UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Contribution to the knowledge of Iron Deficiency and Iron Overload through the study of *TMPRSS6* and *SLC40A1* genes

Vera Clotilde Pico Pessoa

Mestrado em Biologia Humana e Ambiente

Dissertação orientada por:
Doutora Paula Faustino, Instituto Nacional de Saúde Dr. Ricardo Jorge
Prof. Doutora Deodália Dias, Faculdade de Ciências da Universidade de Lisboa

2022

Agradecimentos

Gostaria de agradecer ao Dr. Fernando de Almeida, presidente do Conselho Diretivo do Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA) e à Dra. Glória Isidro, coordenadora do Departamento de Genética Humana (DGH) que permitiram a execução do trabalho laboratorial referente a esta dissertação.

À minha orientadora a Doutora Paula Faustino pela oportunidade de executar este trabalho no seu Grupo de Investigação em Hemoglobinopatias, Metabolismo do Ferro e Patologias Associadas da Unidade de I&D do DGH, e por toda a transmissão de conhecimentos. Agradeço também ao Pedro Lopes por todo o apoio e auxílio prestado no decorrer desta tese, bem como aos colegas da UTI e UMO do INSA, pois sem o seu contributo esta tese não seria possível. Às minhas colegas do mesmo laboratório, à Alexandra Oliveira, Beatriz Leitão, Daniela Santos e Isabel Germano, pelo vosso auxílio e apoio no decorrer deste trabalho.

Este trabalho beneficiou de amostras anteriormente obtidas no âmbito do Inquérito Nacional com Exame Físico (INSEF)*. Assim, agradeço ao Doutor Carlos Matias Dias, coordenador do Departamento de Epidemiologia (DEP) do INSA, ao Doutor Baltazar Nunes, coordenador da Unidade de Investigação do DEP e à Doutora Marta Barreto, coordenadora do biobanco do INSEF*, pela disponibilização das amostras angariadas no âmbito desse projeto, e à Doutora Irina Kislaya, pela cedência da base de dados contendo informação demográfica e de saúde dos indivíduos estudados.

Agradeço também a oportunidade possibilitada pela Comissão Cientifica da Sociedade Portuguesa de Genética Humana (SPGH), aquando da 25ª Reunião Anual, de apresentar um *Poster* relativo a resultados obtidos no decorrer deste trabalho.

Gostaria de agradecer também a três professores do mestrado de Biologia Humana e Ambiente, a Professora Deodália Dias, o Professor Francisco Martins, e a Professora Teresa Rebelo, por todo o vosso apoio pedagógico e psicológico prestado, mesmo quando essa não era a vossa função.

Estou também grata aos meus amigos, colegas, família, e *crafty fam*, obrigada por todo o vosso apoio, por acreditarem em mim apesar de tudo.

**O INSEF, desenvolvido no âmbito do Projeto Pré-definido do Programa Iniciativas em Saúde Pública, foi promovido pelo Instituto Nacional de Saúde Doutor Ricardo Jorge através do Departamento de Epidemiologia e beneficiou de apoio financeiro concedido pela Islândia, Liechtenstein e Noruega, através das EEA Grants.*

Sumário

O metabolismo do ferro centra-se no controlo rigoroso da absorção e manutenção do ferro. É através da disrupção da sua homeostasia que surgem doenças associadas à carência ou ao excesso de ferro. No caso da carência de ferro, pode desenvolver-se anemia microcítica e hipocrómica, enquanto uma absorção excessiva de ferro pode conduzir à hemocromatose, com consequente deposição de ferro em alguns tecidos e perda de função de certos órgãos. Tanto na anemia ferropénica como na hemocromatose, pode haver contribuição genética para o seu desenvolvimento.

Neste trabalho pretendeu-se compreender o papel dos genes *TMPRSS6* e *SLC40A1* para o desenvolvimento de patologias associadas ao défice de ferro (Anemia microcítica e hipocrómica) e excesso de ferro (Hemocromatose Hereditária do tipo IV ou Doença da Ferroportina). Pretendeu-se também estudar a patogenicidade das variantes encontradas nestes genes, estabelecer relações genótipo-fenótipo, e compreender a contribuição genética para o desenvolvimento destas doenças.

Para o défice de ferro, estudámos 100 participantes do Inquérito Nacional de Saúde com Exame Físico (INSEF), estudo realizado anteriormente no INSA, cujos hemogramas revelavam microcitose e/ou hipocromia. Nos DNAs correspondentes, foi realizada uma pesquisa de variantes no gene *TMPRSS* por *Next-generation sequencing* (NGS) e confirmação por sequenciação Sanger.

No caso do excesso de ferro, analisámos do ponto de vista estatístico e bioinformático os resultados de NGS do gene *SLC40A1*, previamente obtidos no nosso laboratório, referentes ao estudo de 110 casos clínicos cujos fenótipos sugeriam a presença de Hemocromatose Hereditária. Realizámos estudos *in silico* para averiguar a patogenicidade das variantes encontradas no referido gene responsáveis pela Doença da Ferroportina ou da Hemocromatose Hereditária de tipo IV.

Na população com microcitose e/ou hipocromia foram identificadas 36 variantes no gene *TMPRSS6*. Entre elas destacam-se três nunca descritas: duas *missense* que vieram a demonstrar-se patogénicas, c.1580T>G (p.Phe527Cys) e c.1585T>C (p.Cys529Arg), e uma benigna, de localização intrónica (c.836+27G>C) sem influência no *splicing*. Quanto aos estudos de associação genótipo-fenótipo, foram encontradas diferenças significativas para a dispersão do volume eritrocitário na presença da Lys244Glu (p=0.028), e para a concentração média de hemoglobina corpuscular na presença da Pro24= (p=0.009). Nesta população, as mutações patogénicas foram encontradas sobre-representadas, assim como os polimorfismos descritos como de risco para o desenvolvimento da patologia, quando comparado com o descrito na literatura para a população em geral.

Na população com excesso de ferro foram estudadas sete alterações no gene *SLC40A1*. Salientando-se três variantes patogénicas raras associadas à Doença da Ferroportina (p.Gly80Ser, p.Val162del, e p.Gly204Ser). Apenas a variante intrónica *SLC40A1*:c.44-21A>C apresentou associações significativas entre os níveis de Ferro Sérico e Saturação de Transferrina (p<0.001), pois os indivíduos com esta variante apresentavam valores inferiores nestes biomarcadores, indicando um efeito protetor.

Em ambas as populações foi possível verificar que os fenótipos mais graves se encontravam associados a variantes patogénicas dos genes *TMPRSS6* e *SLC40A1*. Apesar destas variantes poderem afetar a expressão do gene e poderem causar alterações nas suas respetivas proteínas, dada a complexidade do metabolismo do ferro, é preciso sempre ter uma perspetiva integrada para compreender as implicações que cada uma pode causar.

Palavras-chave: Metabolismo do ferro, Anemia ferropénica, Doença da Ferroportina, Hemocromatose Hereditária, *TMPRSS6*, *SLC40A1*.

Abstract

Iron metabolism is focused on rigorous control of iron absorption and maintenance. It is through disruptions in its homeostasis that diseases associated to iron deficit or overload arise. Iron deficiency can cause microcytic and hypochromic anaemia, while excessive iron absorption can lead to haemochromatosis, with consequent iron deposition in some tissues and loss-of-function in some organs. Both in iron deficient anaemia as in haemochromatosis, there can be a genetic component contributing to their development.

In this work we intended to understand the role of *TMPRSS6* and *SLC40A1* genes for the development of pathologies associated to iron deficiency (microcytic hypochromic anaemia) and iron overload (Hereditary Haemochromatosis type IV or Ferroportin Disease). We also intended to study the pathogenicity of the variants found in these genes, establish genotype-phenotype associations, and understand the genetic contribution for the development of these pathologies.

For the Iron Deficiency study, we analysed 100 participants from the National Survey of Health with Physical Exam (INSEF), a study previously conducted in INSA, whose complete blood count revealed microcytosis and/or hypochromic. In their DNAs, a screening for alterations in the *TMPRSS6* gene was conducted through Next-generation sequencing (NGS) and confirmation through Sanger sequencing.

For the Iron Overload study, we analysed from a statistical and bioinformatic perspective previously obtained NGS results for *SLC40A1*, concerning 110 clinical cases whose phenotype suggested haemochromatosis. We performed *in silico* studies to assess the pathogenicity of the found variants, and if they were causative of Ferroportin Disease or Hereditary Haemochromatosis of type IV.

In the population with microcytosis and/or hypocromia, 36 variants were identified in the *TMPRSS6* gene. From those, three haven't been previously described: two missense that were concluded to be pathogenic c.1580T>G (p.Phe527Cys) and c.1585T>C (p.Cys529Arg), and another benign, intronic (c.836+27G>C) with no impact on splicing. As for the genotype-phenotype association studies, significant differences were found for Red Cell Distribution Width for Lys244Glu (p=0.028), and for Mean Corpuscular Haemoglobin Concentration for Pro24= (p=0.009). In this population, pathogenic alterations were over-represented, as well as reported polymorphisms of risk for anaemia development, when compared to the general population reports.

In the iron overload population, seven alterations in the *SLC40A1* gene were studied. Mainly the three rare pathogenic variants associated to Ferroportin Disease (p.Gly80Ser, p.Val162del, and p.Gly204Ser). Only the intronic variant *SLC40A1*:c.44-21A>C had significant associations to the levels of Serum Iron and Transferrin Saturation ($p<0.001$), as the individuals with this variant presented lower values for these biomarkers, indicating a protective effect.

In both populations, it was possible to assess that the most severe phenotypes were associated to pathogenic variants of *TMPRSS6* and *SLC40A1* genes. Although these alterations might affect gene expression and may cause changes to their respective proteins, given the complexity of the iron metabolism, it is always necessary to have an integrated view to understand the implications that each can cause.

Keywords: Iron Metabolism, Iron-deficiency Anaemia, Ferroportin Disease, Hereditary Haemochromatosis, *TMPRSS6*, *SLC40A1*.

Table of Contents

List of Figures

List of Tables

Abbreviations list

A	Adenine
AA	Amino Acid
AD	Allelic Depth
Ala	Alanine
Alt	Alternative
AR	Autossomal Recessive
Arg	Arginine
Asn	Aspargine
Asp	Aspartate
AuB	Allelic Unbalanced
bp	Base pairs
BMP6	Bone morphogenetic protein 6
BMPR	Bone morphogenetic protein receptor
BrEt	Bromide Ethidium
C	Cytosine
CADD	Combined Annotation-Dependent Depletion
CBC	Complete Blood Counts
CP	Ceruloplasmin
CUB	C1r/C1s, urchin embryonic growth factor and BMP1domain
Cys	Cysteine
DCYTB	Duodenal Cytochrome b
del	Deletion
DGS	*Direção Geral de Saúde*
DMSO	Dimethyl sulfoxide
DMT1	Divalent Metal Transporter 1
DNA	Deoxynucleic Acid
dNTPs	Nucleoside triphosphate
DP	Depth
EDTA	Ethylenediamine tetraacetic acid
FD	Ferroportin Disease
FPN	Ferroportin
fL	femtolitre = 10^{-15} L = μm^3 = cubic micrometer
F_{st}	Fixation index
FT	Ferritin
FTH	Ferritin Heavy chain
FTL	Ferritin Light chain
G	Guanine
GI	Gastrointestinal
Gly	Glycine
GQ	Genotype quality
GT	Genotype
GWAS	Genome-wide association studies
Hb	Haemoglobin
HCP1	Haem carrier protein 1
HEPH	Hephaestin

HFE	gene High Fe
HGVS	Human Genome Variation Society
HH	Hereditary Haemochromatosis
HIF	Hypoxia inducible factor
His	Hystidine
HJV	Haemojuvelin
HMOX	Haem oxygenase
HPX	Haem-haemopexin complex
HRG1	Haem responsive gene-1 protein
Ht	Haematocrit
IBD	inflammatory bowel disease
IBS	Iberian Spanish
ID	Iron deficiency
IDA	Iron-Deficiency Anaemia
INSA	*Instituto Nacional de Saúde Doutor Ricardo Jorge*
INSEF	*Inquérito Nacional de Saúde com Exame Físico*
IO	Iron Overload
IL	Interleukin
IRE	Iron Responsive Element
IRIDA	Iron Refractory Iron-Deficiency Anaemia
IRP	Iron Regulatory Protein
IUPAC	International Union of Pure and Applied Chemistry
IV	Intravenous
IVS	Intervening Sequence
K_s	Solubility constant
Kb	kilo base pairs
kDa	kilo Dalton
LD	Linkage Desiquilibrium
LDL	Low-density lipoprotein
Leu	Leucine
LSEC	Liver sinusoidal endothelial cells
LIC	Liver iron concentration
LIP	Labile iron pool
MAF	Minor Allele Frequency
Max	Maximum
MCV	Mean Corpuscular volume
MCH	Mean Corpuscular Haemoglobin
MCHC	Mean Corpuscular Haemoglobin Concentration
Med	Median
Min	Minimum
MRI	Magnetic Resonance Imaging
mRNA	Messenger RNA
MSA	Multiple Sequence Aligment
MT2	Matriptase-2
mV	milivolt
NBIA	Neurodegeneration with Brain Iron Accumulation
ng	Nanogram
NGS	Next Generation Sequencing

nm	nanometers
NSAIDs	Non-steroidal anti-inflammatory drugs
nt	Nucleotides
OMIM	Online Mendelian Inheritance in Man
p	p-value
PCR	Polymerase Chain Reaction
pg	picogram = 1 x10^{-12} gram
pH	potential of Hydrogen
Phe	Phenylalanine
PLAT	Platelet count
Pro	Proline
RBC	Red Blood Cells
Ref	Reference
RDW	Red blood cell Distribution Width
RNA	Ribonucleic Acid
ROS	Reactive Oxygen Species
RV	Reference Value
SD	Standard Deviation
SEA	Sperm protein, Enterokinase and Agrin domain
Ser	Serine
SLC40A1	Solute Carrier Family 40 Number 1- gene encoding for ferroportin
SMAD	Mothers Against Decapentaplegic
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SPGH	*Sociedade Portuguesa de Genética Humana*
T	Thymine
TBE	Tris-borate EDTA
TF	Transferrin
TMPRSS6	Transmembrane Serine Protease 6 – gene encoding for Matriptase-2
TS	Transferrin Saturation
TfR2	Transferrin Receptor 2
Tyr	Tyrosine
UICB	Unsaturated *iron*-binding capacity
UMO	*Unidade de Genética Molecular*
UTI	*Unidade de Tecnologia e Inovação*
UTR	Untranslated Region
UV	UltraViolet
V	Volt
Val	Valine
VCF	Variant Call Format
VEP	Variant Effect Predictor
VUS	Variable of Uncertain Significance
WBC	White Blood cells
WHO	World Health Organization
µg	microgram
0/0	Wild type
0/1	Heterozygous
1/1	Homozygous

1 Introduction

1.1. Iron, a physicochemical and biochemical perspective

Iron comprises about 5% of the Earth's crust[1], being one of the most widely available metals[2-4]. Due to its nature, it can have different oxidation stages (from -2 to +6 oxidation levels), with a redox potential ranging from 1000 to -550 mV[1,5]. These three features, widespread availability, great redox potential, and easy electron exchange, make it an ideal candidate for biological use, as they allow the flexibility needed in life systems[2,6]. As such, iron is an indispensable metal for a variety of biological functions, as it is used as a cofactor for different proteins and enzymes[2,7]. Iron is present in processes as diverse as: cellular respiration and electron transport (cytochromes), in metabolism (catalase and haem peroxidase), oxygen storing (myoglobin) and transportation (haemoglobin)[3,5-12].

However, these characteristics also imply its greater potential to create chaos if its levels are not held under control. The lack of iron equilibrium, homeostasis, in the cases of deficit or overload in the organism, can cause a variety of pathologies due its disruption[2,3,5,8-14].

1.1.1. Ferrous and Ferric iron

The most common iron species are ferrous divalent (Fe^{2+}) ions and ferric trivalent (Fe^{3+})[4]. These are the iron forms used by organisms. Their redox potential ranges between -500 to +600 mV depending on the conditions in which these reactions occur, exchanging electrons readily[4,7,15].

Ferric iron (Fe^{3+}) is quite abundant, but due to its low solubility (~10^{-18} M), it is much harder to absorb and use in aqueous conditions than Ferrous iron (Fe^{2+}) (solubility of ~10^{-1} M)[4,6,16]. Thus organisms created mechanisms or conditions in which they can increase iron solubility by reducing their oxidation state.

In aerobic aqueous conditions, free Fe^{2+} can reduce O_2, thus creating the superoxide radical $O_2^{\bullet -}$, a Reactive Oxygen Species (ROS) as represented in the following chemical equations (Fenton/Haber-Weiss Reactions)[2,4,7,15]. Other ROS also induced by iron include hydrogen peroxide (H_2O_2) and hydroxyl radical ($^{\bullet}OH$) represented in the equations, among others[2,7].

$$Fe^{2+} + O_2 \rightarrow O_2^{\bullet -}$$

$$2\ O_2^{\bullet -} + 2\ H^{+} \rightarrow H_2O_2 + O_2$$

$$Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + OH^{-} + {}^{\bullet}OH \quad \text{(Fenton)}$$

$$H_2O_2 + O_2^{\bullet -} \xrightarrow{Fe} OH^{-} + {}^{\bullet}OH + O_2 \quad \text{(Haber-Weiss)}$$

These reactions are extremely relevant, as ROS are responsible for damages of varying severity, thus the importance of tight control to avoid their presence, as they create a self-feeding cycle[6,7].

1.1.2. Iron coordinating structures

Iron has different applications in organisms, and depending on the purpose, it can be found within different structures types, like Oxo di-iron, Iron-sulphur clusters and Haem[11,16] (**Figure 1.1**). These coordinating structures increase stability, and diminish its potential nefarious effects on cells[16].

Haem proteins are found in oxygen rich environments, while iron-sulphur cluster proteins are present in anaerobic conditions, but all structures are of great importance as they are a part of a wide range of proteins[7,11,16]. Non-haemic iron proteins are present in processes as diverse as DNA and steroid

synthesis, drug metabolism, gene regulation, cellular proliferation and differentiation[1,3]. These compounds also include flavin-iron enzymes, transferrin and ferritin[1,3].

The haem structure, also known as iron-protoporphyrin IX (**Figure 1.1 C**), has particular significance by enabling oxygen metabolism[16]. It is present in several proteins as a co-factor, with functions as wide as those responsible for both oxygen transportation and storage (haemoglobin and myoglobin, respectively), electron transfer (cytochromes), metabolism of different substances (cytochromes, catalase, peroxidase, so forth), signalling, among others[1,3,4].

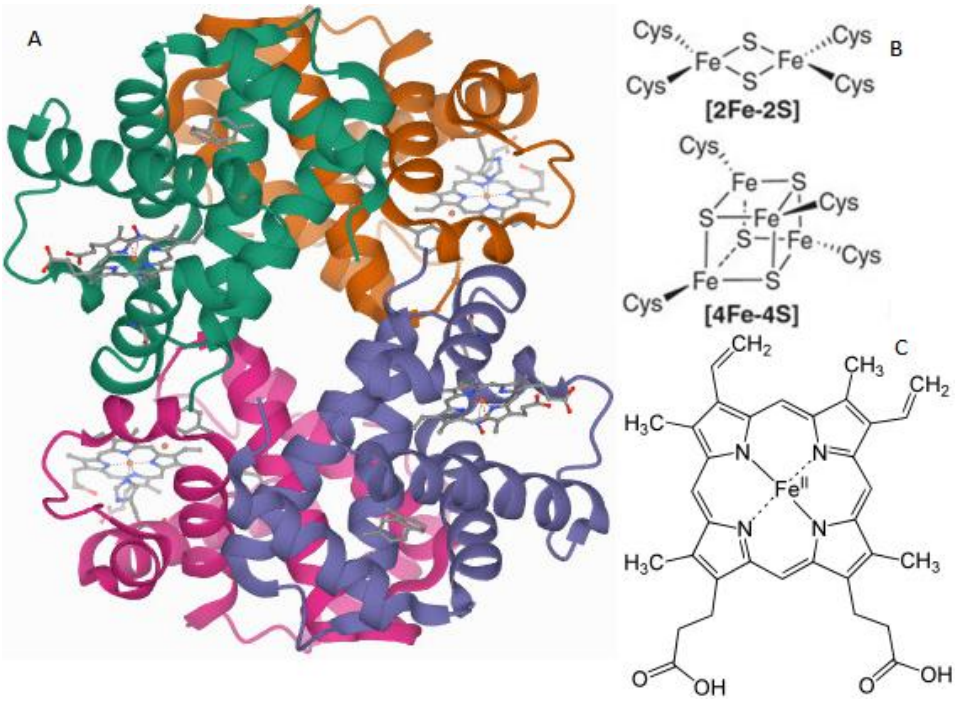

Figure 1.1: Structure of haemoglobin, and iron stabilising structures. A. In this representation, the two β chains of this tetrameric protein are identified by the bonding of a nitrate ligand to them. To each chain one haem is attached, within its core it is possible to visualize the iron centre (the rust coloured ball), to which oxygen molecules bind to in order to be transported throughout the body. Model from PDB via X-ray crystallography[17]. **B.** Here are portrayed two different types of Iron-sulphur clusters[16]. **C.** Haem in detail. Images from Kaplan, and Yi[6,17].

The most well-known haemoprotein is haemoglobin (Hb), responsible for oxygen and carbon dioxide capture, it transports oxygen throughout the body, replacing it by carbon dioxide (cellular waste) and releasing it in the lung's alveoli (being replaced by new oxygen molecules)[6,11].

Hb is the pigmented protein that gives erythrocytes, or Red Blood Cells (RBC), its characteristic colour (which depends on the amount of oxygen bond to it)[11,18]. This protein is a tetrameter, comprised by four globin chains each associated to an iron-protoporphyrin IX centre; Hb A has two different types of chains, two alpha and two beta (**Figure 1.1 A**)[18].

1.2. Iron metabolism

Humans use iron similarly to other organisms, as the iron metabolism is quite conserved among species. It has been established that a male adult has on average 50-60 mg of iron per kg, so for a 70kg body, the equivalent iron presence would be of 3500-4000mg (**Figure 1.2**)[1,6,9].

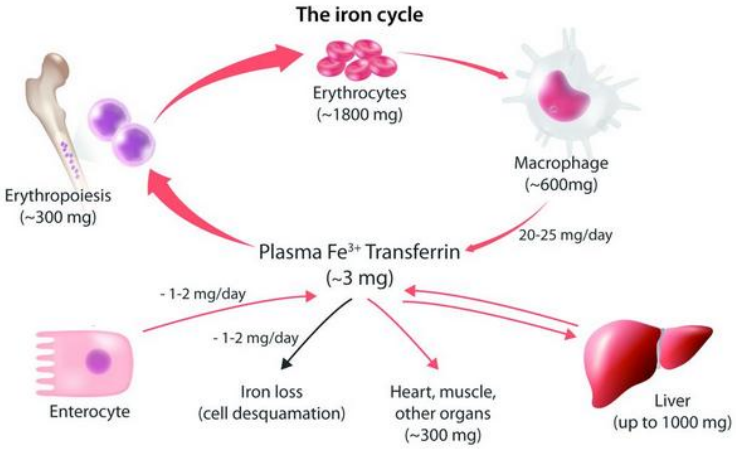

Figure 1.2: Iron distribution throughout the organism. Nearly two thirds (65%) of iron are within haemoglobin in erythrocytes, and about 10% are in the muscle's myoglobin and enzymes of different tissues. The remaining 25% are located within macrophages (mostly present in the bone marrow, spleen and liver), bone marrow, and stored in the liver within ferritin. Image adaptation from Camaschella[8].

In RBC we find the greatest iron concentration, as about 70% of our total iron is bound to haem groups. Tissues using most iron include the muscles and bone marrow (where erythropoiesis, process of RBC production, mainly occurs)[5,12]. Large amounts of iron are stored within hepatocytes, these cells have a high importance in the iron metabolism, as they synthesize transferrin, an iron transporter, as well as hepcidin, an important 25 amino acids (AA) long peptide[2,6].

Although not depicted in Figure 1.2, iron also has an important role in central nervous system. Iron is used in the synthesis of myelin and neurotransmitters, thus its deficiency hinders brain development and functioning, causing neurological deficits[2,7,15].

On the other hand, excessive iron in the brain can contribute to the development of neurodegenerative conditions such as Alzheimer's disease, Parkinson, including less common ones like Neurodegeneration with brain iron accumulation (NBIA), among others[7,15,19]. Due to high energetic requirements and a weak antioxidant system, the brain and neurons are particularly susceptible to damages caused by oxidative stress, thus the presence of high levels of iron exacerbate its impact on these tissues further contributing to damages [15].

Furthermore, iron tends to accumulate with aging, and depending on the areas affected, these deposits might cause more serious outcomes[15,19]. This accumulation can lead to a form of iron-mediated cell death, ferroptosis, due to increased ROS and phospholipidic peroxidation promotion, leading to tissue damage in the long run[15,19].

These processes encompass Iron Metabolism, which is comprised by a series of mechanisms that all contribute for iron homeostasis (**Figure 1.3**). To regulate the delicate iron metabolism, there are two types mechanisms involved, the Systemic, and the Cellular Regulation[2,5,8-10]. Hepcidin is a major player in the regulation of this intricate system, being regarded as the iron metabolism hormone, whose expression is adapted to the organism's iron needs[6,11,20].

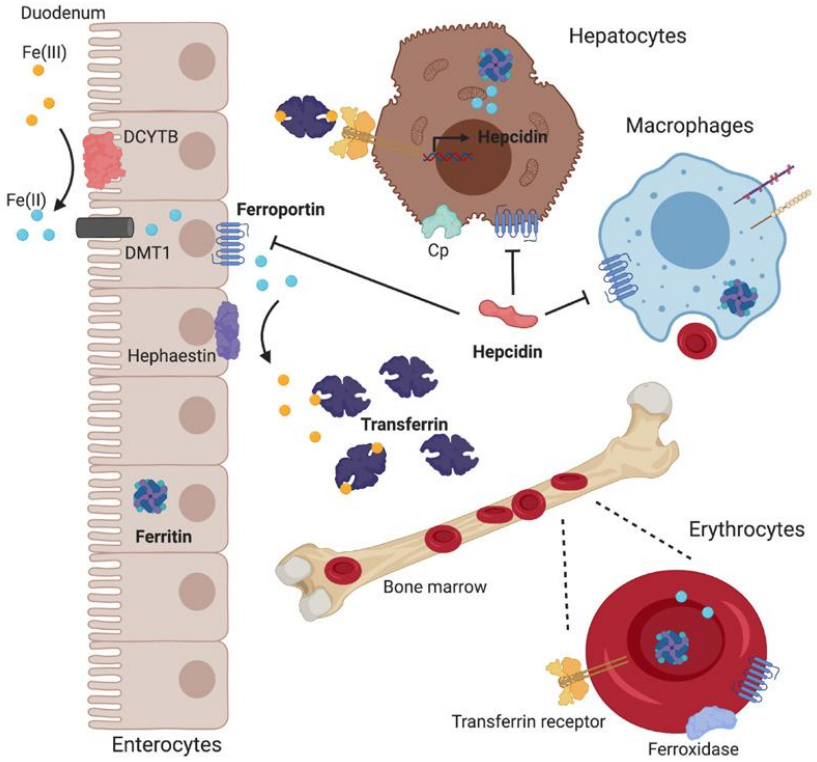

Figure 1.3: Iron Metabolism. Iron absorption occurs in duodenal epithelial cells. Fe^{+3} (orange balls) is reduced to Fe^{+2}(cyan balls) by duodenal cytochrome b (Dcytb). They enter through the Divalent Metal-ion Transporter-1 (DMT1), once inside it can be stored in ferritin, or exported by ferroportin and subsequently oxidized hephaestin (oxidation is also executed by ceruloplasmin (Cp) or ferroxidase). Transferrin transports 2 Fe^{+3} and releases them in its receptors expressed in the surface of cells. RBC are produced in the bone marrow, and in the end of their life cycle, they are phagocytosed by macrophages for recycling, releasing iron back into circulation. All these processes are regulated by hepcidin. Produced in hepatocytes, it regulates the iron metabolism by binding to ferroportin and blocking iron export. Its synthesis is responsive to iron levels (transferrin binding to its receptors, presence of ferritin or labile iron pool (free Fe^{+2} within cells). Image from Roemhild[6]

Hepcidin is produced in the liver and circulates around in the serum; it manages the iron metabolism systemically by inhibiting ferroportin (FPN), the sole cellular iron exporter, causing iron export suppression when it binds to it[20,21]. Hepcidin synthesis can be modulated in order to regulate iron homeostasis[1,6,13]. Hepcidin levels increase when there is iron excess in the organism, cytokines, infections or inflammation. On the contrary, hepcidin levels diminish in cases of iron deficiency, hypoxia, anaemia, erythropoiesis, among other factors[2,5,8,11,20].

Iron homeostasis is mostly controlled at the level of iron absorption, thus an active mechanism for iron elimination wasn't developed[9,11]. However, iron loss can occur, due to blood loss (menstruation or wounds) or epithelial desquamation of the intestinal lumen in which some iron is stored[5,8,12].

In the case of iron metabolism disruptions, iron homeostasis is unbalanced, and this succeeds for both extremes, Iron Deficit (ID) and Iron Overload (IO).

Cases of ID might occur when the iron requirements are too high, for the amount of iron available. On the opposite side, an excessive presence of iron in the body might cause its accumulation in organs, such as the heart and liver, causing diseases; it can also damage proteins, lipids, and DNA through the creation of free radicals (ROS)[7,11,15,19].

Unlike IO, the majority of ID cases are due to malnutrition and alimentary shortages; nevertheless, there is biological and genetic propensity for its development[10,22-26].

1.2.1. Absorption

Iron enters the organism via the gastrointestinal (GI) system, by ingestion, and through its course its absorption occurs[3]. The chime, which contains within it iron in multiple forms, passes through the stomach, and in there it is exposed to low pH levels due to the action of the stomach's acid pumping gastric glands[1,11]. After passing through the stomach, starts the absorption process. Most of it takes place in the villi of the duodenal epithelium, located in the beginning portion of the small intestine[9,12].

The iron can be either of haem or non-haem origin, and its route through the enterocyte's transporters depends of its form[1,2,11]. Haem is the most readily absorbed iron form, followed by ferrous divalent (Fe^{2+}) and then ferric trivalent (Fe^{3+})[11,3].

In the western diet the absorbed iron is about 10% haem and 90% non-haem (mostly in the $Fe(OH)_3$ complex form), on average, which represents a daily intake of about 15-20 mg of iron, although only around 1-2 mg are absorbed per day to maintain the iron homeostasis[2,27]. Depending on the original food source and how it is presented for absorption, iron might be more or less bioavailable[1,3]. Bioavailability is the extent to which an organism absorbs a substance[3,11]. Being dependant on the absorption capacity, the enterocytes absorb iron on different rates depending on the iron form[11].

The majority of ionic iron in food isn't found as Fe^{2+}, the easiest form to be absorbed, but as Fe^{3+}. To allow absorption, reduction from Fe^{3+} to Fe^{2+} has to occur, because Fe^{3+} is very insoluble[12,27]. Fe^{3+} is reduced by ascorbate ferrireductase (Dcytb), a transmembrane cytochrome b enzyme located in the duodenal villi that uses electrons and ascorbic acid (vitamin C) as a substrate to catalyze this reaction[12,27].

Fe^{3+} reduction can also be executed through low pH value exposition[1,11]. It might be due to the lower pH values found in the duodenum, as intestinal pH neutralisation by bile occurs after the duodenum, that iron solubility increases and, consequently, availability for absorption is increased[1,11].

The transporter which absorbs ionic iron into the enterocyte is Divalent Metallic Transporter (DMT1), it absorbs Fe^{2+} working coupled with protons, it can also transport other divalent metals such as, zinc and copper, as it isn't iron specific[1-3,11]. It has been found, in animal studies, that metals such as manganese, zinc or lead, compete with iron, which means, that they are absorbed in detriment of iron[1-3,11]. The representation of these processes can be seen below in detail in Figure 1.4.

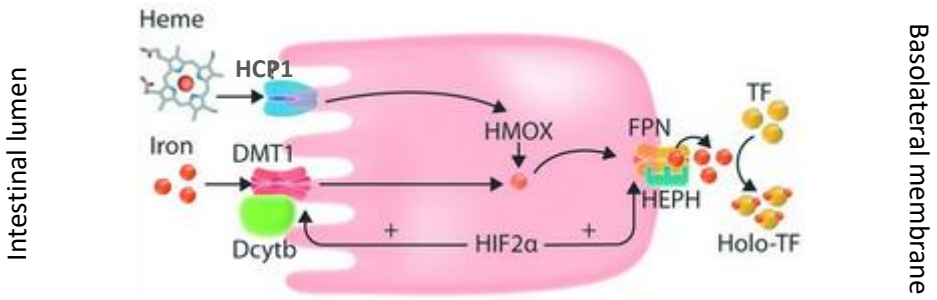

Figure 1.4: Detailed Enterocyte. Haem bound iron is transported into the enterocyte by Haem Carrier Protein 1 transporter (HCP1). Haem oxygenase (HMOX) releases iron from haem. Ionic Fe^{3+} has to be reduced to Fe^{2+} by ascorbate ferrireductase Duodenal cytochrome B (Dcytb), also known as Plasma membrane ascorbate-dependent reductase, before entering. It uses ascorbate as a substract for electron donation for iron reduction. After, it is transported by Divalent Metallic Transporter (DMT1) into the enterocyte's Labile Iron Pool (LIP), from which it can be exported by ferroportin (FPN), located in the basolateral membrane. Once more the iron has to change its oxidation level. This is done by Hephaestin (HEPH), afterwards Fe^{3+} is apt to be transported to other areas by transferring (TF). Hypoxia-inducible factor 2-alpha (HIF2α) stimulates the expression of both transmembranar transporters, DMT1 and FPN. Adaptation from Camaschella[8].

If iron is haem bound, its absorption is via Haem Carrier Protein 1 transporter (HCP1)[8]. Once inside the enterocyte, it can be degraded by Haem oxygenase (HMOX), thus releasing the iron from its core,

or be kept intact and exported into circulation[1,8,27]. There are advantages in absorbing haem; it is more efficient, absorption takes place regardless of the pH level, and since the iron is already bound to protoporphyrin IX there is no absorption inhibition by the formation of organometallic complexes, which are common in iron-rich plant sources[11].

Once within the enterocyte's cytoplasm, and depending on the organism's needs, the absorbed iron can be either stored within ferritin or exported by ferroportin (FPN) to blood plasma for its distribution in the circulatory system[5,6,8,12]. It can also gather to form the Labile Iron Pool (LIP)[1,7]. This free iron aggregate is quite reactive, but low molecular weight compounds are able to chelate the iron ions present in it[1,7]. Due to the ephemeral nature of epithelial cells, the iron stored inside them is lost when the epithelial enterocytes are desquamated and excreted along the GI track[1].

1.2.2. Storage and Export

As previously mentioned, after iron absorption, it can pool in the LIP of the enterocyte, but it cannot remain there for long, thus two options are possible, storage or export. The route chosen depends of systemic iron needs[11]. If there is a large iron demand, there is no necessity for its storage, thus it is exported via FPN present on the enterocyte's basolateral membrane.

Upon FPN release, Fe^{2+} is converted to Fe^{3+} by the ferroxidase Hephaestin (associated to FPN in enterocytes, as seen in **Figure 1.4**) or ceruloplasmin (CP), a plasma ferroxidase[8]. Solely this oxidation state allows transferrin (iron transporting protein with the capacity to bind two Fe^{3+} ions) binding, to transport it to every tissue in need of iron[2,5,6,8,9,27]. FPN controls the availability in which iron circulates throughout the organism as its exporter; its activity is firmly regulated by hepcidin[1,6,11]. FPN is also located in macrophages (**Figure 1.5**), and syncytiotrophoblasts (placental epithelial layer that communicates with the uterus for embryonic and foetal growth)[6,13].

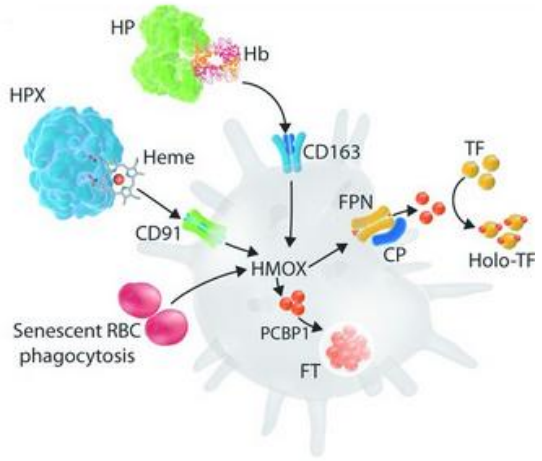

Figure 1.5: Detailed Macrophage. Iron entry has different routes. Within Hb, haptoglobin (HP) or haem-hemopexin complexes (HPX), transporters are used (CD163 (haemoglobin-haptoglobin receptor) and CD91 (LDL-related receptor)). If iron is within senescent RBC, phagocytosis occurs. Regardless of entry type, iron is haem bond, so it has to be processed by Haem oxygenase (HMOX) to free iron, after release it is transported by PBCP1 to be stored in ferritin (FT), or exported directly by ferroportin (FPN), being by oxidized by ceruloplasmin (CP) transferring (TF) uptake, once loaded with two iron ions it becomes Holo-TF. Image adapted from Camaschella[8].

