nov. 2011, doi: 10.1038/nbt.1992.

[61] C. Ferrer-Costa, M. Orozco, and X. de la Cruz, “Characterization of Compensated Mutations in Terms of Structural and Physico-Chemical Properties,” J Mol Biol, vol. 365, no. 1, pp. 249–256, Jan. 2007, doi: 10.1016/j.jmb.2006.09.053.

[62] J. Xu and J. Zhang, “Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis.,” Mol Biol Evol, vol. 31, no. 7, pp. 1787–92, Jul. 2014, doi: 10.1093/molbev/msu130.

[63] L. Gao and J. Zhang, “Why are some human disease-associated mutations fixed in mice?,” Trends in Genetics, vol. 19, no. 12, pp. 678–681, Dec. 2003, doi: 10.1016/J.TIG.2003.10.002.

[64] L. Azevedo, “Compensatory Evolution in Disease-Associated Genes,” in eLS, Chichester, UK: John Wiley & Sons, Ltd, 2015, pp. 1–4. doi: 10.1002/9780470015902.a0022402.

[65] D. M. Jordan et al., “Identification of cis-suppression of human disease mutations by comparative genomics.,” Nature, vol. 524, no. 7564, pp. 225–9, Aug. 2015, doi: 10.1038/nature14497.

[66] N. M. Cruz et al., “Modifier genes as therapeutics: The nuclear hormone receptor rev erb alpha (Nr1d1) rescues Nr2e3 associated retinal disease,” PLoS One, vol. 9, no. 1, 2014, doi: 10.1371/journal.pone.0087942.

[67] G. Suriano et al., “In vitro demonstration of intra-locus compensation using the ornithine transcarbamylase protein as model,” Hum Mol Genet, vol. 16, no. 18, pp. 2209–2214, Sep. 2007, doi: 10.1093/HMG/DDM172.

[68] H. Matsumura et al., “H558R, a common SCN5A polymorphism, modifies the clinical phenotype of Brugada syndrome by modulating DNA methylation of SCN5A promoters,” J Biomed Sci, vol. 24, no. 1, 2017, doi: 10.1186/s12929-017-0397-x.

[69] L. R. Vincent et al., “In vivo-selected compensatory mutations restore the fitness cost of mosaic penA Alleles that confer ceftriaxone resistance in Neisseria gonorrhoeae,” mBio, vol. 9, no. 2, 2018, doi: 10.1128/mBio.01905-17.

[70] K. Ravishankar, X. Jiang, E. M. Leddin, F. Morcos, and G. A. Cisneros, “Computational compensatory mutation discovery approach: Predicting a PARP1 variant rescue mutation,” Biophys J, vol. 121, no. 19, 2022, doi: 10.1016/j.bpj.2022.05.036.

[71] L. Azevedo et al., “Improving the in silico assessment of pathogenicity for compensated variants.,” Eur J Hum Genet, vol. 25, no. 1, pp. 2–7, 2016, doi: 10.1038/ejhg.2016.129.

[72] M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov, “Epistasis as the primary factor in molecular evolution,” Nature, vol. 490, no. 7421, pp. 535–538, Oct. 2012, doi: 10.1038/nature11510.

[73] T. B. Sackton and D. L. Hartl, “Genotypic Context and Epistasis in Individuals and Populations,” Jul. 14, 2016, Cell Press. doi: 10.1016/j.cell.2016.06.047.

[74] L. Azevedo, J. Carneiro, B. van Asch, A. Moleirinho, F. Pereira, and A. Amorim, “Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components.,” BMC Genomics, vol. 10, no. 1, p. 266, Jun. 2009, doi: 10.1186/1471-2164-10-266.

[75] M. Lynch and J. S. Conery, “The evolutionary fate and consequences of duplicate genes,” Science (1979), vol. 290, no. 5494, pp. 1151–1155, 2000, doi: 10.1126/SCIENCE.290.5494.1151.

[76] J. Zhao, A. I. Teufel, D. A. Liberles, and L. Liu, “A generalized birth and death process for modeling the fates of gene duplication,” BMC Evol Biol, vol. 15, no. 1, Dec. 2015, doi: 10.1186/s12862-015-0539-2.

[77] V. E. Prince and F. B. Pickett, “Splitting pairs: the diverging fates of duplicated genes,” Nat Rev Genet, vol. 3, no. 11, pp. 827–837, Nov. 2002, doi: 10.1038/nrg928.

[78] J. A. Birchler and H. Yang, “The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation,” Plant Cell, Mar. 2022, doi: 10.1093/PLCELL/KOAC076.

[79] E. Gonzalez et al., “The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility,” Science, vol. 307, no. 5714, pp. 1434–1440, Mar. 2005, doi: 10.1126/SCIENCE.1101160.

[80] P. Hallast, K. Rull, and M. Laan, “The evolution and genomic landscape of CGB1 and CGB2 genes,” Mol Cell Endocrinol, vol. 260–262, 2007, doi: 10.1016/j.mce.2005.11.049.

[81] A. Henke and J. Gromoll, “New insights into the evolution of chorionic gonadotrophin,” 2008. doi: 10.1016/j.mce.2008.05.009.

[82] T. Ohta, “Evolution of Gene Families,” in Brenner’s Encyclopedia of Genetics: Second Edition, 2013. doi: 10.1016/B978-0-12-374984-0.00497-6.

[83] H. Pöpperl et al., “Segmental expression of Hoxb-1 is controlled by a highly conserved autoregulatory loop dependent upon exd/pbx,” Cell, vol. 81, no. 7, 1995, doi: 10.1016/S0092-8674(05)80008-X.

[84] T. MacCarthy and A. Bergman, “The limits of subfunctionalization,” BMC Evol Biol, vol. 7, no. 1, 2007, doi: 10.1186/1471-2148-7-213.

[85] M. Kimura, “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences,” J Mol Evol, vol. 16, no. 2, 1980, doi: 10.1007/BF01731581.

[86] T. Gojobori, W. H. Li, and D. Graur, “Patterns of nucleotide substitution in pseudogenes and functional genes,” J Mol Evol, vol. 18, no. 5, 1982, doi: 10.1007/BF01733904.

[87] A. Kawashima and Y. Satta, “Substrate-dependent evolution of cytochrome P450: Rapid turnover of the detoxification-type and conservation of the biosynthesis-type,” PLoS One, vol. 9, no. 6, 2014, doi: 10.1371/journal.pone.0100059.

[88] S. Kumar, M. P. Suleski, G. J. Markov, S. Lawrence, A. Marco, and A. J. Filipski, “Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations,” Genome Res, vol. 19, no. 9, pp. 1562–1569, Sep. 2009, doi: 10.1101/GR.091991.109.

[89] T. A. P. de Beer, R. A. Laskowski, S. L. Parks, B. Sipos, N. Goldman, and J. M. Thornton, “Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset,” PLoS Comput Biol, vol. 9, no. 12, Dec. 2013, doi: 10.1371/JOURNAL.PCBI.1003382.

[90] S. Richards et al., “Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.,” Genet Med, vol. 17, no. 5, pp. 405–24, May 2015, doi: 10.1038/gim.2015.30.

[91] P. C. Ng and S. Henikoff, “SIFT: Predicting amino acid changes that affect protein function.,” Nucleic Acids Res, vol. 31, no. 13, pp. 3812–4, Jul. 2003, doi: 10.1093/nar/gkg509.

[92] Y. Choi and A. P. Chan, “PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels,” Bioinformatics, vol. 31, no. 16, pp. 2745–2747, Aug. 2015, doi: 10.1093/bioinformatics/btv195.

[93] H. Mi, A. Muruganujan, and P. D. Thomas, “PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees,” Nucleic Acids Res, vol. 41, no. D1, pp. D377–D386, Nov. 2012, doi: 10.1093/nar/gks1118.

[94] I. A. Adzhubei et al., “A method and server for predicting damaging missense mutations,” Nat Methods, vol. 7, no. 4, pp. 248–249, Apr. 2010, doi: 10.1038/nmeth0410-248.

[95] A. Niroula, S. Urolagin, and M. Vihinen, “PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants,” PLoS One, vol. 10, no. 2, p. e0117380, Feb. 2015, doi: 10.1371/journal.pone.0117380.

