UNIVERSIDADE DE TRÁS-OS-MONTES E ALTO DOURO

Gene discovery in pediatric immune and inflammatory diseases

Master's Thesis nominated to obtain academic degree in Technologic and Comparative Molecular Genetics

Pedro Eduardo Moreno Cardoso

February, 2018

Laboratory of Translational Immunology, VIB Center for Brain and Disease Research

Supervised by: Dr. Vasiliki Lagou, Post-Doc and Erika Van Nieuwenhove, PhD

Promotor: Prof. Adrian Liston, Group Leader

Vila Real, 2018

“Logic will get you from A to B,

imagination will take you everywhere.”

Albert Einstein

UNIVERSIDADE DE TRÁS-OS-MONTES E ALTO DOURO

Gene discovery in pediatric immune and inflammatory diseases

Master's Thesis nominated to obtain academic degree in Technologic and Comparative Molecular Genetics

Pedro Eduardo Moreno Cardoso

Group Leader: Prof. Adrian Liston

Supervised by: Dr. Vasiliki Lagou

Co-supervised by: Erika Van Nieuwenhove, PhD student

JURY MEMBERS:

Dr. Vasiliki Lagou  _____

Dr. Raquel Chaves_____

Dr. Isabel Gaivão_____

Vila Real, 2018

DECLARATION AND CONFIDENTIALITY

I Pedro Eduardo Moreno Cardoso confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

This study is confidential and all the information here present belongs to Translation and Immunology Laboratory, therefore all the information and experiment results here described should be only used for evaluation purposes.

SUMMARY

Primary immunologic diseases are a newly recognized and expanding group of autoinflammatory and autoimmune diseases characterized by a genetic predisposition for immune system dysregulation. Most of them have not been yet genetically characterized as they include a wide range of heterogeneous conditions.

In the current work, we used next generation sequencing, specifically whole-exome sequencing. We applied an established GATK pipeline for processing the data and performed variant filtering to find novel or known genetic mutations in young patients with rare autoimmune/autoinflammatory disorders. In most of the cases, these patients were analyzed as part of a family trio with or without siblings. Through this pipeline and filtering, we were able to propose disease causing mutations and suggest new disease mechanistic pathways in some of the families. Furthemore, we have scrutinated *CECR1* gene in patients with ADA2 deficiency and discovered both new and reported mutations that cause the reduced enzymatic activity. Additionally, we analyzed patients affected with the polygenic disease, Juvenile Idiopathic Artritis, and discovered two rare mutations in the known gene for JIA, *IL6R*. Promising candidate mutations from exome-sequencing were taken forward for confirmation by SANGER sequencing. In this thesis, we describe the pitfalls of exome-sequencing and the important information, apart from biological function, that has to be taken into consideration for choosing a candidate variant. We conclude that although whole exome sequencing is indeed very useful for identifying the genetic cause of Mendelian diseases, such as PID, this process is not always straightforward as very complex and time-consuming functional tests are always necessary for confirming the exact role of the mutation in the disease pathogenesis.

Keywords: Next Generation Sequencing; Whole Exome sequencing; primmary immunodeficiencies; diagnosis; genetic mutation; bioinformatic analysis

RESUMO

Doenças imunológicas primárias são um grupo recentemente reconhecido e em expansão de um conjunto de patologias auto inflamatórias e autoimunes, em que os indivíduos afetados apresentam predisposição genética para a desregulação do sistema imunitário. A maioria destas patologias ainda não foram geneticamente caracterizadas devido à grande diversidade de características heterogéneas que apresentam.

No presente trabalho foi utilizada a nova geração de sequenciação (Next-generation sequencing), especificamente a sequenciação de todo o exoma (Whole-Exome Sequencing: WES). Além disso, utilizamos um protocolo desenvolvido pela GATK (Genome Analysis Toolkit) para processar os dados resultantes da sequenciação e aplicamos uma série de filtros em todas as variantes de forma a encontrar mutações genéticas novas, ou referenciadas em jovens pacientes com doenças autoimunes e/ou auto inflamatórias raras. Na maior parte dos casos foram analisados os pacientes e os pais (designado "trio design"), por vezes também os irmãos e esporadicamente indivíduos isolados. Através do protocolo e da filtragem utilizada, foi possível sugerir novas mutações patogénicas ou, pelo menos, mutações que causam dano na proteína final, e também, sugerir novas vias e mecanismos que podem levar ao aparecimento das doenças. Além disso, escrutinamos molecularmente o gene *CECR1* em pacientes com deficiência na proteína ADA2, identificando mutações novas e também mutações anteriormente referenciadas que causam a redução da atividade enzimática. Adicionalmente, analisamos pacientes afetados com uma doença poligénica, a artrite idiopática juvenil, e descobrimos duas mutações raras no gene *IL6R,* gene implicado no desenvolvimento desta doença. Mutações candidatas promissoras obtidas a partir da WES foram confirmadas por sequenciação de SANGER. Nesta tese descrevemos também algumas falhas da WES e a informação que deve ser levada em consideração para a escolha de uma variante candidata, excetuando a função biológica. Concluímos que a WES é de facto muito pratica e útil na identificação da causa genética das doenças mendelianas, como as doenças primárias imunológicas, mas no entanto, este processo de identificação não é linear sendo, obrigatoriamente, necessários testes funcionais complexos e exaustivos para confirmar o papel exato da mutação na doença.

Palavras-chave: Next-generation sequencing, Whole Exome Sequencing (WES); imunodeficiências primárias; diagnóstico genético; mutações genéticas; análise bioinformática

ACKNOWLEDGEMENTS

A entrega desta tese é o culminar de pouco mais de um ano de trabalho repleto de esforço e dedicação. Quero, em primeiro lugar, agradecer ao Adrian pela oportunidade de poder fazer investigação no seu laboratório, a qual me permitiu conhecer um novo país, e estar em contínuo contacto com o ambiente vivido num laboratório de topo. Agradecer também às minhas orientadoras, Vaso e Erika, que me guiaram durante este ano e me ajudaram a crescer enquanto investigador. Um especial obrigado pela motivação que me conseguiram incutir.

Em segundo lugar, agradecer a todos os meus professores do mestrado de genética molecular comparativa e tecnológica pelo apoio e conhecimento transmitido e em especial, um obrigado à direção deste mestrado.

Um agradecimento a todas as pessoas que me acompanharam neste ano e que me fizeram crescer profissionalmente e também pessoalmente. Um especial obrigado a todos os amigos e colegas novos que conheci nesta aventura mas também aos que já me acompanham desde do início desta caminhada académica.

Quero também exprimir o meu grande agradecimento aos meus pais que tornaram possível este caminho percorrido em terras estrangeiras, pela confiança depositada e suporte emocional e financeiro. Muito Obrigado mãe e pai queridos por todo o amor, carinho e valores que me transmitem. Um obrigado também á minha irmã pela presença e amor. Apesar de não me teres visitado continuo a gostar muito de ti! ;)

Por fim, quero especialmente agradecer à minha namorada, Tatiana, por me ter acompanhado todo este ano em quer estivemos separados, felizmente apenas fisicamente. Pela confiança depositava, pelas palavras de motivação e carinho, pela tua sempre presença, pela tua amizade, um Muitíssimo Obrigado. Amo-te! <3

Este ano foi uma experiência de onde levo muitos bons momentos e alguns menos bons, mas todos eles contribuíram para um profundo desenvolvimento pessoal e profissional.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PRR	Pattern recognition receptors
TLR	Toll-like receptor
RLRs	Retinoid acid-inducible gene –I (RIG-I)- like receptors
NLRs	Nucleotide-binding oligomerization domain (NOD)-like receptors
TCR	T cell receptor
BCR	B cell receptor
DC	Dendritic cells
IFN	Interferon
AIDs	Autoinflammatory diseases
Ads	Autoimmune diseases
MHC	Major histocompatibility complex
FMF	Familial mediterranean fever
CAPS	Cryopyrin-associated periodic syndrome
TNS	Tumor necrosis factor
TRAPS	TNF receptor associated periodic syndrome
UV	Ultraviolet radiation
IL	Interleukin
SLE	Systemic lupus erytomatosis
HIDS	Hyperimmunoglobulinemia D and periodic fever syndrome
NOMID	Neonatal-onset multisystem inflammatory disease
DIRA	Deficiency of the interleukin-1 receptor antagonist syndrome
NGS	Next-generation-sequencing
PID	Primary immunologic diseases
SLE	Systemic lupus erytomatosis
MWS	Muckle–Wells syndrome
FCAS	Familial cold auto-inflammatory syndrome
AR	Autosomal recessive
XR	X-linked recessive
GDI	Gene damage index
AF	Allele frequency
CADD	Combined annotation dependent depletion
SIFT	Sorting intolerant from tolerant
Polyphen2	Polymorphism phenotyping v2
MSC	Mutational significant cutoff
HGC	Human gene connectome
LOF	Loss of function
GAF	Gain of function
CNV	Copy number variable
WGS	Whole genome sequencing
WES	Whole exome sequencing
DNA	Deoxyribonucleic acid

SNP	Single nucleotide polymorphism
BWA	Burrows-wheeler aligner
VCF	variant all format
GTEx	Genotype tissue expression
LCL	Lymphoblastoid cell line
NK	Natural killer
DADA2	ADA2 deficiency
JIA	Juvenile idiopathic arthritis

1. Introduction

1.1 The human immune system

The most important function of our immune system is to protect the organism from pathogens. A healthy immune repertoire is able to identify a wide variety of agents, from parasitic worms to viruses, and simultaneously differentiate these from its own 'self' molecules. Innate and adaptive immunity constitute two classes of the immune system and these are present in the most structural and ancestral forms of life. The innate immune system developed earlier than the adaptive immune system, with the latter first found in jawed vertebrates [1]. The immune system is composed and organized by soluble and cellular receptors and effector mechanisms, using different cell types to signalize the attacker and its defense mechanisms. The pattern recognition receptors (PRR), proteins that recognize pathogens such as bacteria and viruses, are part of the innate immune system sensors [1]. Some better known receptors are: Toll-like receptor (TLR), the retinoid acid-inducible gene –I (RIG-I)- like receptors (RLRs) and the nucleotide-binding oligomerization domain (NOD)-like receptors (NLRs) [2]. The innate immune system is the first to react when the system is attacked by a pathogen and can, in turn stimulate the adaptive immune system [3].

The cells classified as innate immune system effectors are macrophages, dendritic cells, and other antigen-presenting cells. On the other hand, we have B and T cells, constituting the adaptive immune system. The antigen recognition system of T cells (T cell receptor, TCR) functions purely to sense antigen, while the antigen recognition system of B cell (B cell receptor, BCR) has functioned as both a receptor and, in the secreted form as antibody, an effector molecule [3]. The adaptive and innate immune system are tasked to maintain tolerance in the case of foreign molecules that do not endanger the organism [2]. The adaptive defense system responds much slower than the innate but once activated is much more specific and efficient. The process consists of two phases. The first one (initiation phase) starts with the recognition of the self-nucleic acids released during apoptosis and internalization by the dendritic cells (DC) through the TLRs, causing production of IFN-α by these cells. The function of IFN-α is to stimulate the dendritic cells maturation, present autoantigens, recruit B and T cells that result in the production of antibodies which recognize the antigen of the pathogen [2,4]. Antibodies can neutralize the pathogen and/or induce macrophage activation to eliminate the foreign molecule showing that the adaptive and innate system work together to defend the organism against pathogenic agents [5].

1.1.1 Autoimmune and autoinflammatory diseases

Despite the fact that autoinflammatory diseases (AIDs) have only been recently defined, they have been present in organisms for a longer time than autoimmune diseases (ADs). AIDs and ADs are characterized by a chronic activation of the immune system that, as a consequence, gives rise to tissue inflammation and damage in genetically predisposed individuals. In AID the innate immune system is directly responsible for tissue inflammation, while in ADs the main effector of the inflammatory process is the activation and production of autoreactive antigen-specific T cells and autoantibodies [3]. The adaptive immune system is not present in invertebrates and therefore there is no evidence that they have autoimmune phenomena. The case is different for vertebrates, where AD have been found, and strong evidence have been presented, for example, the genetic association of major histocompatibility complex (MHC) variants with disease [6].

In more detail, the AIDs are a group of inflammatory diseases that are recurrent, unprovoked and occur in the absence of infection. Some examples of these disorders are familial mediterranean fever (FMF), cryopyrin-associated periodic syndrome (CAPS) and TNF (receptor associated periodic syndrome (TRAPS). In AID there is frequently inflammation thought to be driven by the inflammasome that is known to promote the maturation of inflammatory cytokines IL-1β and IL-18 [7], but the role of the inflammasome in AD is not yet understood. Although, considering the wide variety of endogenous signals that can activate NLRs, such as ultraviolet radiation (UV) [8,9] and some inflammasome products, like IL-1β, they are thought to also play a role in the adaptive immunity [10,11].

Autoimmune diseases (ADs) can be classified into systemic ADs (disorders that harm multiple organs) or localized ADs (single organ or tissue). The causes of systemic ADs are still unclear due to the implication of a complex immunological mechanism, comprising genetic and environmental factors [3]. However, defective or excessive adaptive immune responses, such as immune reaction against normal and non-pathogenic self-molecules, are the beginning and main reason for the appearance and development of ADs.

The innate immune system can detect pathogens and initiate a response by activating own pathways and molecules to defend the individual. It recognizes foreign molecules through PRR that make a posterior activation of intracellular pathways, such as to induced expression of genes like Interferon (IFN) alpha, IFN-β, TNF and interleukin

(IL)-1 genes. A dysregulated or abnormal synthesis of these receptor proteins may lead to AIDs and also ADs [12]. For example, the AID has been shown to be highly correlated with a specific type of PRR, the NLRs [2,13]. Several patients with AIDs carry mutations in pyrin, cryopyrin or TNF-receptors [14,15]. TLR have also been implicated in AD such in the case of lupus erythematosus (SLE), where production of type I IFN is activated [4,16].

In general, the alteration of the immune homeostasis causes the appearance of AD and AID. Although AID and AD are divided into two groups, making the study of the diseases less complicated, they share pathways and have several similarities. Therefore they might be considered as one single group with a long spectrum of immune pathologies and clinical abnormalities where we would have pure AD on one side and in the far end pure AID [3].

1.1.2 Polygenic and monogenic diseases

Studies have already associated genetic variation to the predisposition and development of AIDs and ADs. Some of the known both monogenic and polygenic autoimmune and autoinflammatory diseases are present in **Table 1.1.2.1**.

The AIDs can show recessive or dominant patterns of inheritance and can be monogenic or complex in nature. The systemic juvenile idiopathic arthritis disease is a polygenic disease where the immune system and in particular 3 cytokines (IL-6, IL-1 and IL-18) are thought to play a role [5]. Some mutation in several genes, such as mutations in IL-1 gene, have been reported to confer a higher susceptibility to the disease [17,18]. Genetic mutations that are involved in the regulation pathways of the immune reaction, such as mutations in the TLR-interferon signaling pathway, more specifically the interferon regulatory factor 5 and signal transducer and activator of transcription 4 (*STAT4)* genes, have been associated with AD considered to be predominantly polygenic such as rheumatoid arthritis, systemic lupus erythematosus (SLE), systemic sclerosis [19]. A single-nucleotide mutation in several inflammasome genes, for example NLRP1, nucleotide-binding oligomerization domain-like receptor protein 1 (rs2670660 SNP), have been suggested to confer an increased risk for the development of SLE as well as for occurrence of lupus nephritis, rash and arthritis [20].

The monogenic AIDs have several known genes, with some encoding proteins of the inflammasome [21]. Mutations in monogenic AID have also been identified. The first and most common are disease-causing mutations in the *MEFV* gene, in patients with FMF

[22]. Several examples of monogenic AD diseases were also reported, such as mutations in *AIRE,* encoding a transcription factor that causes autoimmune polyendocrinopathy syndrome. Pathogenic mutations in this gene, mostly nonsense or frame-shift, lead to a truncated or abnormal protein and dysregulation of self-antigen expression in the thymus [23].

Despite the great contribution of monogenic diseases to the understanding of gene regulation, mechanistic pathways and pathogenic mutations, the interest in disease studies is shifting away from monogenic disorders towards multifactorial disorders [24]. Insights gained through the study of monogenic disorders continues to provide us with relevant information to the understanding of complex diseases and can subsequently aid in developing targeted treatments. Therefore, we generally should invest more time and resources to research in monogenic diseases, especially in inborn errors where it remains vital [25].

Table 1.1.2.1: Classification of autoinflammatory and systemic autoimmune diseases.
FMF: familial Mediterranean fever; TRAPS: TNF receptor-associated periodic syndrome; CAPS: cryopyrin-associated periodic syndrome; FCAS: familial cold auto-inflammatory syndrome; MWS:Muckle–Wells syndrome; HIDS: hyper-immunoglobulinemia Dsyndrome; PAPA: pyogenic arthritis, pyoderma gangrenosumand acne;; sJIA: systemic juvenile idiopathic arthritis; CRMO: chronic recurrent multifocal osteomyelitis syndrome; NOMID: neonatal-onset multisystem inflammatory disease; DIRA: deficiency of the interleukin-1 receptor antagonist syndrome; IL: interleukin APS: autoimmune polyendocrinopathy syndrome; IPEX: immunodysregulation polyendocrinopathy enteropathy X-linked syndrome; ALPS: autoimmune lymphoproliferative syndrome; SLE: systemic lupus erythematosus; UCTD: undifferentiated connective tissue disease; MCTD: mixed connective tissue disease. Adapted from [26].