The main function of transferrin is iron distribution of throughout the organism, releasing it in transferrin receptors (TfR1), present in the cell's surface, furthermore it is responsible for the maintenance of iron serum in an inert state[2,12]. Nearly all nucleated cells have transferrin receptors on their surface, having a higher affinity for transferrin bond to iron to full capacity[1].

After cellular uptake, the iron is relocated in order to suppress the cell's needs. It can be transported into the mitochondria, to synthesize haem or iron-sulphur clusters, required for protein synthesis[19]. If the cells acquire more iron than what they require, the extra iron has to be safeguarded, to avoid damages, preventing its toxic effects[6].

Ferritin secures iron within its structure; it is the best indicator of iron presence in the organism, as less serum ferritin is expressed when there are low amounts of iron[11]. Ferritin is comprised by 24 subunits of heavy and light chains, encoded by *FTH* and *FTL*, respectively[2,11]. These chains are expressed at different rates depending on cellular storage needs, forming a capsule in which up to 4500 iron atoms can be stored[2]. Another iron-storing structure is haemosiderin, however it isn't as flexible/dynamic as ferritin due to its reluctance to release iron when the organism is in need[11].

1.2.3. Recycling

Although there was a strong emphasis put in the section regarding iron absorption, most iron transported by transferrin isn't of absorbed origin, but collected from RBC, more precisely, iron that results from the clearance of senescent RBC by macrophages, and thus recycled (**Figure 1.6**)[2,11].

RBC are the largest iron reservoir in the body, as shown in Figure 1.2. When the iron carried by Transferrin is loaded into TfR1 located in the cells of the bone marrow, it is used for erythropoiesis, which means the creation of new RBC[11]. Alterations in erythropoiesis can simultaneously cause anaemia and IO[1]. A single RBC can contain 280 million Hb molecules, which accounts for over one billion iron atoms, as per haemoglobin each globin chain has a haem associated[2].

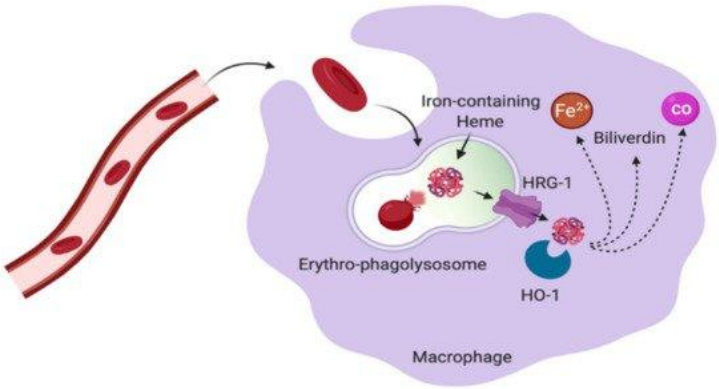

Figure 1.6: Senescent RBC clearance by spleen macrophages. Erythrophagocytosis is here depicted with further detail; the macrophage engulfs the senescent RBC into a lysosome in which it will be digested to release haem, which is then exported to the cytosol by the haem responsive gene-1 protein (HRG-1), where it is degraded by Haem oxygenase (HMOX), releasing iron, along with carbon monoxide and Biliverdin (product of haem degradation). Image from Vogt[2]

An RBC has a lifetime of about 120 days, as they age they are less apt to perform their functions and are thus targeted by spleen macrophages to phagocyte them and recycle their components, iron and haem, that can be reused to form new RBC (**Figure 1.6**)[1,2,27].

After digestion of senescence RBC, macrophages release Fe^{2+} through FPN[5], and then CP converts it to Fe^{3+} so it can bind to TF[5,8,9,12,27]. Despite the majority of senescent clearance being performed by spleen resident macrophages, those present in the liver, also known as Kupffer cells, are also able to perform this task[2].

1.2.4. Regulation

1.2.4.1. Cellular regulation

Cellular regulation is very complex, and a lot is still to be discovered[5]. This regulation occurs at an intracellular level, modulating protein expression accordingly to its individual needs[28]. Cellular iron regulators can be involved in different steps of the intracellular iron metabolism, from cellular entry or exit of iron or haem groups, to its circulation within the cell, to storage or cellular iron balance[2,5].

Cellular regulation can be divided into two categories, pre-transcriptional or post-transcriptional[28,29]. Post-transcriptional regulation includes processes such as polyadenylation, alternative splicing, microRNAs, protein breakdown, and the Iron regulatory protein (IRP)/Iron responsive element (IRE) system[28]. The IRP/IRE system is a mechanism developed by cells for the maintenance of intracellular iron, through the control of its import, export, and storage (**Figure 1.7**)[2,5,28].

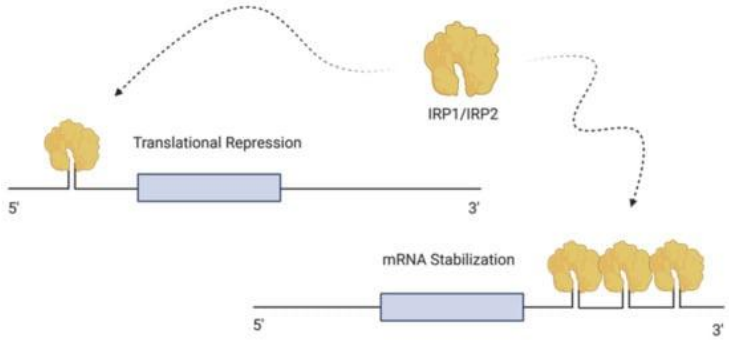

Figure 1.7: Scheme of IRP/IRE system for cellular regulation. Certain genes have an IRE (the hairpin structure depicted), that can be present in both ends. If the IRP binds to the 5'UTR IRE, the translation of that mRNA is suppressed thus that protein won't be produced; this is represented in the upper strand. If the IRP binds to the 3'UTR IRE, that mRNA translation is increased, it also stabilises the transcript, as shown in the lower strand depicted. Image from Vogt[2].

IRP are RNA-binding cytoplasmatic proteins that upon interacting with iron metabolism proteins mRNA, in their IRE, regulating them thus maintaining cellular iron balance[2,5,19]. They are mostly comprised by IRP1 (90kDa) and IRP2 (105kDa)[2]. Depending on the iron concentration they are exposed to, they change their conformation, allowing them to bind or not to IRE, as they have highest IRE affinity when the cells are iron depleted[2]. IRP1 possesses iron-sulphur [4Fe-4S] clusters (structures represented in **Figure 1.1 B**), that when exposed to high iron levels causes structural rearrangement that prevents it from binding into an IRE[2].

IRE are extremely conserved hairpin structures of 25-30 nucleotides located in either extremity of the untranslated regions (UTR) in the mRNA of genes that codify for proteins of the iron metabolism[2]. Genes possessing IREs include *DMT1* and *TRF1/2* (import), *SLC40A1* (FNP, export), and *FTH* and *FTL* (ferritin, storage), it also impact the genes for Haem Carrier protein 1 (HCP1) and Haem oxygenase 1 and 2 (HMOX-1/2)[2,5,19,21]. Depending on which end the IRE is located, IRP binding has different outcomes. In 5'UTR it inhibits translation, while on 3'UTR end it prevents endonucleolytic cleavage and degradation (**Figure 1.7**)[2].

If the cell has iron surplus, IRPs are unable to bind to IRE, so 5'UTR IRE proteins are expressed continuously, while those with 3'UTR IRE are degraded, as they are no longer protected by IRP binding[2]. IRP binds to FPN or ferritin's mRNA 5' UTR IRE, to control iron export or storage, respectively[5,19]. In the cases of intracellular iron deficit, IRP bind to the ferroportin's IRE, halting its synthesis, which means that further iron export will not occur, thus retaining as much iron as possible until it reaches an ideal balance (**Figure 1.7**)[2,19,28].

1.2.4.2. Systemic Regulation

The systemic regulation of iron metabolism mainly revolves around iron absorption, recycling and its primary regulatory agent is hepcidin, whose role is fundamental in the iron homeostasis[2,8,13,11,20]. Other regulators involved include FNP, DMT1, Bone marrow Protein 6 (BMP6), Hephaestin, CP, transferrin receptor 2 (TfR2), Homeostatic iron regulator protein (High Iron, HFE), haemojuvelin (HJV)[5,6]. The last three mentioned, along with TfR1, are considered iron sensors[28].

As previously mentioned, hepcidin is a hepatic peptide hormone expressed by the *HAMP* gene (hepcidin antimicrobial peptide), whose synthesis is modulated mainly through the action of HFE, TfR2 and HJV in accordance to the iron status in the organism (**Figure 1.8**)[2,5,14].

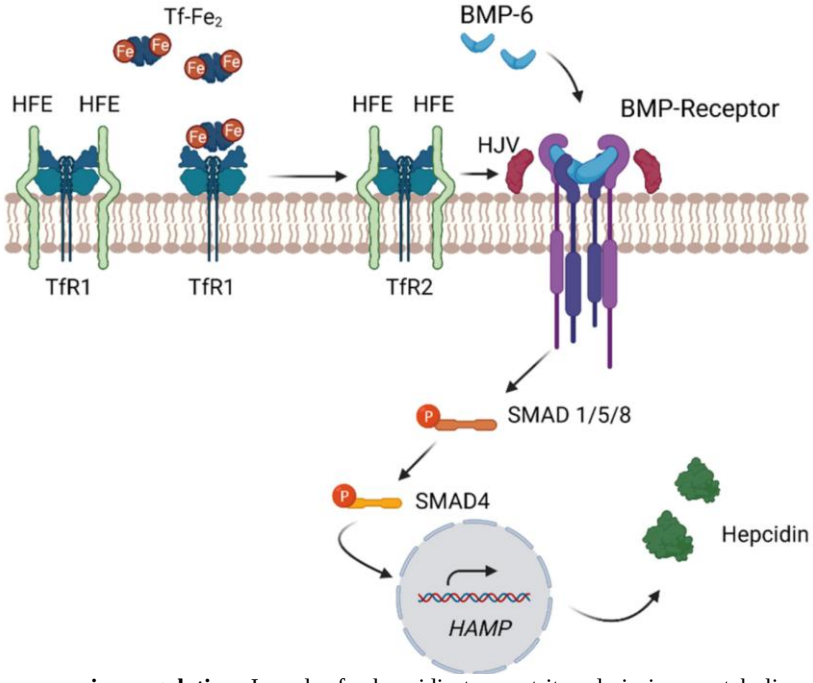

Figure 1.8: Hepcidin expression regulation. In order for hepcidin to exert its role in iron metabolism regulation, its gene expression has to be stimulated. At the surface of hepatocytes, are located several transmembranar proteins (HFE, TfR2, HJV) that regulate hepcidin gene expression through a SMADs signalling pathway. Image from Vogt[2].

The Hepcidin gene, *HAMP*, expression can be modulated. Hepcidin synthesis increases in cases of excess iron in the organism, cytokines, infections or inflammation, and is diminished in cases of iron deficiency, hypoxia, anaemia, erythropoiesis among other factors[5,11,20,28]. It is also possible to inhibit its activity, through matriptase-2 (MT2), which cleaves HJV[11,20]. Other proteins, and respective genes, that impact and impair normal hepcidin functioning, include HFE, TFR2, HJV, and even alteration to *HAMP* itself[5,10-14,21].

The hepcidin-ferroportin relation is interconnected with other mechanisms of iron regulation[13,20,29]. Hepcidin causes alterations in the structural conformation of FPN (**Figure 1.9**)[2,9,14,21], thus repressing its activity in iron export, and controlling the iron metabolism through ferroportin inhibition[5,8,9,10,20].

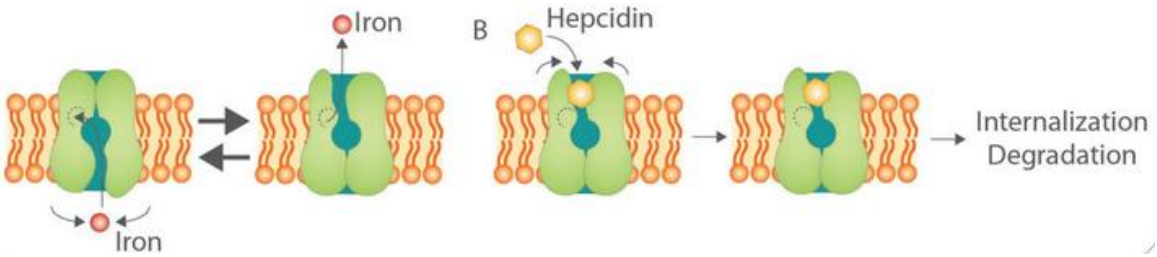

Figure 1.9: Hepcidin regulation of ferroportin under normal conditions. The first 2 images represent FPN iron export. In **B,** after hepcidin binds to FPN, it blocks iron export, also signaling for its destruction. Adaptation from Pietrangelo[13].

Disruptions in hepcidin production are associated with several conditions; its over-expression is associated with anaemia of chronic disease, for example, while its impaired expression is associated with quite a few congenital disorders[2,11,20].

High hepcidin levels induce iron retention in macrophages, high serum transferrin levels, and it can also cause erythropoiesis restriction due to the lack of available iron[5,9,28]. This is the action mechanism in anaemia of inflammation, as iron export is restricted to diminish the growth of pathogenic agents (as they require iron to grow), that cause infections and subsequent inflammation[8,10,13,20].

Through the macrophage's immunological activity and cytokine release, stimulate hepcidin production consequently decreasing iron absorption and storage (**Figure 1.10 A**)[2,3,5,20,22]. Increased expression also occurs with hepcidin producing adenomas (a benign type of glandular cancer) [20].

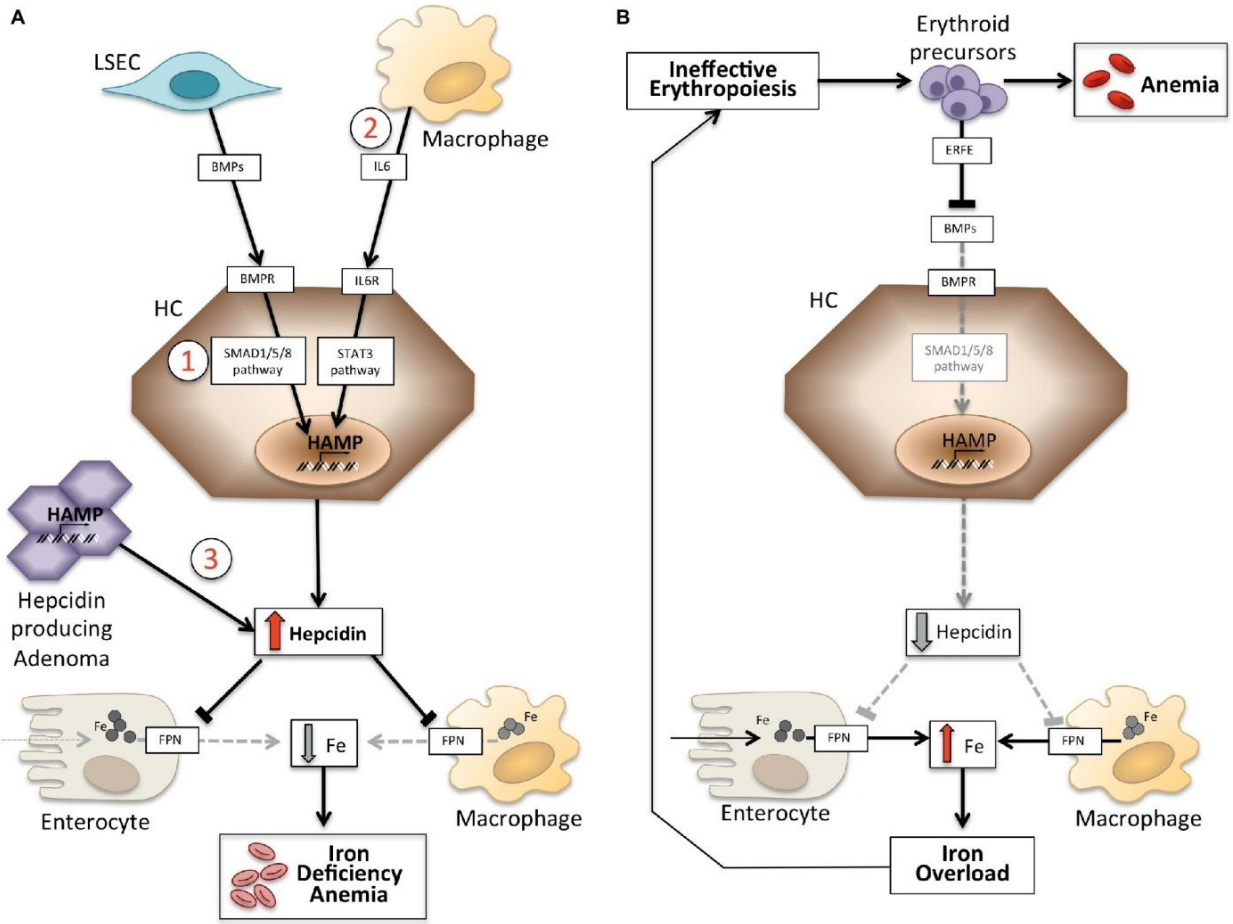

Figure 1.10: Pathophysiological mechanisms caused by different levels of hepcidin. A) Increased hepcidin levels, in **A.1** hepcidin is produced in hepatocytes stimulated by liver sinusoidal endothelial cells (LSEC), through the SMADs pathway; **A.2** its production is increased by macrophage cytokines (IL6) release; in **A.3** a cancer, adenoma, that produces hepcidin. This high level of hepcidin inhibits iron export by FPN, decreasing iron in circulation, conducing to IDA development. **B)** Diminished hepcidin expression due to ineffective erythropoiesis and ERFE production by RBC precursors, which inhibit *HAMP*, leading to increased iron levels in circulation, as there isn't sufficient hepcidin to regulate FNP activity, this leads to iron overload which further disturbs erythropoiesis leading to anaemia. Image from Pagani[20]

Several diseases of the iron metabolism are originated by the disruption of hepcidin functioning[8,13,20,21]. When its levels decrease significantly, it leads to IO, which can be due to alterations in the genes involved in its regulation such as *HFE*, *TFR2*, *HJV* [20]. On the other hand, alterations in other *HAMP* regulators, such as matriptase-2, may originate an abnormal high level of hepcidin expression, giving rise to ID.

1.3. Iron homeostasis disruptions due to Iron Deficiency

1.3.1. Iron deficiency due to nutritional causes

Iron deficiency (ID), also known as sideropenia, is a frequent malnutrition worldwide[8,20,22-25], affecting 25% of the global population[6]. The World Health Organization (WHO) has established risk factors for its development, increased iron demands, insufficient supply, and increased blood losses[10]. In the beginning of a mild ID, only a reduction in iron metabolism serum biomarkers is observed, such as reduced iron storage in ferritin[30]. However, once completely established, anaemia (low Hb levels) is also observed, giving rise to Iron-Deficiency Anaemia (IDA)[11].

Anaemia *per se* means that the amount of RBC or Hb is low[3, 6,10]. It can be caused by various things, from long term inflammatory conditions, to kidney or bone marrow dysfunctions (stimulates RBC production and is responsible for erythropoiesis, respectively), bleedings, poor alimentation, pregnancy, haemoglobinopathies, RBC haemolysis, obesity and alcoholism[3,6,10].

ID occurs most frequently in pre-menopausal women (with the highest iron requirements, ~2.38 mg per day on average[11]), children, individuals with blood loss or whose blood is frequently drawn, or as a consequence of a poor diet[6,8,10]. Between 2 to 5% of male adults and post-menopausal women in developed countries suffer from IDA[30]. IDA in pregnancy is connected to several issues, even leading to maternal or foetal death, and only 25% of anaemia cases in pregnancy are not due to ID[11,31,32].

Both ID and IDA are multilayered conditions associated to many causes (**Figure 1.11**)[3]. They are dependant of social and financial status, being more prevalent among those in precarious living conditions[26]. In countries like Portugal, usually it's due to low intake of iron or lower bioavailability (malabsorption)[3,26].

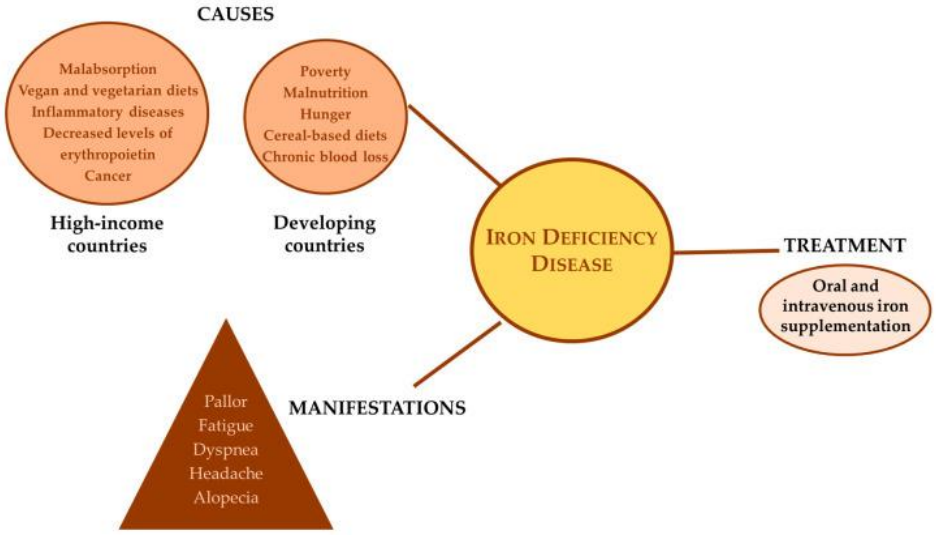

Figure 1.11: Nutritional IDA Summary. The causes behind it might differ, but all factors that contribute for IDA development display the same pathological manifestations and are similarly addressed for treatment. Image from Liberal[3].

Iron malabsorption can be due to alterations in gastrointestinal pH level and mucosa leading to absorption disturbances[30,32,33]. A higher pH level diminishes the optimal conditions for duodenal absorption, this happens in cases of reduced gastric acid production whether from bodily malfunction (inflammation of the stomach lining, partial or total stomach removal) or due to medication (anti-acids or proton-pump inhibitors)[1,32,33]. Several factors impact iron absorption rate, such as an inefficient intake, due to congenital conditions or alterations that impair absorption, as well as inefficient alimentary diets (**Figure 1.12**)[30,33].

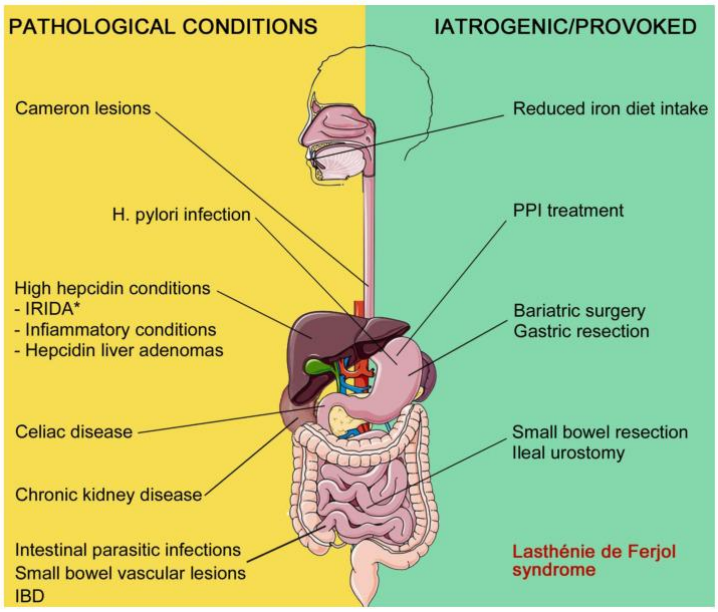

Figure 1.12: Causes of iron anaemia. IDA causes can be intrinsic (pathological) or extrinsic (iatrogenic). Signalled in red is the Lasthénie de Ferjol syndrome, a factitious disease in which patients self-inflict themselves, by bloodletting, to the point of causing IDA. IRIDA- Iron Refractory Iron-Deficiency Anaemia, IBD- Inflammatory bowel diseases, PPI- proton-pump inhibitors, H. pylori – the *Helicobacter pylori* bacteria causes stomach walls inflammation. Image from Migone De Amicis[33].

If the GI epithelium is inflamed, the available superficial area for absorption is diminished, likewise, if the stomach's acid secreting capacity is impaired it also impacts absorption negatively[30,32], these account as physiological limitations. Several chronic gastrointestinal conditions due to malabsorption or blood loss can cause IDA[30]. Conditions like Celiac and Crohn's disease cause villi atrophy due to inflammation, thus reducing in the area available for absorption[1,32,33]. Iron loss is mainly because of haemorrhages throughout the gastrointestinal tract (occult blood loss), due to ulcers, cancers, parasitic infections, inflammation, haemorrhoids, or inflammatory bowel disease[1,30].

In lower income countries, besides factors that afflict higher-income countries, which account for half cases, infections and their consequences, are to blame for ID's high prevalence[3]. Parasitic infections, like Malaria, cause inflammation and internal bleedings which can lead to IDA[3,33].

There are several variables besides the physical constrictions that affect iron absorption, namely diet (discussed in detail ahead, section **1.3.3**). An inadequate and unbalanced diet, can exacerbates the development of other anaemia types due to micronutrients shortage, such as vitamin A, B_2 and B_{12}, and folic acid[3,11,33]. These deficiencies can also disguise IDA, both vitamin B_{12} and folic acid increase RBC size (macrocytic), so their triple deficiency may showcase normal sized RBCs (normocytic), despite IDA being characterized by small sized RBC (microcytic)[33]. Thus, by coexisting with ID they cause a greater dispersion between different RBC sizes (increased RBC distribution width, RDW)[30].

Strategies of fortification or biofortification have been developed to ensure that even people of a lower economic standing, are able to improve ID through the consumption of these improved foods[3]. As they are more accessible and stable to consume than other therapeutic approaches to improve iron levels[3,11].

Depending on the duration of anaemia it can be chronic or acute; if a large blood loss occurs suddenly it can cause acute anaemia, while prolonged inflammation can cause chronic anaemia[1,33]. IDA is mostly due to a mismatch between iron intake and absorption in relation to iron requirements[3]. ID can be functional or absolute. It is functional when despite having sufficient iron stored it isn't mobilised in order to satisfy iron needs, while in the case of absolute deficiency the organism's iron reserves are exhausted[10,30].

1.3.2. Diagnosis

The diagnosis for ID or IDA is done through the use of laboratory tests, such as Complete Blood Counts (CBC) (**Table 1.1**) and by measuring the levels of iron storage (ferritin)[10,31]. By comparing the obtained results to the standard parameters, using the lower end ranges of each parameters, having in consideration that these values shift depending on the age or sex of the individual[10,25,30,31].

Table 1.1: Standard values for haematological parameters in CBC for adults.

Haematological parameters	Unit	Reference values	Critical Value	
			Low	High
Red Blood Cells (RBC)	**10^{12}/L**	3.85-5.20 **F**	-	-
		4.31-6.40 **M**		
Haemoglobin (Hb)	**g/dL**	11.50-16.00 **F**	<6.00	>19.90
		13.60-18.00 **M**		
Haematocrit (Ht)	**%**	34.70-46.00 **F**	<18.00	>61.00
		39.80-52.00 **M**		
Mean Corpuscular volume (MCV)	**fL**	80.00-97.00	-	-
Mean C. Haemoglobin (MCH)	**pg**	26.00-34.00	-	-
Mean C. Haemoglobin Concentration (MCHC)	**g/dL**	32.00-36.00	-	-
Red blood cell Distribution Width (RDW)	**%**	11.50-15.00	-	-
White Blood cells (WBC)	**10^{9}/L**	4.00-10.00	<1.00	>30.00
Platelet count (PLAT)	**10^{9}/L**	140-440	<25	>1000

F - Female; **M** - Male; **C.**-Corpuscular; "- "stands for the absence of pre-established critical values. Adapted from DGS Norm nº 063/2011 updated in 12/09/2012[31].

Several guidelines for exist anaemia diagnosis[31]. WHO established a diagnosis for anaemia, strictly meaning a low RBC number or a low Hb concentration, in adults above the age of 15 for Hb of below 13 g/dL for men, and below 12 g/dL for women or 11 g/dL during pregnancy[10,30,32]. Besides Hb other diagnosis parameters include, Mean corpuscular volume (MCV), or Mean corpuscular haemoglobin (MCH)[31].

In addition, serum iron-related biomarkers are crucial to ID diagnosis: ferritin below <30ng/ml, and Transferrin saturation (TS) below to 20%, but without inflammation signals[30,34-36]. The further below are CBC and biormarkers results, the more serious the condition is[30]. Other indicators include increased total iron-binding capacity and serum transferrin receptors, whose expression is equivalent to erythropoiesis and iron needs, these indicators unlike ferritin aren't influenced by inflammation[11,30]. The ratio between TfR and ferritin, has been used to detect alterations in iron storage and functionality, however it isn't widely used due to its higher cost and lack of a standard TfR assay[11].

The main consequences of ID include microcytosis (lower volume RBC, low MCV) and hypochromia (lower RBC Hb concentration, low MCH), and low level of total Hb. The low levels of these parameters can cause fatigue, a decreased thermoregulation capacity, dyspnea, immunological dysfunction and neuro-cognitive damages, in children it can also cause psychomotor and cognitive anomalies[22,25,26,28,30,32]. It is also possible to find these features in haemoglobinopathies and other anaemia types[18,30,34,35].

Since the early 70s, several mathematical indexes based on haematological parameters (such as Hb, MCV, MCH, and others) were developed to distinguish IDA from other hypochromic microcytic anaemia conditions, such as haemoglobinopathies (mainly Thalassaemia)[25,34-36]. These indices can be promptly applied to CBC parameters individually[34-36], although combining different formulas tends to increase sensitivity, specificity, and performance[35]; however they are not reliable enough for a diagnosis *per se*[36].

1.3.3. Prevention and treatment

One can say that IDA prevention is mainly through adequate nutrition, thus the importance of a well-balanced diet, for both its prevention and treatment[11]. Also, it is important to be aware that iron absorption can be impacted by the action of competitors, like other metals (see section **1.2.1**), inhibitors and enhancers, which will be addressed promptly.

Iron can be inhibited by several substances, like polyphenols, phosphates and phytates, present in plants[3,11,32]. These prevent absorption by forming insoluble complexes with iron[3,11,32]. Proteins like albumin, casein and whey (from milk), and some of soybeans also inhibit its absorption[11]. Calcium is the only iron inhibitor that is also able to limit haem absorption, with a more pronounced effect in single meals than in varied multiple meals diets[11]. Non-haem iron sources are found all food types, however these are less bioavailable, by inorganic iron complexes formation or because most of this iron is Fe^{3+}, thus they are more susceptible to have their absorption potential diminished[1,3,4]. Their bioavailability can be increased through processing[3,11].

In order to enhance absorption, some adjuvants may be used: ascorbic acid, citrate, and some amino acids[3,11]. The several studies have shown that adding myoglobin and Hb rich meats in meals, makes iron more promptly absorbed, also significantly increasing non-haem absorption[3,4]. Ascorbic acid (vitamin C) is able to overcome all of iron's inhibitors; its incorporation reduces inorganic iron complexes formation, preventing their inhibition, but its effect can be greatly diminished by degradation through cooking (as it is thermolabile)[1,11].

So to summarise, to prevent and improve ID one should avoid or reduce intake of drinks that inhibit iron absorption like milk (proteins and calcium), coffee and tea (phytocomponents), and increase the amount of iron-rich foods in the diet, particularly haem, along with vitamin C rich foods[11,32].

IDA treatment is heavily dependent on the underlying conditions that trigger its development, as the way to tackle it to ameliorate IDA depends on its root cause (**Figure 1.13**)[10,30,32].

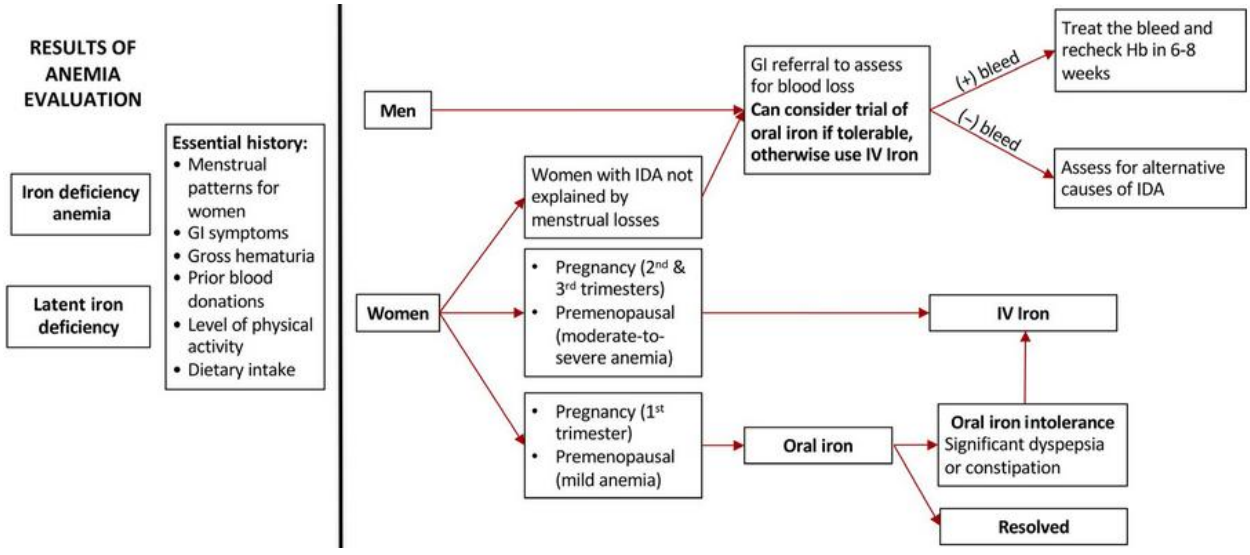

Figure 1.13: Algorithm for Iron-deficiency anaemia diagnosis and treatment. For the diagnosis and management of ID and IDA it is necessary to make an assessment of the individual's history and have in consideration the differences between sexes, as these precede the treatment route to use. IV – intravenous. Image from Elstrott[32].

If ID is due to an insufficient intake, chronic or congenital conditions, malabsorption, blood loss, use of substances like Aspirin (Non-steroidal Anti-inflammatory drugs (NSAIDs)), these conditions need to be addressed first[30,32,33]. To prevent unnecessary exams, Hb electrophoresis should be performed to detect haemoblobinopathies, like Thalassaemia, that have similar symptoms to IDA[10,30,34-36].

If alimentation alone isn't efficient enough to improve ID, then ferrous sulphate ($FeSO_4$) supplementation is required, which can be in oral or intravenous route[3,6,30]. The $FeSO_4$ tables should be taken for a minimum of 3 months to ensure iron stocks, with advised monitoring[30]. To guarantee iron levels and the condition's improvement, these CBC should be performed in 3 month intervals for at least a year[30]. When oral administration isn't adequate (cases of intolerance or unresponsiveness), iron can be supplied intravenously or via intramuscular gluteus injections[30,32]. As last case resort, only to increase Hb (to enable other treatments), blood transfusions might be required[3,6,30].

These strategies may however prove to be ineffective. ID is said to be refractory if after a treatment with the oral supplementation, for a minimum duration of a month, fails to increase Hb levels[32,33]. Refractory anaemia is usually acquired (due to GI track maladies), but more rarely, can be due to genetic variants that limit absorption or condition the iron metabolism's normal functioning[32,33].

1.3.4. Iron deficiency due to genetic causes

1.3.4.1. Iron Refractory Iron-Deficiency Anaemia (IRIDA)

Iron deficiency anaemia may also be a consequence of a genetic defects such as in Iron Refractory Iron-Deficiency Anaemia (IRIDA)[20,32]. It is a rare autosomal recessive condition (OMIM #206200), with an estimated prevalence of 1 in a million, due to alterations in gene that codifies for matriptase-2, MT2 (*TMPRSS6*) that cause protein loss-of-function[20,24,37,38]. This condition expresses its phenotype mainly during infancy[20,22,28,32]. Only through molecular analysis it is possible to diagnose it correctly[33].

IRIDA presents itself with the same classic ID phenotypes, hypochromic microcytic anemia, but additionally low TS (below 10%), normal or augmented ferritin, and increased iron accumulation within macrophages (due to increased hepcidin levels inhibiting iron export)[20,33,37,38]. The main difference regarding IDA is its non-responsiveness to oral iron supplementation, and abnormal or lower rate of IV iron use[5,8,20,33]. These two therapeutic approaches ameliorate IDA symptomology but are ineffective for IRIDA patients, as they suffer from refractory anaemia, due to MT2 malfunctioning[32,37-40].

MT2 is expressed in epithelial cells, mainly in hepatocytes (involved in hepcidin expression regulation)[28,37], but also in the kidneys, spleen, mammary tissues, brain, uterus, among others in lower levels[22,37]. Besides being able to cleave HJV it also is able to cleave various other proteins, like HFE and TfR2[38].