[96] T. A. Peterson, E. Doughty, and M. G. Kann, “Towards precision medicine: Advances in computational approaches for the analysis of human variants,” 2013. doi: 10.1016/j.jmb.2013.08.008.

[97] G. Zhang et al., “Triangulation of the human, chimpanzee, and Neanderthal genome sequences identifies potentially compensated mutations,” Hum Mutat, vol. 31, no. 12, pp. 1286–1293, Dec. 2010, doi: 10.1002/humu.21389.

[98] “Initial sequence of the chimpanzee genome and comparison with the human genome,” Nature, vol. 437, no. 7055, pp. 69–87, Sep. 2005, doi: 10.1038/nature04072.

[99] Y. Xue et al., “Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding,” Science (1979), vol. 348, no. 6231, pp. 242–245, Apr. 2015, doi: 10.1126/SCIENCE.AAA3952.

[100] S. E. Hunt et al., “Ensembl variation resources,” Database, vol. 2018, Jan. 2018, doi: 10.1093/database/bay119.

[101] A. Löytynoja and N. Goldman, “webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser.,” BMC Bioinformatics, vol. 11, no. 1, p. 579, Nov. 2010, doi: 10.1186/1471-2105-11-579.

[102] P. D. Stenson et al., “The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies.,” Hum Genet, vol. 136, no. 6, pp. 665–677, 2017, doi: 10.1007/s00439-017-1779-6.

[103] Z. Yang, “PAML 4: Phylogenetic Analysis by Maximum Likelihood,” Mol Biol Evol, vol. 24, no. 8, pp. 1586–1591, Apr. 2007, doi: 10.1093/molbev/msm088.

[104] A. Stamatakis, “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies,” Bioinformatics, vol. 30, no. 9, p. 1312, 2014, doi: 10.1093/BIOINFORMATICS/BTU033.

[105] A. Niroula and M. Vihinen, “How good are pathogenicity predictors in detecting benign variants?,” PLoS Comput Biol, vol. 15, no. 2, 2019, doi: 10.1371/journal.pcbi.1006481.

[106] M. S. Hassan, A. A. Shaalan, M. I. Dessouky, A. E. Abdelnaiem, and M. ElHefnawi, “Evaluation of computational techniques for predicting non-synonymous single nucleotide variants pathogenicity,” Genomics, vol. 111, no. 4, 2019, doi: 10.1016/j.ygeno.2018.05.013.

[107] C. Dong et al., “Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies,” Hum Mol Genet, vol. 24, no. 8, 2015, doi: 10.1093/hmg/ddu733.

[108] R. Ghosh, N. Oak, and S. E. Plon, “Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines,” Genome Biol, vol. 18, no. 1, 2017, doi: 10.1186/s13059-017-1353-5.

[109] W. S. W. Wong, Z. Yang, N. Goldman, and R. Nielsen, “Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites.,” Genetics, vol. 168, no. 2, pp. 1041–51, Oct. 2004, doi: 10.1534/genetics.104.031153.

[110] Z. Yang, W. S. W. Wong, and R. Nielsen, “Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection,” Mol Biol Evol, vol. 22, no. 4, pp. 1107–1118, Jan. 2005, doi: 10.1093/molbev/msi097.

Appendices

Appendix I: Results from the pathogenicity prediction tools.

Appendix II: Results from site selection analalysis using Bayes empirical Bayes (BEB) method.

Appendix I Results from the pathogenicity prediction tools.

GENE	UNIPROTKB	MUTATION	TYPE	POLYPHEN-2	SCORE	PON-P2	SCORE	FATHMM	SCORE
OTC	**P00480**	A12T	CPD	Tolerated	0	Tolerated	0.05	Damaging	-5.45
OTC	**P00480**	R26Q	CPD	Tolerated	0.002	Tolerated	0.133	Damaging	-5.41
OTC	**P00480**	R40H	CPD	Damaging	0.54	Unknown	0.292	Damaging	-5.62
OTC	**P00480**	R40C	CPD	Damaging	1	Unknown	0.555	Damaging	-5.64
OTC	**P00480**	N47T	CPD	Tolerated	0.07	Tolerated	0.174	Damaging	-4.9
OTC	**P00480**	E52D	CPD	Damaging	0.688	Unknown	0.665	Damaging	-5.52
OTC	**P00480**	T125M	CPD	Damaging	0	Tolerated	0.095	Damaging	-4.96
OTC	**P00480**	A155S	CPD	Damaging	0.492	Unknown	0.466	Damaging	-4.81
OTC	**P00480**	A12S	Polymorphism	Tolerated	0	Tolerated	0.039	Damaging	-5.44
OTC	**P00480**	A12E	Polymorphism	Tolerated	0.041	Unknown	0.08	Damaging	-5.49
OTC	**P00480**	R26W	Polymorphism	Tolerated	0	Tolerated	0.045	Damaging	-5.49
OTC	**P00480**	R26G	Polymorphism	Tolerated	0	Tolerated	0.033	Damaging	-5.45
OTC	**P00480**	N47D	Polymorphism	Tolerated	0	Tolerated	0.045	Damaging	-4.46

OTC	**P00480**	M56I	Polymorphism	Tolerated	0.001	Tolerated	0.104	Damaging	-5.04
OTC	**P00480**	S60E	Polymorphism	Damaging	0.984	Unknown	0.734	Damaging	-5.45
OTC	**P00480**	L63M	Polymorphism	Damaging	1	Unknown	0.71	Damaging	-5.18
OTC	**P00480**	L76F	Polymorphism	Damaging	0.969	Unknown	0.331	Damaging	-5.47
OTC	**P00480**	L77I	Polymorphism	Damaging	0.974	Unknown	0.456	Damaging	-5.38
OTC	**P00480**	G79R	Polymorphism	Damaging	1	Damaging	0.964	Damaging	-5.7
OTC	**P00480**	G83S	Polymorphism	Tolerated	0.003	Unknown	0.288	Damaging	-4.94
OTC	**P00480**	L95M	Polymorphism	Tolerated	0.045	Unknown	0.32	Damaging	-5.11
OTC	**P00480**	T125K	Polymorphism	Tolerated	0.002	Tolerated	0.03	Damaging	-4.79
OTC	**P00480**	A135T	Polymorphism	Tolerated	0	Tolerated	0.033	Damaging	-4.97
OTC	**P00480**	A137V	Polymorphism	Tolerated	0.039	Tolerated	0.172	Damaging	-4.83
OTC	**P00480**	I160V	Polymorphism	Tolerated	0.001	Tolerated	0.123	Damaging	-5.23
OTC	**P00480**	H182R	Polymorphism	Damaging	0.988	Damaging	0.928	Damaging	-4.9
OTC	**P00480**	L215I	Polymorphism	Tolerated	0.011	Unknown	0.346	Damaging	-5.27
OTC	**P00480**	Q216R	Polymorphism	Tolerated	0.003	Tolerated	0.146	Damaging	-5.18
OTC	**P00480**	A217V	Polymorphism	Tolerated	0.003	Tolerated	0.086	Damaging	-5.08
OTC	**P00480**	A233S	Polymorphism	Tolerated	0.337	Unknown	0.524	Damaging	-5.54
OTC	**P00480**	A237T	Polymorphism	Damaging	0.975	Unknown	0.296	Damaging	-5.44
OTC	**P00480**	L244F	Polymorphism	Tolerated	0.003	Tolerated	0.114	Damaging	-5.25
OTC	**P00480**	L244M	Polymorphism	Tolerated	0.007	Tolerated	0.198	Damaging	-5.37
OTC	**P00480**	T247M	Polymorphism	Damaging	0.831	Unknown	0.452	Damaging	-5.51