Autoinflammatory diseases		Autoimmune diseases	
Monogenic	**Polygenic**	**Monogenic**	**Polygenic**
FMF	Still's disease	APS type I	Rheumatoid arthritis
TRAPS	Crohn's disease	IPEX	SLE
CAPS	Behçet's disease	ALPS	Systemic sclerosis
FCAS	Gout	Type I diabetes (also polygenic)	Polymyositis/ Dermatomyositis
MWS	Systemic juvenile idiopathic arthritis	Hypothyroidism (also polygenic)	Systemic vasculitis
HIDS	Psoriasis	Hemophagocytic lymphohistiocytosis	Sjögren Syndrome
Blau's syndrome	-	DiGeorge Syndrome	UCTD
PAPA syndrome	-	SLE monogenic form	MCTD
CRMO	-	-	Addison disease
NOMID	-	-	Omenn Syndrome
Majeed's syndrome	-	-	Primary Biliary Cirrhosis
IL-10 deficiency syndrome	-	-	Celiac disease
DIRA	-	-	-

1.1.3 Disease triggers

Environmental and endogenous factors can also contribute to the development of AD and AID. Pathogenic infections are one of the most common factors that can trigger an excessive reaction of the innate immune system, in individuals genetically predisposed, through binding to TLRs and/or the activation of inflammatory process, as observed in cases of Muckle-Wells syndrome (MWS), hyperimmunoglobulinemia D and periodic fever syndrome (HIDS) [27], NOMID, DIRA, Behçet's disease, gout, and chondrocalcinosis [28].

The adaptive immune system can also be excessively activated by infections through molecular mimicry, activating and expanding the previously activated T cells and activation of new T cells by microbial superantigens [29,30]. There are also point cases where vaccines and adjuvants trigger AD/ AID syndromes [31,32]. Moreover, physical agents can induce an excessive immunologic reaction driving AD and AID as previously demonstrated by Kastner, Aksentijevich, & Goldbach-Mansky (2010) in diseases like MWS, FCAS, and Raynaud's phenomenon by cold exposure; in DIRA, HIDS, TRAPS, and Behçet's disease by physical trauma; and in SLE, by UV light exposure and smoke. As shown before, the UV light can also activate the inflammasome and can stimulate AIDs [8]. Several other eliciting factors in AD and AID are crystal deposition [34] and psychological stress, as well as generic stress [35]. All these factors can disrupt immune system homeostasis.

1.2. Clinical features

1.2.1 Age of onset

Immune diseases that occur in the neonatal period or early infancy are suspicious for monogenic AD or AID, as we can assume that the factors driving the disease are not yet under the great influence of the environment or other factors. Monogenic diseases like FMF, MWS and TRAPS can appear in early infancy or later in life [27]. Polygenic immune-related diseases generally develop in adolescence or adulthood [36], and are rare in early infancy [37]. There are, however, polygenic disorders that have a high incidence in childhood, such as Type I Diabetes [38].

1.2.2 Signs and symptoms

The clinical features associated with autoimmunity and AID are not very well differentiated because certain features are common to both. They also have a wide spectrum of organ involvement and different degrees of disease severity [3].

Some shared symptoms that AD and AID patients can have are recurrent fever that can last for a few days to several weeks, weight loss, fatigue, malaise, flu-like symptoms, lymphadenopathy, and splenomegaly [26]. Although the manifestations are similar, in terms of febrile episodes, they seem to be more persistent in ADs than AIDs [39]. Patients with AID, such as CAPS, TRAPS, HIDS, JIA and Still's diseases, but also AD patients that suffers from SLE, regularly develop skin manifestations such as urticarial-like rash that must be carefully examined for a correct diagnosis [3,40,41].

In addition, musculoskeletal manifestations are frequent and include arthralgias, myalgias and arthritis, among others [42]. AID frequently affects the large joints (ankle, knee), while in AD the there is no specificity, it can affect large as well as small joints [43]. The AD and AID can also affect several systems: hematopoietic [44], gastrointestinal [45,46], respiratory, vascular and nervous [26]. The clinical manifestations can differ within AID and AD depending on the type and severity of the disease and especially by the organ or tissue that is affected. Besides that it is also possible that two different diseases with different clinical phenotype can overlap in terms of molecular genetics and immunology, such in the case of type 1 diabetes and multiple sclerosis [47,48].

1.3 Bioinformatic analysis

1.3.1. Next-Generation sequencing (NGS)

In the last decade, genome-wide association studies have identified hundreds of common risk alleles for complex human diseases [49–52]. These studies have provided biological insights into several diseases like Crohn´s disease, among others [53–55]. However, because the construction of physical maps has been improved, there is the possibility to sequence all subcloned fragments of the genome individually and then join the finished sequencing together, making a good representation of entire chromosomes. When they got to that point, important large genomes, like the human genome, were sequenced [56]. To be able to make a good analysis of the short reads that are formed in

Next-generation-sequencing (NGS) platforms it is mandatory to have a reference genome to be the substrate for the alignment prior to variant detection [57].

The NGS technology has been improving and four major changes have been done. One of them is the preparation of DNA NGS libraries in a cell-free system instead of bacterial cloning of DNA fragments. Two, is that now, several thousand to millions of sequencing reactions are done in parallel. Three, the sequencing output is detected without the need for electrophoresis. Fourth, the improving of general processes of different fields, including microscopy, nucleotide biochemistry, computation and data storage [58]. All these changes made the sequencing much more practical and also decreased its cost and running time [59]. Some of the first new NGS technologies were: 454 pyrosequencing [60]; Illumina/Solexa that perform sequencing with reversible terminator nucleotides; and sequencing by Oligo Ligation Detection (SOLiD) [61]. Other NGS technologies have also been developed, such as Qiagen-intelligent bio-systems sequencing-by-synthesis [62], Polony sequencing [63], and a single molecule detection system [64].

The combination of the availability of large well-characterized cohorts, improved genotyping technologies like NGS and the development of new bioinformatics methods to analyze the output have facilitated several discoveries [65–68]. Especially, the advances in NGS was the principal factor for the increased discovery of novel genetic causes of known and novel Primary Immunologic Diseases (PID) [69,70]. PID is a group of disorders affecting several components of the immune system such as neutrophils, macrophages, dendritic cells, complement proteins, natural killer cells, and T and B lymphocyte [71]. SANGER sequencing is an accurate and effective method for confirming a mutation in a gene. However, it is only useful when there is a small set of candidate genes and the patient's phenotype is typical of a mutation in a specific gene. When the gene is not known, NGS-based gene panels, such as whole-exome sequencing (WES) and whole-genome sequencing (WGS) are the best way for detecting the genetic variant causing the disease. The NGS studies have gained popularity among the scientific community and exome-sequencing technologies are being widely used in genetic studies of Mendelian disorders [72]. There is also an increasing need and interest in extending them to complex traits [73]. For this purpose, new methods for design, analysis and interpretation of exome-sequencing studies are being developed [74–76]. Sequencing is now a powerful approach for clinical diagnosis, therapeutic identification, disease risk and prenatal testing, among others [57].

1.3.2. Generating NGS data

NGS allows the sequencing of hundreds of millions of small fragments in parallel. This technology can be used to sequence the whole genome or a targeted region, like the entire exome (WES). A schematic figure of NGS process is presented in **Figure 1.3.2.1**. One of the major differences between WGS and WES is that, for WES, the DNA template must be rich in exons implying a step to "catch" it. This process is not needed in WGS and is the vulnerable point of gene panels and WES because the capture or enrichment process is not perfectly homogeneous along all genes and so, some bias can be introduced. There are some more errors that can be introduced due to protocols where gene panel, WES and WGS rely on amplification by PCR for sequencing library preparation [25]. This process is associated with guanine-cytosine bias errors, stochastic errors, template switch errors and polymerase errors. Although, PCR-free library preparation is available for WGS. The major differences between WES and WGS are summarized in **Figure 1.3.2.2**.

The result of NGS is a raw sequence in a FASTQ file that contains the nucleotide sequence for each read and information on the sequence quality for every nucleotide sequenced. This data needs to be trimmed and processed to generate high-quality genotypes for each individual. Normally, one of the first things to do is to check if the DNA samples were contaminated or if they were tracked correctly during the sequencing process [77,78]. Then, short reads obtained are aligned to a reference genome producing a BAM file [79–81] which is further processed by performing calibration of base quality scores [82] and removal of duplicates reads [83]. The final step is the calling of variants that identifies and defines all the alleles resultant of the NGS that are different from the reference sequence [82,84]. The file created by this step is the VCF (variant call format). Annotation of all variants is the next typical step and this is done using a wide range of software [85–87], creating a new final annotated VCF file. This VCF file is the one used by researchers or clinicians because they are much smaller in size than FASTQ or BAM files and also very practical to look for the variant allele that can be causing the phenotype. The annotated VCF files have important information like the physical position of the variant, genotype quality (GQ) score, read depth (DP), the frequency of the allele in different populations/databases and the allele that is different from the reference and the reference allele. The GQ score is a value that translates the accuracy and the confidence you can have in a genotype assigned by the sequencer at a certain position. The read depth or per-base coverage is the number of reads for one nucleotide position and can be used to filter

out low quality variants or find technical errors. The GATK (Genome Analysis Toolkit) provides pipelines for both WGS and WES [88].

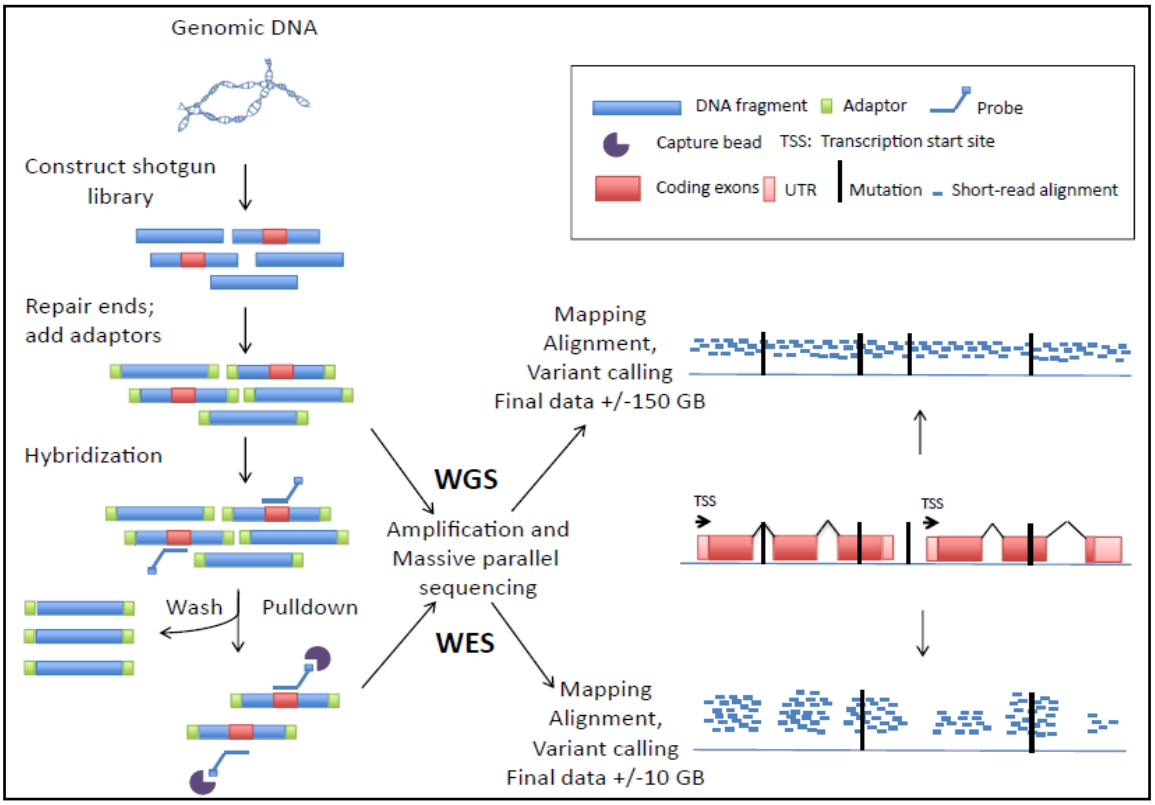


Figure 1.3.2.1: General NGS process [25].

	Targeted sequencing		
	Gene panel	WES	WGS
Pro	• Reduces incidental findings • Allows for higher sequencing depth than WES and WGS for the targeted genes and at a lower cost	• Covers the exome: the most purified region of the genome and the most studied part of the genome • At present, less expensive than WGS • Allows for a gene panel–like analysis or WES	• Introns, regulatory domains, and intergenic regions included • PCR free library preparation possible • More uniform coverage than WES • Superior to gene panels and WES for CNV detection • Computational panel-like/exome-like or entire genome analysis possible
Con	• No identification of novel genes • Dated at onset, frequent updates needed • Might derive attention to a "red herring" • The larger the gene panel, the less cost-effective compared with WES	• Reference bias: unknown exons are not covered • Introns and regulatory domains are generally not included • Coverage is less uniform than WGS and in a well-designed panel • Risk of incidental findings	• More calls ≥ increased analysis complexity • At present, more expensive (sequencing, data storage) • Risk of incidental findings
	• PCR amplification potentially induces guanine-cytosine bias • CNVs not reliably detected Detection of large structural variations (inversions, translocations, and nucleotide repeat expansions) inaccurate with some platforms		
Gained information	Variants in known disease-causing genes	Variants in exome as targeted by the kit	Variants in the entire genome
Possible information	Novel or known disease-causing mutation in a known disease-causing gene (including intronic mutation depending on panel design)	Novel or known disease-causing mutation in known disease-causing genes/novel disease-causing gene within the exome as targeted by the kit	Novel or known disease-causing mutation in known disease-causing genes/novel disease-causing gene/ mutations in deep intronic regions–intergenic regions–regulatory domains

Figure 1.3.2.2: The advantages and disadvantages of gene panel, WES and WGS [25].

1.3.3 Study designs, filtering and selection of candidate mutations

To be able to analyze NGS results there is a need of setting a genetic hypothesis formed based on the clinical phenotype of the patients and other family members, typically parents or siblings. The mode of inheritance, clinical penetrance, and genetic heterogeneity are also taken into consideration [89–91]. In diseases showing an autosomal recessive (AR) pattern of inheritance, homozygous and compound heterozygous variants can be selected. In the case of an autosomal dominant or X-linked recessive disease, heterozygous or hemizygous variants are selected.

In general, we can build up studies based on a single or multiple kindred. The multiple unrelated kindred studies assume that different individuals display the same clinical phenotype caused by a mutation at least in the same gene. This type of study requires the use of healthy controls meaning healthy subjects, ideally of the same ethnic origin as the patients and generated by the same method. This requirements can also be fulfilled by using public databases as control and healthy parents or siblings to determine the segregation of the variants. Comparing the two sets, we can obtain candidate mutations. The variants found in both patients and unaffected subjects are excluded in the case of complete penetrance [92]. The case of incomplete penetrance is more complex because the causing variant might be present in unaffected subjects that do not phenotypically show. In this case, the candidate variants selected after filtering, based on genetic hypotheses and functional criteria, have to be analyzed further at variant and gene levels and all new variants in new or known genes need to be functional validated by experimental studies [25].

In the case of single kindred studies the ideal situation, from the point of view of someone who is doing the analysis, is when the disease present in the patient is received by the consanguineous parents in an AR inheritance pedigree. In this case, the variants selected can be homozygous or compound heterozygous with low AF. If we have several healthy or diseased siblings the accuracy of the variant selection is better and much easier but we need to have in mind that in consanguineous families, the disease can also follow an AD or XR inheritance model or can be due to *de novo* mutations. In a case where the parents are not consanguineous, the genetic hypotheses that can be pursued are XR, AR and AD. In a male that has a sporadic phenotype, it is difficult to define a specific genetic hypothesis. When we have various affected relatives in a single kindred the probability of mapping the potentially disease-causing variants to a specific locus is higher due to

what linkage analysis can give us [25]. In the trio design, the patients and their parents are sequenced. This strategy is normally used in a situation where a single patient suffers from rare, early-onset and, highly deleterious phenotype suggesting the possibility of a heterozygous or hemizygous *de novo* mutation or compound heterozygous mutation. The variant or variants chosen must not be present in parents, given that the parents are not affected, except when considering incomplete penetrance. *De novo* mutation at the genome level occur in each generation (30 to 100 *de novo* mutations) and so at least 1 or 2, are probably detected in each exome [93].

Through WES we can identify more than 50,000 variants, so filtering and prioritization are important steps for finding the causal mutation [94–96]. Filtering criteria can include, among others, allele frequency (AF) of variants in public databases and Gene Damage Index (GDI). These public databases (for example Exac3, 1000 Genomes) provide sequencing data of thousands of individuals of various ethnicities [97]. If we are dealing with a rare disease, it is likely that the driving mutation is also very rare or ideally not present in any of these databases (private mutation) [25]. Moreover, the availability of an additional database of approximately 500 individuals of the same ethnicity as the samples we analyze can be very helpful in filtering out variants that might be more common in our population than others [98,99]. The GDI is a database of accumulation mutation damage of each human gene in healthy humans, based on 1000 genomes and CADD (combined annotation-dependent depletion) score for calculating impact. This is an efficient tool for filtering out false positive variant [74]. This tool is also based on the logic assumption that a highly mutated gene is unlikely to be disease causing so, when a GDI score is low, the phenotypic impact prediction is high [100].

Several tools for functional predictions, such as the Sorting Intolerant from Tolerant (SIFT) [101] and Polymorphism phenotyping v2 (Polyphen2) [102,103] can be useful in prioritizing interesting candidate mutations. The CADD score is an algorithm that integrates multiple annotations and can predict deleteriousness, pathogenicity and molecular functionality by contrasting variants that survived natural selection with simulated mutations. The range of CADD score is 1 to 99 and an assigned score of 10 or more than 10 indicates that the variant we are looking at is among the top 10% most deleteriousness found in the human genome [104]. Limitations of this tool include the lack of data on intronic variants and a high false-negative rate when using a fixed cutoff for all genes. The MSC (Mutational Significant Cutoff) [105] predicts the biological impact of variation in humans integrating information from CADD, SIFT and PolyPhen2 and gives

a score that represents the lowest expected clinically relevant CADD cutoff value for that specific gene. As such, a CADD/ SIFT /Polyphen2 score equal or above the MSC generated is predicted to have a high phenotypic impact. The Human Gene Connectome (HGC) is another tool that can prioritize a list of genes by their biological proximity to defined core genes known to be associated with the phenotype of interest and has the power to predict novel gene pathways [72,106].