In normal cases, when HJV comes to associate to BMP6 that binds to BPM6 receptors (BMPR), MT2 located at hepatocytes surface can cleave the bond between HJV and the complex formed by BMP6 and its receptor[37]. This cleavage leads to decreased hepcidin transcription, inhibiting it through HJV degradation[5,38]. As HJV induces hepcidin expression through the promotion of phosphorylation of transcription factors, SMADs (seen in detail in **Figure 1.8**)[2,5,38]. When this cleavage occurs *HAMP* gene expression isn't stimulated, thus inhibiting it accordingly to the organism's iron needs, allowing adequate iron absorption in the duodenum (**Figure 1.14 A**)[38].

However, in the case of deficient MT2 due to alterations, its inhibitory activity upon *HAMP* expression, through the cleavage of HJV, is decreased or absent, provoking increased hepcidin production, regardless of iron levels[37]. This increased deregulation leads to inadequate iron absorption, at the epithelial level of the duodenum, by decreasing absorption, export and recycling, thus contributing to IRIDA or a similar phenotype regardless of iron status **(Figure 1.14 B)**[5,8,22,24,28,41].

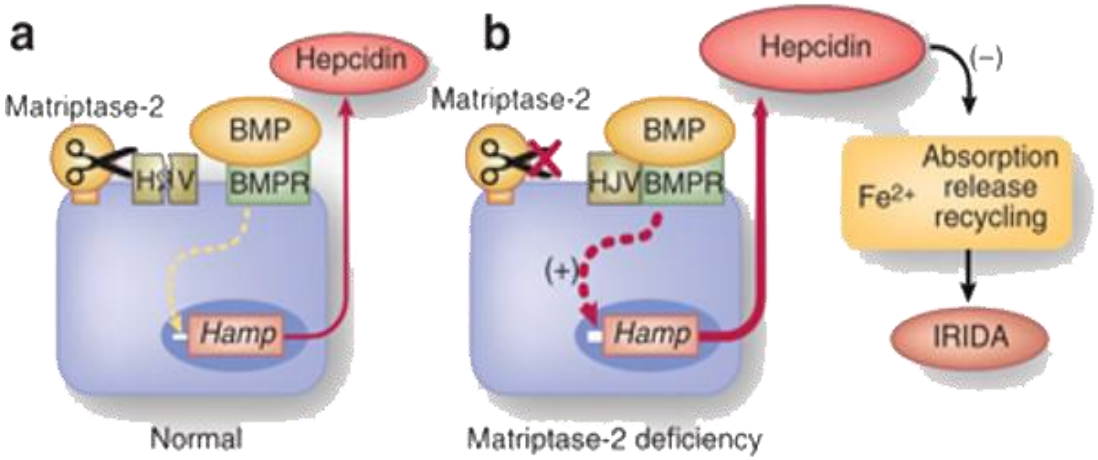

Figure 1.14: Regulation of hepcidin expression by Matriptase-2. **A)** Normal hepatocyte, MT2 acts upon HJV, cleaving it from the complex in which it is bond to BMP-BMPR. **B)** MT2 deficiency, there is an impediment to its inhibitory action upon *HAMP* expression, thus hepcidin is produced regardless of iron requirements. That increase in hepcidin causes decreased iron absorption, release and recycling, thus contributing to IRIDA or an IRIDA-like phenotype. Image from Cui[37].

Hepcidin levels in ID are decreased, in order to maximize iron absorption, and increase iron export by macrophage's FPN[20]. In IRIDA, MT2 has a nonexistent or decreased inhibitory activity over hepcidin, so its levels are abnormally high, preventing iron absorption and causing IO in macrophages as it prevents FPN of exporting iron into circulation[20,33].

As MT2 alterations cause increased hepcidin levels which prevent iron absorption at the GI level, treatments as such $FeSO_4$ supplementation tables are inefficient as DMT1 as FPN have their activities inhibited by high hepcidin levels[30,32,38].

Mice studies have used different treatment approaches to mitigate their IRIDA phenotype, suggested that the most efficient iron form was Sucrosomial iron[38]. This novel supplementation form makes use of ferric pyrophosphate encapsulated in a phospholipidic bi-layered micelle with a sucrester outer layer, this structure enables it to withstand gastric conditions as well as bypass DMT1 absorption via, by being absorbed independently from it[40].

In IDA patients, increasing Hb mitigates their condition, however in IRIDA patients this increase has to be done more carefully[30,32]. As they are at risk for IO, even within the average Hb range, due to the lack of hepcidin inhibition[32].

1.3.4.2. *TMPRSS6* gene

Matriptase-2 (MT2) is a highly conserved transmembrane serine protease, comprised by eight domains[23]. Starting at the N-terminus, the TM-Transmembrane domain, the SEA (Sea urchin sperm protein, Enteropeptidase, Agrin) extracellular domain region, two CUB domains (Complement factor C1s/C1r, Urchin embryonic growth factor, and Bone morphogenetic protein),three LDL Receptor domains, and the Serine protease domain at the C-terminus[23]. The gene that codifies for MT2, *TMPRSS6*, is located in chromosome 22 (22q12-13), and has 18 exons (**Figure 1.15**)[22,23].

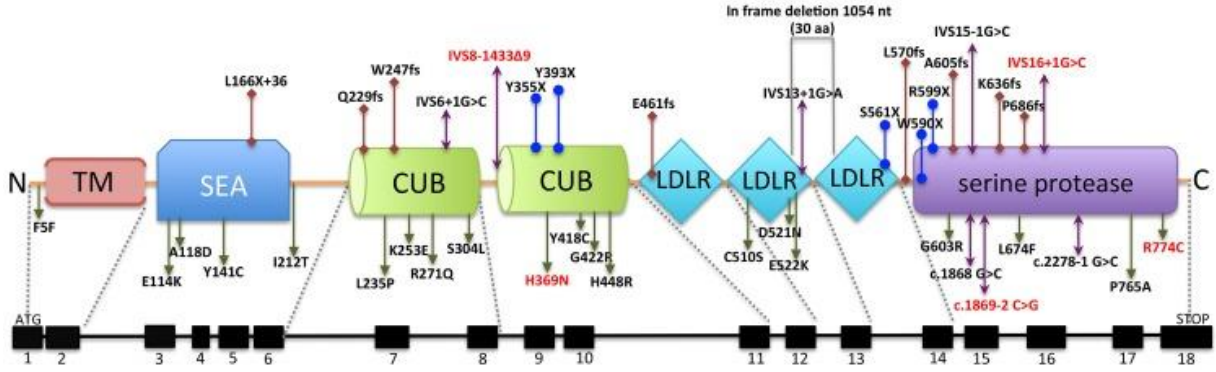

Figure 1.15 –The schematic structure of the *TMPRSS6* gene and the corresponding protein, Matriptase-2. In the first row are represented MT2's structural domains, in the second the gene with exons numbered. The dashed lines represent which exons corresponds to each domain. The arrows represent the different types of alterations. Green are missense, blue are nonsense, purple are splicing variants, and red are frameshift alterations. In-frame deletions are in grey. The alterations marked in red prevent the normal development and functioning of the protein (haploinsufficiency). Image from Wang[23].

MT2 loss-of-function alterations make hepcidin inhibition infeasible[37]. *TMPRSS6* alterations that can cause IRIDA include in-frame deletions, frameshifts, missense, nonsense, and splicing alterations[23,37]. Besides loss-of-function mutations, other less severe alterations cause iron absorption modulation at the duodenal level and in haematological parameters, by changing the levels of hepcidin, thus increasing susceptibility to the development of ID[20,22,37,41]. Thus, genetic variants in *TMPRSS6* can cause iron metabolism alterations with a variable range of severity, ranging from severe in IRIDA (due to homozygous or rare composed heterozygous loss-of-function mutations), to a slight increment in susceptibility for the development of ID or IDA due to the presence of some functional common Single Nucleotide Polymorphisms (SNPs)[20,22-24,42]. Some studies have established that certain polymorphisms in this gene are more frequent in individuals with ID than in healthy controls[20,24].

Some Genome Wide Association Studies (GWAS) have been conducted for this gene in different populations[23,24,42]. In these studies it was possible to identify and establish which SNPs were associated to lower concentrations of ferritin, haemoglobin, and ID increased risk, amongst those, p.Val736Ala and p.Asp521= (rs4820268), are widely known as functional SNPs[23,42].

Functional SNPs are those that have been statistically established as being associated to a phenotype; these SNP can be both in coding and non-coding areas of the gene[43,44]. The alteration of Valine to Alanine, p.Val736Ala (rs855791), is the most widely known polymorphism associated with alterations in iron and haematological parameters[23,24,42].

1.4. Iron homeostasis disruptions due to Iron Overload

1.4.1. Iron Overload

Iron overload (IO) is considered to be a state in which homeostasis is disrupted due to excess iron, owed to an imbalanced absorption or accumulation of it[21,28]. Due to the absence of an iron elimination mechanism, mutations associated with increased absorption or accumulation, on the long run cause damages in different tissues, these lesions due to iron deposition can be visualized via medical imaging in tissues, such as the liver or heart[13,14,21]. IO can be achieved by an increased intake of iron, or due to genetic propensity.

Haemochromatosis is a medical term that means iron overload. IO is mostly systemic but in cases of iron misdistribution, some affected tissues present IO while others have normal iron homeostasis[10]. It can have a primary or secondary origin, primary IO is due to genetic conditions[21]. Secondary IO is usually due to hepatic affections, hepatitis, cirrhosis, or high alcohol consumption[21,46].

Conditions causing IO due to iron misdistribution include, Neuroferritinopathy (affects mostly the brain with iron accumulation in the basal ganglia), X-linked sideroblastic anaemia and Friedreich ataxia, both are systemic affecting the mitochondria[13]. Other genetic conditions causing IO include Ferroportin Disease (FD), Aceruloplasminemia (CP absence), Atransferrinemia (transferritin absence) [10,12-14,28], DMT1 deficiency, H-ferritin related IO, Hereditary iron-loading anaemia with inefficient erythropoiesis, and different types of Hereditary Haemochromatosis (HH) (**Table 1.2**)[13].

Table 1.2- Hereditary Haemochromatosis.

HH Type	Gene	Gene encoded protein function	Associated phenotype	Onset (years)
I	*HFE*	Regulates HAMP and interacts with TfR1/2	↑sFt, ↑TS, hepatomegaly, fatigue, Fe deposition, moderate severity	40-50
II	*HJV; HAMP*	Promotes hepcidin production (IIa); Regulates ferroportin and Fe absorption (IIb)	↑sFt, ↑TS, cardiomyopathy, reproductive defects, Severe	15-20
III	*TRF2*	Hepcidin regulator, Holo-TF receptor	↑sFt, ↑TS, fatigue, Fe deposition, moderate severity	30-40
VI	*SLC40A1*	Regulates Fe export	2 phenotypes associated to gain or loss-of-funtion *, variable severity	*

↑- increased, sFt – serum ferritin, TS – transferrin saturation. HH type II also called juvenile is associated to 2 genes with similar phenotype (IIa e IIb). The (*) marked segments will be disclosed with more detail. Adapted from[13,21,28,41].

Hereditary Haemochromatosis (HH) tends to be associated to certain ethnic groups, the most common, HH type I, affects mostly people of Northern European ancestry, the other types affect people of broader background but aren't as common[10,13,21, 45-47].

Addressing IO, means prevent further development, and ideally mitigate its effects, the latter is of upmost importance, as excessive iron is conducive to cellular damage or even death, as ferropoptosis[2,3,6,7,15,19]. Unlike ID, the genetic contribution for IO development is much more significant, as cases due to an excessive alimentary intake are very rare in comparison.

The measures previously mentioned to mitigate ID can be implemented in an opposite fashion for resolving IO. Vitamin C increases iron absorption, so it should be avoided, while calcium intake should be increased as it inhibits iron absorption[3,11,32]. However, more drastic therapeutics options may need implementation. The main therapy for HH is phlebotomy, the medical term for blood drawing, which is the easiest to implement, and causes the least side effects, comparing to the next alternative option, which is iron chelating treatments[10,46].

1.4.2 Hereditary Haemochromatosis

HH are characterised by excessive absorption and cellular iron accumulation, in parenchymal cells, and several organs as the liver, heart, and pancreas[10,12-14,21,28,48]. Depending on the affected gene, the clinical manifestations, clinical onset, and mode of inheritance might differ (see **Table 1.2**)[13,28,46,48].

Diagnosis guidelines include high ferritin levels (>300 ng/mL in males/>200 ng/mL in premenopausal women) and transferrin Saturation (TS) above 45%[48]. Besides the testing of those biomarkers HH can also be diagnosed by genotypes, presence of certain phenotypes, medical imaging tools (as MRI), or through biopsies, of the liver, for example[10,46]. In order to avoid unnecessary tests and invasive diagnosis tools, algorithms have been developed (**Figure 1.16**)[10,13].

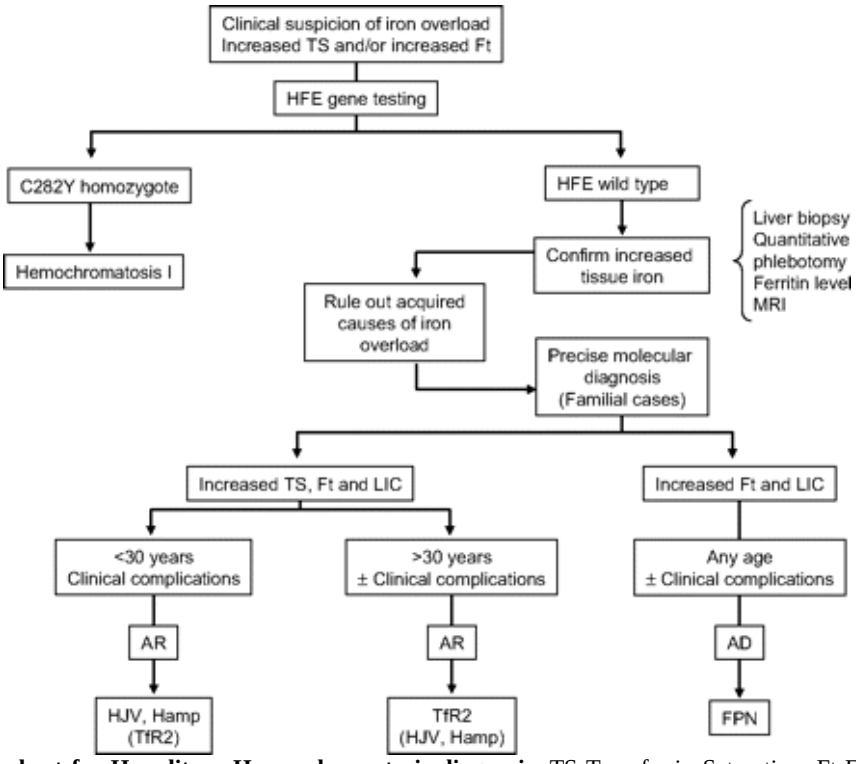

Figure 1.16: Flowchart for Hereditary Haemochromatosis diagnosis. TS-Transferrin Saturation, Ft-Ferritin, LIC- liver iron concentration, AR- autosomal recessive, AD- autosomal dominant. Image from Muñoz[10].

Most HH have clinical consequences ranging from mild to a severe impacts on the individuals health, however those with affected *HAMP* always have severe implications[13,41]. In HH, hepcidin levels are too low for the amount of iron present in the organism, due to alterations in genes related to hepcidin production, such as *HFE, HJV* (juvenile HH IIa), and *TRF2*[10]. However in alterations that directly influenced by hepcidin (juvenile HH IIb) or FPN, the same isn't observed[10]. They are mainly caused by loss-of-function mutations, only for HH type IV the reverse is true, but in all cases their pathogenicity rests on the impaired functional activity of hepcidin[13,28,41].

Unlike other types, HH type I can be diagnosed solely on a genetic basis, p.Cys282Tyr homozygous, or composite heterozygous alterations combining the former with p.His63Asp or p.Ser65Cys[10,21,41]. A previous Portuguese population study established an increased northern presence (5.8% vs. 0.9% in the south) for the p.Cys282Tyr altered allele, and a homogeneous p.His63Asp presence (17-20%)[45].

1.4.2.1 Hereditary Haemochromatosis type IV and Ferroportin Disease

The majority of HH are of autosomal recessive transmission conditions except for those affecting the *SLC40A1* gene (**Table 1.2**)[21,28,41,47]. Alterations in this gene are responsible for HH type IV and Ferroportin disease (FD), both are referenced by the same OMIM number, #606069 (**Table 1.3**)[13,49].

Table 1.3 – Differences between the pathologies associated to the *SLC40A1* gene.

Hereditary Haemochromatosis type IV	Ferroportin disease
Ferroportin gain-of-function	Ferroportin loss-of-function
Fe accumulation due to increased export	Fe retention due to decreased export
Hepatopathy, cardiomyopathy, arthropathy, endocrinopathy	Hepatic malady, slight anaemia
Onset between 40 to 50 years	Affects individuals of every age
Clinical expression ranges from mild to severe	Mild clinical expression

Adaptation from Pietrangelo and Silva[13,28]

As seen in Table 1.3, both conditions are due to FPN alterations, but with a differentiated expression[13]. FD is among the most common forms of genetic IO (only surpassed by HH type I), being present in all human populations[47]. FD's loss-of-function in FPN is due to increased hepcidin sensibility, which causes iron export incapacitation in macrophages, Kupffer cells, as hepcidin more easily inhibits its activity, but FPN export capacity remains preserved and functional in enterocytes[13,14]. IO is systemic with a macrophage preference[13], lower iron in circulation causes lower TS, higher serum ferritin levels, even leading to a slight anaemia[29]. As FD is caused by increased hepcidin sensitivity, depending on the pathological variant affecting an individual, the severity range can vary greatly. The most serious outcome would be an alteration that completely inhibits iron export upon hepcidin binding to FPN blocking its activity[47]. It can also be due to alterations that impact FPN expression[13]. Putting it briefly, in FD the loss-of-function alterations block iron export by macrophages after RBC recycling, while gain-of-function alterations of HH type IV make hepcidin's regulation ineffective, thus the iron is exported without control (**Figure 1.17**)[49].

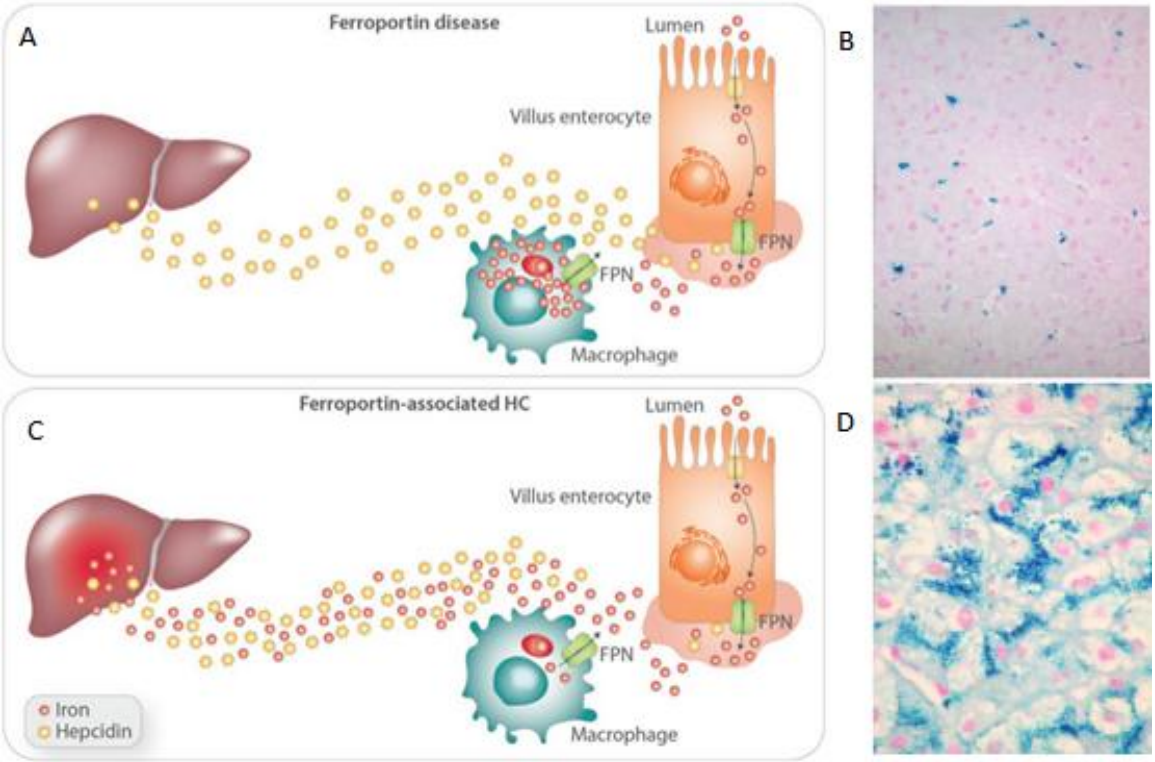

Figure 1.17: Comparison between FD and HH type IV mechanisms with histological samples. **A** and **B** represent FD, while **C** and **D** represent HH type IV. **A**) In FD, IO doesn't affect the liver as seriously; iron accumulates in macrophages, retaining it, and decreasing available iron in circulation. **C)** In HH type IV, occurs hepatic IO, uncontrolled iron export as hepcidin no longer inhibits FPN activity, iron is thus apt to accumulate in more tissues. **B** and **D** are liver histological pictures with Prussian blue stain, the blue represents iron accumulation. **B)** The blue dots represent the affected macrophages/Kupffer cells. **D)** Extensive hepatocyte IO; the spared pink bits are unaffected Kupffer cells. Adaptation from Pietrangelo[13].

Another difference in regards to HH type IV is their TS values, as previously said, FD in general has normal or low TS values while HH type IV has high TS values[47], as there is more iron in circulation.

When hepcidin binds to FPN it causes a conformational alteration that prevents iron export from the cytoplasm to plasma[5,13,28]. Alterations that enable FPN to withstand hepcidin regulation, increasing its resistance to inhibition, origin HH type IV[13,47]. Through unhindered iron exportation, due to a FPN gain-of-function alterations, makes it is more similar to other HH types in symptomatology than FD[13].

The increased resistance to hepcidin, through diminished binding capacity, causes increased GI iron absorption and uninhibited macrophage iron export, leading to the development of IO in multiple tissues as it can be seen in Figure 1.17[13,29].

1.4.2.2 The *SLC40A1* gene

The Solute Carrier family 40, *SLC40A1* gene encodes for the highly conserved transmembranar protein ferroportin (FPN)[13]. FPN is primarily expressed in the basolateral membrane of enterocytes, spleen macrophages, hepatocytes, bone marrow and placenta[5,12-14,28,29]. Its regulation is preformed both before and after transcription, through the IRP/IRE system (as stated in the section **1.2.4.1**)[2,29].

The *SLC40A1* gene is located in chromosome 2 (2q32), it has 1713 nucleotides and eight exons (**Figure 1.18**)[14,28,29]. Mutations in *SLC40A1* gene are the most common cause of hyperferritinemias, with some mutations associated to specific populations[13,14]. Pathogenic homozygous mutations can be incompatible with life[13], as this gene is responsible for the synthesis of the sole iron export protein in mammals[2,29].

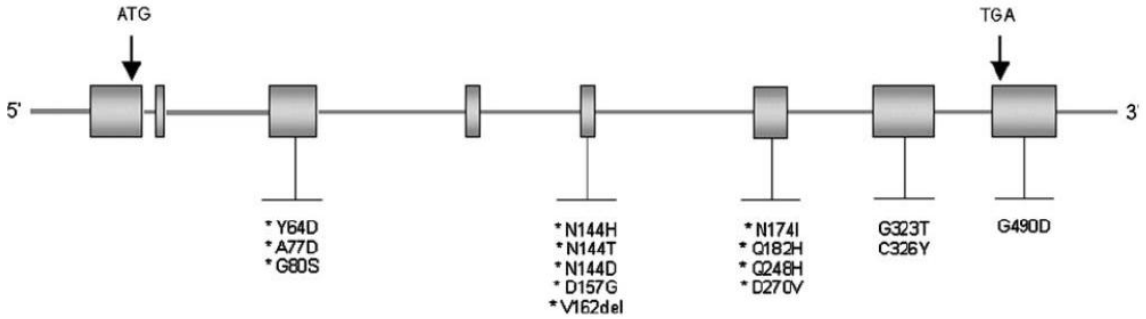

Figure 1.18 – Scheme of the *SLC40A1* gene. The squares represent exons while introns are represented by lines. Some of most common variants are also represented, as well as the start (AGT) and stop codons (TGA). Image from Le Gac[14]

Depending on the alterations in gene, it can cause gain or loss-of-function of FPN, with these causing HH type IV or FD, respectively[13,47]. By having a autosomal dominant transmission, individuals with an heterozygous alteration are automatically affected, with each alteration having a differentiated pathological expression even within the same condition[5,10,12-14,28,29].

The most common alterations are p.Val162del, p.Ala77Asn and p.Gly80Ser, causing FD, as they are associated with loss-of-function[13,47]. Alterations causing gain-of-function, and thus HH type IV, include p.Asn144Asp, p.Asn144His, p.Cys326Tyr, p.Cys326Ser, or Tyr64Asn[13,14,29,48].

Mutations like p.Tyr64Asn and p.Cys326Ser annul FPN's sensitivity to hepcidin, leading to high TS levels[13], as the body is trying to avoid the nefarious effects of unbound Fe^{3+} in circulation. While mutations that only restrain hepcidin regulation, as p.Asn144Asp or p.Asn144His, allow TS to remain within the normal range, thus being alterations associated with less severe IO outcomes[13].

The p.Cys326Ser variant causes the most severe consequences, by completely blocking hepcidin access to the FPN catalytic centre, leading to uncontrolled iron export and consequently haemochromatosis[29].

Thus the study of *SLC40A1* gene mutations and their corresponding association genotype/phenotype is of utmost importance in order to understand the complex mechanisms of their pathogenecity.

2 Aims

The main goal for this research was to increase the knowledge in regards to iron metabolism through the study of the *TMPRSS6* and *SLC40A1* genes. In order to do so this work was divided into two parts, one related to iron deficiency, and another to iron overload, each part has its corresponding population, with different tasks to reach our goals:

1. Study individuals suspected of suffering from Iron-Deficiency Anaemia or Iron Deficiency through:

I. Sequencing the *TMPRSS6* gene;
II. Classification of the pathogenicity of the novel or rare genetic variant found;
III. Assess genotype and phenotype associations;
IV. Characterise the role of functional polymorphisms in susceptibility to iron deficiency development.

2. Interpreted the genetic screening previously performed in patients suspected of either Hereditary Haemochromatosis type IV or Ferroportin Disease through:

I. Statistical analyses and *in silico* analyses of previously obtained *SLC40A1* NGS results
II. Assessment of genotype and phenotype associations;
III. Classification of the pathogenicity of the novel or rare genetic variant found;
IV. Characterisation of the pathogenic mechanism subjacent to Hereditary Haemochromatosis type IV or Ferroportin Disease

3 Materials and Methods

3.1 Biological samples

3.1.1 Iron Deficit samples

In a study previously conducted in INSA, the National Survey of Health with Physical Exam (INSEF), 4812 individuals representative of the Portuguese population were selected[50]. From those individuals, biological samples were obtained, consisting in serum and peripheral blood collected in EDTA tubes (both frozen and stored at INSA's biobank by the INSEF study)[50].

The blood collected in EDTA tubes was used to obtain Complete Blood Counts (CBC). After CBC analysis, we selected those revealing microcytosis and/or hypochromic RBC to partake in our study. As the limits for these vary depending on the source, we have chosen DGS criteria: MCV below 80 fL, and MCH below 27 pg[10,31]. Thus, from the whole individuals only 204 samples matched the selection criteria suggestive of iron deficiency. From these, one hundred were randomly selected for our study to deepen our knowledge, as analysing all samples would be impractical.

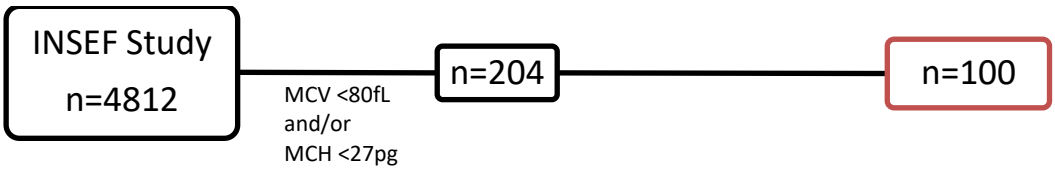

Figure 3.1: Algorithm for sample selection. Within the black boxes are the samples numbers from which we selected our population, the red box represents our population. Between the primary sample of the INSEF study and the 204 group, is the criteria used to filter the samples with at least one of those characteristics.

The samples were screened for the presence haemoglobinopathies, as these genetic pathologies can give rise to an IDA-like phenotype. Unfortunately, iron biomarkers analyses were not possible.

3.1.2 Iron Overload samples

One hundred and ten samples of peripheral blood were collected in EDTA tubes from patients that had a diagnosis of hyper-ferritinemia with unknown causes, with their consent, from the department of immunohaemotherapy of *Hospital Santa Maria*. For sample inclusion, serum ferritin above 300ng/mL or TS over 60 %, were used. Samples were excluded if risk factors for secondary IO were present, like being alcoholic or suffering from hepatitis[21].

The samples were previously screened for mutations in several genes related with non-classic hereditary haemochromatosis. Since our aim is to study HH-type IV and Ferroportin disease, we performed statistical analysis and *in silico* analyses of their genotyping results for the *SLC40A1* gene.

3.2 DNA extraction from the biological samples

The DNA used in this study was extracted from peripheral blood of the INSEF individuals in study. The extraction took place in the *Unidade de Genética Molecular* (UMO) of INSA, in MagNa Pure, a nucleic acid extractor by Roche. The quantity and quality control of the extracted DNA were accessed via the NanoDrop one Spectrophotometer by Thermo Fisher Scientific.

DNA quantity and quality is important due to the sensitivity of the techniques to perform. To ensure the best results possible in the next steps, the quality control assessment used specific ratios, 260/280 and 260/230. Nucleotides absorb UV light at 260 nm while contaminants absorb at other wavelengths. Thus a ratio of 1.8 for 260/280 nm or 2.0 for 260/230 nm are considered good measures of pure DNA.

3.3 Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction is a molecular biology technique that allows to exponentially increase sequences of interest in a laboratory setting. In our case, we were interested in increasing the amount of DNA of our samples for *TMPRSS6*.

In order to amplify the DNA fragments, besides our DNA template we also required specific primers[37], dNTPs, and buffers to ensure the best medium possible for the correct functioning of Taq polymerase, thus optimizing the reaction's yield.

All PCRs apply this same principle, but different nuances are added depending on the aim. Besides the standard PCR we also performed Long Range PCR, and Nested PCR. For the implementation of these techniques different thermocyclers by Biometra were used (T Gradient Thermocycler, T1 Thermocycler, T Professional Thermocycler).

3.3.1 Nested PCR

The principle of this technique is amplifying a previously amplified DNA, to obtain a higher DNA amount on the second amplification. This method can be applied by using different primers or the same primers as the starting PCR, the latter approach is the one used in our study.

We applied it in cases in which the first amplification was devoid of bands in the agarose gel. If in the DNA was amplified in the first PCR but in a low amount, this technique allowed to further increase DNA quantity, thus enabling band appearance on the gel after the second PCR.

3.3.2 Long Range PCR

Depending on the length of the interest fragment, the classic PCR technique might not be adequate. Long Range PCR allows amplification of long DNA fragments, even over 5000 base pairs (bp)[51]. In order do Next-generation Sequencing, amplification of the whole *TMPRSS6* gene is required. With 18 exons, it is a long gene, so it was amplified in three amplicons of 8266 bp, 5831 bp, and 9421 bp (**Table 3.1**).

Table 3.1: Amplification of *TMPRSS6* gene subdivided in three long fragments.

Gene	Chromossome	Location	Amplicon's gene location	Size (bp)
TMPRSS6	22q12.3	NC_000022.11	c.-173 (5'UTR) to c.658+104 (int 6)	8266
			c.659-240(int 6) to c.1223+103(int 10)	5831
			c.1224-138 (int 10) to c.* 83 (3'UTR)	9421

This information is in accordance with the genome coordinates assembled in GRCh38.p13.

Different primers were used in three different long range PCRs in order to obtain the three amplicons (**Table 3.2**).

Table 3.2: Primers used for Long PCR.

Amplicon	Primers	Primer Size	Sequence	Fragment
1	Ex1_Fw	21 bp	5' –CTGAGACCTCCGTCTGTCCTC– 3'	8266 bp
	Ex6_Rv		5' –CCCTGCACACACAACAGAAGC– 3'	
2	Ex7_Fw		5' –AGGCGTGAAGCTCAGTGTGTG– 3'	5831 bp
	Ex10_Rv		5' –GAGATTGGGGACTTGGGCTTC– 3'	
3	Ex11_Fw		5' –AGGGAGAAATCAGGGCAGAGG– 3'	9421bp
	Ex18_RV		5' –CCCAGTCAATTCCCAACAGTC– 3'	

Owing this technique's amplification extent and higher sensitivity, it is more prone to errors with DNA concentrations over 30 ng/µL. It is also processed more slowly, due to increased fragment size, but it is more accurate than a regular PCR due to the use of a more specialized Taq, the LA Taq Hot Start version kit by TAKARA[51]. The master mix used (**Table 3.3**), was the same for the three amplicons, except for the primers, that are described in the Table 3.2[36].

Table 3.3:Master Mix for Long Range PCR.

Reagent	Concentration	Per reaction (µL)
ddH_2O	-	15.9
BSA	10 mg/mL	0.35
10XLA PCR Buffer II (Mg^{2+} plus)	25 mM	2.5
dNTP'S Mix	2.5 mM per dNTP	4
ExN_Fw	25 µM (25pmol/ µL)	0.5
ExN_Rv	25 µM (25pmol/ µL)	0.5
TaKaRa LA Taq Hs	1.25 U/µL	0.25
Volume per tube	-	24.0
DNA	30ng/µL	1

N stands for the number of the exon.

In the Table 3.4 are depicted the conditions for each amplicon amplification. They differed slightly in times and temperatures, and although the differences are minor, the time differences made one amplification take a bit over 6 hours while the shortest took only 4 hours.

Table 3.4: Conditions for the amplification of each amplicon.

		1		2		3	
	Cycles	**T (^oC)**	**Δt**	**T (^oC)**	**Δt**	**T (^oC)**	**Δt**
Initial denaturation		98	4 m	98	4 m	94	4m
Denaturation		98	30 s	98	30 s	94	30s
Hybridisation	**30**	63	30 s	64	30 s	65	30s
Extension		72	8 m	72	6 m	72	10m
Final extension		72	10 m	72	10 m	72	10 m
Pause		4	15 m	4	15 m	4	15 m

T- Temperature; **Δt-** time variation, s- second, m-minute.

3.3.3 PCR Semi-Quantity and Quality Control

The fragments quantity and quality appraisal was performed via agarose gel Electrophoresis. Agarose concentration changed according to the fragment size, 1% was used most times, a 0.5% concentration was used for the amplicons, or 2% for smaller fragments, in all cases TBE buffer 1X was used.

Ethidium bromide was the fluorescent agent used to stain our DNA samples under UV light; it works by intercalating itself between bases. As it is a highly mutagenic substance, special security measurements had to be taken to decrease risks associated with handling it. After the gel was ran, it was viewed with UVITEC Cambridge, Gel Documentation System, FIRE-READER XS, and had its photograph printed by Mitsubishi P93D Medical Thermal Monochrome Ultrasound Printer if needed.

3.4 Next-generation Sequencing (NGS)

This method has been developed in the last decade and vastly contributed for furthering knowledge in several genetic domains[52,53]. It is able to process a high amount of information, and be used in several ways, such as whole genome and exome sequencing, targeted panel sequencing, and RNA sequencing[53]. In our study the whole exome sequencing was performed, as only a single gene was of interest. NGS can be summarised in different steps (**Figure 3.2**).

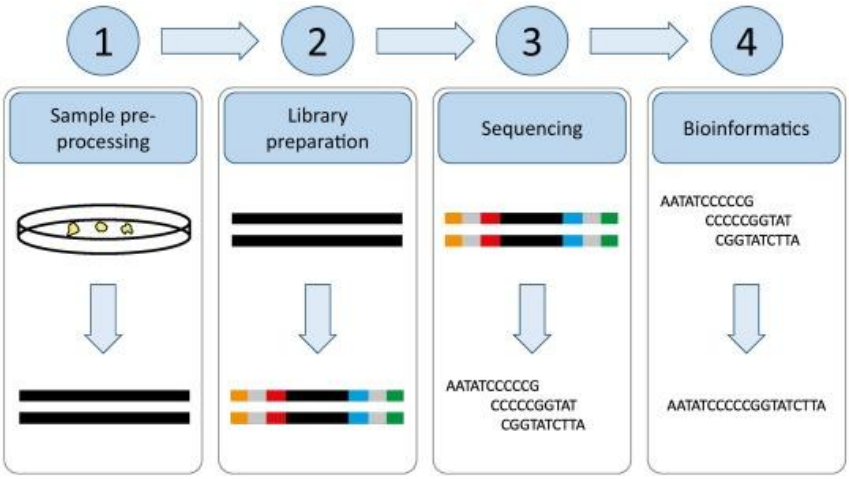

Figure 3.2: NGS workflow. Here are summarily depicted the steps taken in order to sequence a sample through NGS. Starting in sample processing, library preparation, sequencing and bioinformatics. Image from Hess[53].