OTC	**P00480**	D249E	Polymorphism	Tolerated	0	Unknown	0.276	Damaging	-5.57
OTC	**P00480**	H255R	Polymorphism	Tolerated	0	Tolerated	0.062	Damaging	-5.17
OTC	**P00480**	H255S	Polymorphism	Tolerated	0	Tolerated	0.188	Damaging	-5.15
OTC	**P00480**	H255K	Polymorphism	Tolerated	0	Unknown	0.214	Damaging	-4.93
OTC	**P00480**	H255Q	Polymorphism	Tolerated	0	Tolerated	0.076	Damaging	-5.13
OTC	**P00480**	H255C	Polymorphism	Tolerated	0.372	Tolerated	0.169	Damaging	-5.26
OTC	**P00480**	G269R	Polymorphism	Damaging	1	Damaging	0.864	Damaging	-5.31
OTC	**P00480**	Q270H	Polymorphism	Damaging	0.998	Unknown	0.532	Damaging	-5.35
OTC	**P00480**	R320Q	Polymorphism	Tolerated	0.215	Tolerated	0.125	Damaging	-5.27
OTC	**P00480**	R320K	Polymorphism	Tolerated	0.001	Tolerated	0.151	Damaging	-5.22
OTC	**P00480**	L341V	Polymorphism	Tolerated	0.009	Unknown	0.508	Damaging	-5.47
ABCD1	**P33897**	A616T	CPD	Damaging	0.913	Damaging	0.986	Damaging	-8.17
ABCD1	**P33897**	D555N	CPD	Damaging	0.998	Damaging	0.967	Damaging	-3.27
ABCD1	**P33897**	P543L	CPD	Damaging	0.449	Damaging	0.911	Damaging	-7.21
ABCD1	**P33897**	R518W	CPD	Damaging	0.635	Damaging	0.912	Damaging	-7.35
ABCD1	**P33897**	M403K	CPD	Damaging	0.987	Damaging	0.827	Damaging	-4.01
ABCD1	**P33897**	R236H	CPD	Damaging	0.969	Unknown	0.746	Damaging	-6.35
ABCD1	**P33897**	M67V	CPD	Tolerated	0	Tolerated	0.115	Damaging	-6.1
ABCD1	**P33897**	A616D	Polymorphism	Damaging	0.994	Damaging	0.993	Damaging	-8.2
ABCD1	**P33897**	A616V	Polymorphism	Damaging	0.984	Damaging	0.973	Damaging	-8.14
ABCD1	**P33897**	R518Q	Polymorphism	Damaging	0.984	Damaging	0.916	Damaging	-7.29

ABCD1	**P33897**	R518G	Polymorphism	Damaging	0.983	Damaging	0.941	Damaging	-7.29
ABCD1	**P33897**	R236C	Polymorphism	Damaging	0.969	Unknown	0.77	Damaging	-6.37
ABCD1	**P33897**	T693M	Polymorphism	Tolerated	0.416	Unknown	0.779	Damaging	-3.39
ABCD1	**P33897**	G677D	Polymorphism	Damaging	0.999	Damaging	0.924	Damaging	-3.8
ABCD1	**P33897**	D638Y	Polymorphism	Damaging	0.998	Damaging	0.968	Damaging	-7.14
ABCD1	**P33897**	S636I	Polymorphism	Damaging	0.997	Damaging	0.99	Damaging	-7.19
ABCD1	**P33897**	V635M	Polymorphism	Damaging	0.918	Damaging	0.987	Damaging	-7.21
ABCD1	**P33897**	I613N	Polymorphism	Tolerated	0.195	Damaging	0.838	Damaging	-3.45
ABCD1	**P33897**	Q611H	Polymorphism	Damaging	0.997	Damaging	0.932	Damaging	-7.66
ABCD1	**P33897**	L585P	Polymorphism	Damaging	1	Damaging	0.975	Damaging	-7.3
ABCD1	**P33897**	I558T	Polymorphism	Damaging	0.966	Damaging	0.975	Damaging	-3.31
ABCD1	**P33897**	Q556R	Polymorphism	Damaging	0.999	Damaging	0.932	Damaging	-7.14
ABCD1	**P33897**	F540S	Polymorphism	Damaging	0.989	Damaging	0.929	Damaging	-7.04
ABCD1	**P33897**	F540C	Polymorphism	Damaging	0.996	Damaging	0.976	Damaging	-7.15
ABCD1	**P33897**	P508L	Polymorphism	Damaging	0.847	Damaging	0.943	Damaging	-3.38
ABCD1	**P33897**	P484R	Polymorphism	Damaging	0.998	Unknown	0.851	Damaging	-7.17
ABCD1	**P33897**	H420P	Polymorphism	Tolerated	0.105	Unknown	0.758	Damaging	-4
ABCD1	**P33897**	Y415C	Polymorphism	Damaging	0.993	Damaging	0.989	Damaging	-4.1
ABCD1	**P33897**	S404P	Polymorphism	Damaging	0.934	Unknown	0.658	Damaging	-4.01
ABCD1	**P33897**	R389H	Polymorphism	Damaging	0.97	Damaging	0.98	Damaging	-4.04
ABCD1	**P33897**	R389G	Polymorphism	Damaging	0.975	Damaging	0.941	Damaging	-3.3

ABCD1	**P33897**	R389C	Polymorphism	Damaging	0.985	Damaging	0.975	Damaging	-4.05
ABCD1	**P33897**	M329R	Polymorphism	Damaging	0.997	Damaging	0.928	Damaging	-6.37
ABCD1	**P33897**	A314P	Polymorphism	Damaging	0.864	Unknown	0.614	Damaging	-6.32
ABCD1	**P33897**	R259W	Polymorphism	Damaging	0.999	Damaging	0.897	Damaging	-6.46
ABCD1	**P33897**	N214D	Polymorphism	Tolerated	0.375	Damaging	0.924	Damaging	-6.34
ABCD1	**P33897**	V208E	Polymorphism	Damaging	0.99	Damaging	0.845	Damaging	-3.59
ABCD1	**P33897**	A205E	Polymorphism	Damaging	0.958	Damaging	0.778	Damaging	-6.36
ABCD1	**P33897**	S161P	Polymorphism	Damaging	0.653	Unknown	0.814	Damaging	-3.47
ABCD1	**P33897**	V147G	Polymorphism	Damaging	0.94	Unknown	0.57	Damaging	-3.46
ABCD1	**P33897**	A119D	Polymorphism	Damaging	0.752	Damaging	0.814	Damaging	-3.46
ABCD1	**P33897**	A99D	Polymorphism	Damaging	0.449	Unknown	0.794	Damaging	-3.53
ABCD1	**P33897**	A95D	Polymorphism	Damaging	0.673	Unknown	0.728	Damaging	-3.57
AR	**P10275**	Q196R	CPD	Tolerated	0.022	Unknown	Unknown	Tolerated	-1.14
AR	**P10275**	P392S	CPD	Tolerated	0.21	Unknown	Unknown	Damaging	-3.89
AR	**P10275**	G492S	CPD	Tolerated	0.015	Unknown	Unknown	Damaging	-2.22
AR	**P10275**	S176R	Polymorphism	Damaging	0.998	Unknown	Unknown	Damaging	-3.73
AR	**P10275**	N235K	Polymorphism	Tolerated	0.189	Unknown	Unknown	Damaging	-3.31
AR	**P10275**	L272F	Polymorphism	Damaging	0.857	Unknown	Unknown	Damaging	-3.9
AR	**P10275**	E355Q	Polymorphism	Damaging	0.965	Unknown	Unknown	Damaging	-3.76
AR	**P10275**	P380R	Polymorphism	Damaging	0.845	Unknown	Unknown	Damaging	-3.93
AR	**P10275**	P392R	Polymorphism	Damaging	0.952	Unknown	Unknown	Damaging	-3.93