1.3.4 Experimental Validation and diagnosis

NGS can facilitate the identification of specific variants. However, the causal relationship between genotype and phenotype must be validated experimentally. As referred to before, the NGS can display false-positive results associated with the disease, even more when the mode of inheritance is not clear. Single-patients studies have been frequent and very informative allowing the discovery of more than 20% of the PID causing genes. In terms of experimental validation some guidelines have been submitted, especially in single-patient genetic studies [107]. The experimental studies must demonstrate that a mutation or two (when compound heterozygous) in a specific location change the final protein in terms of function, structure or expression. One of the first tests that can be done is testing the protein expression of a disease-causing variant. Altered protein expression is a good indication that a variant is disease causing but it is not sufficient to establish causality because the loss of some proteins are inoffensive [91,108,109]. Thereafter testing the functional effect of the variant in a cell type and an assay relevant to the clinical phenotype must be demonstrated. If we find ourselves with a de novo mutation, it is good to test it in several cell types to show that the mutation is germline and not somatic but, in both cases, it can still be a disease causing variant [110]. The mechanism of the disease-causing variant effects should be demonstrated by loss of function (LOF) or gain of function (GOF) for a minimum of one biological function related to the phenotype. The new molecular biology tools like CRISPR/CAS9 editing have revolutionized the way we can prove the causality of a determined variant [107,111].

The use of NGS for molecular diagnosis is increasing and they base their screening on known PID-causing genes and mutations that were published by research groups [112]. Not everyone agrees with this molecular diagnosis and some have raised questions about the benefit of given a molecular diagnosis to PID patients [113]. Affected PID patient show clinical heterogeneity which makes the clinical phenotype

characterization very important for the molecular diagnosis, because the same phenotype can be the result of different genes or vice versa [25]. The molecular diagnosis is essential but by who and how this information should be transmitted to the patients requires more discussion. This type of diagnosis can provide prognostic information in cases with a strong genotype-phenotype correlation. It can also be used in neonatal diagnosis or in early onset disorders identifying potentially fatal PIDs enabling, for example, the prevention of infection or indication for transplantation. Finally, the molecular diagnosis promotes new forms of treatments that directly target the modified signaling pathway where the mutated protein acts or allows gene therapy or gene editing [114,115].

Whole genome sequencing has a number of benefits compared to WES because it can detect the intronic and intergenic mutations, (Copy number variable) CNVs and give uniform coverage. However, it remains very costly compared to WES due to the sequencing quantity while in WES the most expensive step is the exome capture. For now, WES is the most widely used approach to diagnose patients that have suspicion of a PID because it is cheaper and there are a lot of functional annotation available for the exome. However, WGS is being used despite its cost and problems with data storage, because it can be more informative in the future and presents an higher coverage of exon than WES [112]. To confirm diagnosis using NGS, the monoallelic or biallelic genotype must be strongly associated with the disease and have a causational relation that has already been shown in previous studies. Experiments aiming to validate new variants are of supreme importance because we cannot trust blindly on *in silico* predictors of pathogenicity. The phenotype description by both clinicians and researcher in diagnostic and research setting is very important because there are known PID genes with disease-causing mutations not yet covered by WES [25].

1.4 Aim

Autoimmune and autoinflammatory diseases are an expanding group of heterogeneous conditions characterized by a genetic predisposition which affects people worldwide. On the other hand, the Whole Exome Sequencing is an uprising tool being used in several research centers for the discovery of the causes and pathways that lead to the disease [25,99,116].

In this thesis, we aim to discover new genes or new mutations that cause the autoimmune and autoinflammatory diseases in young patients, using Whole Exome Sequencing and subsequent bioinformatic analysis.

2. Materials and methods

2.1 Patients

Several families with three or more members were analyzed. Patients were diagnosed with severe primary immunologic diseases. Details about each of the family cases are described in Table 6.1, shown in the Supplementary data.

2.2 Whole Exome Sequencing

Whole-exome sequencing was performed in all healthy or affected individuals. Genomic DNA samples for whole-exome sequencing were prepared from heparinized peripheral blood using the FlexiGene kit (Qiagen). Exome sequence libraries were prepared using a SeqCap EZ Human Exome Library version 3.0 kit (Roche NimbleGen). The paired-end sequencing was performed on the Illumina HiSeq 2000 (Genomics Core Facility, University of Leuven, Belgium).

2.2.1 Processing NGS data

The raw data generated by sequencing was processed using several bioinformatics programs. All computationally-demanding analyses were performed on the Vlaams supercomputer centrum (VSC) that is able to store and process large amounts of data. To be able to work with this platform and submit unix commands, a Linux terminal, in this case PuTTY (http://www.chiark.greenend.org.uk/~sgtatham/putty/), was used. To be able to transfer data to the platform, WinSCP (https://winscp.net/eng/index.php) was used.

For each exome available, the Genome Analysis ToolKit (GATK) best practice for germline SNP and indels in WES pipeline developed by Broad Institute was followed. This pipeline consists of three main steps: pre-processing, variant discovery and preliminary analyses as shown in **Figure 2.2.1.1**. During pre-processing, mapping of the raw reads to the human genome reference (hg19 build) using BWA [79] is carried out. Next, duplicates are marked using Picard tools (https://broadinstitute.github.io/picard/index.html) to reduce the bias that can be introduced by sequencing process, pulling out uninformative reads. This process marks duplicates based on the starting point of the reads and doesn't eliminate the reads but only flags them to not be counted in downstream processes. Then, local realignment of reads around insertions/deletions (indels) takes place, such that the number of mismatching

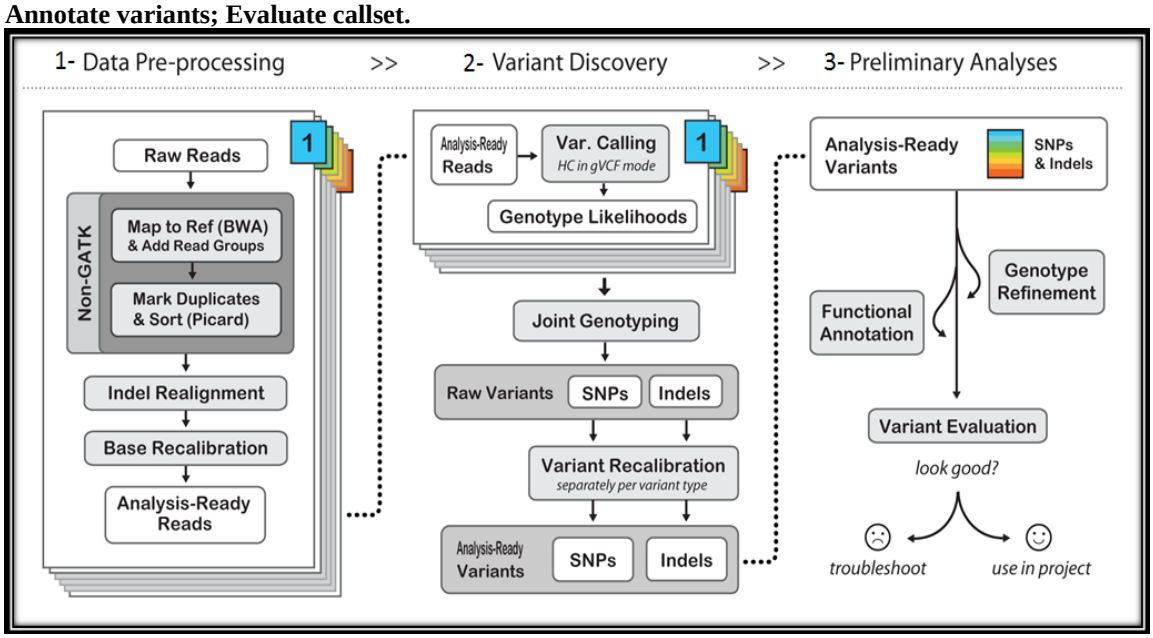
bases is minimized across all the reads. This is because the presence of an indel in the individual's genome with respect to the reference genome can result in alignment artifacts.

Local realignment transforms regions with misalignments due to indels into clean reads containing a consensus indel. The last step of pre-processing is base recalibration during which GATK recalibrates the scores given by the sequencer to each base. This process allows to get more accurate base qualities and improve variant calling. Pre-processing is followed by the variants discovery step, during which HaplotypeCaller identifies the position where our data is different with respect to the reference genome. For each potentially variant site, the program applies Bayes' rule, using the likelihoods of alleles given the read data to calculate the likelihoods of each genotype per sample given the read data observed for that sample. The most likely genotype is then assigned to the sample generating a genomic VCF (gVCF) file per sample containing single nucleotide variants and indels.

These intermediate files can then be used for joint genotyping of multiple samples (i.e. all family members of a trio). Variant recalibration of each variant is done afterward. This process is done by a decomposing and normalization step. The decomposing step separates the variants present in the same position because, before this step, they are together. The normalization process normalizes the way that the representation of the variants is presented in the VCF file. This process follows the laws of parsimony, representing the variant in as few nucleotides as possible, and by left alignment, what means shifting the start position of specific variants to the left, until is no longer possible and without resulting in an empty allele.

All variants are then annotated using ANNOVAR. Annovar is a software that functionally annotates genetic variants, providing a lot of important information based on actualized data, making the downstream filtering possible. It uses the human reference genome (build hg19) to determine the chromosome, start position, end position, the reference and alternative allele and the gene where a called variant is. It identifies protein coding changes, like splicing, and aminoacid changes that happen due to a specific variant. It also has the power to give us information on the specific genomic regions where the variants are located, including conservation or regions of segmental duplications. Furthermore, it is able to identify the variants that are registered in variant databases like dbSNP, and the allele frequency of each in public databases, in our case, popfreq_all_20150413 (database containing all allele frequency of 1000G, ESP6500, ExAc and CG46).

Figure 2.2.1.1: Best Practices for Germline SNP & Indel Discovery in whole genome and exome sequence. 1- Pre-processing: mapping to reference genome using BWA; Mark duplicates using Picard tools; Base quality score recalibration. 2- Variant discovery: generate GVCF (genome VCF) per-sample with HaplotypeCaller; performe Joint Genotyping; Filter Variants. 3- Refine genotype; Annotate variants; Evaluate callset.



2.2.2 Filtering and candidate selection

Variant filtering was based on the following criteria: 1) keep variants overlapping a coding exon or within 2 base pairs of a splicing junction; 2) exclude synonymous variants; 3) exclude those present in segmental duplications; 4) keep variants not present or only present in less than 0.5% of the individuals in public databases like 1000 Genomes (1000 Genomes Project), ESP6500 (alternative allele frequency in ALL subject in the NHLBI-ESP project with 6500 exome, including the indel calls and Chromosome Y calls), Exac3 (exome aggregation consortium 6500 allele frequency data); 5) exclude variants not present in the patient and 6) keep variants with a GDI score less than 12.40551, as proposed by Yuval Itan et al. (2015)[74] to get variants that are probably disease-causing genes with high rates of true positives and low rates of true negatives.

The remaining variants were then grouped by their type: *de novo*, compound heterozygous and homozygous depending also on the mode of inheritance suggested by the clinical information on all family members.

2.2.3 Variant prioritization

Prioritization of filtered variants was based on: 1) whether they are within known PID genes or other genes biologically close to PID genes based on the Human Gene

Connectome server (http://hgc.rockefeller.edu/); 2) the probability of a functional impact on the protein based on functional prediction programs, such as CADD, Polyphen2 and SIFT predictors. The higher the CADD score, the most deleterious the variant. A score of 20, means the variants is among the 1% most deleterious variants. The Polyphen2 predicts the impact of aminoacid subtitutions and the scores range from 0 to 1. A score ranging from 0.85 to 1 is confidently predicted to be a damaging variant. SIFT predictor also predict the impact of aminoacid mutations but is more specific for missense mutations. The SIFT score range from 0 to 1, but a score ranging from 0 to 0.05 means the variants is confidentially deleterious; 3) The MSC (mutational significance cutoff) was also used. The MSC server estimates the impact of genetic variants by gene-specific cutoff significance. The MSC score represent the lowest expected clinical relevant CADD cutoff value for a specific gene. It includes CADD, Polyphen2 and SIFT scores and if a score of these predictors are equal or above the MSC score generated then the variants are predicted to have a high phenotypic impact.

Several other aspects were assessed: (1) function of the gene (2) protein and mRNA expression profiles available in GeneCards (http://www.genecards.org/) [117,118] and GTEx (Genotype-Tissue expression) [119], (3) search on literature for Human/ Animal studies in the gene, (4) protein structure to see which domain might be affected by the mutation using UniProt [120] and (5) homologies and conservation patterns in other species. All this information was obtained by assessing public databases, such as OMIM [121] and NCBI [122]. Variants were also screened through the Human Gene Mutational Database [97] to identify published disease-causing variants.

2.3 SANGER sequencing confirmation

Variants selected following the approach described above were SANGER sequenced for confirmation. Prior to SANGER sequencing, the specific region containing the mutation was amplified.

Primers for PCR amplification were designed with the help of NCBI tool, the Primer-Blast (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). This tool provides specific primers and also reveals to us the intended and the unintended products of chosen primers by doing a blast against the Human reference genome. The primers were ordered from IDT (Integrated DNA technologies). PCR reactions were performed in SimpliAmp Thermal Cycler machine (Applied Biosystems®). For the reaction, Q5 polymerase (New

England BioLabs®), Kappa polymerase (KAPABIOSYSTEMS) or KOD polymerase (EMD Millipore) were used. The standard reaction of the three enzymes are in **Table 2.3.1**. All primers used can be found in Supplementary Data, **Table 6.2**. Some of the PCR reactions needed to be optimized. The optimization was achieved by adapting up and down the volume of Magnesium and/or annealing temperatures.

Agarose gel electrophoresis was performed to confirm the size of the PCR product. The 1% agarose (BioExpress™) in 1X Tris-AcetateEDTA (TAE) Buffer (50X TAE: 248g of Tris base, 57mL of Acetic acid, 100mL EDTA at 0,5M) gel was prepared and Midori Green DNA gel stain (BulldogBio™) was added (5µl). After solidification, the gel was placed in a chamber with 1X TAE. Then, 7µl or 18µl of PCR product, if the primers amplified only one or more fragments, respectively, were added with 6X DNA Loading dye (Thermo Scientific™) (1,4µl or 3,6µl, respectively) and further loaded into the gel. Also, 6µl of GeneRuler 100bp DNA ladder, ready-to-use (Thermo Scientific™) was loaded onto the gel. Gels were run by using Enduro™ Power Supplies (Labnet international, inc), that was set to 140V and 500mA for 40 minutes. Gels were imaged with G: Box Chemi-XRQ system (SYNGENE) and GENESys software (SYNGENE™).

To prepare PCR product with only one fragment, the excess of PCR primer and dNTPs that were not consumed in PCR reaction were clean-up by 1µl of Exonuclease I (New England BioLab®) and 1µl of Shrimp Alkaline Phosphorylase, (rSAP) (New England Biolabs®) digestion directly in 5µl of PCR product. Then, the reaction was incubated for 5 minutes at 37^0C and 10 minutes at 80^0C [123]. At this point, PCR products are ready for downstream SANGER sequencing.

When the PCR product had more than one fragment we excised the correctly sized band from the electrophoresis gel and proceeded to gel clean-up. This gel clean-up was done using NucleoSpin® Gel and PCR Clean-up kit (MACHEREY-Nagel) following the manufacturer's protocol. Then, the samples were sent for SANGER sequencing.

Processed samples were then sent to the LGC group company (https://www.lgcgroup.com/), placed in Germany, for SANGER sequencing. The sequence received from LGC was aligned to a reference sequence using SnapGene software (http://www.snapgene.com/). We use this software to be able to easily manage sequence data, import transcripts from NCBI with all the features of the gene such as exons, active and regulatory sites, position and name of restriction sites in the sequence and, to perform the alignment of the raw SANGER sequence data with the intended transcript sequence so we can easily confirm or not the mutation.

Table 2.3.1: Standard protocols for Q5, KOD and Kappa polymerases.

Q5 polymerase protocol		KOD polymerase		Kappa Polymerase	
PCR Reagents (ul)	(1X)µl	PCR Reagents	(1X)µl	PCR reagentes	(1X)µl
dH_20	15,75	dH_20	14,25	dH_20	16,5
5X Q5 Buffer	5	10X KOD Buffer	2,5	5X KAPA HiFi Buffer	5
10mM dNTP	0,5	25 mM MgSO4	1,5	10mM dNTP	0,5
Mg	0	10mM dNTP	2,5	Mg	0
Q5 high-fidelity DNA polymerase	0,25	KOD hot start polymerase	0,5	Q5 high-fidelity DNA polymerase	0,5
DNA (65ng/µl)	1	DNA 65ng/uL	1	DNA (65ng/µl)	1
Rev primer (10µM)	1,25	DMSO	1,25	Rev primer (10µM)	0,75
FW primer (10µM)	1,25	Rev primer (10uM)	0,75	FW primer (10µM)	0,75
Total	25	FW primer (10uM)	0,75	Total	25
		Total	25		

2.4 RNA extraction and cDNA synthesis

The extraction of RNA using trizol is very efficient in cell lysis and the resulting RNA is generally with good quality to perform cDNA synthesis. In this protocol is important to keep the RNA samples always on ice to minimize the possibility of activation of RNAses. Along all protocol filter-tips were used to minimize contamination. We started the protocol by getting PBMCs in 500µl of Trizol from the freezer. Trizol was added (500µl) to these samples after the isolation of PBMCs from blood and were incubated 15 minutes at RT and then stored at -80Cº. We wait for the samples to defrost on ice. When we got defrosted samples we added 100µl of Chlorophorm (1/5 Volume of Trizol added). The samples were vortexed for 30 to 60 seconds and incubated at RT for 2 to 3 minutes. Then we centrifuge 15 minutes at full speed at 4Cº. After this, we carefully removed the aqueous phase into a fresh 2ml tube. It is very important to not transfer the white liquids under the aqueous phase because it is DNA. When we get to this point we use a kit called RNA clean & Concentrator™-5 produced by Zymo Research Company to obtain a better quality and concentration of RNA. After the process the concentration of RNA was tested using the BioPhotometer D30 machine from Eppendorf™.

The cDNA was prepared using the GoScript Reverse Transcription System from Promega™ (https://be.promega.com/-/media/files/resources/protcards/goscript-reverse-transcription-system-quick-protocol.pdf). The extracted RNA was used to synthetize cDNA. The resulting cDNA was tested with primers designed for *CECR1* transcript gene to evaluate the concentration and quality of the cDNA. Normally is assumed that the concentration of cDNA is similar to the concentration of RNA used for synthesis.