First in preparation, pre-processing, we assembled the previously amplified amplicons of *TMPRSS6* obtained in long range PCR, in 200 µL eppendorfs, prefacing a minimum sample volume of 10 µL per tube. The different amplicons had to have similar concentration between them, so a prior agarose gel band intensity comparison was performed. The Amplicon mix was purified using paramagnetic beads (Agencourt AMPure Purification kit by Beckman Coulter) followed by quantification in a Qubit 3.0 fluorometer (Life Technologies).

After this, library preparation was required; this process involves the binding of adapters (synthetic DNA similar to the one found in the walls of the cartridges to the used further ahead) to fragments of our DNA of interest, this is called Tagmentation (see 2 of Figure 3.2, in which the coloured portions are the adapters). Several workflows exist for library preparation. Iin *Nextera* XT kit by Illumina used in this work, it is comprised by quantification, Tagmentation, amplification, clean-up and a final quantification[53]. It makes use of paramagnetic beads to normalise, denature and dilute the libraries to load into MiSeq cartridges for sequencing. Each sample was associated to one MiSeq cartridge within the flow cells[52]. Once within the cartridge the library binds to the fragments in its surface forming clusters, which allow fragment sequencing. This is one of the most sensitive points in NGS, as it can be more prone to errors in this phase as it is a complex process[53].

The 100 samples were loaded into two flow cells of 300 cycles, as these had 96 cartridges, and targeted to high-throughput sequencing in the MiSeq bench top sequencer (Illumina). This sequencing technique uses small and parallel platforms, allowing massive amounts of short reads per run in a flow cell. The best advantages of the Miseq system are increased sequencing accuracy, elimination of repetitive regions, its quick processing and library preparation, 4 hours and 90 minutes respectively[52].

After Sequencing and alignment, data analysis proceeded using several tools: Sequencing Analysis Viewer (Illumina) and FastQC (Babraham Bioinformatics) for quality assessment; MiSeq Reporter software package (Illumina) for read mapping, variant calling and filter; FastQ (Babraham Bioinformatics) to screen for contamination between samples; Variant Effect Predictor (VEP)[54] to annotate variants, and Integrative Genomics Viewer for reads and variants visualization. The resulting data was then compiled and presented in Excel spreadsheets and Variant Call Format (VCF) text files. From the given information, we select those of interest (pathogenic, rare or novel variants, which have not been reported in databases) for further studies or confirmation.

The NGS results for each variant were given in the following format, GT:AD:DP:GQ. With each standing for Genotype, Allelic Depth, total Depth, and Genotype Quality, respectively. AD represents

the number of readings for each allele, being represented as the reference and the altered allele, in this order and separated with a comma. The following example (0,756) represents a homozygous variant case. Depending on the ratio number between alleles it gives a different GT, but regardless of the number of reads for each allele, the sum of both their reads in AD gives the DP, which is the total number of reads. GQ value is representative of the quality of the genotype, ranging from 0 to 99, however only cases with a GQ>90 are considered.

The samples should always come back with PASS, which means that the variant is well balanced and its GT is clear. However this isn't always the case, in cases in which the reads aren't even, an error message is given, Allelic unbalance (AuB). GT could give four possible results, (0/0), (0/1), (1/1), and AuB, the first three correspond to wild type, heterozygous, and homozygous for the alteration. AuB is an error that occurs when both alleles give of reads in AD, but in an unequal ratio, as heterozygous alterations assume a ratio of 50/50 between each allele. If the ratio between the alleles is below 30/70 this error message shows up, and either the reference or the altered allele can be in either end of it. The AuB doesn't mean that the sample does not have that variant, it simply showcases that there were reads for each alleles that are outside the presupposed 50/50 ration of balance. Each AuB was evaluated individually to ensure if it is possible to accept it as a true heterozygous case or not.

3.5 Sanger Sequencing

In order to validate genetic variants detected by NGS, the specific exon/intron where the variant is located was amplified by conventional PCR. After adequate control through gel electrophoresis, the high quantity and quality amplified DNA was purified. DNA purification can take two different approaches; one is chemical while the other is biophysical (See in Supplementary Data).

In the chemical procedure, the DNA is purified via neutralization of the contaminant components that might interfere with its further analysis. ExoSAP-IT from Thermo Fisher was used for chemical purification, in the majority of DNA samples.

The biophysical purifying approach makes use of classic biochemical interactions, in a chromatographic column, and physical force, through centrifugation. It was required in *TMPRSS6*'s 16 exon amplification, which included DMSO in the mix (GenoMed JetQuick purifying kit).

Sanger Sequencing is used in order to determine the nucleotide sequence in DNA fragments, it requires an amplification with fluorescent-labelled ddNTPs (kit BigDye Terminator v1.1 cycle sequencing, Applied Biossystems), followed by electrophoretic separation and chromatogram acquisition. In this methodology, solely one primer is used, either reverse or forward, depending on how upstream or downstream the alterations of interest for confirmation are.

DNA is denatured and amplified and when one of the four ddNTPs is added to the sequence it terminates amplification. The Sanger reaction gives rise to fluorescent-labelled DNA chains with different fragments sizes. The samples are loaded into the capillary gel electrophoresis matrix, and the fragments separation is according to size, through their run in the matrix. Larger fragments move more slowly remaining on top, while smaller fragments move further down accordingly. This separation occurs in a Genetic Analyser ABI 3500 by Applied Biossystems. The capillary reading starts from bottom to top, in an increasing fragment size fashion, since each nucleotide is associated to a particular fluorescent label it is possible to pinpoint each nucleotide in the DNA sequence represented by the chromatogram. For the analysis of the obtained chromatograms we used the Geospiza Finch TV 1.5.0 software.

The confirmation of the genetic alteration comes with the comparison between our obtained sequences with the canonical sequence. This comparison can be performed with the assistance of software tools. Any deviation from the latter is a possible indication of a variant.

3.6 *In silico* analyses

For each genetic variant identified, a search in different public data bases such as Ensembl (https://www.ensembl.org/index.html), UniProt (https://www.uniprot.org/), NCBI or ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) is performed. Considering the possible consequences for each alteration (e.g. *missense*, affecting splicing, frameshift), we used several programs to predict their pathogenic effect. Ahead are promptly described some of these tools.

3.6.1 PolyPhen-2

This software (http://genetics.bwh.harvard.edu/pph2/) predicts the functional effect of missense alterations, it stands for Polymorphism Phenotyping[55]. It works from an user point-of-view by inserting the protein sequence in the FASTA format (text-based representation of the protein's sequence with the AA being represented in sequence by their single lettering code). Then selecting the canonical and altered AA, and the alteration location. The prediction is based on features that characterize the substitution, like phylogeny, its placement in the sequence, possible structural alterations induced by it, and classifies them in a probabilistic manner[55].

The results are represented by HumDiv, HumVar and the Multiple sequence alignment[55]. The latter compares different species (at most 75), regarding the protein's polypeptide chain sequence in the same position, to establish the rate of variation among them, to see if the site is conserved or not[55].

HumDiv and HumVar are predictive model algorithms which work in slightly different ways[55]. HumDiv can be used to access rare variants while HumVar has in consideration pathogenic alterations and benign SNPs with a Minor Allele Frequency (MAF)>1%[55]. Both give a prediction score for sensitivity (a damaging alteration is correctly classified, a true positive) and specificity (chance that a benign alteration is classified as such)[55]. Their results range from a score of 0.0 (benign) to 1.0 (damaging), along with probably damaging (high confidence prediction) and possibly damaging (lower confidence, but it doesn't necessarily mean that it has a milder effect) and benign predictions[55].

3.6.2 MutPred2

This tool also has in consideration missense alterations, and provides previsitions on their possible consequences. Which can range from altered mechanisms in the protein's structure and dynamics, metal and macromolecular binding, catalytic, and post transcriptional modifications[56]. Its performance is higher when compared to other platforms (self-reported)[56].

Its web server version (http://mutpred.mutdb.org/) works in the following way: after providing a valid e-mail, in the information box insert the protein's FASTA, with the alterations in its header. After submitting, the processed data and results are sent to the given e-mail. Each submitted variant is studied through 6 featured categories which include, sequence, homology conservation, or changes in structure or function[56]. After this extraction's data is assembled and worked in a neural network, which then gives two scores[56]. The general score indicates the pathogenicity of the given variant, ranging from 0 to 1, the higher the score the more pathogenic it is[56]. The property score is given to the 53 properties predicted by this tool, also ranging from 0 to 1, thus its latter score is a probability of gain or loss of the given property because of the alteration, the higher the score, the more likely is the alteration of the property[56].

3.6.3 PROVEAN

It stands for Protein Variation Effect Analyzer, this tool unlike most of other tools can also make predictions for deletions and insertions (http://provean.jcvi.org/seq_submit.php)[57,58]. It allows for the insertion of several alterations, through the insertion of the single letter AA, in its canonical version and the altered with their position in between, and the transcript for the protein, which can be retrieved from Ensembl, UniProt or NCBI RefSeq.

Its predictions have a cut-off score of -2.5. If an alteration has a value below it implies that it is damaging with a high accuracy[57]. This tool has been established as comparable to SIFT or PolyPhen2, as it consistent with the majority of the results given by those tools[57,58]. It can also study genes (http://provean.jcvi.org/genome_submit_2.php?species=human), and its possible application are still increasing[58].

3.6.4 VarSeak

This tool allows the study of splicing alterations, (https://varseak.bio/). Its use requires gene identification, its transcript, and the variant with HGVS nomenclature. It provides some information automatically (chromosome, strand, start and end positions as well as exon number and cDNA length, exon/intron, cDNA position, genomic position, rs identification). It has a scaling grade of increasing impact in splicing, of 1 up to 5. In grade 1 no splicing effect is detected, 2 is likely no splicing effect, 3 is unknown splicing effect, 4 is likely splicing effect, and grade 5 there is splicing effect predicted.

3.6.5 CADD

Combined Annotation Dependent Depletion (https://cadd.gs.washington.edu/snv), or CADD, is a tool based on machine learning that allows for the scoring for countless possible variants in the human genome[59]. It gives predictive scores about the pathogenicity of SNVs, through raw or scaled scores, we choose scalled scores as the pathogenicity of variants is more clearly seen[59]. Variant scores above 10 are in top 10% most deleterious substitutions possible in the human genome, over 20 is the top 1%, and over 30 are the top 0.1% of the most deleterious variants[59].

CADD integrates information derived from several other tools to provide its score. It accounts for allelic diversity, functionality, pathogenicity, severity of the alteration, effects on regulation, so forth[59]. This permits it to be more precise when compared to other tools[60].

3.7 Statistical analysis

Genotype phenotype association studies are a derivative of Genome-wide association studies (GWAS) in a smaller scale[61,62]. GWAS allow the identification of SNPs associated to a phenotype[61,62]. We are doing the opposite, as they test various SNPs throughout the genome in search for an association to a phenotype[61], while here, given the phenotypes, we wanted to find genotypes associated to them, through *TMPRSS6* and *SLC40A1* variants study, and establish if there was a relation between them.

In order to do so, we started by conducting a study of CBC parameters of the ID population, each parameter was tested for normality, thorough the Shapiro-Wilk test. If their distribution was normal, they would be tested with parametric tests, if not, non-parametric tests would be performed. These same statistical methods were used for the IO population and their iron biomarkers.

The U Mann-Whitney test, also known as Wilcox test for independent samples, was used for skewed hematological parameters and iron biomarkers. One way ANOVA was used for parameters and biomarkers which had normal distributions. All tests were performed with a level of significance of 0.05. For this genotype/phenotype association study we used SPSS Statistics software by IBM Corporation (launched in 2021, for Windows, version 27.0).

4 Results and Discussion

With this work, we intended to study alterations in iron metabolism caused by genetic variants in *TMPRSS6* and *SLC40A1*, and their impact on affected individuals. Thus we divided this chapter into two parts; one regards ID genetic modulation (4.1), and the other regards two IO genetic pathologies of dominant transmission (4.2).

4.1 Iron Deficit

4.1.1 Characterization of the haematological phenotype of the ID studied population

The haematological parameters of this population, adults with microcytic and/or hypochromic RBC, are depicted in Table 4.1. The mean age of this population was 50 years. Certain parameters were divided according to sex as there are biological differences due to it[31].

Table 4.1: Haematological parameters in our ID population

Units			Mean	SD	Med	Min	Max
RBC*	$x10^{12}$/L	F	4.90	0.67	4.83	**3.75**	**7.30**
		M	4.92	0.69	4.84	4.30	**6.70**
Hb*	g/dL	F	12.14	1.75	12.10	**6.50**	15.30
		M	**12.23**	1.79	**12.20**	**11**	16.7
Ht	%	F	38.03	4.96	37.70	**23.00**	**48.7**
		M	**38.26**	5.08	**37.95**	**35.50**	52.00
MCV	fL		**77.90**	6.15	**79.60**	**56.9**	85.9
MCH	pg		**24.86**	2.25	**25.7**	**16.25**	28.10
MCHC	g/dL		**31.90**	1.04	**31.80**	**28.30**	**36.10**
RDW	%		**15.51**	1.80	**15.30**	12.30	**20.30**
WBC	$x10^9$/L		7.14	2.27	6.43	4.00	**16.40**
PLAT	$x10^9$/L		263.12	62.35	255.50	153	415

Bold - values outside the reference range[31]; **SD**- standard deviation; **Med -** Median; **Min -** Minimum value in the population; **Max -** Maximum value in the population. **RBC**- Red Blood Cells**; Hb -** haemoglobin**; Ht –** haematocrit**; F –** Female**; M -** Male**. MCV -** Mean Corpuscular volume**; MCH-** Mean Corpuscular Haemoglobin**; MCHC**-Mean Corpuscular Haemoglobin Concentration**; RDW -** Red blood cell Distribution Width**; WBC -** White Blood cells**; PLAT -** Platelet count.

Most individuals in our population are women (80%), as out of the 100, only 20 are men. The sex ratio discrepancy can easily be explained as women are more prone to have ID than men[32,33]. For women their haematological parameters mean values were always within the Reference Value (RV) range. Regarding the minimum and maximum, for RBC, Hb, and Ht both values were outside the RV range. For Hb, the minimum of 6.50 g/dL is nearly half of the lower RV limit, and in Ht the minimum is also much lower than the RV (23% vs. 34.7%)[31].

For men, however, only the average value for RBC was within the normal range, as for Hb and Ht values, they were below the RV inferior limit (<13 for Hb and <39.8 for Ht)[31]. The minimum value of each parameter of the population was always under the normal range for both genders, except for the RBC value in women and RDW. As for the gendered parameter's maximum, only in RBC it surpassed for both genders, while Ht was only surpassed in the female lot.

Not all of the haematological parameters are sex dependent, so for those, the entire population was studied as a whole. Regarding MCV, MCH, and MCHC parameters, they all had an average below the RV lower limit. This was expected, attending to the haematological ID-like phenotype selection criteria of the participants. Inversely, RDW is observed above the RV, as it calculates the dispersion of

RBC sizes, the higher the value, the greater the discrepancy between RBC sizes. Iron shortages in ID result in microcytic RBC, but they can co-exist with normocytic RBC, contributing to a higher RDW.

WBC and PLAT were presented in Table 4.1 to ensure that their values were normal. Only the WBC maximum is abnormal, a leukocytosis case (WBC>10 $x10^9$/L)[31]. All haematological parameters, except WBC and PLAT (as they aren't relevant for ID study), were tested regarding their normality, through the Shapiro-Wilk test. Only Hb and Ht had a normal distribution, the rest were skewed.

4.1.2 Preparing samples for variants screening in *TMPRSS6* gene

In order to screen for *TMPRSS6* variants in our 100 samples presenting a suggestive ID phenotype, we amplified them by long range PCR in three amplicons as described in Material & Methods section. However, the third fragment had been already amplified by our colleague Daniela Santos, thus this work consisted of the first two fragments amplification, of 8266 bp and 5831 bp, respectively.

The two amplicons of each sample were jointed together in a 200 µL eppendorf prefacing a minimum volume of 10 µL per tube. Each fragment required similar band intensity in the final mix, thus similar concentrations, to ensure that NGS processed as correctly as possible. To compare the intensity of each fragment, and ensure equivalent ratios of each amplicon, all samples were analysed in a 0.5% agarose gel electrophoresis, as exemplified in Figure 4.1.

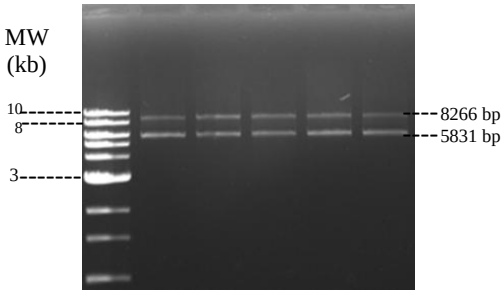

Figure 4.1: Calibration of the concentration of each long range PCR fragment by gel electrophoresis. Each well has the 2 first amplicons, the first amplicon is on top and the second amplicon is below it. Agarose gel concentration of 0.5%, the molecular ladder used was 1kb DNA Ladder from New England Biolabs Japan (measures from 10000 bp down to 500 bp).

After obtaining a gel in which both bands presented a similar intensity under UV light, consequently a similar concentration, we labelled the samples for NGS and sent it to UTI (*Unidade de Tecnologia e Inovação*) to perform purification, library preparation, cluster generation and sequencing.

4.1.3 Variants in the *TMPRSS6* gene detected by NGS

TMPRSS6 is the sole gene we studied to access the variant's impact in ID, IDA, and IRIDA development. Regarding NGS results, a total of 229 alterations were found (**Figure 4.2**)[54].

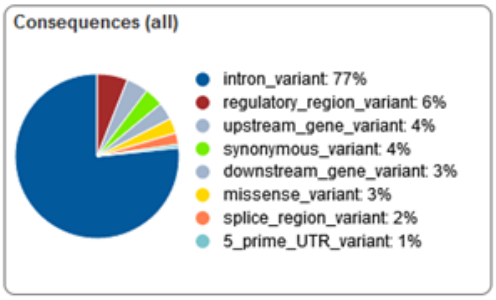

Figure 4.2: Graphic representation of the variants found in NGS. This representation was performed by Ensembl's VEP tool, which takes NGS provided VCF results, and analyses it through different perspectives[54]. This graphic encompasses the results from the first 2 *TMPRSS6* amplicons, which had 139 variants, all represented here[54]. Only 7% affect coding areas, which are 55% synonymous, and 45% missense.

Despite the large number of found alterations, not all of them corresponded to real variants, as some might be due to NGS processing errors, especially in deletions, or alterations induced in long range PCR[51,53]. We only took in consideration, alterations that occurred in coding regions and those that were 50 nucleotides (nt) upstream or downstream from the exon/intron junction, as these might affect splicing. Deep intronic variants, were unreported as they are unlikely to be associated to pathogenic effects. After applying the exclusion criteria, we were left with 36 different alterations (**Table 4.2**).

Table 4.2: *TMPRSS6* variants detected by NGS from our ID population.

ID	Position	HGVSc	HGVSp	Ref/Alt	Exon	Intron
rs11704654	22:37103346	c.72G>A	p.Pro24=	ccG/ccA	2	-
rs150521260	22:37103235	c.183G>A	p.Val61=	gtG/gtA	2	-
rs763214117	22:37098548	c.204G>T	p.Gly68=	ggG/ggT	3	-
rs732756	22:37098380	c.336+36A>G	-	A/G	-	3
rs143244650	22:37096693	c.359C>T	p.Thr120Ile	aCc/aTc	4	-
rs147700428	22:37096662	c.390C>T	p.Ser130=	tcC/tcT	4	-
rs142971835	22:37096661	c.391G>A	p.Val131Ile	Gtc/Atc	4	-
rs79013645	22:37089591	c.579A>C	p.Leu193=	ctA/ctC	5	-
rs2743824	22:37095505	c.631+46A>G	-	A/G	-	6
rs2235324	22:37089684	c.730A>G	p.Lys244Glu	Aag/Gag	7	-
rs5995378	22:37089578	c.836C>T	p.Ser279Leu	tCg/tTg	7	-
rs764379026	22:37089571	c.836+7C>T	-	C/T	-	7
rs2235326	22:37089555	c.836+23A>G	-	A/G	-	7
Novel	22:37089551	c.836+27G>C	-	G/C	-	7
rs201148397	22:37086418	c.838G>T	p.Val280Leu	Gtg/Ttg	8	-
rs113287112	22:37086249	c.973+34G>A	-	G/A	-	8
rs2111833	22:37084757	c.1056G>A	p.Ser352=	tcG/tcA	9	-
rs9610642	22:37075298	c.1197-18G>T	-	G/T	-	10
rs79816125	22:37074600	c.1441+10C>T	-	C/T	-	12
rs111807510	22:37074595	c.1441+15C>T	-	C/T	-	12
rs2072860	22:37074564	c.1441+46C>T	-	C/T	-	12
rs755795871	22:37075254	c.1223C>T	p.Pro408Leu	cCc/cTc	11	-
rs881144	22:37075250	c.1227C>T	p.Tyr409=	taC/taT	11	-
rs901290763	22:37075189	c.1288T>C	p.Ser430Pro	Tgg/Cgg	11	-
rs117576908	22:37075168	c.1309C>T	p.Arg437Trp	Cgg/Tgg	11	-
rs4820268	22:37073551	c.1536C>T	p.Asp512=	gaC/gaT	13	-
Novel	22:37071008	c.1580T>G	p.Phe527Cys	tTc/tGc	14	-
Novel	22:37071003	c.1585T>C	p.Cys529Arg	Tgt/Cgt	14	-
rs76970337	22:37070961	c.1627G>A	p.Asp543Asn	Gat/Aat	14	-
rs111813777	22:37070890	c.1672+26C>T	-	C/T	-	14
rs145814440	22:37070525	c.1800C>T	p.Asp600=	gaC/gaT	15	-
rs115310908	22:37069081	c.2105G>T	p.Arg702Leu	cGc/cTc	16	-
rs377498210	22:37069032	c.2113+41G>A	-	G/A	-	16
rs855791	22:37066896	c.2180T>C	p.Val727Ala	gTc/gCc	17	-
rs2235321	22:37066886	c.2190C>T	p.Tyr730=	taC/taT	17	-
rs73886915	22:37066170	c.2319C>T	p.Ser773=	agC/agT	18	-

ID- Reference SNP identification code; **Position-** chromosomal location according to GRCh38; **HGVSc** – coding DNA variant location; **HGVSp-** variant location in regards to the translated protein; **Ref/Alt-** comparison between the Reference and the Altered nucleotide, in coding alterations the codon is showcased, with the nucleotide implied capitalized.

Out of the 36 variants, 24 were located in coding regions, with 13 being missense and 11 synonymous. The remaining 12 were intronic but located less than 50nt away from of the intron/exon junctions. Three were novel (not described in public databases), two were missense located in close proximity in the 14th exon, and one intronic located in the 7th intron. It is important to notice that not all of these alterations are present for every individual, some of these only occurred in one individual, whereas others affected the majority of the population.

The variants were named according to HGVS rules, with the **ENST00000676104.1** transcript used for the coding DNA, and the **ENSP00000501573.1** transcript for the protein sequence[63]. It is important to clarify that in the middle of this thesis the canonical reference for this gene was altered in Ensembl. This caused a misallignment between our results and literature. The new transcript has the same number of exons and introns, but as the 27 nucleotides located in the first exon no longer account for the first nine AA of the protein, this represents a difference of nine AA in regards to the previous transcript. The functional SNP Val736Ala became Val727Ala, for example, and this is true for the protein's entire extention.

The NGS results were in the GT:AD:DP:GQ format, standing for Genotype, Allelic Depth, total Depth, and Genotype Quality, respectively. GT was given based on AD, three results were possible, 0/0, 0/1, and 1/1, which correspond to wild type, heterozygous, and homozygous for the variant, respectively. In some cases the AD given did not respect the 30:70 ratio, thus no GT would be attributed to those samples, which received the Allelic Unbalance (AuB) error result.

The AuB error appeared in some of our population's most common alterations. Some samples had various AuB while the vast majority had none, it could be that those DNA samples had poorer quality. Only eight of the 36 alterations, presented AuB. There were three coding alterations affected by AuB, p.Asp512=, p.Asp600=, p.Try730=, each with six, one, and four cases, respectively.

The 5 intronic alterations with AuB were, IVS3+36, IVS7+23, IVS8+34, IVS10-18, and IVS12+46, with one, two, four, one, and five cases each, respectively. Some AuB sample cases in coding variants were sequenced through Sanger, and were all confirmed with ratios below the 30/70 mark, like 25/75 or 20/80, so we accepted them as heterozygous. In the case of intronic variants the confirmations weren't performed, thus their AuB were all placed within the wild-type category.

4.1.4 Sanger Sequencing for NGS Variants Validation

Due to the sheer amount of data obtained via NGS, and the possibility that some of the found alterations might not be real, validation is required, to confirm their veracity. Sanger sequencing was only performed for the coding variants.

As some of the intronic variants found, were located close to the primers that would used for their sequencing, their viewing and subsequent confirmation in the Sanger's electropherograms wouldn't be possible. Due to their proximity to the start of the fragment by the primers, and overlaping with the background noise, which makes it harder to discern between the peaks of interess from unwanted noise peaks.

For the confirmation of each variant, their corresponding exon was amplified by simple PCR, followed by purification and Sanger sequencing. In Figure 4.3 are shown 11 PCR fragments prior to purification and sequencing by Sanger. In the shown preparation samples from seven exons were amplified: two samples of exon 2 (395bp), one sample of exon 3 (323bp), three samples of exon 4 (301bp), one sample of exon 5 (378 bp), two samples of exon 7 (584bp), one sample of exon 8 (364 bp) and one sample of exon 9 (458bp).

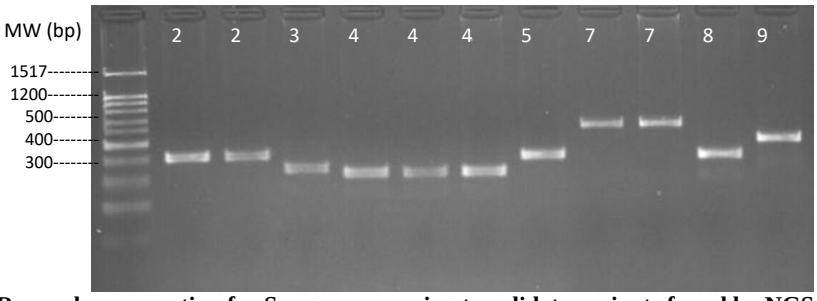

Figure 4.3: PCR sample preparation for Sanger sequencing to validate variants found by NGS. The numbers below the wells represent each specific exon of *TMPRSS6* gene. This gel has 2% agarose, the molecular ladder used was 100 bp DNA Ladder from New England Biolabs Japan (measures from 1500bp down to 100bp), and it was run under 70V for 40 minutes.

The particular variants prepared for validation showcased previously were: c.72G>A (p.Pro24=) and c.183G>T (p.Val61=) present in exon 2; c.204G>T (p.Gly68=) in exon 3; in exon 4, c.359C>T (p.Thr120Ile), c.390C>T (p.Ser130=) and c.391G>A (p.Val131Ile); c.579A>C (p.Leu193=) in exon 5; c.730A>G (p.Lys244Glu) and c.836C>T (p.Ser279Leu) in exon 7; in exon 8 c.838G>T (p.Val280Leu); c.1056G>A (p.Ser352=), all in their respective order in the gel.

Every coding alteration given by NGS was confirmed via Sanger sequencing. The majority of the samples used were heterozygous, as they are easier to spot than a homozygous alteration. In Figure 4.4 are represented some electropherograms of variants in heterozygous individuals, with the codon with the alteration circled in black.

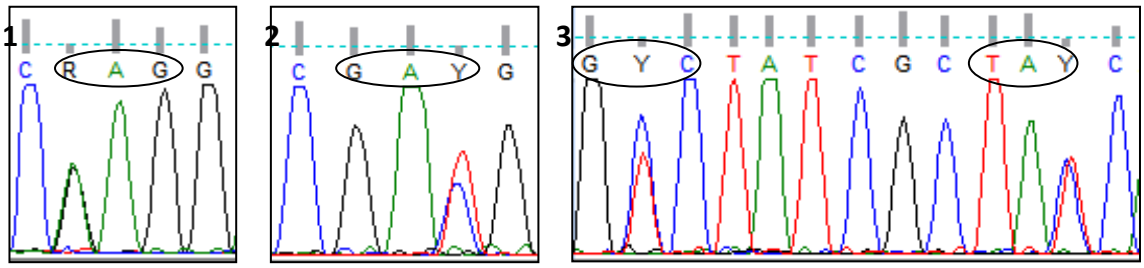

Figure 4.4: Electropherograms of Sanger confirmation of variants. The IUPAC nucleotide code R stands for adenine or guanine, and Y stands for thymine or cytosine. Within the black circles are the alterations. **1)** Heterozygous alteration in sample 66 for c.730A>G (p.Lys244Glu); **2)** Validation of the heterozygosity for the c.1536C>T (p.Asp512=) in sample 69. **3)** Validation of the double heterozygosity for c.2180T>C (p.Val727Ala) and c.2190C>T (p.Tyr730=), in sample 70.

In all variants we chose to show the electropherogram of heterozygous individuals. However, the variant c.836C>T (p.Ser279Leu) only had homozygous individuals with it. In such cases, the development of an algorithm for their quicker detection was ideal. The employed algorithm was a code that compared the canonical exon sequence to the sequence given by our sequenced samples. After applying this code into R, it automatically identified the differences between sequences, improving the detection efficiency enormously.

4.1.5 Variant Analyses

Given the size of our population, we decided to focus more on the alterations, rather than each individual; however, in certain cases the individuals are discussed in more detail. We have chosen to study the alterations, by separating them according to their pathogenicity. To determine to which group each variant belonged to we used *in silico* tools.

Missense alterations contribute to the most serious alterations in the protein caused by a single nucleotide variant (SNV), which causes the switch from one AA to another, thus it can have a major impact on the protein's viability depending on how drastic the alteration is. Synonymous alterations are SNV that, on the other hand, maintain the same AA in the protein.

The missense variants were assessed mainly with three different *in silico* tools, PolyPhen-2[55], MutPred2[56], and PROVEAN[57]. Besides these tools, data bases, such as NCBI's ClinVar, Ensembl, were also used. They were also studied in regards to the possible impact of the switch in AA, from a biochemical point of view[64]. As depending on the AA replaced, the impacts of the alteration might be negligible or severe.

Another factor we took in consideration was how prevalent were the variants within the population, as if it were a functional SNP, they would appear more frequently. Ensembl has a feature that showcases the variant allelic presence in different populations. That data was mostly obtained from the 1000 Genomes Project Phase 3 allele frequencies (https://www.internationalgenome.org/), giving, for each population present in that study, a Minor Allele Frequency (MAF)[65]. Since the Portuguese population hasn't included in that study, we choose to compare our data to the closest geographic population to ours, the Iberian Spaniard Population (IBS)[65,66]. In the cases of unreported variants in the IBS population, we addressed to which part of the globe it appeared more frequently.

However in rare variants in which no data from the 1000 Genomes Project was available, we used the data regarding the Genome Aggregation Database (gnomAD)[67], also available in Ensembl. This database (https://gnomad.broadinstitute.org/) can also used as a standalone tool. Although gnomAD is more comprehensive and provides more information than the 1000 Genomes Project, the main reason why we did not use it as a primary MAF source, lays on the way it studied populations. In gnomAD the European (non-Finnish) populations studied were grouped in six classes: Swedish, Estonian, Bulgarian, Southern, North-western and other non-Finnish Europeans. With its broader classes it makes harder to discern our population in it.

Some frequent genetic variants were also studied from a statistical perspective. We compared individuals with the variants to those without, in regards to the haematological parameters differences.

4.1.5.1 Pathogenic Alterations

Out of the 24 coding alterations found in the *TMPRSS6* gene, six were classified as being probably pathogenic (**Table 4.3**).

Table 4.3: Probably pathogenic variants in *TMPRSS6*.

ID	HGVSp	HGVSc	Exon	Ref	Alt	CADD	HT	HM
rs143244650	p.Thr120Ile	c.359C>T	4	aCc	aTc	24.1	1	0
rs117576908	p.Arg437Trp	c.1309C>T	11	Cgg	Tgg	24.7	3	0
Novel	p.Phe527Cys	c.1580T>G	14	tTc	tGc	29.0	1	0
Novel	p.Cys529Arg	c.1585T>C	14	Tgt	Cgt	29.4	1	0
rs76970337	p.Asp543Asn	c.1627G>A	14	Gat	Aat	24.9	1	0
rs115310908	p.Arg702Leu	c.2105G>T	16	cGc	cTc	22.2	1	0

ID- Reference SNP identification code; **HGVSp**- variant location in regards to the translated protein; **HGVSc** – coding DNA variant location; **Ref** and **Alt**- stand for Reference and Alteration respectively, the nucleotide impacted in the codon is capitalized; **CADD**- Represents the PHRED score for each variant; **HT**- Heterozygote, represents the number of individuals with both the wild type and the alteration; **HM**-Homozygote, stands for the number of individuals with both alleles altered.

As these variants affected very few people, some more information about the affected individuals is shown, along with the information regarding each alteration.

4.1.5.1.1 p.Thr120Ile

In this missense alteration (c.359C>T, rs143244650), threonine is replaced by isoleucine in the 120th position of the protein, it was found in only one heterozygous individual (**Table 4.4**).

Table 4.4: Detailed information regarding the heterozygous individual affected by p.Thr120Ile.

		RBC	Hb	Ht	MCV	MCH	MCHC	RDW	AD	DP
ID	Sex	$x10^{12}$/L	g/dL	%	fL	pg	g/dL	%	Ref, Alt	
12	F	4.87	13	39.8	81.7	**26.7**	32.7	12.3	2046,2786	4833

Bold - values outside the reference range[31] or our criteria; **ID-** is the number of the individual; **RBC-** Red Blood Cells**; Hb -** haemoglobin**; Ht –** haematocrit**; MCV -** Mean Corpuscular volume**; MCH-** Mean Corpuscular Haemoglobin**; MCHC-**Mean Corpuscular Haemoglobin Concentration**; RDW -** Red blood cell Distribution Width**; AD-** Allelic Depth, **Ref** and **Alt-** stand for Reference and Alteration respectively; **DP-** total Depth. The Genotype quality was 99%.

Most haematological parameters in this individual were within the normal RV range, with only MCH being lower than the criteria. This MCH value is, however higher than the mean for our population, thus this phenotype isn't so serious.

This individual also presented the following variants: IVS6+46 (0/1), Lys244Glu (0/1), IVS7+23 (1/1), Ser352= (0/1), IVS12+15 (0/1), IVS12+46 (1/1), Asp512= (1/1), IVS14+26 (0/1), Val727Ala (1/1), some of these are considered to be functional SNPs.

This variant has not been reported on ClinVar yet. It has a MAF<0.01 globally, and it was undetected in the IBS population. Only two distinct African populations, the African Caribbean in Barbados and the Lunya of Kenya, had individuals with this variant, each with only one individual showcasing it. The MAF for our population is 0.5%. Since this variant was only found in individuals of African ancestry in literature[65], we decided to check the ancestry of this individual. This particular individual wasn't born in Portugal, so we might speculate that she was born in or at least has African ancestry.

Threonine is an AA with an neutral-polar side chain, and is replaced by isoleucine, that has a hydrophobic side chain[64]. Isoleucine is one of the most hydrophobic, thus apolar AA, whereas Threonine is fairly neutral and slightly smaller in size, as it has a shorted aliphatic chain[64]. Regarding the analyses with *in silico* tools they are summarised in Table 4.5.

Table 4.5: *In silico* analyses performed for p.Thr120Ile.

Thr120Ile	CADD	PROVEAN	MutPred2	HumDiv	HumVar
Score	24.1	-2.999	0.227	0.999	0.966
Prediction	Top 1%	deleterious	*	Prob D	Prob D

Prob D- Probably Damaging; *- more detailed information in the text.

CADD scored it among the top 1% most deleterious variants. MutPred2 didn't predict any considerable effects associated to it. The Effects predicted through PolyPhen-2 (HumDiv and HumVar) can be seen below in Figure 4.5. According to its Multiple Sequence Alignment (MSA), we could state that this is a highly conversed site, as the comparison between 67 species, only eight had a different AA (Serine) and none of these species was mammalian. Such high predictive score values indicate that this is might be a highly damaging alteration.

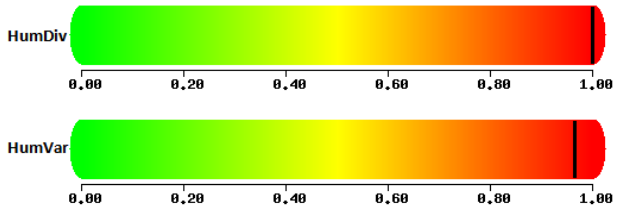

Figure 4.5: PolyPhen-2 heat bar for p.Thr120Ile. For parameters predicted it to be probably damaging, with sensitivity: **0.14** and specificity: **0.99** for HumDiv, and with sensitivity: **0.61** and specificity: **0.93** for HumVar.

This alteration takes place in the SEA domain of matriptase-2 (MT2)[68]. This domain is important for MT2 trafficking from the cytosol to the cell's membrane, where it performs its cleaving activity[68].

If this variant's impact is as similar to the alteration in the close position reported by McDonald[68] (Try141Cys, position 132 in the current nomenclature), it means that the protein's activity might be compromised[68]. As the protein would fail to be brought up to the surface of the hepatocytes, and placed between the phospholipidic layers of the cell[68]. This would consequently mean that the capacity to inhibit hepcidin, throught HJV cleavage, would be compromised, as this activity happens extracellularly, and that mutation causes MT2 to fail to be transported to the surface of the cell[68]. Thus this variant could have an impact in hepcidin reduction, creating an IRIDA like phenotype[20,37].