AR	**P10275**	S413N	Polymorphism	Tolerated	0.075	Unknown	Unknown	Damaging	-3.73
AR	**P10275**	A475V	Polymorphism	Tolerated	0.178	Unknown	Unknown	Damaging	-2.39
AR	**P10275**	G507D	Polymorphism	Damaging	0.911	Unknown	Unknown	Damaging	-2.67
AR	**P10275**	L548F	Polymorphism	Unknown		Unknown	Unknown	Damaging	-2.32
AR	**P10275**	A646D	Polymorphism	Unknown		Unknown	Unknown	Damaging	-2.27
AR	**P10275**	S651G	Polymorphism	Unknown		Unknown	Unknown	Damaging	-2.36
BMP15	**O95972**	R61Q	CPD	Tolerated	0.001	Tolerated	0.15	Damaging	-1.65
BMP15	**O95972**	R76H	CPD	Damaging	0.959	Tolerated	0.192	Damaging	-1.68
BMP15	**O95972**	R138H	CPD	Damaging	0.854	Tolerated	0.153	Tolerated	-1.26
BMP15	**O95972**	R206H	CPD	Tolerated	0.003	Tolerated	0.08	Tolerated	-1.33
BMP15	**O95972**	Y235C	CPD	Damaging	0.95	Unknown	0.416	Damaging	-1.52
BMP15	**O95972**	I243V	CPD	Tolerated	0	Tolerated	0.039	Tolerated	-1.07
BMP15	**O95972**	S5R	Polymorphism	Tolerated	0.368	Tolerated	0.051	Damaging	-2.15
BMP15	**O95972**	R61W	Polymorphism	Tolerated	0.001	Tolerated	0.146	Damaging	-1.71
BMP15	**O95972**	R68W	Polymorphism	Damaging	0.877	Unknown	0.486	Damaging	-1.7
BMP15	**O95972**	R76C	Polymorphism	Damaging	0.959	Unknown	0.396	Damaging	-1.71
BMP15	**O95972**	W221R	Polymorphism	Tolerated	0.174	Tolerated	0.116	Tolerated	-1.13
BMP15	**O95972**	R329C	Polymorphism	Damaging	0.916	Unknown	0.282	Damaging	-1.86
COL4A5	**P29400**	G545R	CPD	Damaging	1	Damaging	0.859	Damaging	-6.29
COL4A5	**P29400**	G595R	CPD	Damaging	0.688	Unknown	0.349	Damaging	-3.16
COL4A5	**P29400**	P739S	CPD	Tolerated	0.007	Tolerated	0.238	Damaging	-3.71

COL4A5	**P29400**	S1659N	CPD	Damaging	0.597	Unknown	Unknown	Damaging	#REF!
COL4A5	**P29400**	P88S	Polymorphism	Damaging	0.558	Unknown	0.717	Damaging	-3.14
COL4A5	**P29400**	G264R	Polymorphism	Damaging	0.999	Damaging	0.841	Damaging	-4.63
COL4A5	**P29400**	G264D	Polymorphism	Damaging	0.999	Unknown	0.759	Damaging	-4.61
COL4A5	**P29400**	G331V	Polymorphism	Damaging	1	Damaging	0.911	Damaging	-6.29
COL4A5	**P29400**	G545V	Polymorphism	Damaging	1	Damaging	0.868	Damaging	-6.29
COL4A5	**P29400**	G545S	Polymorphism	Damaging	1	Damaging	0.908	Damaging	-6.27
COL4A5	**P29400**	P739A	Polymorphism	Tolerated	0.083	Unknown	0.289	Damaging	-3.71
COL4A5	**P29400**	M898V	Polymorphism	Tolerated	0.018	Unknown	0.34	Damaging	-2.41
COL4A5	**P29400**	S916G	Polymorphism	Tolerated	0.089	Tolerated	0.238	Damaging	-3.27
COL4A5	**P29400**	G1182R	Polymorphism	Damaging	1	Damaging	0.904	Damaging	-4.76
COL4A5	**P29400**	G1244D	Polymorphism	Damaging	0.885	Damaging	0.956	Damaging	-6.28
COL4A5	**P29400**	Q1308H	Polymorphism	Tolerated	0.333	Unknown	0.203	Damaging	-3.43
COL4A5	**P29400**	P1329S	Polymorphism	Tolerated	0.234	Tolerated	0.162	Damaging	-4.06
COL4A5	**P29400**	L1352I	Polymorphism	Tolerated	0.029	Unknown	Unknown	Damaging	-3.23
COL4A5	**P29400**	R1410C	Polymorphism	Damaging	0.999	Unknown	Unknown	Damaging	-3.4
COL4A5	**P29400**	R1422C	Polymorphism	Tolerated	0.209	Unknown	Unknown	Damaging	-3.38
COL4A5	**P29400**	S1659G	Polymorphism	Damaging	0.526	Unknown	Unknown	Damaging	-3.43
EDA	**Q92838**	R69L	CPD	Tolerated	0.141	Unknown	0.358	Damaging	-3.93
EDA	**Q92838**	R153H	CPD	Tolerated	0	Unknown	0.215	Tolerated	-0.67
EDA	**Q92838**	G2C	Polymorphism	Damaging	0.86	Unknown	0.634	Damaging	-4.11

EDA	**Q92838**	S125C	Polymorphism	Tolerated	0.053	Tolerated	0.214	Damaging	-3.66
EDA	**Q92838**	R153C	Polymorphism	Tolerated	0.183	Unknown	0.547	Tolerated	-0.69
EDA	**Q92838**	R155C	Polymorphism	Tolerated	0.001	Damaging	0.816	Tolerated	-0.64
F9	**P00740**	R449W	CPD	Tolerated	0.294	Unknown	0.777	Damaging	-2.49
F9	**P00740**	V374I	CPD	Tolerated	0.116	Unknown	0.285	Damaging	-2.49
F9	**P00740**	S350A	CPD	Tolerated	0.009	Tolerated	0.212	Damaging	-2.46
F9	**P00740**	S329R	CPD	Tolerated	0.002	Unknown	0.53	Damaging	-3.17
F9	**P00740**	E323K	CPD	Tolerated	0	Tolerated	0.096	Damaging	-2.44
F9	**P00740**	H289Y	CPD	Tolerated	0.128	Unknown	0.248	Damaging	-2.39
F9	**P00740**	N283D	CPD	Tolerated	0.013	Unknown	0.592	Damaging	-3.06
F9	**P00740**	E270K	CPD	Tolerated	0.002	Tolerated	0.155	Damaging	-2.35
F9	**P00740**	V257I	CPD	Tolerated	0	Tolerated	0.113	Damaging	-2.32
F9	**P00740**	F224L	CPD	Tolerated	0	Unknown	0.238	Damaging	-3.47
F9	**P00740**	E142K	CPD	Tolerated	0.005	Tolerated	0.074	Damaging	-4.14
F9	**P00740**	G105D	CPD	Tolerated	0.25	Unknown	0.671	Damaging	-3.15
F9	**P00740**	T84A	CPD	Damaging	0.651	Unknown	0.627	Damaging	-6.04
F9	**P00740**	R75Q	CPD	Damaging	0.999	Unknown	0.853	Damaging	-5.78
F9	**P00740**	C18F	CPD	Tolerated	0.002	Unknown	0.352	Damaging	-2.86
F9	**P00740**	R3H	CPD	Tolerated	0.001	Tolerated	0.113	Damaging	-2.88
F9	**P00740**	R449Q	Polymorphism	Damaging	0.513	Unknown	0.7	Damaging	-2.37
F9	**P00740**	I443T	Polymorphism	Damaging	0.992	Damaging	0.922	Damaging	-2.53