2.5 Quantitative PCR (qPCR)

The qPCR was performed to evaluate the gene expression of *CECR1* transcript. For the amplification of cDNA extracted from family 18 (1/10, derived from PBMC´s) we used a primer mix (10mM, containing forward and reverse primers presented in Table 2.5.1) and Fast SYBR Green Master Mix (Applied Biosystems™) through StepOnePlus™ Real-time PCR system (Applied Biosystems). For the normalization of the target transcript, two different housekeeping genes were used: *GAPDH* and *HPRT1*. We were interested to analyze the expression level of *CECR1* transcript that suffered exon skipping (without exon 7) and thus, the primers bound to exon 7 and exon 8.

Table 2.5.1: Primers used for qPCR.

Primer	Sequence (5´ to 3´)
CECR1 FORWARD	TGCCTTACTTCTTCCACGCC
CECR1 REVERSE	GCAAATCCATGGCCGATTCT
HPRT1 FORWARD	CCTGGCGTCGTGATTAGTGA
HPRT1 REVERSE	CGAGCAAGACGTTCAGTCCT
GAPDH FORWARD	GAAAGCCTGCCGGTGACTAA
GAPDH REVERSE	GCCCAATACGACCAAATCAGAG

2.6 Protein extraction

Cell pellets obtained from LCL's (Lymphoblastoid cell line) were resuspended in lysis buffer (1M Tris-HCl, 2M NaCl, 5% glycerol, 1M dithiothreitol (DTT), protease inhibitor (by ThermoFisher Scientific), phosphatase I Roche (by Roche Applied Science)) and sonicated during 3 intervals of 10 seconds at a 50% amplitude. Afterwards samples were incubated on ice for 30 minutes with added DNase (1µg/ml). The same lysis buffer was enriched with triton-X 2% and added following an incubation on ice for

10 minutes. Through centrifugation (8 minutes on 100,000g at 4°C) a protein rich supernatant was obtained and quantified with Qubit™ fluorometric quantitation.

2.7 Western Blotting

Proteins obtained from LCL's were ran through a 4-12% bis-tris acrylamide gel in a manner that each well contained 50µg of protein dissolved in a solution of LDS [1:4] and xt-reducing agent [1:20]. The gel ran at 200V for 20 minutes. Proteins were transferred from the gel to a blotting membrane with Bolt™ transfer buffer (by ThermoFisher Scientific with added methanol (10%)) and ran for one hour on 28V. Afterwards the membrane was blocked for one hour on room temperature with 5% BSA added to a NCP-wash buffer with 0,01% Tween. Primary antibodies in a NCP solution with 2% BSA and 0,01% Tween were added according to **Table 2.7.1** and incubated overnight at 4°C. Afterwards, the membrane was washed and secondary antibodies were added and incubated for an hour on room temperature (**Table 2.7.1**). Results were acquired by addition of ECL Prime (GE healthcare) creating a detectable chemiluminescence signal through oxidation of luminol. The density of each band was analyzed via ImageJ version 1.51 (National Institutes of Health, Bethesda, USA).

Table 2.7.1: Antibodies used for western blotting and their respective concentrations.

Primary antibody	Concentration	Secondary antibody	Concentration
Mouse anti-Histone H3K4Me3 (provided by Abcam)	1/ 500	Goat anti rabbit IgG (provided by Abcam)	1/ 20000
Rabbit anti-vinculin (provided by Abcam)	1/ 2000	Goat anti mouse IgG (provided by ThermoFisher scientific)	1/ 10000
Mouse anti-GAPDH (provided by ABcam)	1/ 5000		

3. Results

3.1. Identification of the monogenic cause for severe immune diseases

Family Case 1

This family is constituted of 5 Caucasian individuals. The mother and father are healthy and their 3 children are affected (**Figure 3.1.1**). The three children have different clinical phenotypes although some characteristics are shared. The first patient has a Schwachman-like syndrome phenotype, that is characterized by exocrine pancreatic insufficiency, bone marrow dysfunction, skeletal abnormalities and short stature. The second patient has autism and delayed development. The third patient also has autism and is suspected to have pancreas insufficiency.

In all family members, whole-exome sequence was performed and the obtained data was processed through our pipeline. The number of variants obtained at each filtering step is presented in the Supplementary data, **Table 6.4**.

After filtering, the variants were grouped into *de novo*, compound heterozygous and homozygous variants based on the inheritance patterns analyzed. We were interested in finding mutations that were shared among all patients, shared by two or only present in one individual (**Table 3.1.1**).

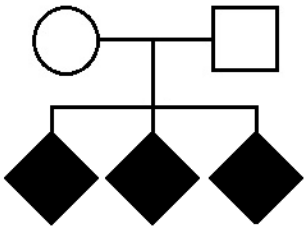


Figure 3.1.1: Pedigree of a family 1 composed by two parents and three children. Affected family members are shown in black.

Table 3.1.1: Different inheritance patterns and respective possible mutated genes in family case 1 patients (obtained after filtering process).

Inheritance pattern	Patient 1	Patient 2	Patient 3
de novo	*LNP1; KRTAP5-1; SELPLG; FAM177A1; MESP2; KRTAP4-1;SHROOM2; TRMT2B; PLXNA3; SPRR3; TMEM201; ALDH4A1; FCAMR; ABCB6; GPBAR1; SEC24D; RGS14; SLC22A23; MTCH2; TNKS1BP1; CAPN5; CNTN5; RAPGEF3; TYRO3; CDC27; ZC3H12D; ANKMY2; CLDN3; SLC4A2; IL33; TPM2; PHF21A;* ***POLE****; SLC24A5;KLHL25;ERN2;FMNL1;SCN4A;USH1G;FBF1;LPIN2;ARRDC2;COPE;C20orf96;CEP250;MAPK11;*	*LNP1; MESP2; IHH; ZAN; MGAT5B; PLPPR3; C5orf38; GNLY; TMEM201; FCAMR; HOXD12; ABCB6; LHFPL4; LHFPL4; GPBAR1; RGS14; SLC22A23; PABPC1; MTCH2; TNKS1BP1; KRT18; TYRO3; BCLAF1; IL33; TPM2; ANO9; FOLR3; ATN1; LPIN2; TXNDC2; DYNAP; TLE6; ARRDC2;* ***SHANK1****; ZNF525; LILRB5; CEP250; PLPPR3;*	*FAM177A1; MESP2; JRK; SSPO; TMEM201; MIB2; ALDH4A1; FCAMR; BCL11A; GPBAR1; SEC24D; RGS14; SLC22A23; MTCH2; TNKS1BP1; CAPN5; TYRO3; SLC4A2; TPM2; MRGPRX1;* ***POLE****; TMEM255B; KLHL25; LPIN2; TXNDC2; ARRDC2; ZNF208; PLEKHG2;* ***SHANK1****; ZNF525; LILRB5; MAPK11;*
Compound Heterozygous	***DIDO1***	*SSPO*	/
Homozygous	*FAM120B; TRPM3; FAM177A1*	*OVGP1; PLPPR3*	*FAM120B*

No biologically relevant mutations were found by looking at the genes that were shared by all the patients. Two relatively good variants could be found when looking for gene variants shared by P1 and P3 (*POLE* gene, more information can be found below) and by P2 and P3 (*SHANK1* gene, which has been associated with autism [124,125]). Moreover, in P1, we could find one compound heterozygous mutation in *DIDO1*.

POLE is a known PID gene and the mutation was annotated as being private (absent in ESP6500, Exac_ALL and 1000G _EUR databases) and with a high impact prediction (**Figure 3.1.2a**). The genotype quality and the filtered coverage depth of this mutation is 61 and 10, respectively for patient 1 and, for patient 3, 99 and 17, respectively. According to GTEx, the mRNA of this gene in normal cells is overexpressed in the cerebellum and cerebellar hemisphere. The protein expression in normal cells is overexpressed in lung alveolar lavage and in monocytes, according to HIPED and other databases (**Figure 3.1.2b**).

a)

Chromome	19	12
Position	51172480	133226080
ID	-	-
Reference_Allele	A	G
Alternative_Allele	G	A
Gene	SHANK1	POLE
CADD_Score	23.600	24.800
MSC-CADD_Score	3.313	12.420
MSC-CADD_Impact_Pred	high	high
PolyPhen2_Score	0.991	0.339
PolyPhen2_Pred	probably damaging	benign
MSC-PolyPhen2_Score	0.350	0.999
MSC-PolyPhen2_Impact_Pred	high	low
SIFT_Score	0.010	0.020
SIFT_Pred	D	D
MSC-SIFT_Score	0.243	0.995
MSC-SIFT_Impact_Pred	high	high

b)

Figure 3.1.2: Impact prediction and protein expression of POLE. a) Mutational Significance Cutoff (MSC) prediction tool that includes CADD, Polyphen2 and SIFT scores. This table was generated on the MSC server. b) This image was taken from GeneCards and represents the protein expression in normal tissues and cells lines from ProtemoicsDB, PaxDb, MaxQC and MOPED of *POLE* gene.

The expression of this genes is not especially high in the pancreas but it is in the monocytes. The monocytes are a type of white blood cell and can differentiate into, for example, macrophages. The monocytes can travel around the body through the circulatory system, get to the pancreas and affect the immunologic system. A

homozygous mutation in exon 34 of *POLE* gene has been associated with immunodeficiency, facial dystrophy livedo and short stature in a consanguineous French family [126]. Patient 1, that present a Schwachman-like syndrome phenotype, also has short stature and immunodeficiency and, patient 3 only present a small part of patient 1 phenotype (pancreas insufficiency). To consider the mutation as a cause of the phenotype, we need to extrapolate the phenotype of patient 3. Patient 3 might have the same problem as patient 1 but respond to *POLE* mutation in a different way, evading the rest of the phenotype which can also mean incomplete penetration (short stature and immunodeficiency). The founded mutation in *POLE* gene, unlike the substitution mutation in the French family that caused a severe decrease in the expression, is a heterozygous mutation that might cause a protein gain of function. This gain of function can change gene expression or influence protein functional state and change the normal response of several cells, for example, T and B lymphocytes, as reported by Pachlopnik Schmid et al. (2012).

The mutation in *SHANK1* has a high impact prediction in MSC (**Figure 3.1.2a**) and was also annotated as a private mutation. The sequencing quality of this mutation can be analyzed by the genotype quality score and coverage depth that is 62 and 6 for Patient2 and 19 and 6 for Patient3, respectively. For *SHANK1* gene, according to HIPED, the mRNA is overexpressed in several regions of the brain: anterior cingulate cortex, amygdala, hippocampus, frontal cortex and cortex. The *SHANK1* protein is thought to be overexpressed in cerebrospinal fluid, urine and breast, according to HIPED and other databases (**Figure 3.1.3**). *SHANK1* mRNA is highly expressed in several tissues of the brain suggesting a regulatory function that can respond to several stimuli and cause a shift in synaptic morphology [127]. Besides that, the disruption of *SHANK1* gene function by both gene deletion in mice or exon deletion in humans has been associated with autism [124,125]. Patient 2 and 3 share autism as phenotype and the same mutation in *SHANK1*. Autism is a neurologic disorder characterized by impairments in communication, social interaction, repeated and stereotyped pattern of behavior. The role of *SHANK1* in autism is a little controversial. Several authors reported that mice lacking Shank1 did not reveal social interaction deficits while a wide study made in human shows that a hemizygous deletion of several exons (1 to 20) is present in males with autism and in their asymptomatic mothers, suggesting a gender bias in autism [125,128]. Besides that, these authors suggest that SHANK1 CNV is one of the first events that lead to autism supported by the discovery

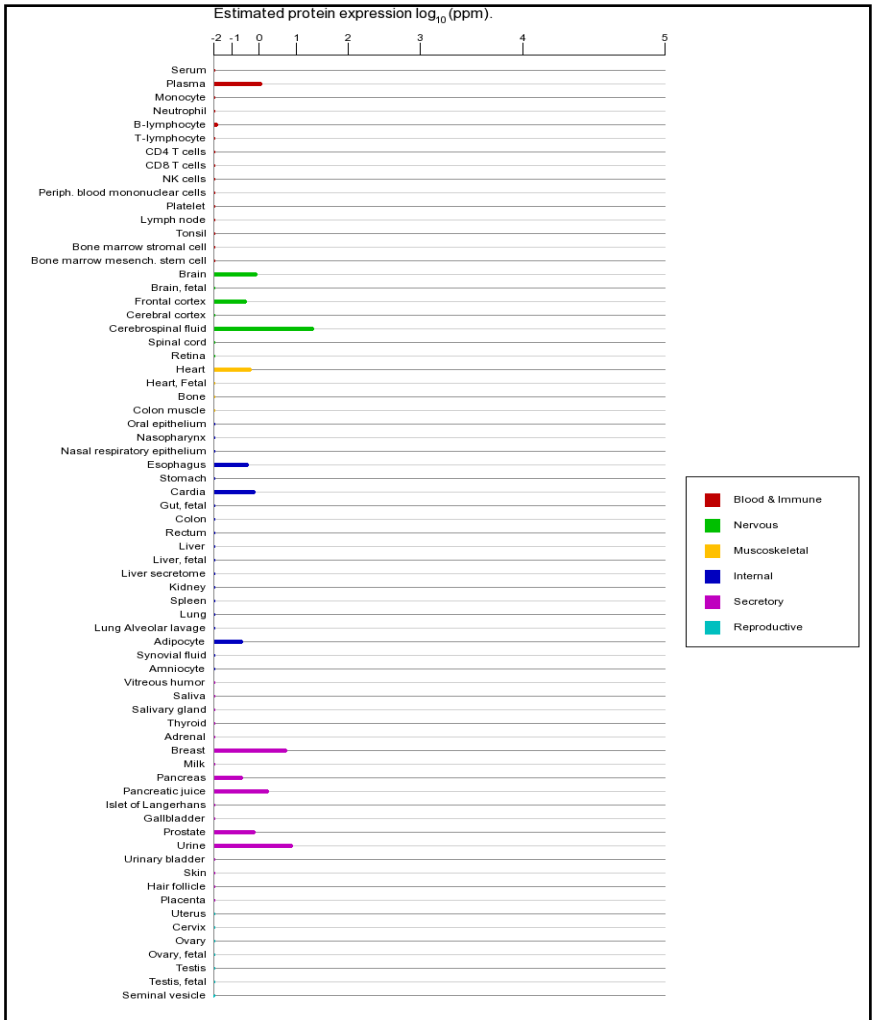


Figure 3.1.3: Protein expression of SHANK1. **This image was taken from GeneCards and represents the protein expression in normal tissues and cells lines from ProtemoicsDB, PaxDb, MaxQC and MOPED of *SHANK1* gene.**

of a *de novo* deletion of *SHANK1* in the same locus of the previously referred patients, but in an unrelated individual [124]. The substitution mutation we found, although it is predicted to have a high impact on protein function, is unlikely to cause such as strong impact as segmental deletions but can be increasing the susceptibility to develop autism in this two patients.

The compound heterozygous mutations in *DIDO1* is constituted by two substitution mutations. One is inherited by the mother (exon16:c.G3994A:p.A1332T) and had a frequency of 0.3% in 1000G_EUR, 0.07% in ExAC_ALL and ESP6500siv2_ALL. Mutational significant cut-off applied to CADD show that this variant has an high impact, while MSC applied to PolyPhen2 and SIFT show that this mutations has low impact prediction. The genotype quality is 99 for patient 1 and the mother and the allele depth supporting the variant is 18 in the mother and 41 in patient 1. The mutation inherited by

the father (exon16:c.G3975T:p.K1325N) was completely absent in all databases. The MSC predicted that this variant has a high impact when applying to CADD, PolyPhen2 and SIFT. Regarding the sequencing quality, the genotype quality is 99 for both patient and father and the allele depth is 31 and 37 for father and patient, respectively. The mRNA expression of this gene is present in almost all types of cells however, the protein is overexpressed in peripheral blood mononuclear cells, plasma, colon muscle and CD8 T cells, according to GeneCards. Fütterer et al. (2005), showed that *DIDO1* knockout mice had abnormalities in the spleen, bone marrow and peripheral blood. The *DIDO1* gene, dead inducer-obliterated 1, is thought to be a transcription activator of caspases 3 and 9 in response to the induction of apoptosis by for example IL-3 starvation or *Myc* upregulation in B lymphocytic cells [130]. *DIDO1* has also been associated with the regulation of hematopoietic processes and to be capable, by its downregulation, to induce hematological myeloid neoplasm [130,131]. The reason this gene is considered a candidate is that the phenotype of *DIDO1* knockout mice revealed abnormalities in the bone marrow, which patient 1 also presents, and abnormalities in the spleen and peripheral blood. This knockout resulted in aberrant levels of erythroid, granulocytic or monocyte cells both in spleen and bone marrow [129]. The recessive inheritance of our variant goes in line with the suspected loss of function mutations that will probably cause deficiencies in the bone marrow. However, this suggestion takes into account that the mutation might not have complete penetrance or be affecting more pathways in humans.

No suggestion or hypothesis relating one of the gene variants with the complete phenotype could be found. No functional tests were done on these three potential candidate genes, because they could only explain one or two different characteristics of the affected individuals. These results could suggest that some of the disease characteristics might be driven by more than one gene.

Family Case 2

Exome-sequence analysis was done in this Caucasian family (**Figure 3.1.4**). The son is affected with colitis disease and pyogenic infection. The colitis disease is characterized by inflammation of the inner lining colon and the pyogenic infections by local inflammation with formation of pus, generally caused by pyogenic bacteria.

The number of *de novo*, compound heterozygous and homozygous variants obtained after filtering was 30, 2 and 2, respectively (**Table 3.1.2**).

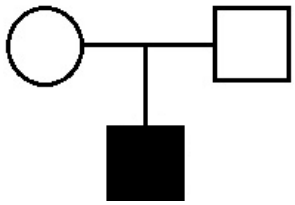


Figure 3.1.4: Family 2 pedigree composed by two parents and one child. The affected family member is shown in black.

Table 3.1.2: Different inheritance patterns and respective possible mutated genes in family case 2 patient (obtained after filtering process).