4.1.5.1.2 p.Arg437Trp

In this alteration (c.1309C>T, rs117576908), arginine is replaced by tryptophan in 437th residue of MT2, this alteration was found in three heterozygous individuals (**Table 4.6**).

Table 4.6: Detailed information regarding the heterozygous individual affected by p.Arg437Trp.

		RBC	**Hb**	**Ht**	**MCV**	**MCH**	**MCHC**	**RDW**	**AD**	**DP**
ID	Sex	x10^{12}/L	g/dL	%	fL	pg	g/dL	%	Ref, Alt	
86	F	**3.75**	**8.6**	**28.3**	**75.4**	**22.9**	**30.3**	**19.1**	153,142	295
113	M	4.9	**11.1**	**35.5**	**72.5**	**22.7**	**31.4**	**17.3**	82,100	182
129	F	**5.6**	14.4	44.8	**79.5**	**25.6**	32.2	**17.1**	34,52	86

Bold - values outside the reference range[31] or our established criteria; **ID-** is the number of the individual; **RBC-** Red Blood Cells**; Hb -** haemoglobin**; Ht –** haematocrit**; MCV -** Mean Corpuscular volume**; MCH-** Mean Corpuscular Haemoglobin**; MCHC-**Mean Corpuscular Haemoglobin Concentration**; RDW -** Red blood cell Distribution Width**; AD-** Allelic Depth, **Ref** and **Alt-** stand for Reference and Alteration respectively; **DP-** total Depth. The Genotype quality was 99%.

In regards to haematological parameters, ID86 and ID113 have severe presentations of the IDA phenotype, while ID129 only has an evidently low MCH. All of them had lower MCV and MCH levels when compared to the rest of the population, and conversely, they had higher RDW, whose normal values are between 11.5-15%[31].

In our population RDW has a skewed distribution, with the majority of individuals having their values below the mean value for this population (15.51 ± 1.80%). The individuals with this variant however, had a much higher percentage, mean of 17.83%. This indicates an abnormally high discrepancy between their RBC sizes, which is suggestive of iron deficiency anaemia.

The three individuals were homozygous for IVS12+46, Asp512=, and Val727Ala; Try730= was only heterozygous for ID12, the others had it as homozygous. ID129 with the milder phenotype, besides those variants, only had homozygous alterations, for the following variants, ivs6+46, Lys244Glu, IVS7+23, Ser352=. All of these variants, except for IVS7+23, were also presented by individual ID86 but as heterozygote. Although they possessed the same variants, except for IVS-7+23, ID129 always presented their shared variants as a homozygous instead of heterozygous. Being homozygous for those variants, might have had a protective effect, decreasing the severity of her phenotype.

The ID phenotype in these individuals is quite apparent, however given their gravity it was important to access if they had comorbidities. Namely, individual ID86 with the most severe ID phenotype, which could be explained by the co-existence of other pathogenic factors. Regarding comorbidities, only ID113 had a haemoglobinopathy due to an alteration on *HBB*, causing Haemoglobin D. The presence of this alteration might help explain the severity of his phenotype. Variants that were only presented by him were p.Pro24= and IVS3+36, in both heterozygous, and he had none of variants shared by the two females.

Another perspective to take in consideration, as ID86 and ID129 are both females, is the menstrual cycle, as it increases the likability of ID development due to blood loss[10,11,69]. The Val727Ala SNP has been implied in ID severity in TT homozygous women, but both individuals possess the CC variant which is protective[69]. So, depending on their age, they might be either pre or post-menopausal, which could explain the differences in severity. In this case, ID86 is a 48 years old and while ID129 is 58 years old. A study establishing the menopausal transition in Southern European women took in consideration the Portuguese population[70]. Their sample consisted of a 1003 Portuguese women with ages between 40 and 65, with the average age of 49.67 years old, with 57.8% reporting being pre-menopausal[70]. Thus it is quite likely that ID86 is still experiencing her menstrual cycle, as she is younger than the mean in the previous mentioned study. While ID129 is most likely in post-menopausal already, explaining her mitigated phenotype in regards to ID86. Arg437Trp

Although Arg437Trp is unreported in the IBS population, it is present in other European populations, having an overall MAF of 1.2%. The Finnish population presented it most frequently, with a 4.5% MAF. In our sample, the MAF was 1.5%, which is in line with the European MAF[65]. The three individuals with this variant were all born in Portugal. In regards to the *in silico* analyses performed, their results can be seen in Table 4.7.

Table 4.7: *In silico* analyses performed for p.Arg437Trp.

Arg437Trp	**CADD**	**PROVEAN**	**MutPred2**	**HumDiv**	**HumVar**
Score	24.7	-1.487	0.319	0.876	0.153
Prediction	Top 1%	neutral	*	Poss D	Benign

Poss D- Possibly Damaging; *- more detailed information in the text.

MutPred2 did not predict consequences for it. PolyPhen-2 gave different predictions for this variant (**Figure 4.6**). HumDiv predicted it to be possibly damaging, while HumVar predicted it to be benign. The difference between HumDiv and HumVar in this variant can be due to the human canonic AA, associated to this position in the protein, being different from the majority of the 68 species whose sequence was used in the MSA. The common AA for this site was Glutamine, present in 64 species, followed by Arginine, present in Humans, Sumatran orangutans and naked mole rats.

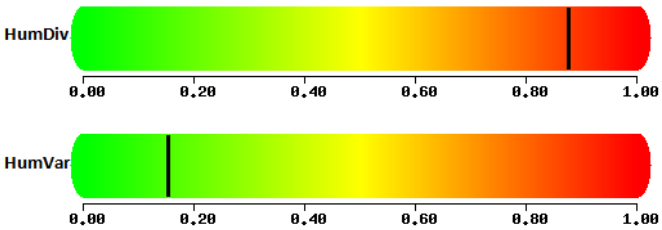

Figure 4.6: PolyPhen-2 heat bar for p.Arg437Trp. This mutation has a HumDiv score of **0.876** (sensitivity: **0.83**; specificity: **0.93**), and a HumVar score of **0.153** (sensitivity: **0.89**; specificity: **0.72**).

Regarding change in the AA, this variant causes arginine to be replaced by tryptophan. Arginine is a positively charged AA, with longest R side chain and also having the highest isoelectric point (10.76), by far, out of all AA[64]. Tryptophan is an aromatic AA comprised by an indole ring. This ring is a bicyclic structure (C_8H_6N) formed by pyrrole (a 5 pointed cyclic structure, $C_4H_4NCH_3$) fused with a benzene (six pointed cyclic structure, C_6H_6), so it a quite voluminous structure[64]. This missense alteration leads to an alteration on the stereochemical environment in the proximity of this position. Although it is not as hydrophobic as other apolar AA, tryptophan is the most voluminous AA, displacing, by replacement, the longest and most polar AA[64]. This increased space occupied might create disruptions in the other AA around this position.

Arg437Trp has been described as being overrepresented in anaemic populations[71]. When using the mice model affected by this variant, the corresponding MT2 could suppress HJV, but slightly less efficiently than wild type[71]. That study suggested that this variant could partially cause MT2 function loss, through the reduction of MT2 cleaving activity[71]. Although this variant, in the heterozygous state, might not produce the phenotype, if associated with other variants, especially if pathogenic, might conduce to IDA or a mild IRIDA like case[71].

In ClinVar, Arg437Trp has been described as likely benign, and Ensembl states that it has an uncertain significance. We have decided to discuss it in this section of "Pathogenic alterations" due to the severity of the phenotype shown by the individuals, as well as its HumDiv score. Its score was much higher than the scores for the variants included in the Variants of Uncertain Significance section (further ahead). Concerning all the information above explained, we can conclude that the pathogenicity of this variant remains to be unveiled. Thus, functional studies should be performed to discover its effects.

4.1.5.1.3 p.Asp543Asn

This missense variant (rs76970337) occurred once in our population, in a heterozygous individual (**Table 4.8**). It is caused by an alteration c.1627G>A, from a guanine to an adenine.

Table 4.8: Detailed information regarding the heterozygous individual affected by p.Asp543Asn.

		RBC	Hb	Ht	MCV	MCH	MCHC	RDW	AD	DP
ID	Sex	$x10^{12}$/L	g/dL	%	fL	Pg	g/dL	%	Ref, Alt	
109	M	**6.4**	16.4	51.2	**79.7**	**25.5**	**32**	14	56,57	113

Bold - values outside the reference range[31]; **ID-** is the number of the individual; **RBC-** Red Blood Cells**; Hb -** haemoglobin**; Ht –** haematocrit**; MCV -** Mean Corpuscular volume**; MCH-** Mean Corpuscular Haemoglobin**; MCHC-**Mean Corpuscular Haemoglobin Concentration**; RDW -** Red blood cell Distribution Width**; AD-** Allelic Depth, **Ref** and **Alt-** stand for Reference and Alteration respectively; **DP-** total Depth. The Genotype quality was 99%.

This individual, interestingly, has Hb in the upper limit of RV range, it also is on the top RV for RBC, Ht and nearly also for RDW, it is on the bottom for MCHC. He presents both microcytic and hypochromic RBC. Other genetic variants were presented by this individual, as a heterozygote (Pro24=, IVS6+46, Lys244Glu, Ser352=, Try730=) and as homozygote (IVS7+23, Asp512=). This individual has some of the most well known functional SNPs of MT2[72]; however he doesn't present the common alteration Val727Ala.

In ClinVar p.Asp543Asn is reported to be Benign. It is unreported in all but African populations[65], with the Yoruba of Nigeria having the highest prevalence with a 2.8% MAF. In our sample its MAF is 0.5%, and the individual 109 wasn't born in Portugal, so he might be of West African origin.

Aspartic acid is a negatively charged AA, and it is among the most polar AA, aspargine on the other hand is a polar uncharged AA whose amide group contributes to its polarity[64]. This substitution occurs in the LDLR_A3 domain of MT2[37]. The bioinformatics predictions can be seen in Table 4.9.

Table 4.9: *In silico* analyses performed for p.Asp543Asn.

Asp543Asn	CADD	PROVEAN	MutPred2	HumDiv	HumVar
Score	24.9	-4.525	0.191	1.000	0.999
Prediction	Top 1%	Deleterious	*	Prob D	Prob D

Prob D- Probably Damaging; *- more detailed information in the text.

MutPred2 didn't attribute it with significant consequences. Probably damaging was the prediction made by both of PolyPhen-2 predictors (**Table 4.9** and **Figure 4.7**), as in its MSA all of the 67 species present had Aspartic Acid in this position, demarking an extremely conserved site.

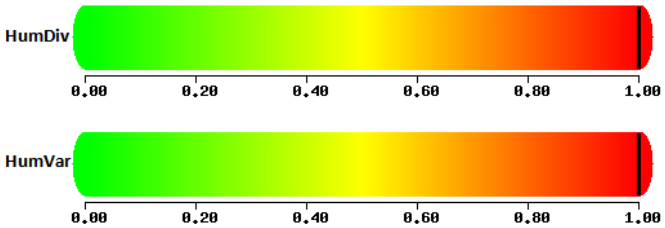

Figure 4.7: PolyPhen-2 heat bar for p.Asp543Asn. Both HumDiv and HumVar have predicted this variany to be probably damaging, with sensitivities of 0.00 and 0.09, and specificities of 1.00 and 0.99, respectively.

We could only find one article that mentioned this variant[73], in a sample of 21 individuals living in the Netherlands with an IRIDA like phenotype, it affected a 44 year old heterozygous women with low Hb and MCV (7.9 g/dL and 62fL, respectively), 10 µg/L of ferritin and 2.7% TS. She was the only individual that had to receive blood transfusions as treatment[73]. They also performed *in silico* studies for this variant, namely Sorting Intolerant from Tolerant (SIFT) which gave it a score of 0.00, which means that the variant was predicted to be damaging (for variants with a score below 0.05)[73,74]. This study also conducted genetic screening for relatives of the enrolled subjects[73]. The propositus only had a sister, and although both shared this variant, her sister did not present the same phenotype[73]. This highlights complexity of the genetics contributing to IDA or IRIDA inheritance.

Another variant, involving the same AA, p.Asp521Asn (c.1534G>A, rs137853120, current nomenclature p.Asp512Asn) has been reported[75]. This variant is located in the LDLR_A2 domain, and has already been implicated in IRIDA, causing MCV, MCH, and RDW modulation[75].

This individual presents both microcytic and hypochromic RBC without anaemia, as he has high levels of Hb and RBC, this indicates that he doesn't have standard IDA[76]. In order to validate the pathogenicity of this variant, further *in vivo* or *ex vivo* studies should be performed.

4.1.5.1.4 p.Arg702Leu

This missense variant (c.2105G>T, rs115310908) takes place in the 16th exon, and it affected one heterozygous individual (**Table 4.10**).

Table 4.10: Detailed information regarding the heterozygous individual affected by p.Arg702Leu.

		RBC	Hb	Ht	MCV	MCH	MCHC	RDW	AD	DP
ID	Sex	x10^{12}/L	g/dL	%	fL	pg	g/dL	%	Ref, Alt	
46	F	4.35	**10.9**	**34.6**	**79.5**	**25.2**	**31.7**	14.9	16,7	23

Bold - values outside the reference range[31]; **ID**- number of the individual; **RBC**- Red Blood Cells**; Hb** - haemoglobin**; Ht** – haematocrit**; MCV** - Mean Corpuscular volume**; MCH**- Mean Corpuscular Haemoglobin**; MCHC**-Mean Corpuscular Haemoglobin Concentration**; RDW** - Red blood cell Distribution Width**; AD**- Allelic Depth, **Ref** and **Alt**- Reference and Alteration respectively; **DP**- total Depth. The Genotype quality was 99%.

This alteration involves the replacement of arginine for leucine in the 702th position of the protein, and it occurs within the Serine protease domain of MT2. This variant was detected by NGS with a low AD and DP, but it was validated by Sanger sequencing.

As previously mentioned, arginine is an AA with a positive charge and it is among the largest AA. On the polar opposite we have leucine, which is a considerably smaller AA, as well as being one of the most hydrophobic AA[64]. The difference between these AA makes this alteration rather disruptive as they are very different from one another.

In ClinVar it is stated as likely benign. It is unreported in European populations, only appearing in African populations, with the Esan of Nigeria presenting it most frequently with a 3.5% MAF. In our population its MAF was 0.5%, and ID46 wasn't born in Portugal, so it is possible that she has African ancestry. In regards to its *in silico* analyses, they are summarised in Table 4.11.

Table 4.11: *In silico* analyses performed for p.Arg702Leu.

Arg702Leu	CADD	PROVEAN	MutPred2	HumDiv	HumVar
Score	22.2	-3.898	0.589	0.820	0.385
Prediction	Top 1%	Deleterious	*	Poss D	Benign

Poss D- Possibly Damaging; *- more detailed information in the text.

Unlike the previous variants, MutPred2 has predicted this alteration to cause several significant molecular consequences for MT2 structure and functioning[56], with a p-value below 0.05. It causes an order alteration in the protein's interface (probability of 0.33, $p=8.0x10^{-3}$), an altered transmembrane function (probability of 0.23, $p=2.3x10^{-3}$), and loss of strand (probability of 0.28, p=0.01). A loop gain (probability of 0.27, p=0.03), relative increase in solvent accessibility (probability of 0.26, p=0.03), and a new catalytic site at Trp698 (probability of 0.24, $p=5.5x10^{-3}$). Implying the high level of negative consequences associated to the shift of AA in this variant.

PolyPhen-2 has given two different previsions, in the HumDiv and HumVar scores (**Table 4.11** and **Figure 4.8**), this can be explained by having the 69 species MSA in consideration. There was considerable AA variation in this protein site. Besides Arginine other AA included Glutamine (ten species), and only Rhesus macaque had Serine. As it is highly variable site, the likability of a pathogenic change is predicted to be lower, even if it is not the case, generally speaking.

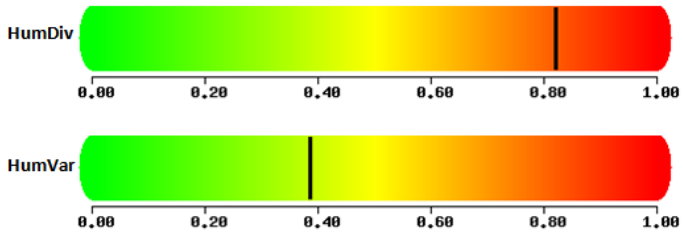

Figure 4.8: PolyPhen-2 heat bar for p.Arg702Leu. HumDiv had a **0.84** sensitivity and **0.93** specificity, while HumVar score of **0.385** (sensitivity: **0.85**; specificity: **0.79**).

Despite ClinVar and Ensembl classify this alteration as benign or likely benign, the consensus given by all *in silico* tools used was that it is a possibly damaging variant instead. Thus, due to the distinct classification of its pathogenicity, this variant should be evaluated by functional studies.

4.1.5.1.5 Novel Variants

These two novel variants, c.1580T>G and c.1585T>C (p.Phe527Cys and p.Cys529Arg, respectively), are missense alterations. They are located in 14th exon of the *TMPRSS6* gene, in the third low-density-lipoprotein receptor class A domain, LDL_A3, of the protein[37].

Due to their extreme proximity, being only separated by one AA in the protein, they are here addressed together. Each variant was found in a single heterozygous individual (**Table 4.12**).

Table 4.12: Detailed information regarding the heterozygous individuals affected by the novel variants

			RBC	Hb	Ht	MCV	MCH	MCHC	RDW	AD	DP
Variants	ID	Sex	$x10^{12}$/L	g/dL	%	fL	pg	g/dL	%	Ref, Alt	
Phe527Cys	38	F	4.97	12.8	38.7	**77.9**	**25.8**	33.1	**18.4**	49,59	108
Cys529Arg	80	M	5.99	15.6	48.5	81	**26.1**	32.2	14.2	34,47	81

Bold - values outside the reference range[31]; **ID**- number of the individual; **RBC**- Red Blood Cells**; Hb** - haemoglobin**; Ht** – haematocrit**; MCV** - Mean Corpuscular volume**; MCH**- Mean Corpuscular Haemoglobin**; MCHC**-Mean Corpuscular Haemoglobin Concentration**; RDW** - Red blood cell Distribution Width**; AD**- Allelic Depth, **Ref** and **Alt**- stand for Reference and Alteration respectively; **DP**- total Depth.

After receiving the NGS results and realizing that no rs ID were associated to these variants, we proceeded to confirm them via Sanger sequencing (**Figure 4.9**).

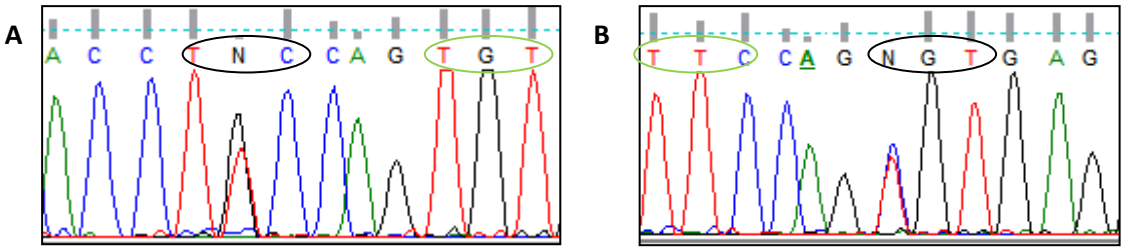

Figure 4.9: Sanger confirmation of the Novel alterations found. A) c.1580T>G (p.Phe527Cys, tTc/tGc); **B)** c.1580T>C (p.Cys529Arg, Tgt/Cgt). Both are in heterozygous individuals. Black circle: altered codon; Green circle: reference codon.

After having their existence confirmed through Sanger sequencing, we ensued with *in silico* studies, to determine their possible pathogenicity. Since these are new and unreported variants, some additional bioinformatics tools were used for their study, which are mentioned further ahead.

Since these are unreported alterations thus far, no MAF has been established for them. But having in consideration that they are only being found now, we have to assume that are rare, with a MAF inferior to 1%, in our population it was 0.5%, for each.

In regards to the alterations from a biochemical point of view, both involve cysteine, which is a very special AA. Cysteine has a thiol group on its end, capable of forming covalent disulphide bonds[64]. These bridges between different cysteine residues in the polypeptide chain are very hydrophobic and serve a structural role in proteins, whose importance can be felt up to the tertiary and quaternary protein structures[64]. So Cys involving alterations can certainly cause several problems for a protein[64].

In p.Phe527Cys, a phenylalanine is replaced by a cysteine. Phenylalanine is an aromatic AA, which means that it possesses the aromatic benzyl group[64]. It is the simplest aromatic AA, and the most hydrophobic. Despite Cys being considered a polar uncharged AA, it is hydrophobic to a similar extent to Phe, however, the latter is far more stable. This replacement has a rather big impact, because Phe is "inert" for the most part while Cys is capable of forming a disulphide covalent bond with another Cys residue, which wasn't meant to occur or form weak hydrogen bonds to oxygen or nitrogen in proximity[64]. The formation of the disulphide bridges is by far the most severe consequence of this alterations as it can cause disturbances in the macromolecular level of the protein and not only on the polypeptide chain[64].

In p.Cys529Arg, cysteine is replaced by arginine, which is the most hydrophilic AA, by being the positively charged with a long aliphatic chain[64]. This replacement is more severe than the previous as here a potential disulphide bond is lost and because the hydrophobicity between these AA is completely opposite[64], unlike p.Phe527Cys in which both AA were relatively apolar. Matriptase-2 has 38 conserved cysteine residues, with three pairs bonding in the catalytic protease domain[77]. Given their location in the 14th exon, p.Cys529Arg is not one of those residues as the Serine protease domain starts in the 15th exon[37,77].

These alterations were tested thoroughly with *in silico* tools (**Table 4.13**). The images below are from the tool Missense3D (http://missense3d.bc.ic.ac.uk/missense3d/) (**Figure 4.10**)[78]. This tool besides giving the structural alterations caused by the variants, also predicts their consequences[78].

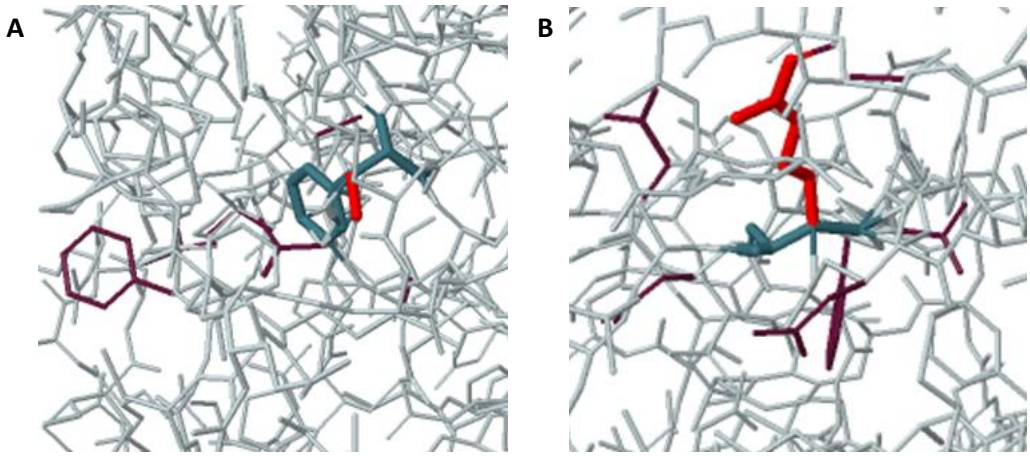

Figure 4.10: Missense3D structural predictions. A) p.Phe527Cys; **B)** p.Cys529Arg. In red and green are the altered and canonic AA, respectively, in purple are possible altered chain interactions with other residues in proximity.

For p.Phe527Cys, Missense3D detected no structural damages, while for p.Cys529Arg the detected damages included cavity alteration, and the introduction of a hydrophilic buried charge[78]. These inner structures tend to be apolar and hydrophobic, so these damages are due to replacement of cysteine by arginine[64,78]. Since arginine is hydrophilic, it causes disruption in the inner side of the protein, the cavity, due to its positive charge, as it can promote a soft denaturation of the site by attracting water into the cavity[79].

Table 4.13: *In silico* analyses performed for the novel variants.

		CADD	PROVEAN	MutPred2	HumDiv	HumVar
Phe527Cys	Score	29.0	-5.795	0.695	0.993	0.870
	Prediction	*	Deleterious	*	Prob D	Poss D
Cys529Arg	Score	29.4	-10.846	0.817	1.000	0.999
	Prediction	*	Deleterious	*	Prob D	Prob D

Prob D- Probably Damaging; *- more detailed information in the text.

When testing both variants through CADD, both had PHRED scores close to 30, scores above 30 mean that the variant ranks among the top 0.1% most deleterious variants in the human genome[59]. As their values are quite close to 30, we can only assume that these novel variants are extremely damaging for the protein, greatly impairing its function.

MutPred2 predicted p.Phe527Cys with several molecular mechanisms consequences; a gain of disulphide bond with Cys522 (probability of 0.33, $p=2.5x10^{-4}$), a Loop Loss (probability of 0.27, p=0.03), and a gain of ADP-ribosylation at Arg532 (probability of 0.19, p=0.05).

For p.Cys529Arg MutPred2 predicted a loss of disulphide bond with the Cys529 (due to the loss of its cysteine) (probability of 0.37, $p=6.6x10^{-4}$), a Loop gain (probability of 0.27, p=0.04), and an altered transmembranar protein (probability of 0.12, p=0.03).

These predictions by MutPred-2 are in line with the biochemical analysis we did from the AA point of view[56,64]. With the gain (in the case of p.Phe527Cys) and loss (in the case of p.Cys529Arg) of disulphide bonds. Along to other unforeseen consequences, like ADP-ribosylation in p.Phe527Cys, which allows a reversible post-transcriptional alteration to the protein which is inexistent with Phe[56,80].

Out of all alterations tested through PROVEAN, p.Cys529Arg had by far the lowest score, therefore being the most deleterious alteration in our population, according to this tool[57]. PolyPhen-2 HumDiv considered both alterations to be probably damaging (**Table 4.13** and **Figure 4.11**). HumVar predicted p.Phe527Cys to be possibly damaging with a sensitivity of 0.72 and a specificity of 0.89, while it was probably damaging for p.Cys529Arg with a sensitivity of 0.09 and specificity of 0.99.

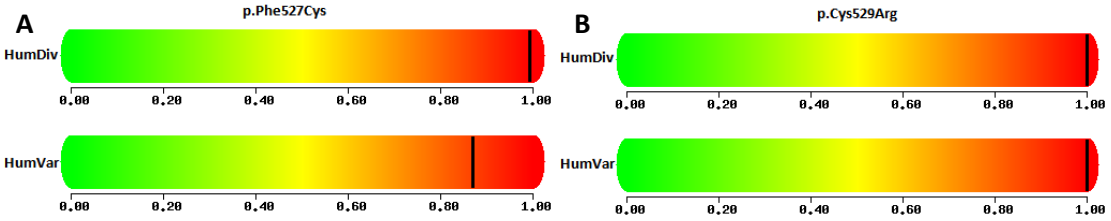

Figure 4.11: PolyPhen-2 heat bar for the novel variants. A) p.Phe527Cys had a HumDiv sensitivity of **0.70** and specificity of **0.97**. **B)** p.Cys529Arg had a HumDiv sensitivity of **0.00** and specificity of **1.00**.

These scores indicate that the alterations might induce damage to the protein's structure. Regarding the MSA of p.Phe527Cys (**Figure 4.12**, within the black box), this position is quite conserved among the 69 species, the human reference AA, Phenylalanine is prevalent in the majority of species. Tyrosine is also present in this position of the protein for 17 species, while other three had histidine, it is possible to see within the black rectangle, in Figure 4.12, Y (tyrosine) in the Brazilian opossum. This variation in the AA between the different species might have contributed to the lower HumVar value of p.Phe527Cys (0.870 vs. 0.999).

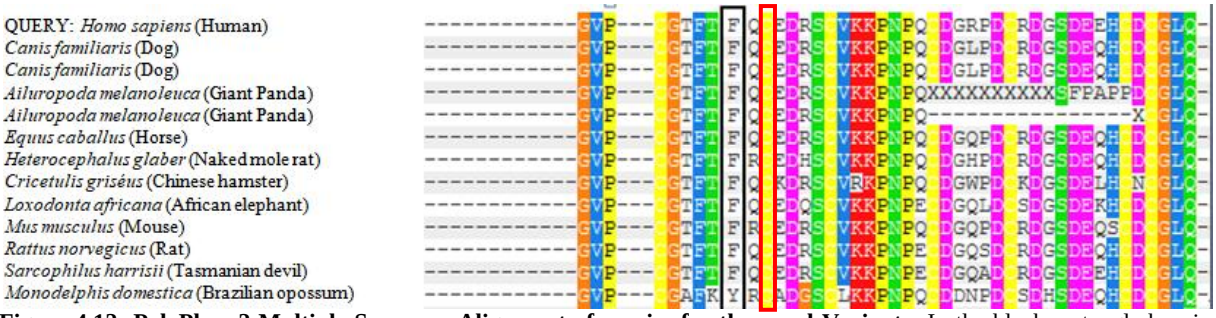

Figure 4.12: PolyPhen-2 Multiple Sequence Alignment of species for the novel Variants. In the black rectangle box is the p.Phe527Cys alteration, while in the red rectangle box is p.Cys529Arg.

Due to the proximity of both alterations, we can also see p.Cys529Arg in Figure 4.12, one column apart, signalled by the red rectangle box (C for cysteine is written in white but it isn't visible due to the yellow background). Regarding p.Cys529Arg MSA, this position is extremely conserved, as all 69 species had cysteine as the reference AA in this position of the protein.

They were also studied with additional tools. SIFT predicted both to be 100% damaging[74], as did the Functional Analysis Through Hidden Markov Models (FATHMM)[81], which gave predicting scores of -5.25 to p.Phe527Cys and -6.75 to p.Cys529Arg. Once more the latter is the most severe of the two[81].

The consensus among all tools was that both alterations are predicted to be damaging, with p.Cys529Arg as the most serious out of the two novel missense alterations. However the phenotype displayed by ID80 with this alteration is slightly less serious than ID38 affected by p.Phe527Cys, thus the assessment of other variants affecting these individuals is needed.

Neither individuals (ID38 and ID80) had co-inherited haemoglobinopathies, nor had homozygous alterations in *TMPRSS6*. Thus their phenotype might be due completely to their variants. Heterozygous variants affecting both individuals were Lys244Glu, Ser352=, IVS-12+46, Try409=, Asp512=, and p.Val727Ala. Heterozygous variants that only affected ID80 were Pro24= and Tyr730=.

It could be that the two variants that are missing in ID38 might have a protective effect, mitigating the phenotype. But another thing that we have to consider is these individuals, besides having different alterations, is that they also have opposite genders. The severity of the phenotype presented by ID38

might just be exacerbated because it is affecting a female in her fertile years[69,70]. On the other hand, ID80 might have a mitigated phenotype because he is male[32,33].

4.1.5.2 Variants of Uncertain Significance (VUS)

Variants of Uncertain Significance (VUS) are considered in this manner when they are located in genes associated to pathologies, but there is uncertainty in regards to their contribution to a condition's phenotype[82]. They are classified as VUS as there is insufficient data to either deny or confirm their pathogenicity[82].

In this category were placed all alterations whose pathogenicity was dubious (**Table 4.14**), as they were classified as benign in some tools but pathogenic in others.

Table 4.14: Variants of Uncertain Significance in the ID population.

ID	HGVSp	HGVSc	Ex	Ref	Alt	CADD	HT	HM	MAF_E	MAF_s
rs5995378	p.Ser279Leu	c.836C>T	7	tCg	tTg	25.6	0	2	1.4	2
rs755795871	p.Pro408Leu	c.1223C>T	11	cCc	cTc	22.3	1	0	<1	0.5
rs901290763	p.Ser430Pro	c.1288T>C	11	Tgg	Cgg	22.7	1	0	<1	0.5

ID- Reference SNP identification code; **HGVSp**- location of the variants in regards to the translated protein; **HGVSc** – coding DNA location of the variant; **Ex**- Exon; **Ref** and **Alt**- Reference and Alteration respectively, the impacted nucleotide is capitalized; **CADD**- Represents the PHRED score; **HT**- Heterozygote, represents the number of individuals with both the wild type and the alteration; **HM**-Homozygote, stands for the number of individuals with both alleles altered; $\mathbf{MAF_E}$- Minor Allele Frequency reported in Ensembl for the global population; $\mathbf{MAF_S}$- Minor Allele Frequency in our study in percentage.

In order to assert the possible pathogenicity of a variant, series of studies have to be performed[82]. To reach an informed and concise conclusion in regards to their contribution to the pathologies associated to the genes in which they appear. Neither of the variants here reported have been mention in literature, which further cements the uncertainty towards them, as they cannot be classified as benign when accounting for most *in silico* tools in here used.

4.1.5.2.1 p.Ser279Leu

This variant (c.836C>T, rs5995378) affected two individuals, both homozygous (**Table 4.15**).

Table 4.15: Detailed information regarding the homozygous individuals affected by p.Ser279Leu.

		RBC	Hb	Ht	MCV	MCH	MCHC	RDW	AD	DP
ID	Sex	$x10^{12}$/L	g/dL	%	fL	pg	g/dL	%	Ref, Alt	
67	F	4.02	**6.5**	**23**	**57.2**	**16.2**	**28.3**	**20.3**	374,2044	2418
118	F	5.4	12.1	37.3	**69.8**	**22.7**	32.5	**19.1**	343,2012	2355

Bold - values outside the reference range[31] or our criteria; **ID**- is the number of the individual; **RBC**- Red Blood Cells**; Hb** - haemoglobin**; Ht –** haematocrit**; MCV** - Mean Corpuscular volume**; MCH**- Mean Corpuscular Haemoglobin**; MCHC**-Mean Corpuscular Haemoglobin Concentration**; RDW** - Red blood cell Distribution Width**; AD**- Allelic Depth, **Ref** and **Alt**- stand for Reference and Alteration respectively; **DP**- total Depth.

The individuals with this variant displayed phenotypes more severe than our population's average, in regards to Hb, Ht, MCV, MCH, and RDW. Both individuals had lower MCV and MCH, and higher RDW, when comparing to our population that had a mean MCV value of 77.90 ± 6.15 fL, a MCH mean of 24.86 ± 2.25 pg, and a mean RDW of 15.51 ± 1.80 %.

The Ser279Leu variant was reported in ClinVar as benign, but also associated with microcytic anaemia (RCV000282501.2). When having in consideration the MCV of both of the affected individuals and that of our population, this previous statement rings as true. As both individuals have MCV values far below the average for our population (77.90 ± 6.15 fL).

All the selection criteria for our samples in studies were presented by sample ID67; it is also clear the severity of her phenotype when compared to the other person's. Both are unaffected by haemoglobinopathies, thus their phenotype might be due to different MT2 variants.

Besides this variant, these individuals also had in common Val727Ala (1/1) and Try730= (0/1). The ID118 individual was also heterozygous for IVS3+36, Ser130=, IVS6+46, Asp512=. Sample ID67 was heterozygous for Leu193= and Ser430Pro (discussed further ahead), and homozygous for IVS12+46. The simultaneous presence of this variant and Ser430Pro in ID67, could have contributed to the severity of their presented phenotype, when compared to the other person's, who has no other possibly pathogenic variant associated.

This in an unreported variant in the IBS population, it is associated with African populations, with a MAF of 5.1% for the pool for all the represented populations of the continent. With it being highest in the Esan population of Nigeria, which had a MAF of 7.6%. Both individuals were born outside Portugal, but the exact location is unreported, so we might assume that they have African ancestry. Its bioinformatics analyses can be seen in the table below.

Table 4.16: *In silico* analyses performed for p.Ser279Leu.

Ser2709Leu	CADD	PROVEAN	MutPred2	HumDiv	HumVar
Score	25.6	-2.412	0.438	0.263	0.032
Prediction	Top 1%	Neutral*	*	Benign	Benign

*- more detailed information in the text.

PROVEAN considered it to be neutral, although this value is quite close to its cut-off value of -2.5[57]. MutPred2 predicted no significant effects associated. Below are the PolyPhen-2 heat bar results (**Figure 4.13**). This is a highly conversed site, according to the MSA, as in the comparison between 66 species, only two did not have serine in this position. Out of the three VUS reported, this was the only PolyPhen-2 gave scores that were not close to zero for both HumDiv and HumVar.

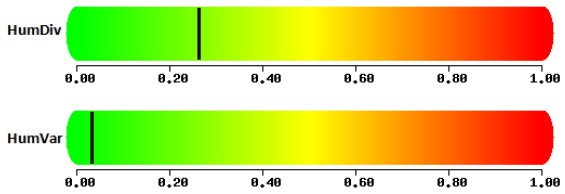

Figure 4.13: PolyPhen-2 heat bar for p.Ser279Leu. This mutation is predicted to be benign by HumDiv and HumVar, with a sensitivity of **0.91** and specificity of **0.88,** and a sensitivity of **0.94** and specificity of **0.60**, respectively.