F9	**P00740**	I443L	Polymorphism	Tolerated	0.373	Unknown	0.48	Damaging	-2.45
F9	**P00740**	T422N	Polymorphism	Damaging	0.999	Unknown	0.726	Damaging	-3.19
F9	**P00740**	S406L	Polymorphism	Damaging	0.562	Unknown	0.675	Damaging	-3.2
F9	**P00740**	E401K	Polymorphism	Tolerated	0.066	Unknown	0.251	Damaging	-3.15
F9	**P00740**	N393I	Polymorphism	Damaging	0.87	Unknown	0.732	Damaging	-3.29
F9	**P00740**	N392D	Polymorphism	Tolerated	0.029	Unknown	0.274	Damaging	-2.97
F9	**P00740**	Y391C	Polymorphism	Damaging	0.897	Unknown	0.53	Damaging	-2.42
F9	**P00740**	K387N	Polymorphism	Tolerated	0.026	Unknown	0.364	Damaging	-3.07
F9	**P00740**	K387E	Polymorphism	Tolerated	0.028	Unknown	0.713	Damaging	-3.14
F9	**P00740**	T381K	Polymorphism	Tolerated	0.023	Damaging	0.798	Damaging	-2.41
F9	**P00740**	T381A	Polymorphism	Tolerated	0.042	Unknown	0.347	Damaging	-2.48
F9	**P00740**	V377D	Polymorphism	Damaging	0.988	Damaging	0.936	Damaging	-2.72
F9	**P00740**	V377A	Polymorphism	Damaging	0.988	Unknown	0.793	Damaging	-2.64
F9	**P00740**	V377F	Polymorphism	Damaging	0.826	Damaging	0.898	Damaging	-2.69
F9	**P00740**	L376P	Polymorphism	Damaging	0.988	Damaging	0.903	Damaging	-2.64
F9	**P00740**	V374F	Polymorphism	Damaging	0.958	Damaging	0.906	Damaging	-2.67
F9	**P00740**	R373I	Polymorphism	Damaging	0.526	Unknown	0.323	Damaging	-2.41
F9	**P00740**	Y371S	Polymorphism	Tolerated	0.091	Unknown	0.35	Damaging	-2.32
F9	**P00740**	K362E	Polymorphism	Tolerated	0.01	Unknown	0.588	Damaging	-3.08
F9	**P00740**	F360I	Polymorphism	Tolerated	0.075	Unknown	0.508	Damaging	-2.43
F9	**P00740**	F360S	Polymorphism	Damaging	0.479	Unknown	0.566	Damaging	-2.34

F9	**P00740**	V359D	Polymorphism	Damaging	0.853	Unknown	0.626	Damaging	-3.12
F9	**P00740**	V359G	Polymorphism	Damaging	0.944	Unknown	0.344	Damaging	-3.17
F9	**P00740**	R358G	Polymorphism	Damaging	0.658	Unknown	0.356	Damaging	-2.55
F9	**P00740**	G351D	Polymorphism	Damaging	0.886	Damaging	0.955	Damaging	-3.22
F9	**P00740**	G351V	Polymorphism	Damaging	0.947	Damaging	0.938	Damaging	-3.05
F9	**P00740**	I344N	Polymorphism	Tolerated	0.013	Tolerated	0.209	Damaging	-2.29
F9	**P00740**	I344F	Polymorphism	Tolerated	0.001	Unknown	0.365	Damaging	-2.39
F9	**P00740**	T342K	Polymorphism	Damaging	0.487	Damaging	0.914	Damaging	-2.38
F9	**P00740**	T342M	Polymorphism	Damaging	0.882	Damaging	0.96	Damaging	-2.51
F9	**P00740**	T342P	Polymorphism	Damaging	0.977	Unknown	0.744	Damaging	-2.47
F9	**P00740**	T342A	Polymorphism	Tolerated	0.009	Unknown	0.785	Damaging	-2.33
F9	**P00740**	K339E	Polymorphism	Damaging	0.471	Unknown	0.512	Damaging	-3.1
F9	**P00740**	L319Q	Polymorphism	Damaging	1	Damaging	0.924	Damaging	-3.9
F9	**P00740**	L319P	Polymorphism	Damaging	1	Damaging	0.953	Damaging	-3.9
F9	**P00740**	L319R	Polymorphism	Damaging	0.998	Damaging	0.945	Damaging	-3.9
F9	**P00740**	H302R	Polymorphism	Damaging	0.98	Unknown	0.811	Damaging	-3.95
F9	**P00740**	R298L	Polymorphism	Tolerated	0.268	Unknown	0.435	Damaging	-3.27
F9	**P00740**	I297N	Polymorphism	Damaging	0.889	Unknown	0.716	Damaging	-2.36
F9	**P00740**	I284N	Polymorphism	Damaging	0.737	Unknown	0.427	Damaging	-2.57
F9	**P00740**	H282Q	Polymorphism	Tolerated	0.409	Unknown	0.565	Damaging	-2.56
F9	**P00740**	H282R	Polymorphism	Tolerated	0.028	Unknown	0.76	Damaging	-2.53

F9	**P00740**	V277D	Polymorphism	Damaging	0.999	Unknown	0.733	Damaging	-4.06
F9	**P00740**	V277F	Polymorphism	Damaging	0.951	Unknown	0.529	Damaging	-4
F9	**P00740**	I262T	Polymorphism	Damaging	0.987	Unknown	0.434	Damaging	-3.43
F9	**P00740**	I262M	Polymorphism	Damaging	0.993	Unknown	0.309	Damaging	-3.44
F9	**P00740**	I262F	Polymorphism	Damaging	0.797	Unknown	0.376	Damaging	-3.44
F9	**P00740**	N258K	Polymorphism	Tolerated	0.295	Damaging	0.789	Damaging	-3.22
F9	**P00740**	V257D	Polymorphism	Damaging	0.504	Damaging	0.76	Damaging	-3.38
F9	**P00740**	V257F	Polymorphism	Tolerated	0.033	Unknown	0.571	Damaging	-3.36
F9	**P00740**	A250V	Polymorphism	Tolerated	0.004	Unknown	0.332	Damaging	-3.09
F9	**P00740**	F238V	Polymorphism	Tolerated	0.109	Tolerated	0.226	Damaging	-3.28
F9	**P00740**	A233D	Polymorphism	Tolerated	0.437	Damaging	0.927	Damaging	-3.5
F9	**P00740**	A192P	Polymorphism	Tolerated	0.003	Unknown	0.66	Damaging	-3.56
F9	**P00740**	A173G	Polymorphism	Tolerated	0.002	Unknown	0.291	Damaging	-3.46
F9	**P00740**	A173V	Polymorphism	Tolerated	0.007	Unknown	0.354	Damaging	-3.5
F9	**P00740**	Q167H	Polymorphism	Tolerated	0.025	Unknown	0.633	Damaging	-4.01
F9	**P00740**	N166Y	Polymorphism	Damaging	0.615	Unknown	0.742	Damaging	-4.08
F9	**P00740**	R162L	Polymorphism	Tolerated	0.004	Unknown	0.592	Damaging	-3.98
F9	**P00740**	V154A	Polymorphism	Tolerated	0.017	Unknown	0.353	Damaging	-3.97
F9	**P00740**	V153M	Polymorphism	Tolerated	0.435	Unknown	0.535	Damaging	-4.05
F9	**P00740**	V153A	Polymorphism	Tolerated	0.003	Unknown	0.294	Damaging	-4.07
F9	**P00740**	Q143K	Polymorphism	Damaging	0.703	Unknown	0.538	Damaging	-4.13

F9	**P00740**	Q143P	Polymorphism	Damaging	0.934	Damaging	0.811	Damaging	-4.19
F9	**P00740**	Q143R	Polymorphism	Tolerated	0.159	Unknown	0.57	Damaging	-4.17
F9	**P00740**	Q143E	Polymorphism	Damaging	0.67	Unknown	0.557	Damaging	-4.11
F9	**P00740**	R140S	Polymorphism	Tolerated	0.019	Damaging	0.783	Damaging	-3.93
F9	**P00740**	T133I	Polymorphism	Tolerated	0.001	Unknown	0.262	Damaging	-2.93
F9	**P00740**	S114P	Polymorphism	Tolerated	0.114	Unknown	0.502	Damaging	-3.09
F9	**P00740**	N113K	Polymorphism	Tolerated	0.063	Unknown	0.824	Damaging	-3.01
F9	**P00740**	I112N	Polymorphism	Damaging	0.912	Unknown	0.82	Damaging	-3
F9	**P00740**	I112T	Polymorphism	Tolerated	0.335	Unknown	0.643	Damaging	-3.01
F9	**P00740**	I112S	Polymorphism	Damaging	0.845	Damaging	0.96	Damaging	-2.98
F9	**P00740**	G105S	Polymorphism	Damaging	0.864	Damaging	0.905	Damaging	-3.4
F9	**P00740**	G105V	Polymorphism	Damaging	0.972	Damaging	0.926	Damaging	-3.43
F9	**P00740**	G105C	Polymorphism	Damaging	0.992	Damaging	0.921	Damaging	-3.44
F9	**P00740**	N104K	Polymorphism	Damaging	0.96	Damaging	0.949	Damaging	-3.56
F9	**P00740**	V92A	Polymorphism	Tolerated	0.063	Unknown	0.739	Damaging	-5.51
F9	**P00740**	K89E	Polymorphism	Tolerated	0.063	Unknown	0.559	Damaging	-5.45
F9	**P00740**	T84R	Polymorphism	Damaging	0.993	Damaging	0.896	Damaging	-6.06
F9	**P00740**	T84I	Polymorphism	Damaging	0.986	Damaging	0.94	Damaging	-5.96
F9	**P00740**	R75P	Polymorphism	Damaging	0.999	Damaging	0.917	Damaging	-5.8
F9	**P00740**	Q57E	Polymorphism	Tolerated	0.068	Tolerated	0.193	Damaging	-5.45
F9	**P00740**	F55I	Polymorphism	Tolerated	0.019	Unknown	0.445	Damaging	-5.34