Inheritance pattern	Patient
de novo	*SPEN; SPRR3; IQCF3; DCLK2; PCSK1; AP4M1; MUC3A; DENND2A; SSPO; PHF19; ADAMTS13;* ***TRPM5****; ZNF143; PYGM; CNTN5; SIK3; CCDC92;* ***PSMB11****; TMEM63C; JAG2; MPI; SH2D7; SRRM2; ZSCAN10; ADGRG1; CNOT1; CDT1; ZNF626; ZNF83;* ***CST7****;*
Compound Heterozygous	*COL7A1; THEG;*
Homozygous	*TRAK1; EBLN2*

Through the analysis of this family, *de novo* mutations in three promising genes were found: 1) A heterozygous variant in the *TRPM5* gene, 2) A heterozygous and private mutation in the *PSMB11* gene and 3) A heterozygous mutation that creates a stop loss codon in the *CST7* gene.

The mutation found in *TRPM5* has a very low frequency among the databases analyzed. More specifically, it is absent in 1000G_EUR and ESP6500siv2_all databases and has a reported frequency of 0,0001 (0.01%) in the Exac_All database. The quality of the genotype call provided by the machine is 61 along with an allele depth and filtered coverage depth of 6 and 2, respectively. The read depth seem to be very low as well as the score for the same position allele in the mother and father (Supplementary data: **Table 6.3**). The results of MSC show that this mutation is predicted to cause damage in the protein (**Figure 3.1.5a**). Based on the STRING Interaction Network, this gene is known to interact with *ITPR3,* a PID gene predicted by Yuval Itan et al. (2015).

a)

Chromome	11	14	20
Position	2436589	23511670	24940348
ID	-	-	-
Reference Allele	C	C	A
Alternative Allele	T	0	C
Gene	TRPM5	PSMB11	CST7
CADD Score	28.600	NA	18.180
MSC-CADD_Score	3.313	3.313	2.279
MSC-CADD Impact_Pred	high	NA	high
PolyPhen2 Score	0.493	NA	NA
PolyPhen2 Pred	possibly damaging	NA	NA
MSC-PolyPhen2 Score	0.239	0.239	0.239
PolyPhen2	high	NA	NA
SIFT_Score	0.050	NA	0.000
SIFT_Pred	D	NA	D
MSC-SIFT Score	0.243	0.000	0.191
MSC-SIFT Impact_Pred	high	NA	high

b)

Figure 3.1.5: Impact prediction and protein expression of *TPRM5*. a) Generated on MSC server using as input the specific mutation in *TRPM5, SHANK1* and *CST7* genes. b) This image was taken from GeneCards and represents the protein expression in normal tissues and cells lines from ProtemoicsDB, PaxDb, MaxQC and MOPED of *TPRM5* gene.

The mRNA expression of *TRPM5* in normal human tissues is high in small intestine-terminal ileum, colon–transverse, pancreas and thyroid, according to GTEx. UniProt/SwissProt also notes for an overexpression of the mRNA in the colon and peripheral blood leukocytes. The protein expression of this gene, according to ProteomicsDB, PaxDb, and MOPED is high in CD8-T cells and in the pancreas (**Figure 3.1.5b**). The overexpression of *TRMP5* gene in leukocytes and T cells and its interaction with *ITPR3* suggest that this gene in operating in immunologic responses. Furthermore, Mutations in *ITPR3* have been associated with high risk for type 1 diabetes and systemic lupus erythematosus. The mutation in *TRPM5* was inherited by the father instead of *de novo* (**Figure 3.1.6**). The probability of this variant to be diseasing causing is now very reduced because the father also has the mutation and he is healthy.

Figure 3.1.6: SANGER sequence result for *TRPM5* variant in the father, mother and patient. SANGER sequence aligned with a reference sequence (REF) using SnapGene, demonstrating the presence of the mutation in exon9 in the patient and father. The mutation is a substitution of a guanine for adenine in position 1241 of the transcript NM_014555 that results in a change of protein sequence from proline to leucine in position 414 (G1241AT:p.R1131Q).

The mutation in *PSMB11* is a deletion in codon 98 of the protein. There is no information about the impact of this mutation (**Figure 3.1.5A**).

The sequencing quality of this mutation is represented with a genotype quality of 99, a filtering coverage depth of 32 and an allele depth of 15. The sequencing quality of the mother and father have lower scores (Supplementary data: **Table 6.3**). This gene generates peptides that are presented by major histocompatibility complex to other cells of the immune system and is thought to play a pivotal role in the development of CD8-positive T cells [132,133]. According to GTEx, the mRNA is overexpressed in testis. Furthermore, the protein is overexpressed in the retina and natural killer cells according to HIPED. *PSMB11* encodes a proteasome that produces peptides presented by Histocompatibility Complex I to other immune cells. The presence of *PSMB11* variant by SANGER sequencing was not validated and we considered it as an artifact because is not present in the family individuals (**Figure 3.1.7**).

Figure 3.1.7: Result of the SANGER sequence of the *PSMB11* gene in the father , mother and patient. SANGER sequence aligned with a reference sequence using SnapGene, confirming that the expected mutation in exon1 is not present. The mutation was a deletion of a cytosine in position 293 of the transcript NM_001099780.

The missense mutation present in *CST7* is a substitution of adenine by cytosine in position 438, turning a stop codon to a cysteine. The quality scores of the exome sequencing for this mutation in the patient can be assessed by the genotype quality that corresponds to 83 and by the number of reads supporting the reference allele, 25, and the alternative, 12. Based on HGC, this gene is biological close to *CTSC*, a known PID gene. Furthermore, this mutation is predicted to be damaging based on CADD and SIFT with scores of 18 and 0, respectively. The MSC-CADD and SIFT also predict that this mutation has a high probability of causing a protein function shift. Accessing GTEx and GeneCards databases for this specific gene, we obtained the information that in normal tissues the mRNA is overexpressed in whole blood. According to HIPED, the protein is overexpressed in NK cells, peripheral blood mononuclear cells and CD8 T cells. It is also expressed in the spleen (**Figure 3.1.8**). The *CST7* gene is thought to be involved in the regulation of antigen presentation and in immune responses and is very important for

eosinophils surviving [134,135]. The expression of this gene is high in cells related to the immune system, such as T cell and natural killer cells, but its exact function is not known. The mutation in *CST7* was also not confirmed by SANGER sequencing (**Figure 3.1.9**).

None of the mutations in *TRPM5*, *PSMB11* or *CST7* were confirmed by SANGER sequence and the whole genome sequencing should be performed in order to find new candidate genes that could be causing colitis disease and the invasive pyogenic infections seen in the patient.

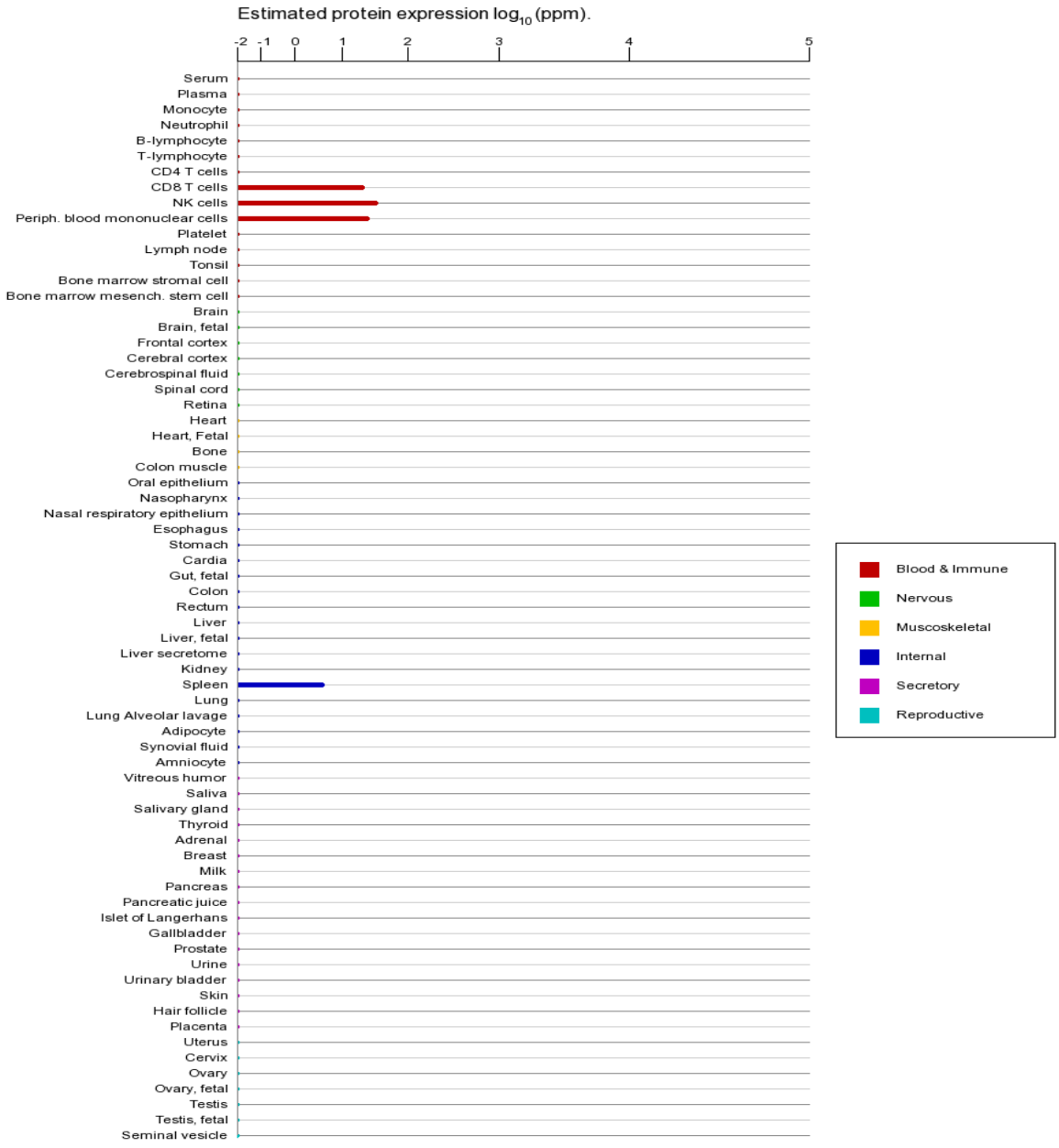


Figure 3.1.8: Protein expression of *CST7*. This image was taken from GeneCards and represents the protein expression in normal tissues and cells lines from ProtemoicsDB and MOPED of *CST7* gene.

Figure 3.1.9: SANGER sequence result for the *CST7* variant in the father, mother and patient. SANGER sequence aligned with a reference sequence with SnapGene, showing that the expected mutation in exon1 is not present in the patient.

Family Case 3

In this family trio (family trees will not be shown for family trios throw-out the document), the patient is affected with PID and medulloblastoma and the parents are healthy. Through exome-sequence analysis and filtering, we got 59 *de novo*, 5 compound heterozygous and one homozygous mutation (**Table 3.1.3**).

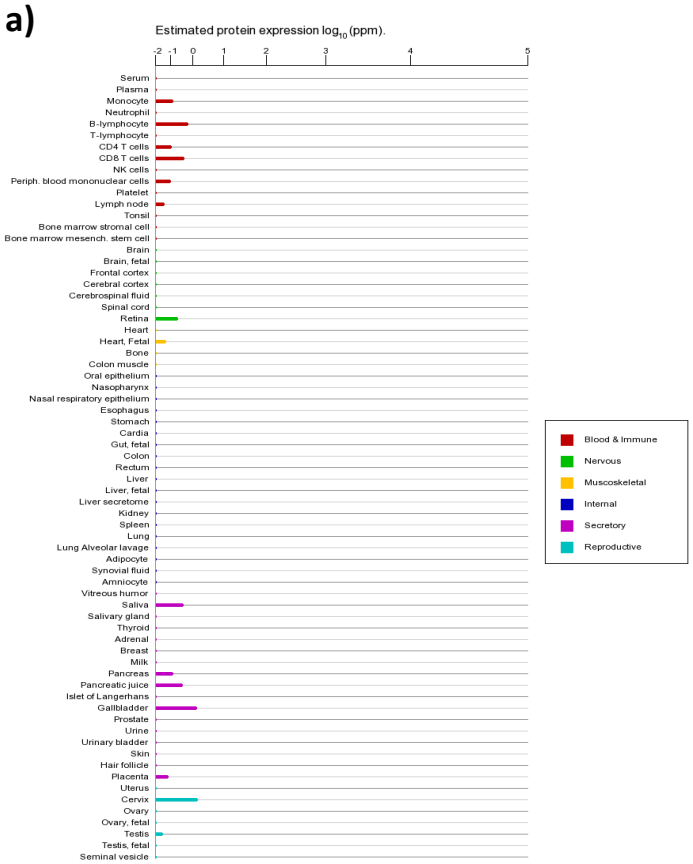
The patient (son) was found to be compound heterozygous for mutations in a biologically relevant gene, *KMT2D* (alias MLL2, MLL4, ALR, and kabuk1). The patient inherited a variant in exon 39 from the mother and a second mutation, in exon 11, from the father.

Table 3.1.3: Different inheritance patterns and respective possible mutated genes in family case 3 patient (obtained after filtering process).

Inheritance pattern	Genes with variants
de novo	*HNRNPCL3,HNRNPCL4; HNRNPCL2; PRAMEF10,PRAMEF33P; UBL4B; NBPF10; SPRR2E; POTEF; POTEE; SATB1; TRAK1; PIK3R4; FGFRL1; HLA-DRB1;HLA-DRB1; ITPR3; TBP; HOXA1; GTF2IRD2; STEAP1; STEAP1; PRKDC; SPATA31A3; ANKRD20A2,ANKRD20A3; RASEF; AGAP4; WT1; OR5L1; TAS2R46; TAS2R43; GXYLT1; KRT18; SSTR1; GOLGA8H; CSPG4; COPS3; RPL19; FASN; ZNF519; CNN2; VCX; SLC25A5*
Compound Heterozygous	*COL14A1; OR10A3;* ***KMT2D****; TRAFD1; FBN1*
Homozygous	*OR2T12*

The first mutation (C12542G:p.S4181C) rs776127830)) is absent in esp6500siv2_all database and the frequency in exac03 is 0,000008332 (≈0,0008%). In Kaviar database the frequency is reported to be $1.3x10^{-5}$. This mutation has good sequencing quality scores with a genotype quality of 99 and an allele depth of 15 for the reference allele and 5 for the alternative allele in the patient. For the second mutation (C3392T:p.P1131L) (rs201623566)) the frequencies are 0,0004, 0,0013 and 0.001 in esp6500siv2_all, exac03 and Kaviar databases, respectively. The genotype quality of this mutation is lower than the previous mutation, presenting a score of 43 and an allele depth of 3 for the reference allele and 9 for the alternative allele in the patient. The *KMT2D* protein expression seems to be high in cervix, gallbladder and B lymphocyte according to HIPED but also relatively high expression in secretory, blood and immune systems (**Figure 3.1.10a**). The mRNA *KMT2D* in normal tissues of human cells is expressed in the brain and thyroid, according to GTEx Portal (**Figure 3.1.11**). Regarding the damage prediction, only CADD was able to provide a score, 12, which indicates that this mutation belongs to the 10% most deleterious mutations of the human genome. The MSC-CADD predicted that both mutations are highly damaging for the protein (Figure, 3.1.10b).

a)



b)

Chromome	12	12
Position	49425946	49443979
ID	-	-
Reference_Allele	G	G
Alternative_Allele	C	A
Gene	KMT2D	KMT2D
CADD_Score	12.650	12.950
MSC-CADD_Score	0.001	0.001
MSC-CADD_Impact_Pred	high	high
PolyPhen2_Score	NA	NA
hvar_prediction	NA	NA
MSC-PolyPhen2_Score	0.046	0.046
PolyPhen2_Impact_Pred	NA	NA
SIFT_Score	NA	NA
SIFT_Pred	NA	NA
MSC-SIFT_Score	NA	NA
MSC-SIFT_Impact_Pred	NA	NA

Figure 3.1.10: Impact prediction of the compound heterozygous mutation and protein expression of *KMT2D*. a) This image was taken from GeneCards and represents the protein expression in normal tissues and cells lines from ProtemoicsDB, PaxDb, MaxQB and MOPED for *KMT2D* gene. b) Table provided by the MSC prediction tools server for the two *KMT2D* mutations.

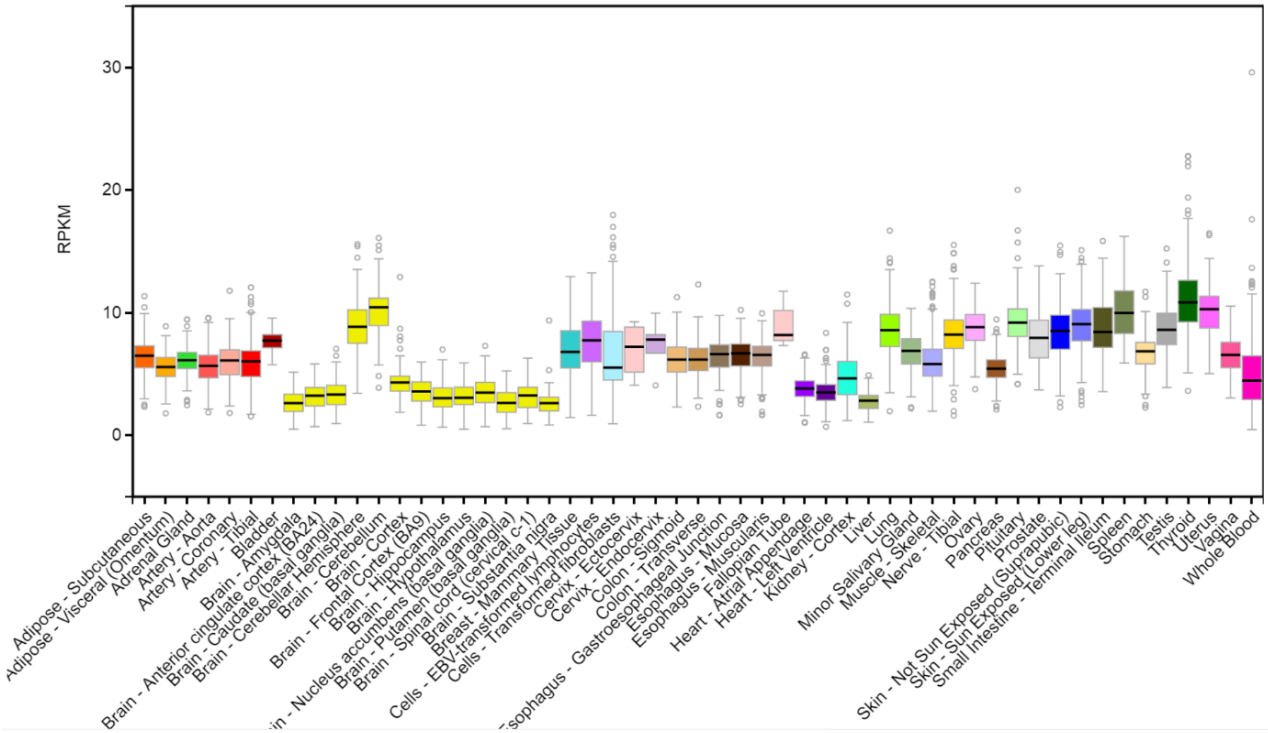


Figure 3.1.11: mRNA expression of *KMT2D*, extracted from GTEx.