Serine is polar uncharged AA with a hydroxyl in the end of its side chain[64]. It is considerably smaller and less voluminous, in comparison to leucine, which has two methyl groups branching off its side chain[64]. Besides these differences in size and volume, they are quite different in regards to their hydrophobicity. As leucine is among the most hydrophobic AA, while serine is among the uncharged AA one of the most hydrophilic[64]. These differences contribute to a great shift in their surroundings when the alteration occurs, which might signal its pathogenicity, although it is considered benign. Having in consideration in which African populations it is found, one might speculate that it might be a potential pathogenic alteration that might bring advantages. As for example heterozygous Hb S, which causes sickle cell anaemia, has a protective effect against malaria infections[83].

4.1.5.2.2 p.Pro408Leu

This missense alteration (c.1223C>T, rs755795871) occurs in the 11th exon of *TMPRSS6*, which is located within the protein's CUB_2 domain[37,77]. It affected one heterozygous woman (**Table 4.17**).

Table 4.17: Detailed information regarding the heterozygous individual affected by p.Pro408Leu.

		RBC	Hb	Ht	MCV	MCH	MCHC	RDW	AD	DP
ID	Sex	$x10^{12}$/L	g/dL	%	fL	pg	g/dL	%	Ref, Alt	
42	F	5.1	13.1	39.5	**77.6**	**25.7**	33.2	**15.1**	59,58	117

Bold - values outside the reference range[31] or our criteria; **ID**- is the number of the individual; **RBC**- Red Blood Cells**; Hb** - haemoglobin**; Ht –** haematocrit**; MCV** - Mean Corpuscular volume**; MCH**- Mean Corpuscular Haemoglobin**; MCHC**-Mean Corpuscular Haemoglobin Concentration**; RDW** - Red blood cell Distribution Width**; AD**- Allelic Depth, **Ref** and **Alt**- stand for Reference and Alteration respectively; **DP**- total Depth. The Genotype quality was 99%.

This variant switches proline by leucine, both AA are apolar, but leucine is more hydrophobic[64]. But instead of possessing an aliphatic chain like leucine, proline's aliphatic chain forms a cyclic structure, which reduces the flexibility of the polypeptide chain thanks to its rigidity[64]. The increased mobility provided by leucine, might disrupt this protein site.

This individual presents hypochromic microcytic RBC, without anaemia, so they might only have ID[76]. However since this individual possesses α-Thalassemia as well, the phenotype might be due to it[18,25]. Other variants affecting *TMPRSS6* included, IVS6+46, Lys244Glu, IVS7+23, IVS12+46, Asp512=, and Val727Ala, all of which were homozygous.

ClinVar does not report it, and in the gnomAD study only non-Finnish Europeans and Ashkenazi Jews presented this variant[67,84]. This individual was born in Portugal which is in line with previously mentioned populations. We also did not find literature associated to this variant. In regards to its bioinformatics analyses (**Table 4.18**).

Table 4.18: *In silico* analyses performed for p.Pro408Leu.

Pro408Leu	CADD	PROVEAN	MutPred2	HumDiv	HumVar
Score	22.3	-2.795	0.346	0.003	0.002
Prediction	Top 1%	Deleterious	*	Benign	Benign

*- more detailed information in the text.

MutPred2 predicted no significant consequences, as did PolyPhen-2's HumDiv and HumVar. In the MSA, among 68 species 15 presented an AA other than proline, eleven had alanine, two had serine and the remaining two had leucine .

4.1.5.2.3 p.Ser430Pro

This missense alteration (c.1288T>C) occurs in the 11th exon of MT2, like the previous variant. It affected one heterozygous woman (**Table 4.19**), and its reference ID is rs901290763.

Table 4.19: Detailed information regarding the heterozygous individual affected by p. Ser430Pro.

		RBC	Hb	Ht	MCV	MCH	MCHC	RDW	AD	DP
ID	Sex	$x10^{12}$/L	g/dL	%	fL	pg	g/dL	%	Ref, Alt	
67	F	4.02	**6.5**	**23**	**57.2**	**16.2**	**28.3**	**20.3**	69,90	159

Bold - values outside the reference range[31]; **ID**- is the number of the individual; **RBC**- Red Blood Cells; **Hb** - haemoglobin; **Ht** – haematocrit; **MCV** - Mean Corpuscular volume; **MCH**- Mean Corpuscular Haemoglobin; **MCHC**-Mean Corpuscular Haemoglobin Concentration; **RDW** - Red blood cell Distribution Width; **AD**- Allelic Depth, **Ref** and **Alt**- Reference and Alteration respectively; **DP**- total Depth. The Genotype quality was 99%.

This variant also involves proline, but unlike the previous, this is the altered AA. Serine is polar thanks to its hydroxyl group, and is smaller and more hydrophilic than proline[64]. This shift causes an increase in rigidity and hydrophobicity in this site[64].

The Ser430Pro variant is correlated to decreased Hb and Ht, as this individual presents severe hypochromic and microcytic anaemia. As this individual also possessed the Ser279Leu VUS their phenotype was been discussed previously.

It is an extremely rare variant, as only 4 individuals presented the altered allele out of the 143250 individuals of all populations present in the gnomAD genomes r3.0 project present in Ensembl[67,84]. This variant is unreported in ClinVar, and all *in silico* tools used, except for PolyPhen-2, predict it to be deleterious (**Table 4.20**).

Table 4.20: *In silico* analyses performed for p.Ser430Pro.

Ser2709Leu	CADD	PROVEAN	MutPred2	HumDiv	HumVar
Score	22.7	-3.663	0.663	**0.000**	**0.001**
Prediction	Top 1%	Deleterious	*	Benign	Benign

*- more detailed information in the text.

MutPred2 predicted this variant to cause an alteration on the transmembranar level of MT2 ($p=9.8 \times 10^{-3}$). PolyPhen-2 classified this variant as benign both in HumDiv and HumVar, with sensitivities of 1.00 and 0.99, and specificities of 0.00 and 0.15), respectively. Regarding its MSA, among 69 species whose sequence has been established only eight presented a different AA, three had threonine, while five had proline. The AA differences between species might have contributed to the attribution of the benign prediction by PolyPhen-2. For this variant, also no information about it was available in literature.

4.1.5.3 Benign coding alterations

Within this category were placed 15 variants, all the synonymous alterations along with several missense (**Table 4.21**).

Table 4.21: Benign coding variants.

ID	HGVSp	HGVSc	Ex	Ref	Alt	CADD	HT	HM	MAF_E	MAF_S
rs11704654	p.Pro24=	c.72G>A	2	ccG	ccA	0.112	24	1	17.8	13
rs150521260	p.Val61=	c.183G>A	2	gtG	gtA	2.689	1	0	0.5	0.5
rs763214117	p.Gly68=	c.204G>T	3	ggG	ggT	8.710	1	0	<1*	0.5
rs147700428	p.Ser130=	c.390C>T	4	tcC	tcT	1.396	1	0	0.7*	0.5
rs142971835	p.Val131Ile	c.391G>A	4	Gtc	Atc	12.86	1	0	1*	0.5
rs79013645	p.Leu193=	c.579A>C	5	ctA	ctC	11.65	2	0	1.7*	1
rs2235324	p.Lys244Glu	c.730A>G	7	Aag	Gag	0.002	44	17	47.7	39
rs201148397	p.Val280Leu	c.838G>T	8	Gtg	Ttg	14.49	1	0	0.5	0.5
rs2111833	p.Ser352=	c.1056G>A	9	tcG	tcA	0.428	44	14	42.5	36
rs881144	p.Tyr409=	c.1227C>T	11	taC	taT	0.051	17	0	7	8.5
rs4820268	p.Asp512=	c.1536C>T	13	gaC	gaT	0.109	40	42	64.5	62
rs145814440	p.Asp600=	c.1800C>T	15	gaC	gaT	10.14	2	0	1.2*	1
rs855791	p.Val727Ala	c.2180T>C	17	gTc	gCc	20.3	35	47	65	64.5
rs2235321	p.Tyr730=	c.2190C>T	17	taC	taT	8.299	44	18	45.3	40
rs73886915	p.Ser773=	c.2319C>T	18	agC	agT	9.975	1	0	1.5*	0.5

ID- Reference SNP identification code; **HGVSp**- variant location in the translated protein; **HGVSc** – coding DNA variant location; **Ex**- Exon; **Ref** and **Alt**- Reference and Alteration respectively, the impacted nucleotide is capitalized; **CADD**- PHRED score; **HT**- Heterozygote; **HM**-Homozygote, for the alteration; **MAF_E**- Minor Allele Frequency for the IBS population; **MAF_S**- MAF in our population, ***-** The global population was used (undetected variants in the IBS population).

ClinVar either reported them as benign or they were unreported. The variants that were present in most individuals, as well as SNPs held as functional, and those with the lowest CADD scores, were also placed in this category.

Unlike the previous sections, each variant won't be given as much emphasis. In the following sections are addressed in general this group's CADD scores, MAF, *in silico* analyses or statistical analyses.

Only functional SNPs will be highlighted individually. Several variants found in our study were considered to be functional SNPs[39,41,71,72,85-90], but only a few will be addressed, as these were the most commonly mentioned in literature.

4.1.5.3.1CADD scores

In Table 4.21 are represented the CADD scores for all the benign alterations. Nearly half (7/15) were below 5, with 5 of them scoring below 1. The lowest value belonged to Lys244Glu, 0.002 with a raw score of -1.264480. Between scores of 5 to 10, were 3 variants (Tyr730=, Gly68=, Ser773=). Asp600=, Leu193=, Val280Leu had scores below 15.

CADD scores above 20 mean that the variant ranks among the top 1% most deleterious variants in the human genome[59], and only Val272Ala surpassed this threshold. It is interesting to notice that the highest and lowest CADD scores both belong to missense variants, Val727Ala and Lys244Glu, respectively.

4.1.5.3.2 Minor Allele Frequency

The MAF for each coding alteration was reported in Table 4.21, thus only the variants without data for the IBS MAF, will be mentioned in this section[65]. Out of the 12 coding benign variants, six are unreported for the IBS population, all with MAF below 2%.

Ser130= (rs147700428), Leu193= (rs79013645), and Ser773= (rs73886915) were mostly found in African population or populations of African descent, the MAF percentage stated was for its prevalence on a global level.

Gly68= (rs763214117) was only found in African individuals of the gnomAD genomes r3.0 project[67,83]. Only in gnomAD were found individuals with Val131Ile (rs142971835), among Latinos and non-Finnish Europeans.

Asp600= (rs145814440) was prevalent in populations with European ancestry, although not appearing in the IBS population, it had a MAF of 2% the Toscanini, Finns and Americans of western European ancestry.

4.1.5.3.3 Benign missense variants *in silico* analyses

The majority of tools only assess missense alterations, so this section is much smaller in comparison to the previous. PROVEAN was used in all coding variants, but only gave real results to the missense variants, as all synonymous were classified as neutral with a score of 0.000 (**Table 4.22**).

Table 4.22: PROVEAN analysis for Benign coding variants.

PROVEAN	Val131Ile	Lys244Glu	Val280Leu	Val727Ala
Score	-0.466	0.144	0.179	0.259
Prediction	Neutral	Neutral	Neutral	Neutral

PROVEAN considers variants with scores below -2.5 to be deleterious, the more positive the value the more likely to be benign is a variant, although he only gives predictions of deleterious or neutral[57]. Val727Ala is predicted as the least deleterious variant.

MutPred-2 predicted them not have any considerable effects associated. PolyPhen-2 predicted them all to be benign with scores close to zero in most missense variants. PolyPhen-2 gave p.Val131Ile (**Figure 4.14 A**) a HumDiv score of 0.343 (sensitivity: 0.90; specificity: 0.89) and a HumVar score of 0.102 (sensitivity: 0.91; specificity: 0.69), this was the only in this category whose HumDiv and HumVar scores were not close to zero. All the others had heat bars like Lys244Glu (**Figure 4.14 B**).

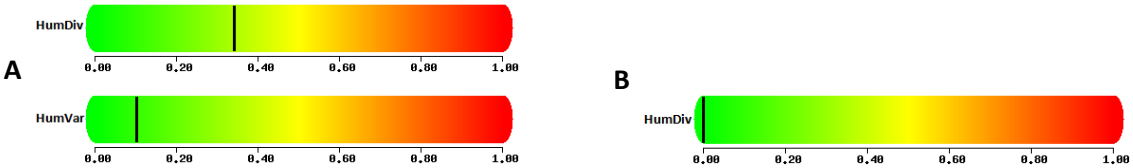

Figure 4.14: PolyPhen-2 heat bars of benign variants. A) p.Val131Ile **B)** p.Lys244Glu, only HumDiv is represented because HumVar's heat bar is equal to the one depicted.

4.1.5.3.4 Statistical analyses

The only significant relations found between the genotype and phenotype were presented by Pro24= and Lys244Glu. Individuals with Pro24= were significantly different in regards to their MCHC (p=0.009). This haematological parameter was considerably lower than the mean for rest of the population, being below the RV[31].

The individuals that possessed the Lys244Glu variant were significantly different from the rest of the population in regards to their RDW values (p= 0.028). Their average value was 15.27% vs. 15.95% in those without it, making it significantly smaller than in individuals without it. The RDW value was 15.51±1.8% for the whole population, since this is a skewed parameter, the majority of individuals RDW fell above the mean value for the population. This indicates that this variant might have a protective effect in the individuals that have it.

4.1.5.3.5 Functional SNPs

Functional SNPs are polymorphic variants that are most likely to be associated to diseases[87]. These are SNPs that tend to appear more frequently in populations, and are associated with disease phenotypes[39,41,71,72,85-90]. The study conducted by Poggiali[24] was the most comprehensive, as most studies only reported two or three SNPs, most commonly Asp512= and Val727Ala[24, 39,41,72, 85,86,91].

In the case of our population, we found four functional SNPs that appeared in more than 35% of our population. These variants were Lys244Glu, Asp512=, Val727Ala, and Try730= (**Table 4.23**). These variants were all included in the Poggiali study, along with Pro24=, Ser352=, and Tyr409=[24].

Table 4.23: Functional SPNs Genotype frequency and MAF comparison.

	Lys244Glu (A>G)				Asp512= (C>T)				Val727Ala (T>C)				Tyr730= (C>T)			
GT	**MAF**	**AA**	**AG**	**GG**	**MAF**	**CC**	**CT**	**TT**	**MAF**	**TT**	**TC**	**CC**	**MAF**	**CC**	**CT**	**TT**
IBS	47.7	27.1	50.5	22.4	64.5	11.2	48.6	40.2	65	11.2	47.7	41.1	45.3	29.9	20.6	49.5
OS	39	39	44	17	62	18	40	42	64.5	18	35	47	40	38	44	18

GT-Genotype; **IBS**- Iberian Spanish population; **OS**- Our study population; **MAF**- Minor Allele Frequency.

In table 4.23 are represented the allelic frequencies for the functional SNPs, in top line besides the HGVSp name are the nucleotides that switch to cause the variant. The GT for each variant is also represented, after the MAF, starting from the wild type and ending in the homozygous variant. It can be observed that, the MAFs in our population are always smaller than the IBS's but never too apart. The same cannot be said about the genotype frequency.

Lys244Glu has a MAF of 43% for the overall European populations, our sample presents a value slightly below that, but comparing to the IBS, its prevalence is 8.7% lower. When comparing the genotype frequency between the two populations, it can be perceived that this variant is under represented in our population. As only 27.1% of the IBS population present the wild type (AA) for this variant vs. 39% in our population. Although the MAF value wasn't too dissimilar in Asp 512=, a 2.5% difference, the prevalence of this variant in our population is lower than in the healthy IBS population, with a bigger wild type percentage (+6.8%) and lower heterozygotes (-8.6%).

Val727Ala affects 60.5% of the population on a global level. The MAF between the IBS population and ours is similar, but the genotype differs, with the wild type genotype (TT) being more apparent in our sample (18% vs. 11.2%). The T allele is seen as a risk factor for IDA[72,92].

Addressing Tyr730=, despite both populations have a similar MAF, their genotype distribution is quite different. The percentage of individuals solely possessed the wild type allele (CC) for this variant in our population (38% vs 29.9%). Unlike Val727Ala, the wild type C allele is not associated to poorer IDA prospects[72]. However the biggest genotype difference is for TT, which is underrepresented in our population.

Jallow *et al.* studied three of these coding variants, Asp512=, Val727Ala and Tyr730=, to assess their impact in oral iron absorption response and hepcidin levels in a healthy sample of the Gambian population, over a period of 5h[72]. Except for Val727Ala, they found significant differences in serum hepcidin levels before and after iron treatment, but not in plasma iron nor in the comparison between reference individuals (without any altered allele) and those with the variants[72]. It is important to state that although this was a healthy population, upon CBC testing, 31.9% of the sample was found to be anaemic, from which 11.9% had IDA. Their most interesting result was for the TT/TC/TC group (homozygous for Asp512=, heterozygous for Val730Ala and Tyr730=, respectively)[72]. This group maintained lower mean hepcidin levels even after the 5h, even while the serum iron had increased, unlike most of the other groups' individuals and reported literature[72]. Hepcidin levels should increase in response to an increase in serum iron, due to its absorption, in order to prevent over-absorption. The lack of this mechanism, indicates a fault in its expression or regulation, thus further studies are required. Jallow's study illustrates the importance of the study of functional SNPs, it is relevant to state that the study was in an African population, Gambian Adults, so it could be that they have other factors influencing it. Despite being the most genetically diverse population group[65,67], African populations present a lesser amount of risk iron alleles, for both iron deficit and overload[85].

The following section will address the functional SNPs in a more individual manner, although references to each other might appear, as they are frequently intertwined.

4.1.5.3.5.1 p.Lys244Glu

Lys244Glu is missense variant that occurs in the 7th exon (c.730A>G , rs2235324), which corresponds to the CUB_1 Domain of MT2[23]. All *in silico* tools predicted it to be a benign variant (**Figure 4.15 C**). This residue is highly variable, considering the MSA, this contributed to the low PolyPhen-2 scores.

This alteration causes the replacement of Lysine by Glutamate in the 244th position of the protein. Both AA are amongst the most hydrophilic AA, as both are polar and charged[64]. Lysine has a positive charge and is slightly more polar with a longer aliphatic chain, while Glutamate is negatively charged due to the carboxyl group on the end of its R group chain[64]. So while their hydrophobicity is similarly low, their sizes are different, so the low impact of the alteration might be due to it[64].

From the functional SNPs, this was the only that had an impact on a haematological parameter, RDW, discussed before. The individuals with it had a lower RDW, being closer to the normal value, below 15%[31] than individuals without it.

Studies have associated it to increased risk of developing ID, IDA and IRIDA when associated with other variants[85,91], however this isn't among the most widely studied functional SNPs. A 2012 study found it to be significantly associated to TS above 10% (p=0.0001)[88]. Beutler found it more prevalent in controls than in men with IDA (a 15.7% difference)[71], which further reinforces its benign effect on preventing IDA. This is in line with our findings, as individuals with it had a significantly lower RBC size dispersion (RDW) than those without this variant.

Poggiali however found this variant more frequently in IDA individuals than in controls (p=0.005)[24]. Controls had the wild type genotype above 90% and the heterozygous was less than 10% with no homozygous individuals. In comparison, the IDA individuals were 10% homozygous for the variant with nearly 40% heterozygous[24]. These values of their IDA cohort are aligned with our obtain data, 17% and 44%, for homozygous and heterozygous respectively.

4.1.5.3.5.2 p.Asp512=

This synonymous variant is widely accepted as a risk factor for ID (c.1536C>T, rs4820268)[39,41,72,86,90]. It is associated to lower Hb, microcytic RBC, increased RDW, lower serum iron concentration, and lower hepcidin levels in urine[39].

This AA residue of MT2 is also associated to a missense variant, p.Asp512Asn (c.1534G>A, rs137853120), that causes IRIDA[75], that affects the first nucleotide in the same codon (GAC), previously mentioned in the pathogenic variants section.

A familial study with 3 siblings with refractory IDA, in which the genes that encode for MT2 and DMT1 were studied, only this variant along with Pro24= and Val727Ala were found[39]. This study also included a population based analysis, and out of the 7 SNPs studied for both genes, only this variant and Val727Ala had significant associations to iron ($p=3.9x10^{-6}$ and $2.0x10^{-6}$, respectively) and ferritin levels (p=0.001 and 0.002, respectively)[39]. These results were replicated for the mostly severely affected child, for iron (p=0.0128) and ferritin ($p=9.57x10^{-5}$)[39].

In our study, no significant differences were found between the individuals with this variant in comparison with those without, in regards to hematological parameters. Several studies have associated the C allele, which corresponds to the wild type, to lower Hb concentrations in ethnically different populations (Europeans and Asians)[42,72]. This aligns with the genotype frequencies in our population, as the wild type in our sample was nearly 9% greater than the IBS population, this overrepresentation is in line with the phenotype shown. In European samples higher ferritin concentration and reduced serum TfR have been associated to Asp512=[42,72], It has also been found to be associated to lower serum iron and TS[86].

4.1.5.3.5.3 p.Val727Ala

This missense alteration is located in the 17th exon (c.2180T>C, rs855791), occurs in MT2's serine protease domain, which is its catalytic domain[89], close to the catalytic site residues[86,92]. ClinVar reports it as benign, however it was been several times implicated on its contributing role in IDA development. It is by far the most studied SNP associated to decreased iron levels[39,41,71,72,85-90], with several studies only taking it in consideration for ID study, but also other conditions[91-94].

Valine and alanine are both nonpolar aliphatic AA, and have a fairly similar structure[64]. With valine being more complex, by having an isopropile group as R chain, instead of the simple methyl like alanine[64]. This increased complexity of the aliphatic chain of valine makes it more hydrophobic in regards to alanine[64]. So their common hydrophobicity might not induce stereochemical environment alterations on its surroundings, but these can occur due to their size difference, as alanine is more compact, thus this is a rather conservative alteration[39,64].

The majority of studies concerning this variant are from the point of view of haematological conditions. But it has also been studied for conditions as diverse as type 2 diabetes[90], breast cancer[91], to end stage renal disease patients undertaking haemodialysis[93], and also in less conspicuous studies like its influence in endurance athletes[94].

Chambers described per each copy of the T allele a 0.13 g/dL decrease in haemoglobin concentration ($p=1.6x10^{-13}$) with this effect being more pronounced in Indian Asians than Europeans ($p=3.5x10^{-33}$)[86]. It was also strongly associated to MCV, MCH, and MCHC: the TT genotype has been associated to lower TS and serum ferritin, higher hepcidin and ratios of hepcidin-iron, in European populations[86, 92].

The homozygous CC has been significantly less reported in an IDA population in a study about the susceptibility to its development by menstruating women[69], which points to its protective role against IDA and reinforces the risk for IDA development in the presence of the T allele.

A Taiwanese study focused on iron supplementation in women, only considered CC or TT Val272Ala homozygous individuals, excluding heterozygous[92]. They found TT associated to significantly lower iron absorption, TS, serum iron, and increased hepcidin to TS ratios[92]. They also suggested a deficient iron absorption and regulation in the TT individuals, as they appeared less apt to regulate iron absorption, both with low iron stored and when they were iron replenished[92]. The individuals with the CC genotype presented strong correlations between iron absorption and serum ferritin, and iron absorption and hepcidin, as well as a better absorption regulation[92]. Significant differences ($p<0.02$) between the CC and TT groups were in MCV, MCH, serum iron, TS, and menstrual blood loss assessment, with the TT group presenting lower mean values for all parameters except for the later[92]. The perceived decreased menstrual blood loss in the TT group might be due to a systemic mechanism to prevent further iron decrease, through blood loss.

In the study that considered *TMPRSS6* expression in Breast cancer patients, no significant relations were found to the variants reported in our work and their haematological parameters[91]. But *TMPRSS6* expression was significantly increased in the tumour tissues (1.88 times higher than in normal tissues)[91]. This hints to a possible link between MT2 and the oestrogen receptor and the HER2 oncogene, which play important roles in breast cancer[91].

End stage renal disease patients, have non-functional kidneys, thus require haemodialysis. Another side effect of this condition is Erythropoietin deficiency, which can cause anaemia, as the kidneys are responsible for the production of this hormone[6,20,93]. Renal malfunctioning is associated to great disturbances in the iron metabolism[10,93]. So in this study, 199 haemodialysis dependant patients and 188 controls were genotyped (C282Y and H63D in *HFE*, and V727A) and had their hepcidin levels evaluated[93]. The genotypes did not significantly differ between the groups, but Val727Ala was shown to modulate the effect of HFE, increasing hepcidin ($p=0.017$), and influence erythropoiesis[93].

When we tested this variant from a statistical point, no significant relationships were detected between individuals with and without it for the haematological parameters, these results were also observed in other studies[89]. In the particular case of our study, this might be due to our study design, as we only contemplated individuals showcasing an ID phenotype. If we had included a control group and compared the results between the controls and the ID individuals, there might have been significant differences between the two groups.

A meta-analysis study however, reported it to be associated to lower Hb, in populations with different ethnicities, and also with decreased serum ferritin[41]. A familial genetic association study found it to be significant for iron and ferritin levels, along with Asp512=[39].

All *in silico* tools predicted it to be benign. Missense3D predicted it as being structurally Neutral, with no structural features disrupted. PolyPhen-2 gave it scores of 0.000 (sensitivity: 1.00; specificity: 0.00) for HumDiv and HumVar (see **Figure 4.14 B** for the heat bars representation). Its MSA however is interesting and different from the majority of the variants in this study (**Figure 4.15**).

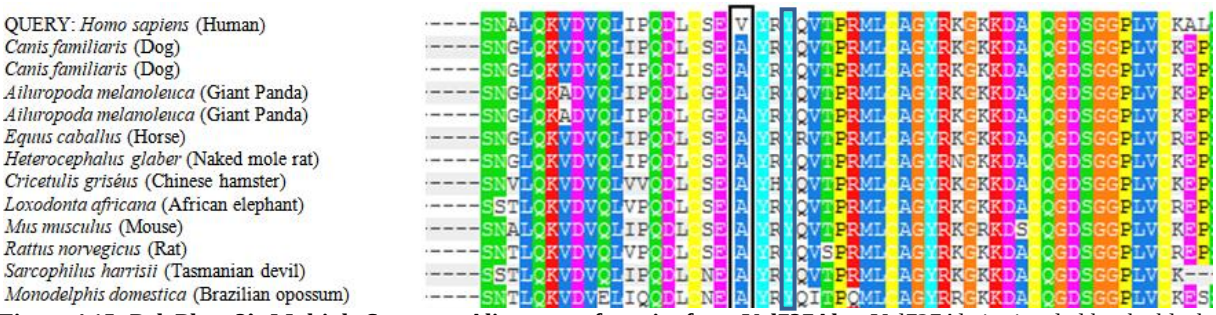

Figure 4.15: PolyPhen-2's Multiple Sequence Alignment of species for p.Val727Ala. Val727Ala is signaled by the black box, while within the blue box is highlighted p.Tyr730 to demonstrate their proximity.

Only humans had valine in this site, while most species had alanine (eight had serine and two had leucine). This indicates this alteration in humans corresponds to an approximation to a conserved residue in other species. This might explain why the canonical T allele corresponds to a risk allele[72,86,92].

As this is among the most common *TMPRSS6* SNPs, we found it interesting to see if there were other variants reported with linkage disequilibrium associated to it. Genetics assume a random rearrangement of inherited alleles, however when they associate in a non-random ways, this is considered to be a linkage disequilibrium (LD)[95].

The IBS population in Ensembl reported LD for 16 variants with a score above r^2=0.90, the majority were deep intronic, but it also included some of our population's variants, like IVS-12+46 (r^2=0.958), Asp512= (r^2=0.958), and Try730= (r^2=1.000). Although we didn't perform the mathematical analyses required to also establish this level of relationship between the variants, it was possible to observe that these variants were very frequently present simultaneously in most individuals in our population.

Having solely the homozygous for this variant in consideration, 47 out of 100 individuals, only in 11 did not possess the altered allele for Try730=, and as for Asp512= and IVS12+46, all homozygous individuals for Val727Ala had at least one allele for these variants. From an opposite perspective, the homozygous for the wild type for Val727Ala, only presented alleles for the other variants once, for IVS12+46 and Asp512=, and twice for Try730=. LD between Val727Ala and Asp512= is strongest among European populations[39,85], with it being alike to Indian Asians (r^2=0.83 vs. 0.63)[86], but the same association strength is not present in West African populations[85].

4.1.5.3.5.4 p.Try730=

This synonymous variant (c.2190C>T, rs2235321), and Val727Are are in the same exon and domain of MT2. Their proximity can be visualized in Figure 4.4 (3), in a Sanger confirmation for both variants, and in Figure 4.15 (within a dark blue box).

This variant is reported as benign in ClinVar, and frequently included in studies of common SPNs for genes correlated to iron status (*TMPRSS6*, *HAMP*, *TF*, *TFR2*, *SLC40A1* and *HFE*) according to Jallow's literature review[85,96].

A study found it significantly associated to Transferrin Saturation below 10% (p=0.0135) but not to haematological parameters[88]. When we tested this variant statistically, no associations were detected between the individuals with it to those without it.

4.1.5.4 Intronic Alterations

In Table 4.24 are presented our population's intronic variants which are 50 nt up or downstream of the intron/exon junction.

Table 4.24: *TMPRSS6* Intronic variants detected by NGS from our ID population, within 50nt from Exons.

ID	Position	HGVSc	IVS	Ref	Alt	HT	HM	MAF_E	MAF_S
rs732756	22:37098380	c.336+36A>G	3+36	A	G	28	1	15.4	15
rs2743824	22:37095505	c.631+46A>G	6+46	A	G	46	18	49.5	42.5
rs764379026	22:37089571	c.836+7C>T	7+7	C	T	1	0	<1*	0.5
rs2235326	22:37089555	c.836+23A>G	7+23	A	G	1	39	64.5	39.5
Novel	22:37089551	c.836+27G>C	7+27	G	C	1	0	-	0.5
rs113287112	22:37086249	c.973+34G>A	8+34	G	A	0	1	1.9	1
rs9610642	22:37075298	c.1197-18G>T	10-18	G	T	1	0	<1*	0.5
rs79816125	22:37074600	c.1441+10C>T	12+10	C	T	2	0	1.4	1
rs111807510	22:37074595	c.1441+15C>T	12+15	C	T	2	0	2.5*	1
rs2072860	22:37074564	c.1441+46C>T	12+46	C	T	36	41	64.5	59
rs111813777	22:37070890	c.1672+26C>T	14+26	C	T	2	0	2.6*	1
rs377498210	22:37069032	c.2113+41G>A	16+41	G	A	1	0	<1*	0.5

ID- SNP identification code; **Position-** chromosome location according to GRCh38; **HGVSc** – coding DNA variant location; **IVS**- InterVening Sequence, the first number is the intron, the second is the position in regards to the exon**; Ref** and **Alt**-Reference and Alteration respectively, the impacted nucleotide is capitalized; **HT**- Heterozygote; **HM**-Homozygote, for the alteration; **MAF_E**- Minor Allele Frequency reported in Ensembl for the IBS population, in percentage; **MAF_S**- Minor Allele Frequency in our study in percentage, ***-** The global population was used (undetected in the IBS population).

The majority of these variants were unreported on ClinVar, or were classified as benign. Since these are non-coding benign alterations, their study is less substantial. In regards to their MAF, those which are unreported in the IBS population, were mostly found in African populations (c.836+7, c.1197-18, c.1441+15, c.1672+26, c.2113+41). There were four intronic variants prevalent in our population and that were reported in the IBS population, IVS3+36, IVS6+46, IVS7+23, and IVS12+46 (**Table 4.25**).

Table 4.25: Genotype frequency in our common intronic variants.

	IVS3+36A>G			IVS6+46A>G			IVS7+23A>G			IVS12+46C>T		
GT	**0/0**	**0/1**	**1/1**	**0/0**	**0/1**	**1/1**	**0/0**	**0/1**	**1/1**	**0/0**	**0/1**	**1/1**
IBS	70.1	29	0.9	24.3	52.3	23.4	12.1	46.7	41.1	11.2	48.6	40.2
OS	71	28	1	36	46	18	60	0.5	39.5	23	36	41

GT- Genotype; **IBS –** Iberian Spanish population; **OS-** our study population; **0/0-** wild type; **0/1-** heterozygous; **1/1-** Homozygous for the alteration.

The MAF for IVS7+23 (c.836+23A>G, rs2235326) is quite different between ours and the IBS population (39.5 vs. 64.5%, respectively). This difference can be explained by the genotype frequency between the two populations, which is significantly different as it can be seen in the table above. The wild type genotype prevalence is five times greater in our population, in comparison to the IBS population, with barely any heterozygous individuals. Unfortunately no literature was found in regards to this variant.

IVS12+46 (c.1441+46C>T, rs2072860) also had different MAF values between both populations, but not as drastic (59 vs. 64.5%, respectively). This is also one of those variants whose presence is strongly correlated to Val727Ala (p-value=$7.7x10^{-3}$, LD r^2=0.90)[97], as mentioned in the Val727Ala section.

All intronic variants were tested with VarSEAK, in order to assess their possible splicing effect. All were classified within the Class 1 (**Figure 4.16**), which means they had no splicing effect. Thus none of them, including the novel variant, had a deleterious impact on the protein, so they are all benign.

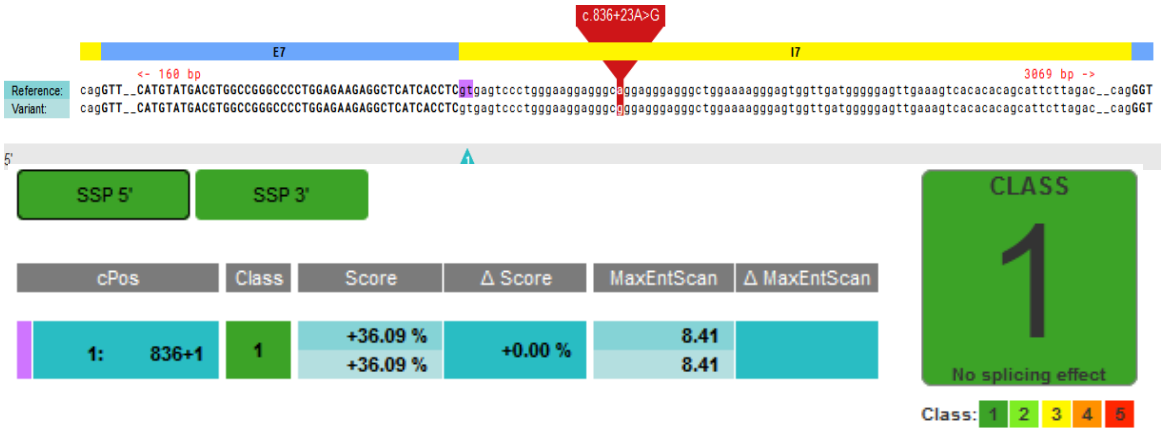

Figure 4.16: Novel intronic Variant c.836+23A>G (22:37089555) analysis through VarSeak. In the image above the alteration is visible within the intron in red, the 7th exon is in blue and the 7th intron is in yellow. If the alteration was in the intronic purple area, it would cause alterations in splicing, since it has no impact it was graded in Class 1.

In regards to statistical analyses, only the individuals that presented the IVS3+36 variant had a significant difference to the rest of the population. This difference was in regards to their RDW (p=0.025). The individuals with this variant had a higher mean RDW (16.06%), meaning that they had a wider range of different sizes between their RBC.

4.1.6 Concluding ID remarks

Throughout this chapter we have accessed the impact of *TMPRSS6* variants found in a population presenting an ID-like phenotype, through different types of analyses (*in silico*, statistics, and literature review). Most variants were found to be benign, as only six were considered pathogenic, and three were VUS.

The most damaging variant, through the different predictions given by our analyses, was the novel variant p.Cys529Arg. Out of the pathogenic variants, p.Arg437Trp seems to be the less severe in this category, but it was more severe than the variants reported in the VUS section.

We expected to have an enriched range of different variants in our studied population, and also some differences in the prevalence of functional SNPs in regards to a healthy population. Both hypotheses were confirmed.

The protective genotypes of functional SNPs were underrepresented in our sample, while the risk alleles frequency increased. Although few statistical associations were significant, if we had had a control group sample to compare this population with, certainly more relationships would be noticed.

4.2 Iron overload

4.2.1 Population Characterization

The iron overload population consisted of 110 patients, 28 females and 82 males, with the mean age of 53 years, sent by the department of immunohemotherapy of *Hospital Santa Maria*, in order to be screened for non-classic Hereditary Hemochromatosis.

The molecular analysis for the iron-related genes *HFE*, *TFR2*, *HJV*, *HAMP* and *SLC40A1* were performed by Ricardo Faria and Rita Simão in the field of their MSc thesis. In this dissertation, we focused only in cases suspected of HH type IV (HH-IV) or Ferroportin Disease (FD), in which the *SLC40A1* gene is involved.

The iron biomarkers used for patients' inclusion were serum ferritin above 300ng/mL or Transferrin Saturation (TS) over 60 %[21,48]. Samples were excluded if IO was of secondary cause, and if they were diagnosed with HH type I. Due to the criteria to partake in our study, the presence of a pathological phenotype previously described, the abnormally high mean values for the iron biomarkers, particularly Ferritin and TS, were expected (**Table 4.26**).

Table 4.26: Serum Iron Biomarkers levels observed in the 110 patients with IO.

	Units	N	Mean	SD	Med	Min	Max
Ferritin	ng/mL	103	**1396.95**	999.709	**1077.00**	56	**4842**
TS	%	108	**77.31**	19.658	**80.00**	**6**	**119**
Iron	μg/dL	24 F	**195.80**	66.15	**197.00**	**50**	**345**
		65 M	**197.43**	66.69	**197.50**	**58**	**374**

Bold- values outside the reference range; **N-** number of biomarker results; **F –** Female**; M –** Male; **SD** - standard deviation; **Med** - Median; **Min** - Minimum value; **Max** - Maximum value; **TS –** Transferrin Saturation.