F9	**P00740**	F55Y	Polymorphism	Damaging	0.601	Unknown	0.424	Damaging	-5.45
F9	**P00740**	F55S	Polymorphism	Damaging	0.477	Unknown	0.744	Damaging	-5.39
F9	**P00740**	F55C	Polymorphism	Damaging	0.895	Damaging	0.905	Damaging	-5.46
F9	**P00740**	N38H	Polymorphism	Tolerated	0.015	Unknown	0.516	Damaging	-2.9
F9	**P00740**	N36T	Polymorphism	Tolerated	0.019	Unknown	0.269	Damaging	-2.85
F9	**P00740**	T29I	Polymorphism	Tolerated	0.091	Unknown	0.311	Damaging	-2.94
F9	**P00740**	A26D	Polymorphism	Tolerated	0.036	Damaging	0.853	Damaging	-2.96
F9	**P00740**	L20S	Polymorphism	Tolerated	0.022	Tolerated	0.149	Damaging	-3.04
F9	**P00740**	I17N	Polymorphism	Tolerated	0.243	Damaging	0.944	Damaging	-3.02
F9	**P00740**	I7F	Polymorphism	Tolerated	0.002	Unknown	0.246	Damaging	-2.93
F9	**P00740**	V227D	Polymorphism	Tolerated	0.392	Unknown	0.819	Damaging	-3.28
F9	**P00740**	V227A	Polymorphism	Tolerated	0.068	Unknown	0.536	Damaging	-3.24
F9	**P00740**	V227F	Polymorphism	Tolerated	0.152	Unknown	0.77	Damaging	-3.26
GLA	**P06280**	M1I	CPD	Tolerated	0.001	Damaging	1	Damaging	-8.49
GLA	**P06280**	L14P	CPD	Tolerated	0.364	Unknown	0.672	Damaging	-8.79
GLA	**P06280**	A15G	CPD	Tolerated	0.003	Unknown	0.244	Damaging	-8.67
GLA	**P06280**	F18S	CPD	Tolerated	0.001	Tolerated	0.115	Damaging	-8.59
GLA	**P06280**	L19P	CPD	Tolerated	0.005	Unknown	0.461	Damaging	-8.92
GLA	**P06280**	W24C	CPD	Tolerated	0.001	Tolerated	0.17	Damaging	-8.59
GLA	**P06280**	A31V	CPD	Damaging	0.617	Unknown	0.799	Damaging	-8.88
GLA	**P06280**	L36F	CPD	Damaging	0.979	Unknown	0.358	Damaging	-8.4

GLA	**P06280**	P40L	CPD	Damaging	1	Damaging	0.938	Damaging	-9.89
GLA	**P06280**	M72I	CPD	Tolerated	0.007	Unknown	0.327	Damaging	-8.15
GLA	**P06280**	E103Q	CPD	Tolerated	0.001	Tolerated	0.067	Damaging	-6.96
GLA	**P06280**	S304N	CPD	Tolerated	0.005	Tolerated	0.206	Damaging	-8.26
GLA	**P06280**	R356Q	CPD	Tolerated	0.006	Tolerated	0.158	Damaging	-6.45
GLA	**P06280**	R363C	CPD	Tolerated	0.308	Unknown	0.222	Damaging	-8.17
GLA	**P06280**	V390M	CPD	Tolerated	0.202	Tolerated	0.096	Damaging	-6.02
GLA	**P06280**	E398A	CPD	Tolerated	0.044	Tolerated	0.097	Damaging	-6.06
GLA	**P06280**	E418G	CPD	Tolerated	0.001	Tolerated	0.124	Damaging	-8.18
GLA	**P06280**	L32P	Polymorphism	Damaging	0.998	Damaging	0.93	Damaging	-9.03
GLA	**P06280**	G35R	Polymorphism	Damaging	0.989	Unknown	0.813	Damaging	-8.22
GLA	**P06280**	M51K	Polymorphism	Tolerated	0.143	Damaging	0.792	Damaging	-8.22
GLA	**P06280**	M51I	Polymorphism	Tolerated	0.044	Unknown	0.396	Damaging	-8.25
GLA	**P06280**	N53D	Polymorphism	Damaging	0.513	Unknown	0.479	Damaging	-6.36
GLA	**P06280**	E59K	Polymorphism	Tolerated	0.047	Unknown	0.65	Damaging	-8.23
GLA	**P06280**	S65T	Polymorphism	Damaging	0.804	Unknown	0.39	Damaging	-8.15
GLA	**P06280**	S65I	Polymorphism	Damaging	0.991	Damaging	0.776	Damaging	-8.43
GLA	**P06280**	A73E	Polymorphism	Damaging	0.962	Damaging	0.825	Damaging	-8.68
GLA	**P06280**	A73V	Polymorphism	Tolerated	0.326	Damaging	0.849	Damaging	-8.54
GLA	**P06280**	M76T	Polymorphism	Damaging	0.997	Damaging	0.852	Damaging	-8.38
GLA	**P06280**	W81R	Polymorphism	Damaging	1	Damaging	0.966	Damaging	-8.2

GLA	**P06280**	W81S	Polymorphism	Damaging	1	Damaging	0.944	Damaging	-8.2
GLA	**P06280**	W81C	Polymorphism	Damaging	1	Damaging	0.951	Damaging	-8.22
GLA	**P06280**	G85D	Polymorphism	Damaging	0.989	Damaging	0.93	Damaging	-8.77
GLA	**P06280**	L89H	Polymorphism	Damaging	0.997	Unknown	0.765	Damaging	-8.48
GLA	**P06280**	L89P	Polymorphism	Damaging	0.998	Unknown	0.859	Damaging	-8.48
GLA	**P06280**	L89R	Polymorphism	Damaging	0.893	Unknown	0.819	Damaging	-8.48
GLA	**P06280**	G104V	Polymorphism	Damaging	0.984	Unknown	0.726	Damaging	-8.78
GLA	**P06280**	A121P	Polymorphism	Damaging	0.965	Unknown	0.515	Damaging	-8.46
GLA	**P06280**	N139S	Polymorphism	Tolerated	0.018	Tolerated	0.127	Damaging	-6.42
GLA	**P06280**	A143T	Polymorphism	Damaging	0.853	Unknown	0.622	Damaging	-8.23
GLA	**P06280**	A143P	Polymorphism	Damaging	0.975	Unknown	0.713	Damaging	-8.27
GLA	**P06280**	G163V	Polymorphism	Damaging	0.946	Unknown	0.845	Damaging	-9.22
GLA	**P06280**	V164G	Polymorphism	Damaging	0.998	Unknown	0.577	Damaging	-8.82
GLA	**P06280**	P210L	Polymorphism	Tolerated	0.356	Unknown	0.528	Damaging	-8.09
GLA	**P06280**	P210S	Polymorphism	Tolerated	0.213	Tolerated	0.239	Damaging	-8.08
GLA	**P06280**	I219T	Polymorphism	Damaging	0.974	Unknown	0.731	Damaging	-8.22
GLA	**P06280**	H225R	Polymorphism	Tolerated	0.412	Unknown	0.605	Damaging	-7.17
GLA	**P06280**	H225D	Polymorphism	Damaging	0.976	Unknown	0.725	Damaging	-7.18
GLA	**P06280**	F229L	Polymorphism	Tolerated	0.226	Tolerated	0.231	Damaging	-7.78
GLA	**P06280**	A230T	Polymorphism	Tolerated	0.006	Unknown	0.461	Damaging	-7.1
GLA	**P06280**	I232T	Polymorphism	Damaging	0.958	Unknown	0.62	Damaging	-8.4