The *KMT2D* (Histone-lysine N-methyltransferase 2D) is a methyltransferase that targets Histone 3 and lysine 4 [136]. The *KMT2D* gene is an epigenetic regulator capable of changing DNA expression, affecting genes involved in development, differentiation and tumor suppression. Furthermore, *KMT2D* is thought to play a role in immune homeostasis supported by the finding of *KMT2D* mutations in developmental diseases, such as Kabuki syndrome or congenital heart disease [137–139]. Patients with Kabuki syndrome have similar clinical characteristics to common variable immunodeficiency, including antibody deficiency, decreased memory cells and vaccine response, susceptibility to infections, reduced serum immunoglobulin levels, autoimmune manifestations, vitiligo [137,138]. In studied Kabuki syndrome patients, the truncating mutations prevail, which suggests that a reduced level of functional protein is the major mechanism behind the disease [140]. The function and exact interaction of *KMT2D* during histone methylation are not yet clear but several studies have highlighted that *KMT2D* defect induces a loss in mono, di and trimethylation [141–143]. The monomethylation occurs mostly at enhancer sites while trimethylation is more frequent in promotor sequences and therefore, we can differentiate promotors and enhancer by their methylation profile [144]. The di- methylation of H3K4 is equally present in promotors and enhancers [143,145].

The potencial role of *KMT2D* in cancer has also been shown different types of cancer (gastric cancer, medulloblastoma, etc) and frequently the mutations found are a loss of function, supporting the role in the suppression of cell proliferation [146]. Although, gain of function mutation was also found, for example, in p53 gene in cells derived from patient tumors and posterior experiments showed that *KMT2D* was regulated by the p53 mutated protein. This study suggests that *KMT2D* might be important for tumor progression [147]. Besides that, Dahlin et al. (2015) have correlated *KMT2D* to medulloblastoma. According to COSMIC (Catalogue of somatic mutation in cancer), the mutation C3392T:p.P1131L have been found in children with medulloblastoma [149].

The compound heterozygous mutation was confirmed in the patient by SANGER sequence (**Figure 3.1.12** and **3.1.13**). After that, I followed a laboratory colleague that performed a Western Blot to verify the methylation levels of H3K4me3. The results of this Western Blot show a reduced level in the order of 75% of H3K4me3, confirming that the variant in *KMT2D* can influence the patient methylation profile (**Figure 3.1.14**).

Figure 3.1.12: SANGER sequence results for the *KMT2D* gene in the father, mother and patient. SANGER sequence aligned with a reference sequence with SnapGene, confirming the presence of the mutation in exon11 in the patient and father. The mutation is a substitution of a cytosine for a thymine in position 3392 that results in a change of protein sequence from proline to leucine in position 1131 (C3392T:p.P1131L).

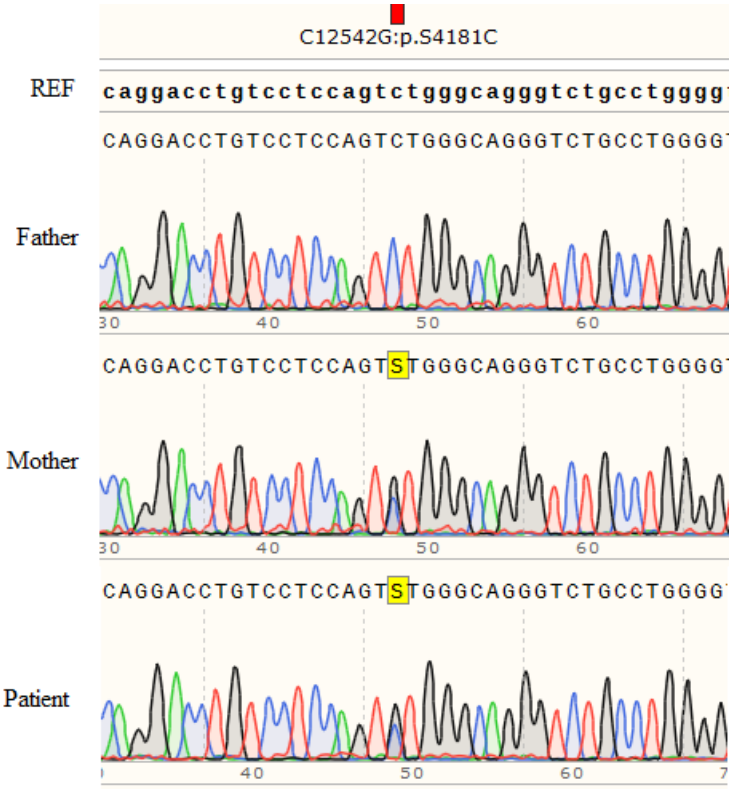


Figure 3.1.13: SANGER sequence results for the second variation in *KMT2D* gene in the mother, father and patient aligned with a reference sequence with SnapGene. This analysis confirmed the presence of the mutation in exon 39 in the patient and mother. The mutation is a substitution of a cytosine for a guanine at position 12542 that leads to an amino acid change from serine to cysteine at position 4181 (C12542G:p.S4181C).

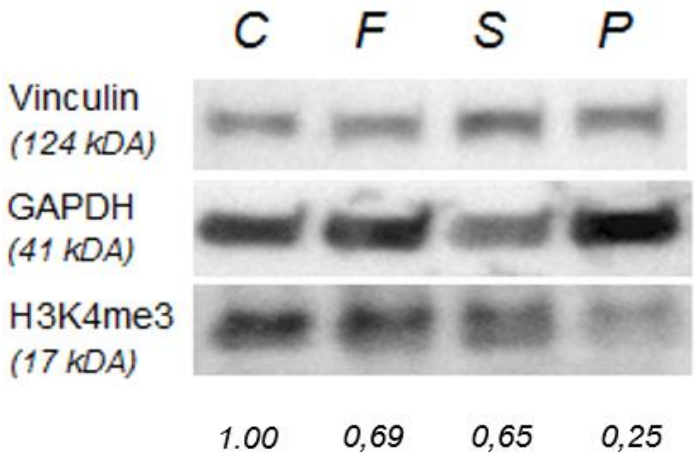


Figure 3.1.14: Immunoblot analysis of H3K4me3 in LCL's from Control (C), Father (F), Sister (S) and Patient (P) respectively. Housekeeping genes sued was Vinculin and GAPDH necessary for standardization of the H3K4me3 bands. The H3K4 level of trimethylation is clearly reduced in the patient as represented by the number below the figure which represent the density of each band.

The loss of function mutations in *KMT2D* have been shown is cases of tumor progression and in Kabuki syndromes and, our mutations that follow a recessive inheritance will probably cause a loss of protein function leading to a decrease in methylation levels. The loss of function of *KMT2D* will cause a general decrease of methylation and might cause a down regulation of more than 300 genes, as reported by Guo et al., (2013), who studied the different gene expression in a cell line with a *KMT2D* knockout. We analyzed the trimethylation of H3K4 due to its importance in the promotor region of actively transcribed genes [150]. The western blot results support that the mutation is affecting the trimethylation level of H3K4. The analysis of the three levels of methylation in a cell line with *KMT2D* knockout showed that the monomethylation level was the most affected and the other methylations levels only suffer minor changes [146]. According to this study, we could predict a difference in methylation not very severe, unlike the results we had. The use of only one control might lead to deviation in the standardization and so our results might be misrepresented. To overcome this problem and to be sure that the trimethylation is indeed reduced in the patient, this western blot should be repeated using at least 3 controls.

To further confirm that his mutation is indeed disease causing we could see which genes are being differently expressed by doing RNA sequencing in cells of the patient, mother, father and control. Besides that, performing an immunophenotyping would of

extreme importance to understand better how the immune system is being affected and perhaps unravel how *KMT2D* can lead to the pathologic phenotype of this patient.

Family Case 4

In this family trio, the patient has PID, retardation and aneurysm. After exome-sequence analysis and filtering, the number of *de novo*, compound heterozygous and homozygous mutations in this family was 18, 3 and 6, respectively (**Table 3.1.4**).

Table 3.1.4: Different inheritance patterns and respective possible mutated genes in family case 4 patient (obtained after filtering process)

Inheritance pattern	Patients genes with mutations
de novo	*ILDR1; **RAET1E**; PABPC1; KCNC1; OR8G1,OR8G5; SKA3; CCDC88C; TYRO3; JMJD7-PLA2G4B,PLA2G4B; TMEM94; CNN2; SF3A2; PPM1F; ASB12; CDX4; **ADAMTSL5**;*
Compound Heterozygous	***PKHD1**; OR8U1; MPRIP;*
Homozygous	*OVGP1; ADAMTSL5; ELP4; SELPLG; PRDM15;*

Two biologically interesting candidate genes related to the immune system could be considered, the *ADAMTSL5* and *PKHD1*.

The mutation found in *ADAMTSL5* was a *de novo* mutation. The patient inherited a mutated allele from the mother (T) and suffered a *de novo* mutation turning a cytosine to a thymine, becoming homozygous for this allele position. *ADAMTSL5* gene has been associated with psoriasis and its stimulation leads to a psoriasis signature cytokine, suggesting that the translated peptide is a psoriatic autoantigen[151]. The new pathway identified shows that ADAMTS- like protein 5 (*ADAMTSL5*) act as a human leukocyte antigen (HLA-C*06:02) of a specific TCR. This TCR becomes active and induces a melanocyte attack, causing lesions in the skin [152]. The HLA-C*06:02 normally presents nonamer peptides with dominant arginine anchors at position 2 and 7 of the protein and has a preference to small hydrophobic residues at the C-terminus [153]. The mutation we found was a homozygous substitution of a glycine to an arginine (p.G275R), around the middle of the protein. This mutation can be creating a new bind site perhaps linked to the activation of more TCR and resulting in skin lesions. If so, this mutation would defined as gain of function (GOF). Although this information, the mutation we found in the

patient had a very low genotype quality (39) and it is unlikely to be confirmed by SANGER sequence. This gene could only be to be causing the skin lesions which is only one of many patient manifestations.

In regard to *PKHD1* gene, which encodes for protein fibrocystin, some mutations have been associated with autosomal recessive polycystic kidney disease and the most common variant is a missense mutation in exon 3 [154,155]. Assessing the mutational database of autossomal recessive polycystic kidney disease, assembled by the department of human genetics, RWTH Aachen University, we found both of our mutations (www.humgen.rwth-aachen.de). One of the mutations (c.G8345C:p.G2782A) was considered disease causing by Mutation Taster and by Polyphen2 but the expected pathogenicity described in the literature characterized this mutation as polymorphic [156]. The second of the compound heterozygous mutation (c.G5134A:p.G1712R) is also present in this database and is predicted to be disease causing by Mutation Taster, with a high confidence, and by Polyphen2 as well. In the literature this mutation is also described as probably pathogenic [157]. The region of the mutations in *PKHD1* had very good sequencing quality in all individuals, and both mutations in the patient had a genotype quality of 99 and allele depth of more than 10. Our mutation indicates that our patient might have polycystic kidney disease but the described phenotype did not point to that. It would be important to imaging the kidney with ultrasound, magnetic resonance imaging (MRI) or computed tomography, despite the ultrasound be more advantageous due to the easiness of use, cost and safety [158]. Our patient, besides PID and retardation, also suffers from an aneurysm. It has been reported that a few patients diagnosed with the autosomal recessive polycystic disease (ARPKD) also presented intracranial or extracranial aneurysms [159,160]. This is one more indication that this patient should be tested for ARPKD. The mutations in these two genes suggest that the patient phenotype might not be induced by only one gene.

We also look at another immune gene, *REAT1E*. This gene presented a *de novo* substitution in the patient that is thought to cause aberrant splicing (exon5:c.623-2A>G). This mutation was private (absent in all databases) and is predicted to be highly damaging by MSC-CADD. In the patient, the genotype quality is 63 and the allele depth for the alternative allele is 4. The sequencing quality in the parents could not be evaluated because the sequencing call was not made. The variant also did not pass all filters of confidence, specifically VQSR Tranche SNP 99.90 to 100. This gene, *REAT1E*, is

predicted to be a candidate PID gene by Yuvan et al. (2015), due to its interaction with a known PID gene, *THBD*. The *REAT1E* gene is a ligand for *NKG2D* which is an activating surface receptor expressed on, for example, natural killer cells, $CD8^+$ T cells and subsets of $CD4^+$ T cells involved in host defense and suspected to drive autoimmune diseases [161]. *REAT1E* is included in retinoic acid early induced transcript-1 (RAE-1) glycoproteins family which are known to be stress molecules activated in pathological condition and are ligands for the immune receptor *NKG2D* [162].

Other authors also found that *RAET1* splice variants, when abnormally expressed lead to an increase of cytotoxic sensitivity of target cells against natural killer cells [163]. According to GTEx, the mRNA of *RAET1E* is overexpressed in esophagus, vagina, cervix and skin exposed and not exposed to the sun. The protein is expression could not be assessed in any database. Although this gene could be a good candidate, the sequencing quality shown is very poor. Besides the unavailable quality score of the parent which indicates bad sequencing, the allele depth of the alternative allele is also very low (4) compared to the reference allele (63). Together, this indicates general low quality sequencing and that this variant is probably an artifact. We exclude this variant as a candidate due to the high probability of being an artifact. However, SANGER sequence confirmation should be executed. The SANGER sequence was not performed in any of the mutations due to the lack of a strong relation with the totality of clinical phenotype.

Family Case 5

In this family trio, the patient is the only affected suffers from PID and EBV (Epstein – Barr Virus)-related lymphoma. In this family, we could find through our exome-sequence analysis pipeline 19 *de novo* mutations but no compound heterozygous or homozygous mutations (**Table 3.1.5**).

Table 3.1.5: Different inheritance patterns and respective possible mutated genes in family case 5 patient (obtained after filtering process).

Inheritance pattern	Patient genes with mutation
de novo	*FAAH2; LAMP2; SMARCA1; PASD1; HLA-DQA2; CFTR; KRTAP9-1; PLEKHN1; SH2D2A; RABL6; IL15RA; MUC5AC; ANO3; DHRS7;* ***SERTAD1****; PTH2; ZNF880;*
Compound heterozygous	/
Homozygous	/

In this family, we found a *de novo* mutation in *SERTAD1*. This mutation is a deletion of 3 base pairs located in exon 2 causing a nonframeshift deletion. This mutation is absent from 1000G_EUR and ESP6500siv2_ALL databases and is reported to have 0,04% frequency in the Exac_ALL database. The genotype quality score of this mutation is 72 and the allele depth is 18 and 3, for the reference allele and alternative allele, respectively. By using HGC, we revealed a biological relation of this gene with *XIAP*. This interaction might be important because *XIAP* gene is thought to play a role in lymphoproliferation and in the immune response to EBV [164]. The mRNA is annotated in LifeMap Discovery as being overexpressed in blood and in peripheral nervous system (peripheral nervous domain). The protein, according to HIPED has a very high expression in only one tissue, the bone marrow mesenchymal stem cells. *SERTAD1* is thought to regulate the cell cycle and to have an anti-apoptotic effect on tumor cells by reducing the degradation of the X-linked inhibitor of apoptosis protein (*XIAP*) [165,166]. As referred before, *XIAP* deficiency has been linked to a lymphoproliferative syndrome (XLP) that is immunodeficiency characterized by hypogammaglobulinaemia, lymphohystiocytosis and lymphomas that are normally developed by the response to infections with EBV. This association was made in a patient with XLP that had mutations in *XIAP* instead of mutations in *SAP*, which caused a deficient expression that ends in the apoptosis of lymphocytes [164]. Besides that, they suggest that *XIAP* is necessary for the differentiation and/or survival of natural killer T-lymphocyte cells. The mutation in *SERTAD1* could fit in the phenotype if this mutation is LOF, leading to more degradation of *XIAP*, what enhances the apoptosis rate. This mutation should be confirmed by SANGER sequence and after that, a qPCR should be done to analyze the level of expression of *XIAP*. If the expression levels are low it supports that the mutation in *SERTAD1* is a loss of function.

Family case 6

This Caucasian family has four individuals and only the father is healthy. The rest of the family, mother and both sons, have invasive pyogenic infections (**Figure 3.1.15**).

Based on the clinical information we focused on mutations in the siblings inherited by the mother or mutations present in both the mother and siblings (**Table 3.1.6**).

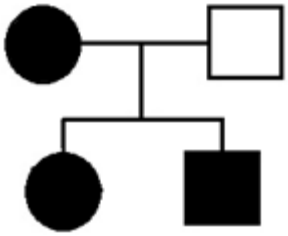


Figure 3.1.15: Pedigree of family 6, composed by two parents and two children. **Affected family members are shown in black.**

Table 3.1.6: Different inheritance patterns and respective possible mutated genes in family case 6 patient (obtained after filtering process).