Unlike the criteria applied to iron deficiency, which makes use of Complete Blood Counts (CBC), here iron biomarkers were used for diagnosis of iron overload. The use of biomarkers isn't as standardised as CBC, because of that, not all iron biomarkers were used for all patients. Serum iron was the least requested biomarker (N=89), while TS encompassed most individuals (N=108).

The normal upper range values for serum ferritin are below 300 ng/mL (20-280 ng/mL for men, and 10-140 ng/mL for women)[98], thus the use of this limit for the IO samples. Only two samples, belonging to two males under 40 years of age, had normal ferritin values (56 and 60 ng/mL), but both had TS>75%.

TS range from 20 to 40% in normal individuals, but based on our inclusion criteria we expected the mean value for this population to be much higher. Only five individuals had TS values within the normal range, and solely one individual had TS below the normal range, which corresponded to the minimum. With a TS mean of 77.31% this value is nearly double of the upper value of the normal range. At times the TS values surpassed 100%, as shown in Table 4.26, this occurred in more than one sample. TS is calculated by dividing serum iron by serum iron plus unsaturated iron-binding capacity (UICB), which corresponds to transferrin's availability to bind to iron[46,85]. Low UICB levels are associated to HH, as most transferrin is bound to iron already, thus unavailable for further binding[46].

Normal serum iron ranges from 70-145 μg/dL in men, and 60-130 μg/dL in women[98], the mean for our sample was considerably higher than that (**Table 4.26**). Only two individuals had values below the normal range, a male and a female with values below 60 μg/dL. With the normal range were found 18 individuals.

4.2.2 Variants in the *SLC40A1* gene detected by NGS

A total of 35 alterations were found in NGS for this gene. Only seven remained after applying the inclusion criteria of "located less than 50nt from a exon/intron junction". The coding variants comprised of four: two missense, one in-frame deletion and one synonymous variant (**Table 4.27**).

Table 4.27: Variants in *SLC40A1* detected by NGS from the IO population.

ID	Position	HGVSc	HGVSp	Ex	In	MAF_E	MAF_s
rs11568351	2:189580468	c.-8C>G	-	-	P	20.6	24.07✡
rs1439816	2:189579904	c.44-24G>C	-	-	1	77.1	77.78✡
rs555193844	2:189579901	c.44-21A>C	-	-	1	0.6*	6.36
rs978427853	2:189575194	c.238G>A	p.Gly80Ser	3	-	-	0.46
rs1172102948	2:189571742	c.485_487del	p.Val162del	5	-	-	0.46
rs387907377	2:189565504	c.610G>A	p.Gly204Ser	6	-	<1*	0.46
rs2304704	2:189565451	c.663T>C	p.Val221=	6	-	60.7	64.82✡

ID- is the number of the individual; **Position**- chromosomal location according to GRCh38; **HGVSc** –coding DNA variant location; **HGVSp**- variant location in protein; **Ex**- Exon; **In**- Intron; **MAF_E**- IBS population MAF; **MAF_S**- our study MAF; *- Global population MAF (undetected in the IBS population); ✡-Not all individuals of our population were accounted for.

Regarding the coding alterations, all but p.Val221= are variants that cause FD[13]. The individuals that had them were heterozygous, but as this is an autosomal dominant pathology, it is enough for them to showcase its phenotype[13]. As such, no case of HH type IV is present in this population[13,99].

This section, in comparison to the previous regarding iron deficiency, is considerably smaller due to the lesser amount of variants found for this gene vs. the gene that encodes for Matriptase-2. This difference in number can be explained due to the different in size between the two genes (8 exons vs. 18 exons) and respective proteins, 571AA in ferroportin and 802AA in matriptase-2. Another perspective is that alterations in *SLC40A1* are of dominant inheritance; in contrast those affecting *TMPRSS6* are recessive even showcasing low penetrance[73]. Genes implied in dominantly inherited diseases, are more conserved than those that cause recessive conditions[100]. As a single heterozygous alteration in a dominant inheritance gene is capable of contributing to the pathology. As this further solidifies why one gene has more variants associated than the other. As a single heterozygous pathogenic variant in *SLC40A1* causes nefarious effects immediately, while besides a pathogenic variant a greater amount of other variants are required in *TMPRSS6* to have an impact on the phenotype, and at times this alone isn't enough[73]. Also, *TMPRSS6* is a highly polymorphic gene, which causes further variant enrichment for variant study (**Figure 4.17**)[77,85].

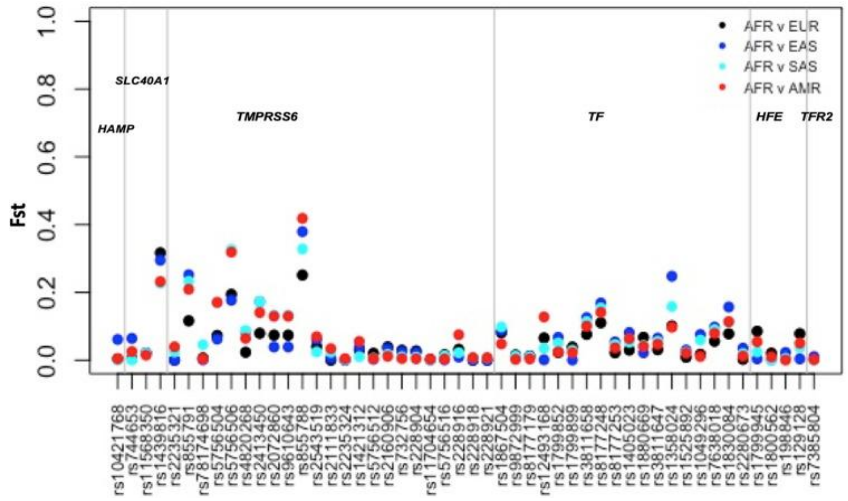

Figure 4.17: SNP comparison between African populations and others for several Iron metabolism genes. F_{st} stands for fixation index. Other populations are European, East Asian, South Asian, and American, respectively. Image from Jallow[85].

Figure 4.17 depicts common SNPs for *HAMP, SLC40A1, TMPRSS6, TF, HFE*, and *TFR2,* whose frequency is compared between the African population vs. Others (European, East Asian, South Asian, and American, respectively)[85]. This image served to illustrate the genetic differences between these populations, and the lower the fixation index, the bigger the difference between populations[85,101]. But to also illustrate the stark difference between the amounts of common SNPs presented by *TMPRSS6* in relation to the other genes, in particular *SLC40A1*. The variants in each gene, present in this work, that showcase the highest and most dispersed F_{st} values are Val727Ala (rs855791) in *TMPRSS6*, and IVS1-24 (c.44-24G>C, rs1439816) in *SLC40A1*[85].

In study conducted with a similarly sized population diagnosed with HH type I (100 Spaniards homozygous for the C282Y variant of *HFE)*, not a single pathogenic alteration for *SLC40A1* was detected[102]. This further highlights the level of conservation of *SLC40A1*. This study studied 5 polymorphisms, three of which are here present, c.-8G>C, IVS1-24, and Val221=[102].

4.2.2.1 Pathogenic Variants:

The presence of a single pathogenic variant allele causes the pathology, as *SLC40A1* alterations are of autosomal dominant nature. This means that if an individual has one alteration, their offspring also inherits the same allele, so they too will be affected by the same variant.

As previously mentioned, ferroportin is the sole iron exporter in humans. In the worst case scenario, if a pathological variant is inherited as a homozygous, it can produce a condition incompatible with life[13]. So it is of upmost importance for individuals with these types of conditions, to be informed of their reproductive possibilities and associated risks. Also which sort of support they should receive in case that they want to have fit offspring, or at least as healthy as possible.

The three pathogenic variants found belong to heterozygous females of different ages, with different phenotype severity. These differences indicate that even if they are all pathogenic variants, causing the same condition, ferroportin disease, they can have variable severity and outcomes[13,47].

4.2.2.1.1 p.Gly80Ser

Gly80Ser (c.238G>A, rs978427853) is a missense alteration in the third exon of the gene that encodes for Ferroportin (FPN). This variant is unreported in ClinVar, and in Ensembl it is reported to be likely pathogenic and pathogenic. It has not been reported in the 1000 Genomes Project nor in gnomAD, although it has been mostly associated to south Europeans, but it has also been found in a Vietnamese family[103-105]. Although we cannot establish a global population MAF for it, in our population had 0.46% MAF, making it quite rare (**Table 4.28**).

Table 4.28: Detailed information regarding the heterozygous individual affected by p.Gly80Ser.

ID	Sex	Age	Iron (µg/dL)	Ferritin (ng/mL)	TS (%)	Ref	Alt	GT	AD (Ref,Alt)	DP
91	F	41	100	**708**	27	G	A	0/1	100,125	225

Bold - values outside the reference range[31] or our criteria; **ID**- individual identification; **TS-** Transferrin Saturation; **GT**- Genotype; **AD**- Allelic Depth, **Ref** and **Alt**- Reference and Alteration; **DP-** total Depth.

This individual has normal values for iron and TS biomarkers, unlike the other individuals with pathogenic variants. This individual presented only an abnormal value for ferritin, in regards to TS, her values are also not far apart from what has been in reported literature[103-105]. These normal TS values are one of the things that distinguishes FD from HH type IV, as if it were a mutation associated to the later, the TS would be much higher[47].

Other variants in the *SLC40A1* gene found in this individual included the synonymous Val221= and the intronic c.-8 C>G in the promoter region, and IVS1-24. The *in silico* analyses results for p.Gly80Ser can be seen in the Table below.

Table 4.29: *In silico* analyses performed for p.Gly80Ser.

Gly80Ser	CADD	PROVEAN	MutPred2	HumDiv	HumVar
Score	26.9	-5.507	0.903	1.000	0.997
Prediction	Top 1%	Deleterious	*	Prob D	Prob D

*- more detailed information in the text; Prob D- Probaly Damaging.

MutPred2 predicted alterations on the transmembranar protein with a probability of 0.22 ($p=3.1x10^{-3}$). Previous studies though SIFT ranked it as probably damaging. The possible pathogenicity of Gly80Ser was studied in PolyPhen-2, the results can be sumarised beneath in Figure 4.18, and with the scores above in Table 4.29. Regarding its MSA, this an extremely conserved position of FNP, as the comparison between 68 species, only four had a different AA (valine and cysteine).

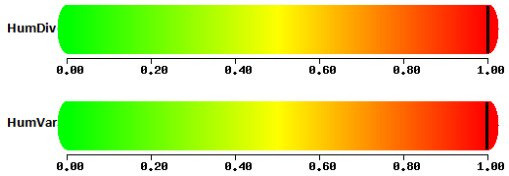

Figure 4.18: PolyPhen-2 heat bar for p.Gly80Ser. Both parameters predict this variant to be probably damaging.

This missense variant causes the replacement of glycine by serine, in the 80th residue of FPN. Glycine is the smallest AA and is apolar, serine on the other hand is a polar uncharged AA with a terminal hydroxyl in its R chain[64,105]. Therefore, besides the increased polarity, this alteration is also able to destabilize the stereochemical environment in that position of the polypeptide chain due to their significant size difference[64].

A French study with 44 HH type IV patients found 18 non-synonymous coding FPN variants[104], Gly80Ser and Val162del were both present in that population. Gly80Ser was present in 4 individuals, a family (mother, daughter and son) and an unrelated man. A milder phenotype was presented by the females, with lower ferritin levels in comparison to males, here it might be due to the protective effect of menstruation in IO, in contrast to ID.

When testing each variant *in vitro*, through a transfected embryonic kidney cell line (HEK293T) expressing FPN with these variants, they induced IO similarly to Ala77Asp[104]. This is a well known variant used as a positive control as it causes FPN loss-of-function, and thus FD (**Figure 4.19 A**)[13,14,47,104]. These results were obtained when, after 48h, they determined the intracellular ferritin levels (through ELISA)[104]. Ferritin is responsive to the levels of intracellular iron, their increased expression is directly associated to an increased iron presence within the cell[4,7].

Further *in vitro* tests, found it to be less efficient in reaching the cell's surface, a requirement for proper FPN functioning, as it is a transmembranar protein[104]. It is possible to see in Figure 4.19, that these two variants are among the most severe as they nearly equate to cells without *SLC40A1* expression[104]. These *in vitro* tests were also performed by McDonald *et al.*, obtaining the same results, and besides the use of the same cell line, they also tested with transfected hepatocytes and enterocytes[105].

The location of Gly80Ser causes conformational changes in the loop into the central helix (**Figure 4.19 B**), possibly affecting its flexibility, thus preventing the natural activity of FPN[104,105]. This goes in line with the MutPred2 findings. In spite of its similarities to Ala77Asp and proximity, reduced iron export and lower cellular superficial expression, Gly80Ser FPN has been reported by De Domenico *et*

al. to be able export iron but with a decreased sensitivity to hepcidin, which is atypical for FD, and in conflict to previous studies[104-106].

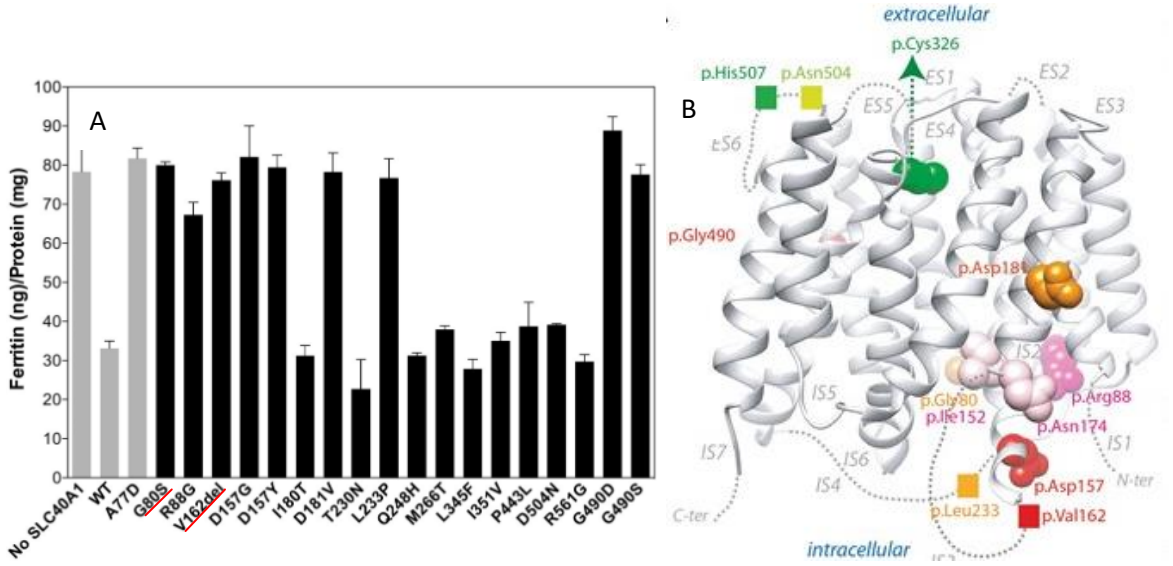

Figure 4.19: Comparison between *SLC40A1* variants effects and their location in FPN. A) Here are depicted the FD variants, and underlined in red are the variants also present in our population. **B)** Gly80 is in the bottom of a helix in the loop in the centre, in light orange, while Val162del is the end of a β sheet, red square. Adaptation from Callebaut[104].

Pietrangelo *et al.* characterised FD as reticuloendothelial macrophages IO[13]. Due to impairing FPN alterations causing loss-of-function, which consequently decrease available iron in circulation, as macrophages don't export the recycled iron[13,106]. This translates to high serum ferritin, to prevent cellular damage (ROS, ferroptosis) from IO, and low TS levels in regards to IO, as iron is retained within macrophages[7]. This explains the types of depositions seen through Magnetic resonance imaging (MRI), and why it can be used as a diagnosis tool[106]. Untreated patients with this variant showcase iron accumulation in the liver, spleen, and bone marrow, like Ala77Asp and Val162del[106]. But after phlebotomy treatment depositions were only visible in the bone marrow, unlike Ala77Asp that also had it in the spleen[106].

4.2.2.1.2 p.Val162del

This variant, c.485_487delTTG (rs1172102948), causes an inframe deletion in the 162th position of the protein[48,104,107], it affected one heterozygous individuals in our population (**Table 4.30**).

Table 4.30: Detailed information regarding the heterozygous individual affected by p.Val162del.

ID	Sex	Age	Iron (µg/dL)	Ferritin (ng/mL)	TS (%)	Ref	Alt	GT	AD (Ref,Alt)	DP
99	F	18	-	**1000**	-	TTG	-	0/1	40,26	66

Bold - values outside the reference range[31] or our criteria; **ID**- individual identification; **TS**- Transferrin Saturation; **GT**- Genotype; **AD**- Allelic Depth, **Ref** and **Alt**- Reference and Alteration; **DP**- total Depth.

It is reported as likely pathogenic both in ClinVar and Ensembl. It is an extremely rare variant in the general population, as neither the 1000 genome project nor gnomAD had a minor allele for it detected[59,65]. However it has been reported as a common genotypes associated to FD[13,47,103,104,106,107]. Its high frequency can be due to the nature of this variant type, a deletion, it has been reported in individuals of different ethnic backgrounds, and it has been linked to a slipped strand mispairing[107]. This variant also exists in other species, like zebra fish, mice and rats[107], once more hinting at being due to an error of molecular machinery due to the presence of three valines in a row.

As mentioned in the previous section, this deletion causes decreased cell surfacing and reduced iron export, which translated to increased ferritin storage (as seen in **Figure 4.19**)[104]. In regards to its position in FPN, its structure is disturbed through the destabilisation of the β-sheet, in which it is located, due to valine loss[104].

In spite of the young age of this individual, one of the characteristics of FD is its ability to affect individuals of variable ages. Unlike HH type I which mostly expresses itself in older individuals, for type IV its onset occurs after between 40-50 years of age[13]. She has high ferritin levels (1000ng/mL), but it is in line to the values presented by other individuals in literature[103,107]. Unfortunately there was no data for iron nor TS. ID99 was one of the two individuals, among the 110, that did not present the values for TS. Thus we don't have the fully displayed extension of her phenotype. Other *SLC40A1*variants found in this individual were IVS1-24G>C, IVS1-21A>C, and Val221=.

In Callebaut's study, Val162del was among the most common variants, affecting seven individuals; their mean ferritin level was 1472.15 ng/mL and most cases had TS below 25%[104]. Other studies also reported high ferritin with normal TS[107].

This variant was studied with PROVEAN's protein tool, which gave it a score of -9.423, thus predicting Val162del to be deleterious. This was the highest score given by PROVEAN to the *SLC40A1* variants. Unfortunately, as this variant is a deletion, and not a SNV, the other *in silico* tools weren't able to conduct their analyses, as they accounted mostly for missense variants. SIFT predicted this alteration to be ranked as damaging.

4.2.2.1.3 p.Gly204Ser

Gly204Ser (c.610G>A, rs387907377) is a missense alteration located in the 6th exon, (**Table 4.31**). It has a reported MAF of below 1%. The analysis, having in consideration the AA, in this alteration wasn't performed as it involves the same AA as Gly80Ser.

Table 4.31: Detailed information regarding the heterozygous individual affected by p.Gly204Ser.

ID	Sex	Age	Iron (µg/dL)	Ferritin (ng/mL)	TS(%)	Ref	Alt	GT	AD (Ref,Alt)	DP
108	F	62	**176**	**1621**	**70**	G	A	0/1	148,121	269

Bold - values outside the reference range[31] or our criteria; **ID-** individual identification; **TS-** Transferrin Saturation; **GT-** Genotype; **AD-** Allelic Depth, **Ref** and **Alt-** Reference and Alteration; **DP-** total Depth.

In literature, individuals with this variant could be split into two groups in regards to their TS; close to 75% and above, or between 20- 50%[103]. For ferritin, its values range from the smallest amounts up to 2000 ng/mL[103]. These values are in line with the phenotype presented by patient ID108.

The patient ID108 presented all the benign variants in *SLC40A1* except for Val221= (c.-8C>G, c.44-24G>C, c.44-21A>C). In regards to Gly204Ser, the results of its bioinformatic analyses are summarised in Table 4.32.

Table 4.32: *In silico* analyses performed for Gly204Ser.

Gly204Ser	CADD	PROVEAN	MutPred2	HumDiv	HumVar
Score	26.5	-4.014	0.603	1.000	0.999
Prediction	Top 1%	Deleterious	*	Prob D	Prob D

*- more detailed information in the text; **Prob D-** Probaly Damaging.

ClinVar has classified Gly204Ser as probably pathogenic variant. MutPred2 predicted it to cause alterations on the transmembranar protein with a probability of 0.27 ($p=3.1x10^{-4}$). Previous studies performed though SIFT have ranked it as probably damaging. Regarding its study by Polyphen-2, in Figure 4.20 are predicted the possible effects of this variant in FPN. According to the MSA, this is a highly conversed site, as the comparison between 69 species, only two had a different AA (cysteine).

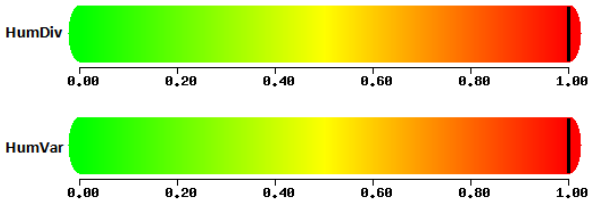

Figure 4.20: PolyPhen heat bar for p.Gly204Ser. For HumDiv the prediction had a **0.00** sensitivity and **1.00** specificity, HumVar had a sensitivity of **0.09** and a **0.99** specificity.

A Brazilian study reported this variant in a 52 year old woman, who was heterozygous for it, and at the time there were no known cases of IO in her family[108]. Her phenotype was 100% TS and 5236 ng/mL of serum ferritin[108], this phenotype is by far much more severe than the one displayed by patient ID108. Unfortunately, further information regarding other variants that could have affected this woman were not displayed. In addition, upon familial studies, it was found that both her daughters had this variant, but without symptoms. The lack of a pathological phenotype in the daughters might be due to their ages, 31 and 33 years old. As menstruation although is a risk for ID, it is protective for IO due to the periodic blood loss, but also because younger people tend to have less severe IO phenotypes, as the regulation in iron metabolism tends do decrease with aging[19].

Gly204Ser was firstly described in a 2011 study that related various factors that had an influence on FD phenotype[103]. This study had 70 participants (members of 33 families), and reported a total of 19 alterations in the *SCL40A1* gene found (**Figure 4.21**)[103], including those in the previous two sections.

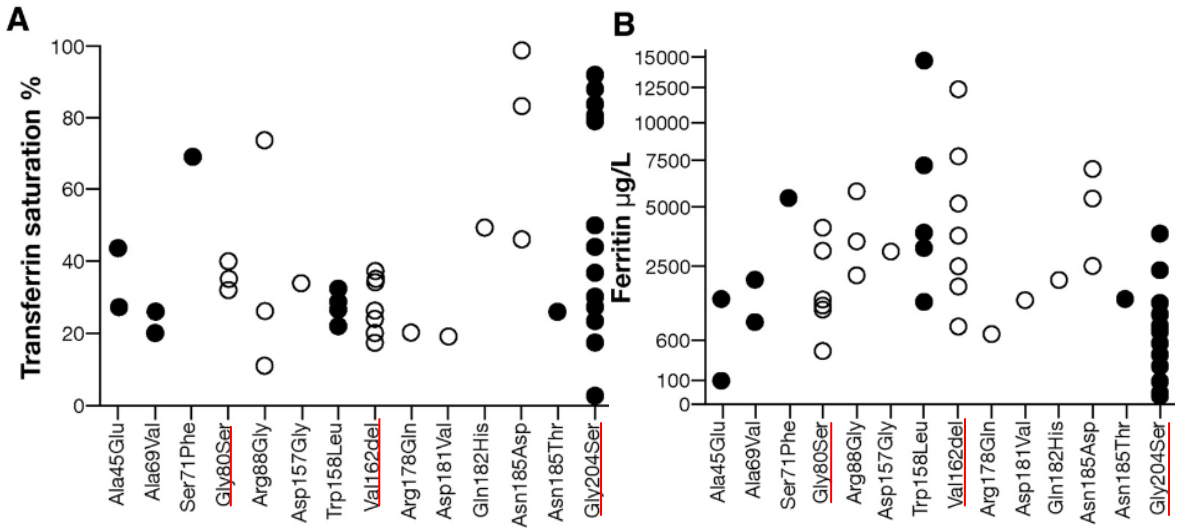

Figure 4.21: Transferrin saturation and ferritin for the variants in the Le Lan *et al* study. The novel variants are represented by the black circles, underlined in red are the variants also found in our study. Images from Le Lan[103].

It is possible to view how common this variant was in comparison to the remaining variants[103]. Figure 4.21 also allows for an easier visualization of the differences between the three pathogenic variants we found in our study. Ferritin values are quite dispersed for Val162del, while for Gly80Ser they are mostly concentrated around 1500 ng/mL mark, the same concentration marks the upper limit of the spectrum of ferritin levels for Gly204Ser (minus two outliners)[103]. As for TS levels, both Gly80Ser and Val162del present TS within the normal range, whereas Gly204Ser presents either very high TS, nearly 80% and above, or ranging around the normal values, 20-40%. These differences in ferritin and TS might indicate a difference between these three variants, although they all cause FD[13]. In liver biopsies of patients with FD, individuals with Gly204Ser present iron deposits more predominantly in hepatocytes, while for individuals with Val162del or Gly80Ser iron has a preferential macrophage deposition (Kupffer cells)[13,103].

4.2.2.2 Benign Variants:

The benign variants here are comprised by intronic variants and the synonymous variant, p.Val221= located in the 6th exon (c.663T>C, rs2304704). ClinVar regards it as benign, as does literature[105]. We tested it through CADD, which gave a predictive score of 9.973. PROVEAN was also used, but since this is a synonymous alteration, a neutral result with a score of 0.000 was given, as there is no change from an AA stand point.

Regarding its presence in populations, it had 60.7% MAF in the IBS population, in a global level it is 55%. In a study with a Brazilian sample of HH type I patients, this variant was one of the reported variants, it was present in that sample with a MAF of 39.2%, which is considerably lower than the one found for the IBS and our population[108]. In our population, this variant was slightly more present, with a MAF of 64.82%. The genotype frequency for Val221= in the IBS population was, 0/0: 11.2%, 0/1: 56.1%, 1/1: 32.7%. While in our population it was 0/0: 11.11%, 0/1: 48.15%, 1/1: 40.74%. Both populations had similar percentages for wild type genotype, however considering the individuals with the homozygous genotype for the variant, they were more frequent in our population.

4.2.2.2.1 Intronic variants:

The intronic variants were all studied though VarSEAK and were classified within Class 1, with no splicing effect predicted. They were also previously studied through SIFT and Human Slicing Finder[74]. For the most part they were predicted to not have any significant effects on splicing, but for IVS1-24 (rs1439816) it was predicted to create an intronic Exonic Splicing Silencer, that might cause alterations in the splicing site.

IVS1-24 (c.44-24G>C, rs1439816) had been previously studied in an Italian cohort diagnosed with HH type I, along with other SNPs belonging to other iron metabolism genes[109]. The MAF in that cohort was similar to ours, they found associations between the presence of at least one C allele to the modulation of biomarkers in males, particularly in serum iron and TS, although without significance ($p=0.09$ and 0.06, respectively)[109]. In females the opposite was seem, serum iron was significantly lower with the GG genotype ($p=0.01$)[109]. Its high presence in our population could thus indicate an increased IO risk, as most individuals are male (75%). However this variant also has been previously reported as being protective against IO, in a study that also mentioned c.-8G>C and Val221=[102].

4.2.3 Statistical analyses:

All the variants that appeared frequently in the IO population (c.-8C>G, IVS1-24G>C, IVS1-21A>C and Val221=), were tested regarding their relation with the serum biomarkers of iron metabolism. It is important to notice that not all iron biomarkers values were available for all samples.

Out of the tested variants, only the IVS1-21A>C variant had statistically significant results for TS (Mann-Whitney Test $p<0.001$) and Iron (ANOVA $p<0.001$). The individuals with this variant had lower mean values for these biomarkers in regards to the rest of the population without it (155.63 µg/dL for iron, and 50.83% for TS). However it isn't possible to establish if this is due to a protective effect of this variant on those biomarkers or not, in regards to ferritin, as most individuals had it above 1000 ng/mL.

4.2.4 Concluding IO remarks:

After analysing the variants found in the *SLC40A1* gene in a group of 110 patients with a phenotype of iron overload, we have found three known pathogenic variants causative of FD: p.Gly80Ser, p.Val162del, and p.Gly204Ser. The iron biomarkers presented by these affected individuals were in line to the phenotypes described in literature for other patients presenting the same mutations.

In regards to the genotype-phenotype study, only for IVS1-21A>C significant differences were found, for a lower serum iron and TS, in regards the rest of the population by the individuals with this variant.

The *SLC40A1* gene revealed a lower number of variants in comparison to *TMPRSS6*, and that can be explained by its high sequence conservation degree in regard to the highly polymorphic nature of the latter. Also, it can be because it encodes for the sole iron exporter, whose variants are of dominant inheritance. If the variants are highly pathogenic, it might produce an unviable individual, preventing its transmission, or in the worse scenario be incompatible with life.

5 Conclusions

The main goal of this research was to study genetic disturbances in the iron metabolism from two different points. Iron deficiency was studied through the investigation of genetic variants in *TMPRSS6* gene in a group of 100 individuals with an ID-like phenotype. On the other hand, iron overload was studied through *SLC40A1* analysis, which is responsible for the Ferroportin Disease and Hereditary Haemochromatosis type IV.

Most of this work was devoted to iron deficiency as it is a far more common problem, affecting millions of people, but also because *TMPRSS6* is highly polymorphic gene, especially when compared to an extremely conserved gene like *SLC40A1*. Despite the differences in the phenotypes presented in the pathologies of both genes, they come together due to the complexity of iron metabolism[109,110].

Through NGS, it was possible to detect and investigate 36 variants for *TMPRSS6*, and 7 variants for *SLC40A1*. While the variants for the latter could be more easily split into benign and pathogenic categories, their analyses were more difficult given that there isn't as much information regarding IO, in comparison to ID.

As expected, an overrepresentation of pathogenic and risk variants were found in our populations, when comparing with what is described in the literature for the general population.

The individuals from our ID population were not part of a clinical study[50]. So it isn't possible to further study their backgrounds, to have a better insight of the variants and their expression among different individuals from the same family. As for the individuals of the IO population, came with a medical referral, the recall for further pedigree investigation could be more easily achieved in the future.

Throughout this study we were able to accomplish all our intended goals. From the classification of the genetic variants found, to the assessment of genotype and phenotype associations. To the characterisation of functional polymorphism's role in iron deficiency development susceptibility. In addition, it was possible to characterize the molecular basis and understand the pathogenic mechanisms subjacent to the three rare cases of Ferroportin Disease, in our iron overload population.

It would be interesting to further study the both populations in regard to other iron biomarkers, especially to assess their hepcidin levels. In the case of the novel pathogenic variants of *TMPRSS6* here described, c.1580T>G (p.Phe527Cys) and c.1585T>C (p.Cys529Arg), if both are found to be associated to high hepcidin levels it will prove, *in vivo*, their pathogenicity. In Ferroportin disease, the pathogenic mechanism does neither involve synthesis nor regulation of hepcidin, but its target protein ferroportin. So in contrast to maptripase-2 affections, hepcidin should be appropriately synthesised and released in response to increased serum iron levels.

In order to complement our study about *TMPRSS6* and *SLC40A1* genetic variants' frequencies, it would be interesting to have a healthy sample control population, to compare to what was found in our two groups of iron-deregulated patients.

Functional studies should be performed as *in silico* tools can help elucidate the potentiality pathogenicity, or not, of different variants, but they do not replace *in vitro* or *in vivo* tests, as these are closer to the physiological conditions in which these pathologies occur.

6 References:

1. Yiannikourides A, Latunde-Dada GO. A short review of iron metabolism and pathophysiology of iron disorders. *Medicines.* 2019; 6(3)85.

2. Vogt A-CS, Arsiwala T, Mohsen M, Vogel M, Manolova V, Bachmann MF. On Iron Metabolism and Its Regulation. On Iron Metabolism and Its Regulation. *Int. J. Mol. Sci. 2021;* 22(9):4591.

3. Liberal A, Pinela J, Vívar-Quintana AM, Ferreira ICFR, Barros L. Fighting iron deficiency anemia: innovations in food fortificants and biofortification strategies. *Foods.* 2020; 9(12)1871.

4. Bou-Abdallah F. The iron redox and hydrolysis chemistry of the ferritins. *Biochim Biophys Acta*. 2010;1800(8):719-731.

5. Pantopoulos K, Porwal SK, Tartakoff A, Devireddy L. Mechanisms of mammalian iron homeostasis, *Biochemistry*. 2012; *51*(29):5705–5724.

6. Roemhild K, von Maltzahn F, Weiskirchen R, Knüchel R, von Stillfried S, Lammers T. Iron metabolism: pathophysiology and pharmacology. *Trends Pharmacol Sci*. 2021;42(8):640-656.

7. Muhoberac BB, Vidal R. Abnormal iron homeostasis and neurodegeneration. *Front Aging Neurosci*. 2013;5:32.

8. Camaschella C, Nai A, Silvestri L. Iron metabolism and iron disorders revisited in the hepcidin era. *Haematologica*. 2020;105(2):260-272.

9. Muñoz M, García-Erce JA, Remacha AF. Disorders of iron metabolism. Part I: molecular basis of iron homoeostasis. *J Clin Pathol*. 2011;64(4):281-286.

10. Muñoz M, García-Erce JA, Remacha AF. Disorders of iron metabolism. Part II: iron deficiency and iron overload. *J Clin Pathol*. 2011;64(4):287-296.

11. Abbaspour N, Hurrell R, Kelishadi R. Review on iron and its importance for human health. *J Res Med Sci*. 2014;19(2):164-174.

12. Pantopoulos K. Inherited Disorders of Iron Overload. *Front Nutr.* 2018;103(5):1-11.

13. Pietrangelo A. Ferroportin disease: pathogenesis, diagnosis and treatment. *Haematologica*. 2017;102(12):1972-1984.

14. Le Gac G, Férec C. The molecular genetics of haemochromatosis. *Eur J Hum Genet*. 2005;13(11):1172-1185.

15. Yan N, Zhang J. Iron Metabolism, Ferroptosis, and the Links With Alzheimer's Disease. *Front Neurosci*. 2020;13:1443.

16. Kaplan J, Ward DM. The essential nature of iron usage and regulation. *Curr Biol*. 2013;23(15)642:646 .

17. Yi J, Thomas LM, Musayev FN, *et al.* Crystallographic trapping of heme loss intermediates during the nitrite-induced degradation of human hemoglobin. Biochemistry. 2011;50(39):8323-8332.

18. Wahed A, Dasgupta A. Chapter 4: Hemoglobinopathies and Thalassemias. *Hematology and coagulation.* Elsevier; 2015;51-58.

19. Ashraf A, Clark M, So PW. The Aging of Iron Man. *Front Aging Neurosci*. 2018;65(10):1-23.

20. Pagani A, Nai A, Silvestri L, Camaschella C. Hepcidin and Anemia: A Tight Relationship. *Front Physiol*. 2019;1294(10):1-7.

21. Faria R, Silva B, Silva C, *et al.* Next-generation sequencing of hereditary hemochromatosis-related genes: Novel likely pathogenic variants found in the Portuguese population. *Blood Cells Mol Dis*. 2016;61:10-15.

22. Camaschella C. Iron deficiency. *Blood*. 2019;133(1):30-39.

23. Wang CY, Meynard D, Lin HY. The role of TMPRSS6/matriptase-2 in iron regulation and anemia. *Front Pharmacol*. 2014;5:114.

24. Poggiali E, Andreozzi F, Nava I, Consonni D, Graziadei G, Cappellini MD. The role of TMPRSS6 polymorphisms in iron deficiency anemia partially responsive to oral iron treatment. *Am J Hematol*. 2015;90(4):306-309.

25. Urrechaga E, Hoffmann JJML. Critical appraisal of discriminant formulas for distinguishing thalassemia from iron deficiency in patients with microcytic anemia. *Clin Chem Lab Med*. 2017;55(10):1582-1591.

26. Samões C, Kislaya I, Sousa-Uva M. *et al.* Prevalence of anemia in the Portuguese adult population: results from the first National Health Examination Survey (INSEF 2015). *J Public Health (Berl.)* 2020

27. Li Y, Huang X, Wang J, Huang R, Wan D. Regulation of Iron Homeostasis and Related Diseases. *Mediators Inflamm*. 2020;2020:6062094.

28. Silva B, Faustino P. An overview of molecular basis of iron metabolism regulation and the associated pathologies. *Biochim Biophys Acta*. 2015;1852(7):1347-1359.

29. Kasvosve I. Effect of ferroportin polymorphism on iron homeostasis and infection. *Clin Chim Acta*. 2013;416:20-25.

30. Goddard AF, James MW, McIntyre AS, Scott BB; British Society of Gastroenterology. Guidelines for the management of iron deficiency anaemia. *Gut*. 2011;60(10):1309-1316.