GLA	**P06280**	S235C	Polymorphism	Damaging	0.775	Unknown	0.735	Damaging	-7.25
GLA	**P06280**	S235F	Polymorphism	Damaging	0.897	Unknown	0.68	Damaging	-7.25
GLA	**P06280**	I239T	Polymorphism	Damaging	0.923	Unknown	0.404	Damaging	-6.99
GLA	**P06280**	I242N	Polymorphism	Damaging	0.988	Damaging	0.894	Damaging	-7.97
GLA	**P06280**	I242F	Polymorphism	Damaging	0.65	Unknown	0.53	Damaging	-7.94
GLA	**P06280**	D244N	Polymorphism	Tolerated	0.025	Unknown	0.684	Damaging	-7.1
GLA	**P06280**	D244H	Polymorphism	Damaging	0.673	Unknown	0.775	Damaging	-7.14
GLA	**P06280**	S247P	Polymorphism	Tolerated	0.389	Unknown	0.619	Damaging	-7.16
GLA	**P06280**	S247C	Polymorphism	Damaging	0.656	Unknown	0.408	Damaging	-7.17
GLA	**P06280**	N249K	Polymorphism	Tolerated	0.004	Unknown	0.28	Damaging	-8.09
GLA	**P06280**	G261D	Polymorphism	Damaging	0.768	Damaging	0.925	Damaging	-7.65
GLA	**P06280**	G261V	Polymorphism	Damaging	0.917	Damaging	0.885	Damaging	-7.65
GLA	**P06280**	Q280K	Polymorphism	Tolerated	0.077	Unknown	0.429	Damaging	-6.54
GLA	**P06280**	Q280H	Polymorphism	Damaging	0.801	Unknown	0.738	Damaging	-6.59
GLA	**P06280**	Q283P	Polymorphism	Damaging	1	Damaging	0.83	Damaging	-6.57
GLA	**P06280**	I289S	Polymorphism	Damaging	0.999	Damaging	0.848	Damaging	-6.93
GLA	**P06280**	I289F	Polymorphism	Damaging	0.952	Damaging	0.87	Damaging	-6.93
GLA	**P06280**	F295C	Polymorphism	Damaging	0.881	Unknown	0.504	Damaging	-6.58
GLA	**P06280**	D313G	Polymorphism	Tolerated	0.377	Unknown	0.335	Damaging	-8.14
GLA	**P06280**	I319T	Polymorphism	Damaging	0.989	Damaging	0.875	Damaging	-6.56
GLA	**P06280**	I319F	Polymorphism	Damaging	0.815	Damaging	0.83	Damaging	-5.6

GLA	**P06280**	S345P	Polymorphism	Tolerated	0.243	Unknown	0.59	Damaging	-8.33
GLA	**P06280**	A348P	Polymorphism	Tolerated	0.322	Unknown	0.584	Damaging	-8.16
GLA	**P06280**	A352D	Polymorphism	Damaging	0.886	Damaging	0.894	Damaging	-6.58
GLA	**P06280**	I354K	Polymorphism	Tolerated	0.011	Unknown	0.258	Damaging	-8.26
GLA	**P06280**	N355K	Polymorphism	Damaging	0.721	Damaging	0.91	Damaging	-6.73
GLA	**P06280**	I384N	Polymorphism	Damaging	0.995	Damaging	0.869	Damaging	-8.2
GLA	**P06280**	Q386P	Polymorphism	Damaging	0.972	Damaging	0.755	Damaging	-8.12
GLA	**P06280**	K391T	Polymorphism	Damaging	0.694	Unknown	0.354	Damaging	-8.08
GLA	**P06280**	K391A	Polymorphism	Tolerated	0.162	Unknown	0.399	Damaging	-8.12
GLA	**P06280**	L403S	Polymorphism	Damaging	0.897	Unknown	0.562	Damaging	-6.14
GLA	**P06280**	I407K	Polymorphism	Damaging	0.469	Unknown	0.633	Damaging	-6.45
GLA	**P06280**	T410K	Polymorphism	Tolerated	0.13	Unknown	0.62	Damaging	-6.07
GLA	**P06280**	T410I	Polymorphism	Damaging	0.548	Unknown	0.669	Damaging	-6.12
GLA	**P06280**	T410P	Polymorphism	Damaging	0.966	Unknown	0.534	Damaging	-6.06
GLA	**P06280**	T410A	Polymorphism	Tolerated	0.14	Unknown	0.575	Damaging	-6.05
GLA	**P06280**	L415P	Polymorphism	Damaging	0.995	Damaging	0.877	Damaging	-6.4
G6PD	**P11413**	R104H	CPD	Tolerated	0.027	Damaging	0.805	Damaging	-4.89
G6PD	**P11413**	D113N	CPD	Tolerated	0.027	Tolerated	0.217	Damaging	-4.84
G6PD	**P11413**	I33M	Polymorphism	Damaging	0.811	Unknown	0.329	Damaging	-5.88
G6PD	**P11413**	R57W	Polymorphism	Damaging	0.995	Damaging	0.839	Damaging	-6.24
G6PD	**P11413**	F66I	Polymorphism	Tolerated	0.011	Unknown	0.489	Damaging	-4.74

G6PD	**P11413**	Y70H	Polymorphism	Damaging	0.999	Unknown	0.463	Damaging	-4.87
G6PD	**P11413**	Y70C	Polymorphism	Damaging	0.997	Unknown	0.733	Damaging	-4.86
G6PD	**P11413**	P92S	Polymorphism	Tolerated	0.031	Unknown	0.234	Damaging	-5.89
G6PD	**P11413**	Y118H	Polymorphism	Damaging	0.993	Unknown	0.691	Damaging	-6.11
G6PD	**P11413**	N126D	Polymorphism	Tolerated	0.001	Unknown	0.2	Damaging	-5.82
G6PD	**P11413**	R136C	Polymorphism	Damaging	0.886	Damaging	0.883	Damaging	-6.9
G6PD	**P11413**	H155N	Polymorphism	Damaging	0.56	Unknown	0.46	Damaging	-5.79
G6PD	**P11413**	H155D	Polymorphism	Damaging	0.675	Unknown	0.397	Damaging	-5.78
G6PD	**P11413**	N165D	Polymorphism	Tolerated	0.089	Unknown	0.704	Damaging	-4.78
G6PD	**P11413**	D181V	Polymorphism	Tolerated	0.026	Unknown	0.697	Damaging	-4.78
G6PD	**P11413**	R182W	Polymorphism	Tolerated	0.331	Unknown	0.628	Damaging	-4.84
G6PD	**P11413**	S188F	Polymorphism	Tolerated	0.039	Unknown	0.555	Damaging	-4.87
G6PD	**P11413**	V303L	Polymorphism	Tolerated	0.009	Unknown	Unknown	Damaging	-7.19
G6PD	**P11413**	V304F	Polymorphism	Damaging	0.96	Unknown	Unknown	Damaging	-7.44
G6PD	**P11413**	D350H	Polymorphism	Damaging	0.83	Unknown	Unknown	Damaging	-7.02
G6PD	**P11413**	I355T	Polymorphism	Damaging	0.986	Unknown	Unknown	Damaging	-7.03
G6PD	**P11413**	M411V	Polymorphism	Tolerated	0.431	Unknown	Unknown	Damaging	-6.76
G6PD	**P11413**	S448G	Polymorphism	Damaging	0.535	Unknown	Unknown	Damaging	-6.94
G6PD	**P11413**	P481R	Polymorphism	Tolerated	0.434	Unknown	Unknown	Damaging	-6.86
MECP2	**P51608**	T196S	CPD	Tolerated	0.006	Damaging	0.743	Damaging	-2.52
MECP2	**P51608**	L328V	CPD	Tolerated	0	Unknown	Unknown	Damaging	-2.46