Inheritance pattern	Gene that contains variants
Inherited by the mother	*KIF17; OLFML2B; SELP; PTPN7; KIF26B; IWS1; MAP3K19;* ***NFKBIZ****; PDIA5; PDGFRA; SH3D19; TARS; DPYSL3; ZNF391; ZBED9; CUL9; MAS1; ZNF680; WBSCR17; STX1A; MTERF1; STRIP2; TMEM209; KAT6A; ZNF704; PNPLA7; PRPF18; MUC5AC; STK33; CAT; LDLRAD3; OR8J1; TNKS1BP1; MARK2; CWF19L2; ARHGAP20; ACSM4; KRT18; MYL6;* ***TYRO3****; MRPS11; IFT140; ABCC12; MT1B; P2RX5; NCOR1; DDX52; CDH7; LILRA4;* ***DEFB132****; KIZ; SUN5;*
Same genotype as the mother	*KRTAP5-1; SKA3; ZFPM1; PIK3R6; SIGLEC12; FAM58A; H15*

The first candidate mutation was a deletion in *TYRO3* gene, absent from public databases. The sequencing quality of this call seems to be very good. The exome quality of all affected individuals (mother and patients) have a genotype quality of 99, allele depth around 98 and around 17 for the reference and alternative allele, respectively. The impact of the mutation was not available in the impact predictor tools (MSC). This gene, according to HGC, interact with a PID gene, *TBX1*. A study reported that the knock out of the TAM receptor (*TYRO3*, Ax1 and Mer) caused a profound dysregulation of the immune system and suggested that this protein functions as an inhibitor of TLR and TLR-induced cytokine-receptor cascade by activating Stat1, which selectively induces production of suppressors targeting cytokine signaling SOCS1 and SOCS3 [167,168]. Another study showed that by causing the genetic ablation of *TYRO3* in mice, the type-2 immunity was enhanced [169]. The mRNA of this gene is overexpressed in the brain but

expression in immune cells also occurs and the protein is overexpressed in the spinal cord, according to GTEx and HIPED, respectively. SANGER sequencing was performed demonstrating that this specific mutation was not present in the affected individuals (**Figure 3.1.16**).

Figure 3.1.16: Results of the SANGER sequence of *TYRO3* gene in mother, father, patient 1 and 2, aligned with a reference sequence using SnapGene. This Figure demonstrates that the mutation in exon 10 is not present in any of the individual. The mutation to be confirmed was a deletion of one guanine at the position 1382 of the transcript NM_006293.

A mutation in *NFKBIZ* gene also seemed to be a great variant fitting very well the phenotype of the affected individuals. This gene is thought to regulate the inflammatory responses to Streptococcus pneumoniae by upregulating *IL6* and *GMCSF* [170]. Kim et al. (2017) generated a *NFKBIZ* knock-out mice reporting the development of spontaneous dermatitis and suggested that this inflammation was rapid due to the expansion of Staphylococcus xylosus. Polymorphisms in *NFKBIZ* were associated with susceptibility to develop invasive pneumococcal disease and psoriasis [172,173]. The mutation in *NKBIZ* is an insertion of several base pairs that creates a splicing mutation. In all affected individuals, this mutation had a genotype quality call of 99 and a coverage depth at least

70 (Supplementary data: **Table 6.3**). In normal tissues, the mRNA expression is upregulated in blood, inner cell mass, ovary, pancreas and epithelium cells according to LifeMap discoveries database and in fallopian tubes according to GTEx (**Figure 3.1.17**). The mutation is localized in the Ankyrin repeated region that is responsible for the interaction with NFKB1/p50 (**Figure 3.1.18**). *NFKB1* has been reported to play a role in the regulation of human natural killer cells maturation and effector functions [174]. A similar homozygous mutation was found in the mother and both patients and a similar heterozygous mutation was also found in the father, as demonstrated in SANGER sequence (**Figure 3.1.19**). This type of variant inheritance looked questionable. It is very unlikely that a rare variant is present in every individuals of a non-consanguineous family which indicates that this variant is an artifact or that the given frequency of the variant is not correct. Although we confirmed the mutation, we observed that the annotation was not precise because the insertion occurs 4 nucleotides after the exon and not 2, as supported by **Figure 3.1.19**. Using the correct annotation of this mutation, the frequency of the mutation turned out to be very different. The new frequency of this mutation is 0.6974% in the Kaviar database. In 1000G and EXAC_ALL databases the frequency is 23% and 21%. We thus realize this mutation was a common variant and had an 18% presence in Exac database.

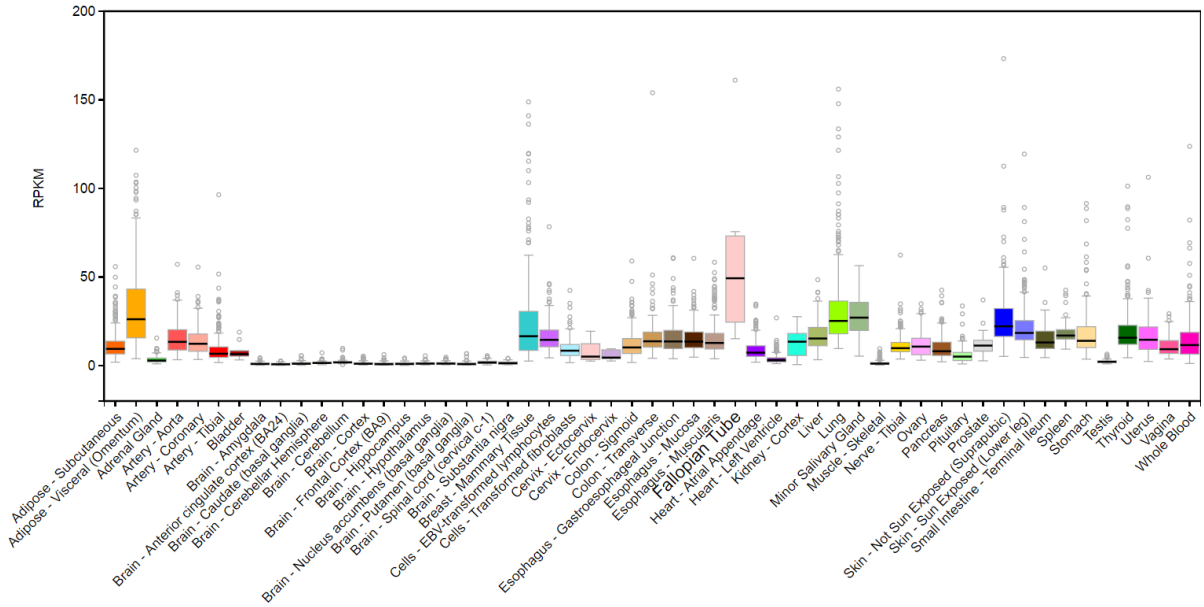
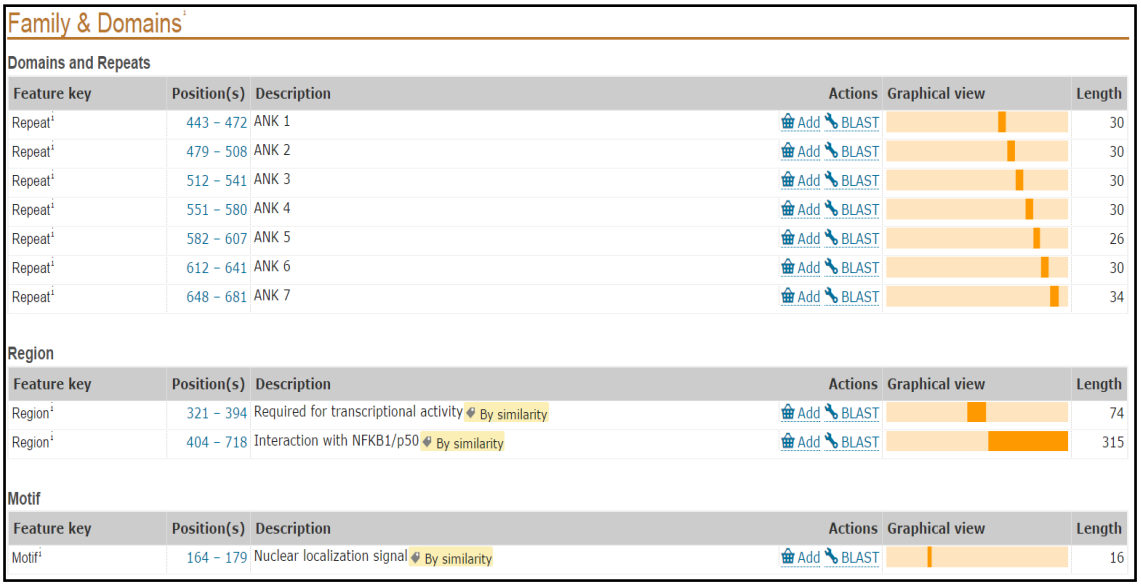


Figure 3.1.17: *NFKBIZ* gene expression in normal human tissues according to GTEx database.



Family & Domains

Domains and Repeats

Feature key	Position(s)	Description	Actions	Graphical view	Length
Repeat	443 – 472	ANK 1	Add BLAST		30
Repeat	479 – 508	ANK 2	Add BLAST		30
Repeat	512 – 541	ANK 3	Add BLAST		30
Repeat	551 – 580	ANK 4	Add BLAST		30
Repeat	582 – 607	ANK 5	Add BLAST		26
Repeat	612 – 641	ANK 6	Add BLAST		30
Repeat	648 – 681	ANK 7	Add BLAST		34

Region

Feature key	Position(s)	Description	Actions	Graphical view	Length
Region	321 – 394	Required for transcriptional activity By similarity	Add BLAST		74
Region	404 – 718	Interaction with NFKB1/p50 By similarity	Add BLAST		315

Motif

Feature key	Position(s)	Description	Actions	Graphical view	Length
Motif	164 – 179	Nuclear localization signal By similarity	Add BLAST		16

Figure 3.1.18: Family and domains regarding *NFKBIZ* gene. Image extracted from UniProt.

Figure 3.1.19: Results of the SANGER sequence of *NFKBIZ* gene in the mother, father, patient 1 and 2. SANGER sequence aligned with a reference sequence using SnapGene. This Figure confirms the presence of a homozygous insertion in the mother and patient 1 and 2 and the presence of a heterozygous insertion in the father, as predicted by the bioinformatic analysis. The mutation is an insertion of several base pairs 4 nucleotides after exon 11 of the transcript NM_001005474.

Another candidate mutation was found in *DEFB132* gene. This mutation is heterozygous and is present in the first amino acid of the sequence. The mutation is a substitution of thymine for adenine and causes an amino acid change from a methionine

to a lysine in the transcript NM_207469. The mutation is not present in the 1000G_EUR database but has low frequency in ExaC_ALL and ESP65000sv2_ALL (0,12% and 0,02%, respectively). This mutation has genotype quality of 99 and an allele depth of 9 and 7 for the reference and alternative allele, respectively. The CADD score for this mutation is 22.2 and all prediction tools like SIFT, Polyphen2 and the MSC included, also predicted to be damaging. Accessing GTEx, we found a differential mRNA expression of this gene. This gene is overexpressed in subcutaneous and visceral adipose tissue and less expression in mammary tissue and prostate. The beginning of the protein is responsible for signaling the peptide and, because our mutation is in the starting codon, the externalization/ secretion of the protein might not occur. Perhaps the protein is not transcribed or translated because the initiation codon is not present. *DEFB132* gene transcribes a protein of the family beta-defensins that are important in the immunologic response to invading microorganisms because they contain antimicrobial peptides. We could associate the mutation in this gene to a lower degree of immunity against pathogens, but even if the mutation caused a loss of function allowing the entrance of pyogenic bacteria, a "normal" protein would still be encoded because is a heterozygous mutation. It is very unlikely for a heterozygous mutation to cause a loss of function and such a severe phenotype as our patient has. Besides that, this gene is known to have a high rate of polymorphism, which supports that the mutation is not disease causing contradicting the impact predictor tools [175]. We have several genes transcribing Defensins and so it is unlikely that the affected individuals would have invasive pyogenic bacteria only because of a mutation in one of the defensins. The SANGER sequence was not done because of the very low confidence on this mutation to cause invasive pyogenic infections.

Family case 7

In this family case, the number of individuals involved was four (**Figure 3.1.20**). The parents are not affected, but their son is affected with hypogammaglobinemia and Burkitt lymphoma (tumor in B cells), streptococcal sepsis (bacterial infection) and severe varicella. Their daughter may be mildly affected with granuloma annulare (inflammatory dermatosis) and low levels of naive T cells.

In this Caucasian family we created a list of *de novo*, compound heterozygous and homozygous mutations for patient one and two. The variant genes obtained after whole

exome sequencing analysis and filtering are presented in **Table 3.1.7**. No good candidate genes were obtained that could fit the observed phenotypes.

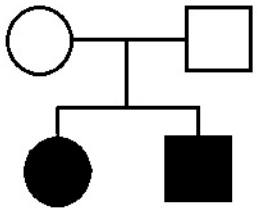


Figure 3.1.20: Family 7 pedigree composed by two parents and two children. **Affected family members are shown in black.**

Table 3.1.7: Different inheritance patterns and respective possible mutated genes in family case 7 patients (obtained after filtering process).

Inheritance pattern	Variant gene in patient 1	Variant genes in patient 2
de novo	*ZAN; ATP13A2; MTR; ATXN1; TECPR1; TFR2;*	*PGM2; AP3S1; CACNA1B; ZAN*
Compound Heterozygous	*WDR63; TRAK1; ASPN; ERCC6L2*	*TRAK1*
Homozygous	*SPRR3*	*ATG3*

Family case 8

In this caucasian family, the index patient has a WHIM like syndrome (congenital immune deficiency) with severe congenital neutropenia (low rate of neutrophils), warts, hyper-IgG, less natural killer cells than normal (warts are indicative of natural killer defect), pneumonia, VZV infection (Varicella zoster virus) and molluscae (viral infection that causes skin rash). After whole exome sequence analysis and filtering, variants were grouped into *de novo* mutations, compound heterozygous and homozygous mutations (**Table 3.1.8**).

Table 3.1.8: Different inheritance patterns and respective possible mutated genes in family case 8 patient (obtained after filtering process).

Inheritance pattern	Variant gene present in patient
de novo	*PLB1; CCDC140; LHFPL4;* ***SEC61A1****; CFTR; CACNA1B; NPM3; VPS37B; SKA3; ADNP2; CNN2;* ***ILF3****; SHANK3; EDA; RYR1; HLA-DQA2*
Compound Heterozygous	*IRX2; CFTR; HLA-DQA2*
Homozygous	*OR6C76*

ILF3 gene seemed to be a good candidate but the inheritance of the mutation was not clear because the sequencing quality of the father was very bad, the machine did not gave any score for the genotype quality. The nuclear export of *ILF3* is known to be important for IL2 mRNA stabilization [176]. Although, an association with the phenotype and this gene was not possible to obtain. The *ILF3* is a gene predicted to be a PID gene by Yuvan et al. (2015) and interacts with *ADAR*, a known PID gene. The mutation is private in 1000g_EUR, Exac_ALL and ESP6500siv2_ALL and has a frequency of 0, 0039% in Kaviar database. The genotype quality was 91 and the number of reads that support the presence of the mutation were 3 and 5 for the reference and alternate alleles in the patient. The probability of this mutation being pathogenic is considered high by the CADD score and MSC-CADD prediction tool unlike the SIFT and Polyphen2 prediction (**Figure 3.1.21**). The expression of the mRNA of this gene has a wide range, according to GTEx, Illumine, BioGPS and SAGE, as well as the protein expression rates, according to ProteomicDB, PaxDB, MaxDB and MOPED. The mutation in *ILF3* gene was confirmed by SANGER sequencing, although it seems to be an insertion and not a substitution and is also present in the father (**Figure 3.1.22**).

Chromome	Position	ID	Reference_Allele	Alternative_Allele	Gene	CADD_Score	MSC-CADD_Score	MSC-CADD_Impact_Pred	PolyPhen2_Score	PolyPhen2_Pred	MSC-PolyPhen2_Score	MSC-PolyPhen2_Impact_Pred	SIFT_Score	SIFT_Pred	MSC-SIFT_Score	MSC-SIFT_Impact_Pred
19	10798096	-	G	A	ILF3	23.200	3.313	high	0.001	benign	0.239	low	0.470	T	0.207	low

Figure 3.1.21: Impact Prediction of the mutation in *ILF3*, using the MSC server tool.

Figure 3.1.22: SANGER sequence of *ILF3* gene in exon 17 aligned with a reference sequence using SnapGene. The mutation was confirmed to be present in the patient but also in the father, although it looks like an insertion. The mutation is a substitution of a guanine by an adenine in position 2134 of the transcript NM_012218 that causes a change of glycine to serine.

Another good candidate gene was the *SEC61A1* gene. This gene presents a *de novo* mutation in exon 5 and is a substitution of an adenine for a guanine at position 275. The mutation is completely private and is predicted to be damaging by all impact predictors used (CADD, Polyphen2, SIFT and MSC). The sequencing call for this mutated allele has a genotype quality of 99 and the allele depth is 19 and 24 for the reference and alternative allele, respectively in the patient. According to LifeMap Discovery, the mRNA expression has a relatively big range such as in bone, lung, kidney, spleen, liver and others. The protein expression is also present in a wide range of organs and tissues but is highly overexpressed in nasal epithelium and bone according to HIPED, MOPED and others (**Figure 3.1.23**). *SEC61A1* gene is known to be crucial in the insertion of secretory and membrane polypeptides into the endoplasmic reticulum [177]. Further studies show that this gene can mediate the immunomodulatory effects of mycolactone and play a role in the regulation of immune cell functions [178]. Heterozygous mutations in *SEC61A1* gene have also been found in two families that presented ADTKD (Autosomal-dominant tubulo-interstitial kidney disease) and congenital anemia with intrauterine growth retardation or neutropenia [179]. These reported mutations are heterozygous loss of function, indicating that this gene can be sensible to any mutations. Our *de novo* mutation is also heterozygous and probably cause the loss of protein function. However, functional tests such as Western Blot must be conducted. Furthermore, Baron et al., (2016) suggest, after analyzing the results of a blockade of Sec61 by mycolactone, that this gene plays a role in the induction of innate and adaptive immune responses to intracellular pathogens, which might explain the viral infection that our patient presents. Despite this information, we could not find any evidence that this gene might influence the expression of natural killer cells that are low in our patient which also contributes to the low capability to protect against pathogens. A deeper analysis of this patient such as immunophenotype could give rise to a new mechanistic hypothesis for *SEC61A1* gene. The SANGER sequence to confirm the mutation should also be performed.