31. *Direcção Geral de Saúde. Norma nº 063/2011, actualizada a 12/09/2013* (https://nocs.pt/prescricao-determinacao-hemograma/)

32. Elstrott B, Khan L, Olson S, Raghunathan V, DeLoughery T, Shatzel JJ. The role of iron repletion in adult iron deficiency anemia and other diseases. *Eur J Haematol*. 2020;104(3):153-161.

33. Migone De Amicis M, Rimondi A, Elli L, Motta I. Acquired Refractory Iron Deficiency Anemia. *Mediterr J Hematol Infect Dis*. 2021;13(1):e2021028.

34. Janel A, Roszyk L, Rapatel C, Mareynat G, Berger MG, Serre-Sapin AF. Proposal of a score combining red blood cell indices for early differentiation of beta-thalassemia minor from iron deficiency anemia. *Hematology*. 2011;16(2):123-127.

35. Jahangiri M, Rahim F, Malehi AS. Diagnostic performance of hematological discrimination indices to discriminate between βeta thalassemia trait and iron deficiency anemia and using cluster analysis: Introducing two new indices tested in Iranian population. *Sci Rep*. 2019;9(1):18610.

36. Hoffmann JJML, Urrechaga E. Verification of 20 Mathematical Formulas for Discriminating Between Iron Deficiency Anemia and Thalassemia Trait in Microcytic Anemia. *Lab Med*. 2020;51(6):628-634.

37. Cui Y, Wu Q, Zhou Y. Iron-refractory iron deficiency anemia: new molecular mechanisms. *Kidney Int*. 2009;76(11):1137-1141.

38. Asperti M, Brilli E, Denardo A, *et al.* Iron distribution in different tissues of homozygous Mask (msk/msk) mice and the effects of oral iron treatments. *Am J Hematol*. 2021;96(10):1253-1263.

39. Kloss-Brandstätter A, Erhart G, Lamina C, *et al.* Candidate gene sequencing of SLC11A2 and TMPRSS6 in a family with severe anaemia: common SNPs, rare haplotypes, no causative mutation. *PLoS One*. 2012;7(4):e35015.

40. Gómez-Ramírez S, Brilli E, Tarantino G, Muñoz M. Sucrosomial® Iron: A New Generation Iron for Improving Oral Supplementation. *Pharmaceuticals (Basel)*. 2018;11(4):97.

41. Nemeth E, Ganz T. Hepcidin-Ferroportin Interaction Controls Systemic Iron Homeostasis. *Int J Mol Sci*. 2021;22(12):6493.

42.Gichohi-Wainaina WN, Towers GW, Swinkels DW, Zimmermann MB, Feskens EJ, Melse-Boonstra A. Inter-ethnic differences in genetic variants within the transmembrane protease, serine 6 (TMPRSS6) gene associated with iron status indicators: a systematic review with meta-analyses. *Genes Nutr*. 2015;10(1):442.

43. Carlton VE, Ireland JS, Useche F, Faham M. Functional single nucleotide polymorphism-based association studies. *Hum Genomics*. 2006;2(6):391-402.

44. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet*. 2020;11:424.

45. Toste S, Relvas L, Pinto C, *et al.* Intragenic haplotype analysis of common HFE mutations in the Portuguese population. *J Genet*. 2015;94(2):329-333.

46. Kowdley KV, Brown KE, Ahn J, Sundaram V. ACG Clinical Guideline: Hereditary Hemochromatosis. *Am J Gastroenterol*. 2019;114(8):1202-1218.

47. Kowdley DS, Kowdley KV. Appropriate Clinical Genetic Testing of Hemochromatosis Type 2-4, Including Ferroportin Disease. *Appl Clin Genet*. 2021;14:353-361.

48. Wu L, Zhang W, Li Y, *et al.* Correlation of genotype and phenotype in 32 patients with hereditary hemochromatosis in China. *Orphanet J Rare Dis*. 2021;16(1):398.

49. Moreno-Carralero MI, Muñoz-Muñoz JA, Cuadrado-Grande N, et al. A novel mutation in the SLC40A1 gene associated with reduced iron export in vitro. *Am J Hematol*. 2014;89(7):689-694.

50. Nunes B, Barreto M, Gil AP, *et al.* The first Portuguese National Health Examination Survey (2015): design, planning and implementation. *J Public Health (Oxf)*. 2019;41(3):511-517.

51. Jia H, Guo Y, Zhao W, Wang K. Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. *Sci Rep*. 2014;4:5737.

52. Ravi RK, Walton K, Khosroheidari M. MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. *Methods Mol Biol*. 2018;1706:223-232.

53. Hess JF, Kohl TA, Kotrová M, *et al.* Library preparation for next generation sequencing: A review of automation strategies. *Biotechnol Adv*. 2020;41:107537.

54. McLaren W, Gil L, Hunt SE, *et al.* The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.

55. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*. 2013;7(7):20.

56. Pejaver V, Urresti J, Lugo-Martinez J, *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat Commun*. 2020;11(1):5918.

57. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31(16):2745-2747.

58. Sandell L, Sharp NP. Fitness Effects of Mutations: An Assessment of PROVEAN Predictions Using Mutation Accumulation Data. *Genome Biol Evol*. 2022;14(1):evac004.

59. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315.

60. Montenegro LR, Lerário AM, Nishi MY, Jorge AAL, Mendonca BB. Performance of mutation pathogenicity prediction tools on missense variants associated with 46,XY differences of sex development. *Clinics (Sao Paulo)*. 2021;76:e2052.

61. Young AI, Benonisdottir S, Przeworski M, Kong A. Deconstructing the sources of genotype-phenotype associations in humans. *Science*. 2019;365(6460):1396-1400.

62. Simonin-Wilmer I, Orozco-Del-Pino P, Bishop DT, Iles MM, Robles-Espinoza CD. An Overview of Strategies for Detecting Genotype-Phenotype Associations Across Ancestrally Diverse Populations. *Front Genet*. 2021;12:703901.

63. Den Dunnen JT, Dalgleish R, Maglott DR, *et al.* HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat*. 2016;37(6):564-569.

64. Nelson DL, Cox MM. Chapter 3: Amino Acids, Peptides, and Proteins. *Lehninger Principles of Biochemistry*. 6th ed. Macmillian Higher Education; 2013;76-81.

65. Genomes Project Consortium, Auton A, Brooks LD, *et al.* A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

66. Hernández CL, Pita G, Cavadas B, *et al.* Human Genomic Diversity Where the Mediterranean Joins the Atlantic. *Mol Biol Evol*. 2020;37(4):1041-1055.

67. Lek M, Karczewski KJ, Minikel EV, *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291.

68. McDonald CJ, Ostini L, Bennett N, *et al.* Functional analysis of matriptase-2 mutations and domains: insights into the molecular basis of iron-refractory iron deficiency anemia. *Am J Physiol Cell Physiol*. 2015;308(7):C539-C547.

69. Pei SN, Ma MC, You HL, *et al.* TMPRSS6 rs855791 polymorphism influences the susceptibility to iron deficiency anemia in women at reproductive age. *Int J Med Sci*. 2014;11(6):614-619.

70. Dotlic J, Pimenta F, Kovacevic N, *et al*. Menopausal transition in Southern Europe: comparative study of women in Serbia and Portugal. *Menopause*. 2017;24(11):1236-1245.

71. Beutler E, Van Geet C, te Loo DM, *et al.* Polymorphisms and mutations of human TMPRSS6 in iron deficiency anemia. *Blood Cells Mol Dis*. 2010;44(1):16-21.

72. Jallow MW, Campino S, Saidykhan A, Prentice AM, Cerami C. Common Variants in the *TMPRSS6* Gene Alter Hepcidin but not Plasma Iron in Response to Oral Iron in Healthy Gambian Adults: A Recall-by-Genotype Study. *Curr Dev Nutr*. 2021;5(3):nzab014.

73. Donker AE, Schaap CC, Novotny VM, *et al*. Iron refractory iron deficiency anemia: a heterogeneous disease that is not always iron refractory. *Am J Hematol*. 2016;91(12):E482-E490.

74. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 2012;40(Web Server issue):W452-W457.

75. Vuckovic D, Bao EL, Akbari P, *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*. 2020 Sep;182(5):1214-1231.e11.

76. Bhatia P, Jain R, Singh A. A structured approach to iron refractory iron deficiency anemia (IRIDA) diagnosis (SAID): The more is "SAID" about iron, the less it is. *Pediatr Hematol Oncol J* 2017;2:48-53.

77. Lee P. Role of matriptase-2 (TMPRSS6) in iron metabolism. *Acta Haematol*. 2009;122(3):87-96.

78. Khanna T, Hanna G, Sternberg MJE, David A. Missense3D-DB web catalogue: an atom-based analysis and repository of 4M human protein-coding genetic variants. *Hum Genet*. 2021;140(5):805-812.

79. Li L, Vorobyov I, Allen TW. The different interactions of lysine and arginine side chains with lipid membranes. *J Phys Chem B*. 2013;117(40):11906-11920.

80. Hoch NC, Polo LM. ADP-ribosylation: from molecular mechanisms to human disease. *Genet Mol Biol*. 2019;43(1):e20190075.

81. Shihab HA, Gough J, Cooper DN, *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat*. 2013;34(1):57-65.

82. Vears DF, Niemiec E, Howard HC, Borry P. Analysis of VUS reporting, variant reinterpretation and recontact policies in clinical genomic sequencing consent forms. *Eur J Hum Genet*. 2018;26(12):1743-1751. doi:10.1038/s41431-018-0239-7

83. Luzzatto L. Sickle cell anaemia and malaria. *Mediterr J Hematol Infect Dis*. 2012;4(1):e2012065. doi:10.4084/MJHID.2012.065

84. Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443.

85. Jallow MW, Cerami C, Clark TG, Prentice AM, Campino S. Differences in the frequency of genetic variants associated with iron imbalance among global populations. *PLoS One*. 2020;15(7):e0235141.

86. Chambers JC, Zhang W, Li Y, *et al.* Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. *Nat Genet*. 2009;41(11):1170-1172.

87. Carlton VE, Ireland JS, Useche F, Faham M. Functional single nucleotide polymorphism-based association studies. *Hum Genomics*. 2006;2(6):391-402.

88. Lee PL, Barton JC, Khaw PL, Bhattacharjee SY, Barton JC. Common TMPRSS6 mutations and iron, erythrocyte, and pica phenotypes in 48 women with iron deficiency or depletion. *Blood Cells Mol Dis*. 2012;48(2):124-127.

89. Al-Jamea LH, Woodman A, Heiba NM, *et al.* Genetic analysis of TMPRSS6 gene in Saudi female patients with iron deficiency anemia. *Hematol Oncol Stem Cell Ther*. 2021;14(1):41-50.

90. Gan W, Guan Y, Wu Q, *et al.* Association of TMPRSS6 polymorphisms with ferritin, hemoglobin, and type 2 diabetes risk in a Chinese Han population. *Am J Clin Nutr*. 2012;95(3):626-632.

91. Mete M, Trabulus DC, Talu CK, *et al.* An investigation of the relationship between TMPRSS6 gene expression, genetic variants and clinical findings in breast cancer. *Mol Biol Rep*. 2020;47(6):4225-4231.

92. Buerkli S, Pei SN, Hsiao SC, *et al.* The *TMPRSS6* variant (SNP rs855791) affects iron metabolism and oral iron absorption - a stable iron isotope study in Taiwanese women. *Haematologica*. 2021;106(11):2897-2905.

93. Pelusi S, Girelli D, Rametta R, *et al.* The A736V TMPRSS6 polymorphism influences hepcidin and iron metabolism in chronic hemodialysis patients: TMPRSS6 and hepcidin in hemodialysis. *BMC Nephrol*. 2013;14:48.

94. Kahraman C, Turgay F, Yigittürk O, Canüzmez AE, Durmaz B, Aşikovali S. Does the TMPRSS6 C > T Polymorphism Modify the Endurance Training Effects on Hematological Parameters? *Biol Trace Elem Res*. 2021;10.1007/s12011-021-02876-y.

95. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9(6):477-85.

96. Pellegrino RM, Coutinho M, D'Ascola D, *et al*. Two novel mutations in the tmprss6 gene associated with iron-refractory iron-deficiency anaemia (irida) and partial expression in the heterozygous form. *Br J Haematol*. 2012;158(5):668-672.

97. Li J, Lange LA, Duan Q, *et al.* Genome-wide admixture and association study of serum iron, ferritin, transferrin saturation and total iron binding capacity in African Americans. *Hum Mol Genet*. 2015;24(2):572-581.

98. Penkova M, Ivanova N. Serum Iron Metabolism Variables in Clinically Healthy Persons. *Open Access Maced J Med Sci*. 2019; 7(3):318-321.

99. Piperno A, Pelucchi S, Mariani R. Inherited iron overload disorders. *Transl Gastroenterol Hepatol*. 2020; 5:25.

100. Furney SJ, Albà MM, López-Bigas N. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. *BMC Genomics*. 2006; 7:165.

101. Jakobsson M, Edge MD, Rosenberg NA. The relationship between F(ST) and the frequency of the most frequent allele. *Genetics*. 2013;193(2):515-528.

102. Altès A, Bach V, Ruiz A, *et al.* Does the SLC40A1 gene modify HFE-related haemochromatosis phenotypes?. *Ann Hematol*. 2009;88(4):341-345.

103. Le Lan C, Mosser A, Ropert M, *et al*. Sex and acquired cofactors determine phenotypes of ferroportin disease. *Gastroenterology*. 2011;140(4):1199-1207.e12072.

104. Callebaut I, Joubrel R, Pissard S, *et al*. Comprehensive functional annotation of 18 missense mutations found in suspected hemochromatosis type 4 patients. *Hum Mol Genet*. 2014;23(17):4479-4490. doi:10.1093/hmg/ddu160

105. McDonald CJ, Wallace DF, Ostini L, Bell SJ, Demediuk B, Subramaniam VN. G80S-linked ferroportin disease: classical ferroportin disease in an Asian family and reclassification of the mutant as iron transport defective. *J Hepatol.* 2011; 54(3):538-544.

106. Pietrangelo A, Corradini E, Ferrara F, *et al.* Magnetic resonance imaging to identify classic and nonclassic forms of ferroportin disease. *Blood Cells Mol Dis*. 2006;37(3):192-196.

107. Roetto A, Merryweather-Clarke AT, Daraio F, *et al.* A valine deletion of ferroportin 1: a common mutation in hemochromastosis type 4. *Blood*. 2002;100(2):733-734.

108. Santos PC, Cançado RD, Pereira AC, *et al.* Hereditary hemochromatosis: mutations in genes involved in iron homeostasis in Brazilian patients. *Blood Cells Mol Dis*. 2011;46(4):302-307.

109. Radio FC, Majore S, Aurizi C, *et al.* Hereditary hemochromatosis type 1 phenotype modifiers in Italian patients. The controversial role of variants in HAMP, BMP2, FTL and SLC40A1 genes. *Blood Cells Mol Dis*. 2015;55(1):71-75.

110. McLaren CE, McLachlan S, Garner CP, *et al.* Associations between single nucleotide polymorphisms in iron-related genes and iron status in multiethnic populations. *PLoS One*. 2012;7(6):e38339.

7 Supplementary material

7.1 Additional Protein and Amino Acids information

- **FASTA format for each protein from UniProt:**

For Matriptase-2:
>sp|Q8IU80|TMPS6_HUMAN Transmembrane protease serine 6 OS=Homo sapiens OX=9606 GN=TMPRSS6 PE=1 SV=3
MPVAEAPQVAGGQGDGGDGEEAEPEGMFKACEDSKRKARGYLRLVPLFVLLALLVLASAGVLLWYFL
GYKAEVMVSQVYSGSLRVLNRHFSQDLTRRESSAFRSETAKAQKMLKELITSTRLGTYYNSSSVYSFGE
GPLTCFFWFILQIPEHRRLMLSPEVVQALLVEELLSTVNSSAAVPYRAEYEVDPEGLVILEASVKDIAAL
NSTLGCYRYSYVGQGQVLRLKGPDHLASSCLWHLQGPKDLMLKLRLEWTLAECRDRLAMYDVAGPL
EKRLITSVYGCSRQEPVVEVLASGAIMAVVWKKGLHSYYDPFVLSVQPVVFQACEVNLTLDNRLDSQG
VLSTPYFPSYYSPQTHCSWHLTVPSLDYGLALWFDAYALRRQKYDLPCTQGQWTIQNRRLCGLRILQP
YAERIPVVATAGITINFTSQISLTGPGVRVHYGLYNQSDPCPGEFLCSVNGLCVPACDGVKDCPNGLDER
NCVCRATFQCKEDSTCISLPKVCDGQPDCLNGSDEEQCQEGVPCGTFTFQCEDRSCVKKPNPQCDGRPD
CRDGSDEEHCDCGLQGPSSRIVGGAVSSEGEWPWQASLQVRGRHICGGALIADRWVITAAHCFQEDSM
ASTVLWTVFLGKVWQNSRWPGEVSFKVSRLLLHPYHEEDSHDYDVALLQLDHPVVRSAAVRPVCLPA
RSHFFEPGLHCWITGWGALREGGPISNALQKVDVQLIPQDLCSEVYRYQVTPRMLCAGYRKGKKDACQ
GDSGGPLVCKALSGRWFLAGLVSWGLGCGRPNYFGVYTRITGVISWIQQVVT

For Ferroportin:
>sp|Q9NP59|S40A1_HUMAN Solute carrier family 40 member 1 OS=Homo sapiens OX=9606 GN=SLC40A1 PE=1 SV=1
MTRAGDHNRQRGCCGSLADYLTSAKFLLYLGHSLSTWGDRMWHFAVSVFLVELYGNSLLLTAVYGL
VVAGSVLVLGAIIGDWVDKNARLKVAQTSLVVQNVSVILCGIILMMVFLHKHELLTMYHGWVLTSCYI
LIITIANIANLASTATAITIQRDWIVVVAGEDRSKLANMNATIRRIDQLTNILAPMAVGQIMTFGSPVIGCG
FISGWNLVSMCVEYVLLWKVYQKTPALAVKAGLKEEETELKQLNLHKDTEPKPLEGTHLMGVKDSNI
HELEHEQEPTCASQMAEPFRTFRDGWVSYYNQPVFLAGMGLAFLYMTVLGFDCITTGYAYTQGLSGSI
LSILMGASAITGIMGTVAFTWLRRKCGLVRTGLISGLAQLSCLILCVISVFMPGSPLDLSVSPFEDIRSRFI
QGESITPTKIPEITTEIYMSNGSNSANIVPETSPESVPIISVSLLFAGVIAARIGLWSFDLTVTQLLQENVIES
ERGIINGVQNSMNYLLDLLHFIMVILAPNPEAFGLLVLISVSFVAMGHIMYFRFAQNTLGNKLFACGPDA
KEVRKENQANTSVV

Table S.1: IUPAC nomenclature for nucleotides and Amino Acids.

Nucleotide codes	Base	Amino Acid code	Three letter code	Amino Acid
R	A or G	A	Ala	Alanine
Y	C or T	C	Cys	Cysteine
S	G or C	D	Asp	Aspartic Acid
W	A or T	E	Glu	Glutamic Acid
K	G or T	F	Phe	Phenylalanine
M	A or C	G	Gly	Glycine
N	any base	H	His	Histidine
. or -	gap	I	Ile	Isoleucine
		K	Lys	Lysine
		L	Leu	Leucine
		M	Met	Methionine
		N	Asn	Asparagine
		P	Pro	Proline
		Q	Gln	Glutamine
		R	Arg	Arginine
		S	Ser	Serine
		T	Thr	Threonine
		V	Val	Valine
		W	Trp	Tryptophan
		Y	Tyr	Tyrosine

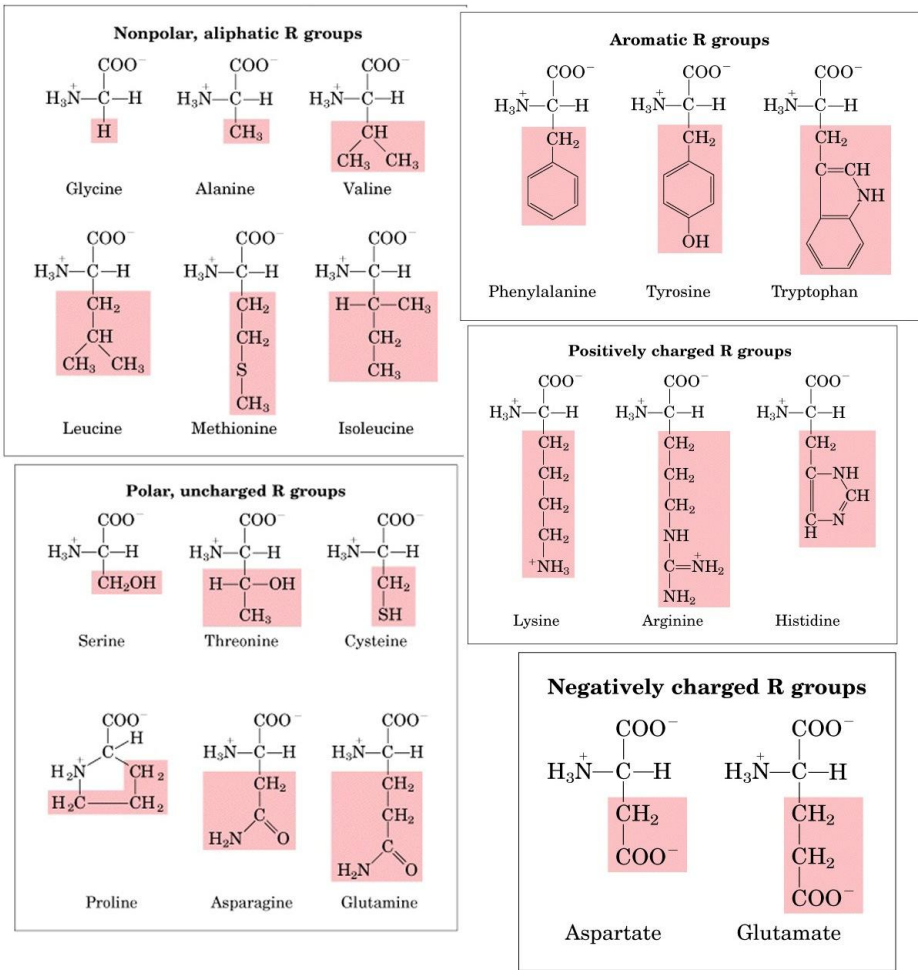

Figure S.1: Amino Acids placed within their respective R group at pH 7, regarding their pK_a. Image from Lehninger[64].

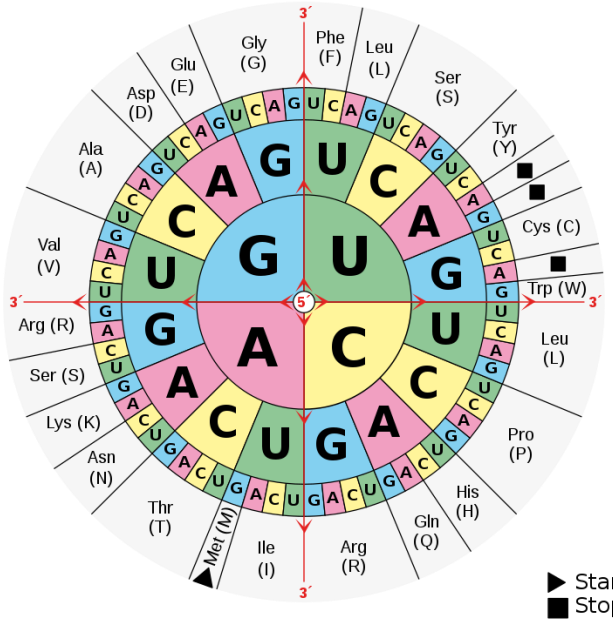

Figure S.2: Amino Acids and nucleotides wheel. Image from https://jgi.doe.gov/proving-codon-genetic-code-flexibility/.

7.2 Additional information regarding the primers and protocols

Table S.2: Primers used for *TMPRSS6* amplification and sequencing[39].

Target	Primer	Sequence	Fragment size (bp)
Exon 1	TMPRSS6_Ex1_Fw TMPRSS6_Ex1_RV	5' – CTGAGACCTCCGTCTGTCCTC – 3' 5' – TGGAAACAGCCTCGCATTTG – 3'	271
Exon 2	TMPRSS6_Ex2_RV TMPRSS6_Ex2_Fw	5' – TGCCGCCTGATGTTGTTACTC – 3' 5' – GCCTGCTACAGTCACCCCAAG – 3'	395
Exon 3	TMPRSS6_Ex3_RV TMPRSS6_Ex3_Fw	5' – GCAGGAGAAGGCATGGAAGAG – 3' 5' – TCCCTGTGAATGCTCCAGATG – 3'	323
Exon 4	TMPRSS6_Ex4_Fw TMPRSS6_Ex4_RV	5' – AGTAGGAGCAAAGGGCACCTC – 3' 5' – GACATGCAGGAAGCCAAGTTC – 3'	301
Exon 5	TMPRSS6_Ex5_Fw TMPRSS6_Ex5_RV	5' – CTTCTGCGTGAAGACGGACAG – 3' 5' – GGCCACACCACAGCTTGTTTC – 3'	378
Exon 6	TMPRSS6_Ex6_RV TMPRSS6_Ex6_Fw	5' – AGACAAGGCTGGCTCCAAGG – 3' 5' – CCCTGCACACACAACAGAAGC – 3'	255
Exon 7	TMPRSS6_Ex7_Fw TMPRSS6_Ex7_RV	5' – AGGCGTGAAGCTCAGTGTGTG – 3' 5' – CTAGCCGTCCTGTCTCCCAGA – 3'	584
Exon 8	TMPRSS6_Ex8_Fw TMPRSS6_Ex8_RV	5' – GATGTCCAGACTCCCGTCCAC – 3' 5' – GAATCTTCCCTCTCCCCATCC – 3'	364
Exon 9	TMPRSS6_Ex9_Fw TMPRSS6_Ex9_RV	5' – ATTTGCTGGCAGAGGTGGTAG – 3' 5' – GGAAACACAGAATCCCAGGTG – 3'	458
Exon 10	TMPRSS6_Ex10_Fw TMPRSS6_Ex10_RV	5' – TGTTGTTAGGGAGGTGGGTTCAC – 3' 5' – GAGATTGGGGACTTGGGCTTC – 3'	287
Exon 11	TMPRSS6_Ex11_Fw TMPRSS6_Ex11_RV	5' – AGGGAGAAATCAGGGCAGAGG – 3' 5' – CCTTGGTGGTTCCAGGGATG – 3'	356
Exon 12	TMPRSS6_Ex12_RV TMPRSS6_Ex12_Fw	5' – GCCACAAGGGTTTGCAGGAAT – 3' 5' – AGTAGGAGCAAAGGGCACCTC – 3'	523
Exon 13	TMPRSS6_Ex13_Fw TMPRSS6_Ex13_RV	5' – GTGATTGGTAACGTGCAATACAGC – 3' 5' – TGAAGCATGTAGCAGGCCTAGA – 3'	285
Exon 14	TMPRSS6_Ex14_Fw TMPRSS6_Ex14_RV	5' – CTCTTCTGGCTCCATCGTTCC – 3' 5' – TGAGATTTCCCTCCAGCTTCC – 3'	295
Exon 15	TMPRSS6_Ex15_Fw TMPRSS6_Ex15_RV	5' – TCTCCCCCTCCATCATTCTCC – 3' 5' – CCCCACCCTTCCCTCTATCTG – 3'	399
Exon 16	TMPRSS6_Ex16_Fw TMPRSS6_Ex16_RV	5' – ACCACCAGCTAGGCGACCTTC – 3' 5' – GCCCAATTTGAATCCCAGCAC – 3'	571
Exon 17	TMPRSS6_Ex17_RV TMPRSS6_Ex17_Fw	5' – GTGGGCAGAGCAGGAGAGAAG – 3' 5' – GATGTGAGCAAAGGGCCAGAC – 3'	337
Exon 18	TMPRSS6_Ex18_RV TMPRSS6_Ex18_Fw	5' – CCCAGTCAATTCCCAACAGTC – 3' 5' – GAATACTTGTCCCCCTGCTTG – 3'	344

Table S.3: Master mix for Exon 16 of *TMPRSS6* gene PCR amplification.

Products Ex 16	Per reaction (µL)	PCR conditions		T(C°)	Δt
Buffer α	2.5	**Initial Denaturation**		95	5m
DMSO	2.5	**Denaturation**		95	20s
BSA	0.425	**Hybridisation**	**32x**	64	20s
MgCl2 (0,1 mM)	0.5	**Extension**		72	20s
dNTPs	0.5	**Final Extension**		72	3m
Ex16_Fw 25 µM (25pmol/ µL)	0.5	**Pause**		4	15
Ex16_Rv 25 µM (25pmol/ µL)	0.5				
Taq Applied Byosystems (5U/µL)	0.1				
ddH2O	16.9				
Volume per PCR:	24				
30ng/µL DNA	1				

- **Protocol for *TMPRSS6* Exon 16 DNA purification in kits (JetQuick® DNA Purification Kit, Genomed):**

1. Heat 30 µL of Elution buffer in a bath between 65-70˚C
2. Into a 1,5ml Eppendorf add the amplified DNA and ddH_2O to obtain a final volume of 100 µL
3. Add to the previous tube, 400 µL of Buffer H1
4. Place a column into a collecting tube, and into the column add 500 µL from the Eppendorf of step 3 → **Centrifuge for 1 min at 12 000 rpm**, throw away the collecting tube
5. Place a column into a new collecting tube, add into the column 500 µL of Buffer H2 →**Centrifuge for 1 min at 12 000 rpm,** discard the collected liquid, and **centrifuge at maximum speed for 1 min,** throw away the collecting tube
6. Place a column into a 1,5mL Eppendorft, carefully add the heated eluition buffer into the center of the column, let it set for 1min → **Centrifuge for 2min at 12 000 rpm**
7. Discard the column, place the DNA at 4˚C for immediate use, or at - 20˚C for storage

Table S.4: Master mix for *TMPRSS6* PCR amplification (all exons except 16).

Products	Per reaction (µL)	PCR conditions		T(C°)	Δt
Buffer β	22.9	**Initial Denaturation**		95	5m
ExN_Fw 25 µM (25pmol/ µL)	0.5	**Denaturation**		95	20s
ExN_Rv 25 µM (25pmol/ µL)	0.5	**Hybridisation**	**32x**	63	20s
Taq Applied Byosystems (5U/µL)	0.1	**Extension**		72	20s
Volume per PCR:	24	**Final Extension**		72	3m
30ng/µL DNA	1	**Pause**		4	15

N stands for the primer of each exon by its number.

- **Protocol for DNA purification for all exons (except 16):**

To the product of each PCR sample, 5 µL are removed to which are added 2 µL of ExoSAP-IT into a 0.2 mL eppendorft. After that, they should go to the thermocycler for purification (37°C for 15m, followed by 80°C for 15m).

Table S.5: Sanger sequencing mix and conditions.

Products	Per reaction (µL)	Conditions		T(C°)	Δt
Big Dye (5x) Buffer	1.75	**Initial Denaturation**		96	4m
Big Dye (2x)	0.5	**Denaturation**		96	10s
H_20	5.25	**Hybridisation**	**25x**	55	5s
Primer NX (2pmol/µL)	1	**Extension**		60	4m
Total volume	8.5	**Final Extension**		60	8 m
PCR purified DNA	1.5	**Pause**		4	15m

N stands for the primer of each exon by its number, for Sanger sequencing it can either be the forward or the reverse.

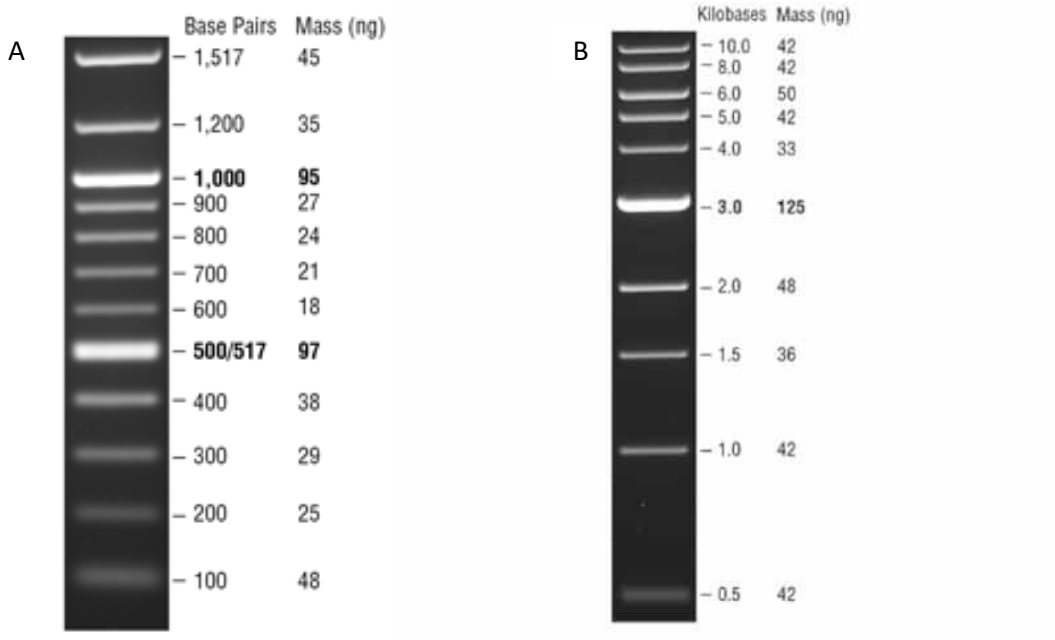

Figure S.3: Molecular weight ladders. A- 100 bp DNA Ladder from New England Biolabs Japan (measures from 1500bp down to 100bp). **B**- 1kb DNA Ladder from New England Biolabs Japan (measures from 10000 bp down to 500 bp)

Table S.6: Composition of Buffers used.

Buffer α (10x)	Buffer β	Electrophoresis loading buffer (Bromophenol Blue)
(NH4)2SO4 166 mM Tris-HCl (pH = 8.8) 670 mM MgCl2 15 mM EDTA 0.67 mM β-Mercaptoetanol 100 mM	KCl2 50 mM Tris-HCl (pH = 8.8) 100 mM MgCl2 15 mM Gelatin – 0.01% (w/v)	Water 8.75 µL Bromophenol Blue 0.22 g Glycerol 18.75 µL EDTA 75 µL NaOH (to turn blue)

7.3 Abstract submitted for poster presentation in SPGH 25th Edition.

WIDENING THE SPECTRUM OF TMPRSS6 GENE PATHOGENIC VARIANTS RELATED WITH HEREDITARY IRON DEFICIENCY

Vera Pessoa[1]; Alexandra Oliveira[1]; Daniela Santos[1]; Joana Mendonça[1]; Miguel P. Machado[1]; José Ferrão[1]; Luís Vieira[2]; Pedro Lopes[1]; Irina Kislaya[3]; Carlos Matias-Dias[3]; Marta Barreto[3]; Paula Faustino[4]

[1]Departamento de Genética Humana, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa; [2]Departamento de Genética Humana, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa; ToxOmics, Faculdade de Ciências Médicas, Universidade Nova de Lisboa, Lisboa; [3]Departamento de Epidemiologia, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa; Centro de Investigação em Saúde Pública, Escola Nacional de Saúde Pública, Universidade NOVA de Lisboa, Lisboa; [4]Departamento de Genética Humana, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa; Instituto de Saúde Ambiental, Faculdade de Medicina, Universidade de Lisboa, Lisboa; Laboratório Associado TERRA, Faculdade de Medicina da Universidade de Lisboa

Iron-Refractory Iron-Deficiency Anemia (IRIDA) is a rare autosomal recessive hypochromic microcytic anemia derived from loss-of-function mutations in the *TMPRSS6* gene, which encodes Matriptase-2, a negative regulator of hepcidin expression. IRIDA patients have high hepcidin levels that prevent iron absorption and recycling. Very few studies concerning this pathology have been carried out in the Portuguese population and its molecular basis is still largely unknown.

In this study, we aimed to identify genetic variants in *TMPRSS6* in a sample of the Portuguese population with a hematological phenotype suggestive of iron deficiency. In addition, we intended to evaluate the performance of NGS for genetic screening of this large gene.

We studied 100 adults with anemia and/or microcytosis and/or hypochromia collected by the Portuguese National Health Examination Survey (INSEF). Other possible genetic causes for these abnormal phenotypes, namely α- and β-thalassemia, were discarded after *HBA1*, *HBA2* and *HBB* genetic screening. The *TMPRSS6* gene (18 coding regions, exon/intron boundaries and regulatory regions) was amplified in 3 long-PCR fragments that were screened by NGS using Nextera XT libraries in a MiSeq platform. The genetic variants found were validated by Sanger sequencing (transcript ENST00000676104.1).

Several known variants were identified along with two unreported mutations, c.1585T>C (p.Cys529Arg) and c.1580T>G (p.Phe527Cys). These novel mutations were classified as pathogenic by *in silico* analyses through Polyphen2, SIFT, and Missense3D. Moreover, Phyre2 software was used to produce a 3D structure of the mutated proteins, based on alignments with known protein structures, as there is no 3D model for Matriptase-2 on online databases. The two novel mutations were found in heterozygosity, explaining the mild abnormal hematological phenotypes and serum iron biomarkers presented by both patients. Functional studies should be performed to validate these findings.

Our results widened the spectrum of *TMPRSS6* pathogenic variants underlying hereditary iron deficiency-related pathologies. In addition, NGS revealed to be an appropriate tool for *TMPRSS6* genetic screening.