MECP2	**P51608**	A358T	CPD	Tolerated	0	Unknown	Unknown	Damaging	-2.58
MECP2	**P51608**	P388T	CPD	Tolerated	0.001	Unknown	0.447	Damaging	-2.61
MECP2	**P51608**	P388L	CPD	Tolerated	0	Unknown	0.548	Damaging	-2.66
MECP2	**P51608**	E397K	CPD	Tolerated	0	Unknown	Unknown	Damaging	-2.4
MECP2	**P51608**	P419S	CPD	Tolerated	0.021	Unknown	Unknown	Damaging	-2.43
MECP2	**P51608**	H51Q	Polymorphism	Tolerated	0.017	Unknown	Unknown	Damaging	-2.59
MECP2	**P51608**	T228S	Polymorphism	Tolerated	0	Unknown	Unknown	Damaging	-2.47
MECP2	**P51608**	R268W	Polymorphism	Damaging	0.99	Unknown	Unknown	Damaging	-3.06
MECP2	**P51608**	R294G	Polymorphism	Tolerated	0.225	Unknown	Unknown	Damaging	-2.64
MECP2	**P51608**	T311M	Polymorphism	Damaging	0.848	Damaging	0.847	Damaging	-3.27
MECP2	**P51608**	S355R	Polymorphism	Tolerated	0.052	Unknown	Unknown	Damaging	-3.15
MECP2	**P51608**	P376R	Polymorphism	Tolerated	0.032	Unknown	Unknown	Damaging	-2.53
MECP2	**P51608**	A378G	Polymorphism	Tolerated	0.16	Unknown	Unknown	Damaging	-2.63

Appendix II. Results from site selection analalysis using Bayes empirical Bayes (BEB) method.

ABCD1

	Pr(w>1)	post mean +- SE for w
443 A	0.992**	1.496 +- 0.064
445 A	0.852	1.389 +- 0.272
447 S	0.967*	1.477 +- 0.130
464 R	0.960*	1.471 +- 0.150

BMP15

	Pr(w>1)	post mean +- SE for w
33 I	0.650	1.287 +- 0.297
46 E	0.629	1.273 +- 0.303
54 G	0.619	1.268 +- 0.302
61 R	0.922	1.455 +- 0.161
80 S	0.695	1.315 +- 0.286
104 V	0.716	1.328 +- 0.279
160 P	0.775	1.365 +- 0.257
168 E	0.624	1.269 +- 0.306
169 G	0.692	1.312 +- 0.289
172 S	0.673	1.299 +- 0.296
200 H	0.613	1.264 +- 0.303
259 R	0.699	1.319 +- 0.281
273 S	0.539	1.212 +- 0.321
281 S	0.852	1.413 +- 0.215

F9

		Pr(w>1)	post mean +- SE for w
6	V	0.658	1.267 +- 0.336
12	L	0.717	1.306 +- 0.321
13	F	0.861	1.412 +- 0.225
16	A	0.524	1.156 +- 0.382
74	R	0.566	1.191 +- 0.373
163	M	0.573	1.203 +- 0.360
188	T	0.877	1.424 +- 0.207
284	E	0.776	1.354 +- 0.279
355	I	0.877	1.425 +- 0.205
361	I	0.687	1.297 +- 0.309
393	M	0.829	1.391 +- 0.247
427	Q	0.536	1.169 +- 0.376
446	P	0.827	1.391 +- 0.244
461	S	0.565	1.205 +- 0.349
473	V	0.890	1.432 +- 0.197
510	F	0.836	1.395 +- 0.243
519	Q	0.832	1.396 +- 0.236
522	A	0.784	1.364 +- 0.265
525	P	0.783	1.359 +- 0.275
550	I	0.672	1.277 +- 0.331
551	L	0.544	1.173 +- 0.378
563	L	0.721	1.309 +- 0.319
565	S	0.730	1.322 +- 0.302
648	I	0.556	1.196 +- 0.355
664	K	0.616	1.232 +- 0.356

670	N	0.510	1.154 +- 0.372
691	L	0.529	1.148 +- 0.399
788	R	0.958*	1.475 +- 0.122
791	L	0.588	1.198 +- 0.383
801	V	0.887	1.431 +- 0.196
810	M	0.825	1.390 +- 0.242
815	L	0.741	1.329 +- 0.298
867	I	0.825	1.388 +- 0.247
870	A	0.561	1.204 +- 0.349
874	I	0.872	1.421 +- 0.210
882	L	0.703	1.303 +- 0.313
888	A	0.762	1.344 +- 0.287
894	T	0.937	1.462 +- 0.150
930	Q	0.689	1.293 +- 0.318
955	M	0.838	1.398 +- 0.237
1010	N	0.605	1.228 +- 0.352
1016	Q	0.666	1.277 +- 0.326
1020	I	0.871	1.421 +- 0.208
1028	T	0.570	1.201 +- 0.361
1040	V	0.568	1.200 +- 0.361
1096	I	0.861	1.412 +- 0.223
1100	V	0.885	1.428 +- 0.202
1136	S	0.851	1.409 +- 0.222
1156	H	0.505	1.163 +- 0.357
1233	V	0.746	1.335 +- 0.290
1307	L	0.554	1.181 +- 0.375
1308	Q	0.542	1.171 +- 0.378
1340	V	0.663	1.282 +- 0.313
1344	A	0.541	1.194 +- 0.345
1353	I	0.826	1.392 +- 0.240
1380	I	0.819	1.384 +- 0.253
1434	T	0.678	1.282 +- 0.328
1455	T	0.709	1.307 +- 0.310
1456	S	0.534	1.173 +- 0.368
1478	Q	0.839	1.399 +- 0.236
1481	L	0.558	1.192 +- 0.363
1483	V	0.762	1.344 +- 0.287
1572	V	0.591	1.217 +- 0.356
1585	H	0.558	1.207 +- 0.342

GLA

		Pr(w>1)	post mean +- SE for w
4	R	0.925	2.387 +- 0.396
6	P	0.999**	2.498 +- 0.054
7	E	0.999**	2.499 +- 0.042
8	L	0.980*	2.471 +- 0.208
9	H	0.857	2.285 +- 0.525
17	R	0.881	2.322 +- 0.485
57	Q	0.680	2.020 +- 0.700
103	E	0.859	2.288 +- 0.522
118	R	0.721	2.082 +- 0.672
177	L	0.881	2.321 +- 0.486
178	E	0.908	2.363 +- 0.433
209	W	0.988*	2.482 +- 0.164
212	Q	0.586	1.879 +- 0.739
241	S	0.629	1.944 +- 0.725
248	F	0.546	1.820 +- 0.747
252	R	0.910	2.365 +- 0.429
330	Q	0.900	2.351 +- 0.449
368	A	0.994**	2.491 +- 0.118
374	K	1.000**	2.499 +- 0.032
392	R	0.986*	2.480 +- 0.174
402	R	0.835	2.253 +- 0.557
406	H	0.925	2.387 +- 0.396
419	N	0.610	1.915 +- 0.732
420	T	0.635	1.953 +- 0.722
421	M	0.686	2.030 +- 0.696
422	Q	0.799	2.198 +- 0.601
426	K	0.753	2.130 +- 0.647
427	D	0.541	1.811 +- 0.748

G6PD

		Pr(w>1)	post mean +- SE for w
115	A	0.768	1.553 +- 0.656
132	S	0.998**	2.103 +- 0.915
516	-	0.958*	2.061 +- 0.959

OTC

	Pr(w>1)	post mean +- SE for w
18 H	0.579	1.191 +- 0.379
20 F	0.702	1.288 +- 0.340
21 M	0.907	1.444 +- 0.195
26 R	0.956*	1.476 +- 0.147
150 T	0.856	1.407 +- 0.241
230 T	0.584	1.195 +- 0.378
255 H	0.995**	1.501 +- 0.077
348 Q	0.631	1.237 +- 0.358

MECP2

	Pr(w>1)	post mean +- SE for w
373 S	0.972*	1.482 +- 0.136
380 V	0.856	1.391 +- 0.273
390 P	0.723	1.269 +- 0.384
400 T	0.571	1.145 +- 0.421