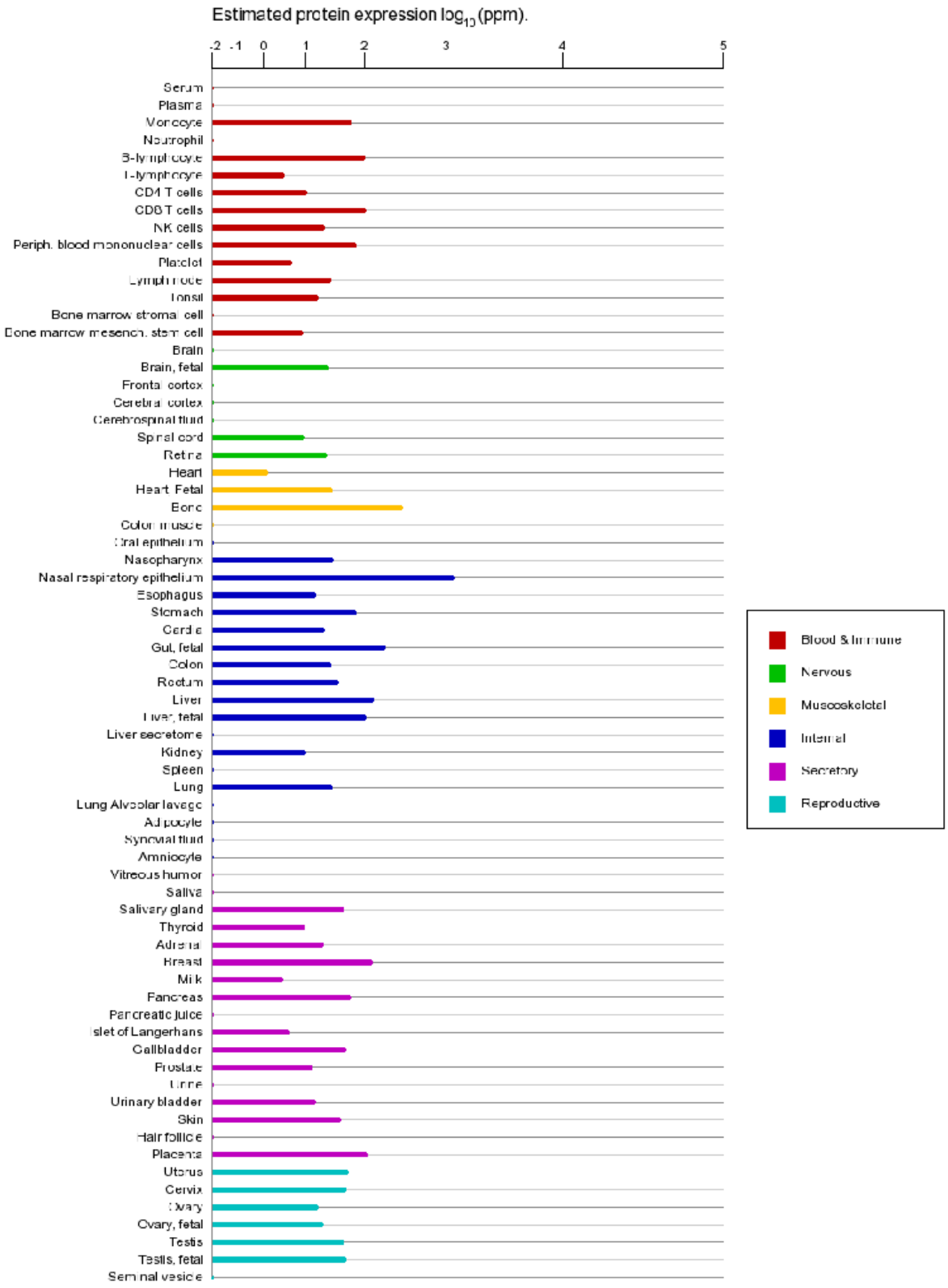


Figure 3.1.23: Protein expression of *SEC61A1* gene in normal cells, according to ProtemoixDb, PaxDb, MaxDb and MOPED.

Family Case 9

In this family trio only the son is affected and the phenotype is characterized by hyper immunoglobulin E and familial systemic sclerosis. In this Caucasian family, we could only find de novo mutations (**Table 3.1.9**). A gene that emerged as candidate after analyzing all variants was *SELPLG* that possesses a heterozygous mutation. This mutation is a nonframeshift deletion of several base pairs. The genotype quality was 99 and the allele depth was 10 for the reference allele and 6 for the alternate. This gene is thought to regulate the expression of the cutaneous lymphocyte-associated antigen in T cells and be involved in the pathogenesis of some infectious agents and in neutrophil migration [180–182]. *SELPLG* gene has been associated with several autoimmune diseases, such as systemic lupus erythematosus. Additionally, *SELPLG* knock out mice have shown to develop an autoimmune syndrome with similar characteristics to multiple sclerosis [183–185]. The mRNA of *SELPLG* is overexpressed in all blood according to GTEx. The protein is overexpressed in peripheral blood mononuclear cells, monocytes, lymph node, cerebral cortex and CD8 T cells (**Figure 3.1.24**). The mutation was checked by SANGER sequence and the result demonstrated that the mutation is an artifact because none of the individuals presented the mutation, although we can see that the father and patient have some mutations in that region (**Figure 3.1.25**).

Table 3.1.9: Different inheritance patterns and respective possible mutated genes in family case 9 patient (obtained after filtering process).

Inheritance Pattern	Gene with variants present in Patient
de novo	*PSORS1C2; HLA-DQA2; MUC5AC; MTCH2; TAS2R46;* ***SELPLG****; SCNN1D; DHRS3; SFXN5; PRKRA; WWC2; MAST4; N4BP3; PM20D2; FDFT1; KRTAP5-1; RAD51AP1; ITGA7; VPS37B; CNOT1; CSNK1D; TGIF1; NEURL2; SLC7A4;*
Compound Heterozygous	/
Homozygous	/

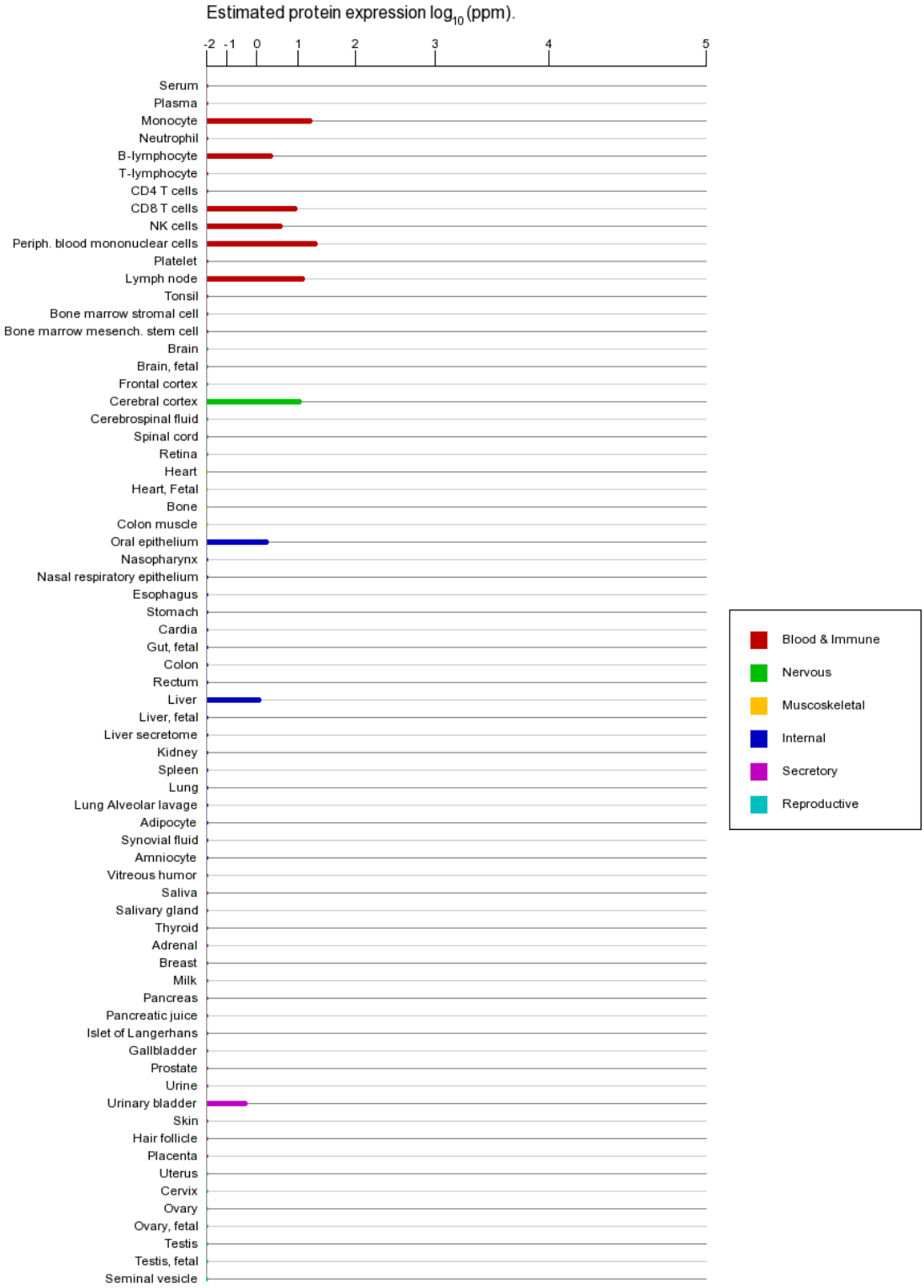


Figure 3.1.24: protein expression in normal tissues and cell lines according to ProteomicsDB, PaxDb, and MOPED for *SELPLG* Gene.

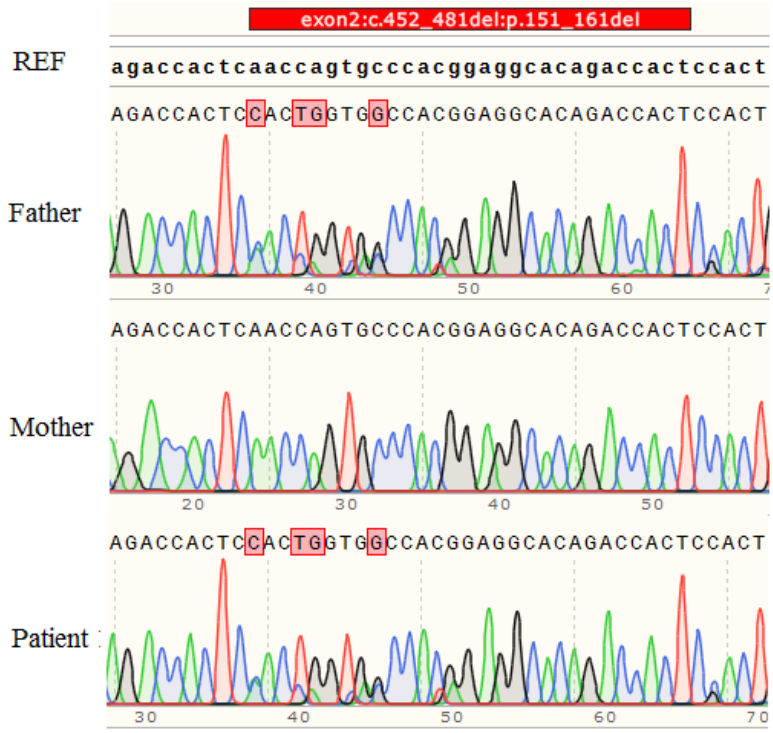


Figure 3.1.25: Aligned SANGER sequence of a region of *SELPLG* gene with a reference sequence. The deletion in this gene has not confirmed. Although, this gene had several mutations present in the father, mother and patient what give us the idea that is a highly mutated gene. This is also a gene that usually shows up as mutated in several other family cases.

Family Case 10

This Caucasian family also followed a trio design and only the son was affected by Combined Immune Deficiency (CID). Filtering the variants from the annotated VCF file, we found 12 *de novo* and four compound heterozygous and two homozygous mutations (**Table 3.1.10**). It was not possible to make an association between the genes and the phenotype and so, no candidate mutations were selected.

Table 3.1.10: Different inheritance patterns and respective possible mutated genes in family case 10 patient (obtained after filtering process).

Inheritance pattern	Patient
de novo	*TTN; NYAP2; MATR3; HMCN2; NEURL1; CHST15; OR51G2; EIF3J; TMED1; ZNF83; ARSD*
Compound Heterozygous	*TULP4; HMCN2; MVP*
Homozygous	*NDUFAF6; KRT4; NEK3; PIBF1*

Family Case 11

In this family trio, the patient has Graves' disease and several skin manifestations such as vitiligo (losing skin pigments), verrucosis/papillomatosis (appearance of wounds) and keloid formation (excess of collagen in wound repair process). The patient also seems to suffer from celiac disease. This disease is an autoimmune disorder where the person affected cannot eat gluten, or risk damage to the small intestine and consequently the absorption of nutrients.

The patient had 16 *de novo* mutations and one compound heterozygous mutation (**Table 3.1.11**). Some of the genes, such as *CNTN5*, *CLEC4M* and *NES*, are related to the immune system, but none of them could be directly linked to the aforementioned phenotype characteristics and so, no further experiments were done.

Table 3.1.11: Different inheritance patterns and respective possible mutated genes in family case 10 patient (obtained after filtering process).

Inheritance pattern	Patient
de novo	***NES**; PDIA6; IWS1; ARMC2; HGC6.3; IQCA1L; MUC2; **CNTN5**; NFATC4; LYSMD4; LHX1; DAZAP1; **CLEC4M**; LILRB5; KCNK15; XIRP1*
Compound Heterozygous	*OR51D1*
Homozygous	/

Family Case 12

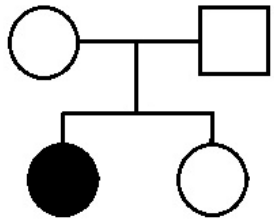


Figure 3.1.26: Pedigree of a family 12 composed by two parents and two children. Affected family member is shown in black.

Whole exome sequencing was done in all individuals of this family. Only one of the siblings is affected with autoimmune neutropenia (low neutrophils) and eczema (inflamed or irritated skin) (**Figure 3.1.26**).

Table 3.1.12 shows the mutated genes present in the patient but not in the sister. We could identify two candidate genes. The first is *TBK1* gene (NM_013254) that has a *de novo* deletion of two nucleotides positioned 1 nucleotide before exon 19 (5´ to 3´), which can cause an alternative splicing. This mutation is completely private in 1000G_ALL, ExAC_ALL, ESP6500_ALL databases. The patient allele depth of reference is 19 and for the deletion is 3, while the genotype quality is 81.

Table 3.1.12: Different inheritance patterns and respective possible mutated genes in family case 11 patient (obtained after filtering process).

Inheritance pattern	Patient
de novo	*PLEKHG5; GRIK3; SYCP1;* ***NMI****; TBR1; MYL1; ZBTB11; PDE6B; FER; AP3S1; ATAT1; OSTM1; NACAD; C10orf95; MUC5AC; KRTAP5-1; SLC22A18AS; C11orf80; OR6C76;* ***TBK1****; SNW1; C15orf40; POLG; GP1BA; CCDC124; ZNF14; ZNF626; ESX1;*
Compound Heterozygous	*SCN2A; ATN1; PRDM15*
Homozygous	*NEB; EBLN2; ASPN; ABO; CHST15; UNKL; MROH8;*

This gene activation, *in vitro*, was previously shown to lead to a type I interferon receptor activation that is necessary for an induction of antigen-specific B and T cells, even without the activation of some innate immune signaling such as TLR9 [186]. *TBK1* is also a regulator of dendritic cells immune response and when mutated can cause autoimmune disorders like autoimmunity [187]. The mRNA expression is present in some cells of the immune system, such as bone marrow, whole blood and lymph node according to BioGPS and, the protein, is overexpressed in peripheral blood mononuclear cells, lymph node and lung, according to HIPED and ProteomicsBD. SANGER sequence to check *TBK1* mutation was performed, demonstrating that any of the individuals had the deletion (**Figure 3.1.27**).

Figure 3.1.27: SANGER sequence of the region of the mutation in *TBK1* gene. The Mutation is not present in family individuals.

The second candidate gene is *NMI*. The *de novo* mutation in *NMI* gene is private and also present only in the patient. The mutation is located in exon5 of the transcript NM_004688 and is a substitution of a thymine for a guanine, changing the amino acid from a histidine to a glutamine (exon5:c.T372G:p.H124Q). In the patient, the genotype quality is 69 and allele depth 42 and 11 for reference and alternative alleles, respectively. According to GTEx, the mRNA expression in normal tissues is overexpressed in whole blood while the protein is overexpressed in peripheral blood mononuclear cells, lymph node and monocytes according to HIPED. An overview of protein expression is in **Figure 3.1.28**. This gene is thought to interact with all STATs, except *STAT2*, by responding to cytokine changes causing the expression of, for example, *STAT3* which encodes transcription factor that will modulate gene expression [188]. The STAT family is known to be involved in several cases of autoimmunity and inflammation, such as skin inflammation [189,190]. The relation among this gene and the phenotype is supported by the functional interaction between *NMI* and several proteins from the STAT family. Given

that *NMI* can enhance the STAT-dependent transcription and cause the augment of IL-2 and IFN γ-dependent transcription we can hypothesize that our mutation is a gain of function that will cause overexpression of STATs and subsequently IL-2 and IFN γ that might lead to a general inflammation and cause eczema [188,189]. To test this hypothesis we could check if there is an overexpression of STATs by quantitative PCR and also perform immunophenotyping to evaluate which immunologic cells are being overexpressed. Although we could hypothesize about the manner that skin inflammation shows up, we did not find any disease causing mutation that could lead to neutropenia. This mutation stills needs to be confirmed by SANGER sequence.

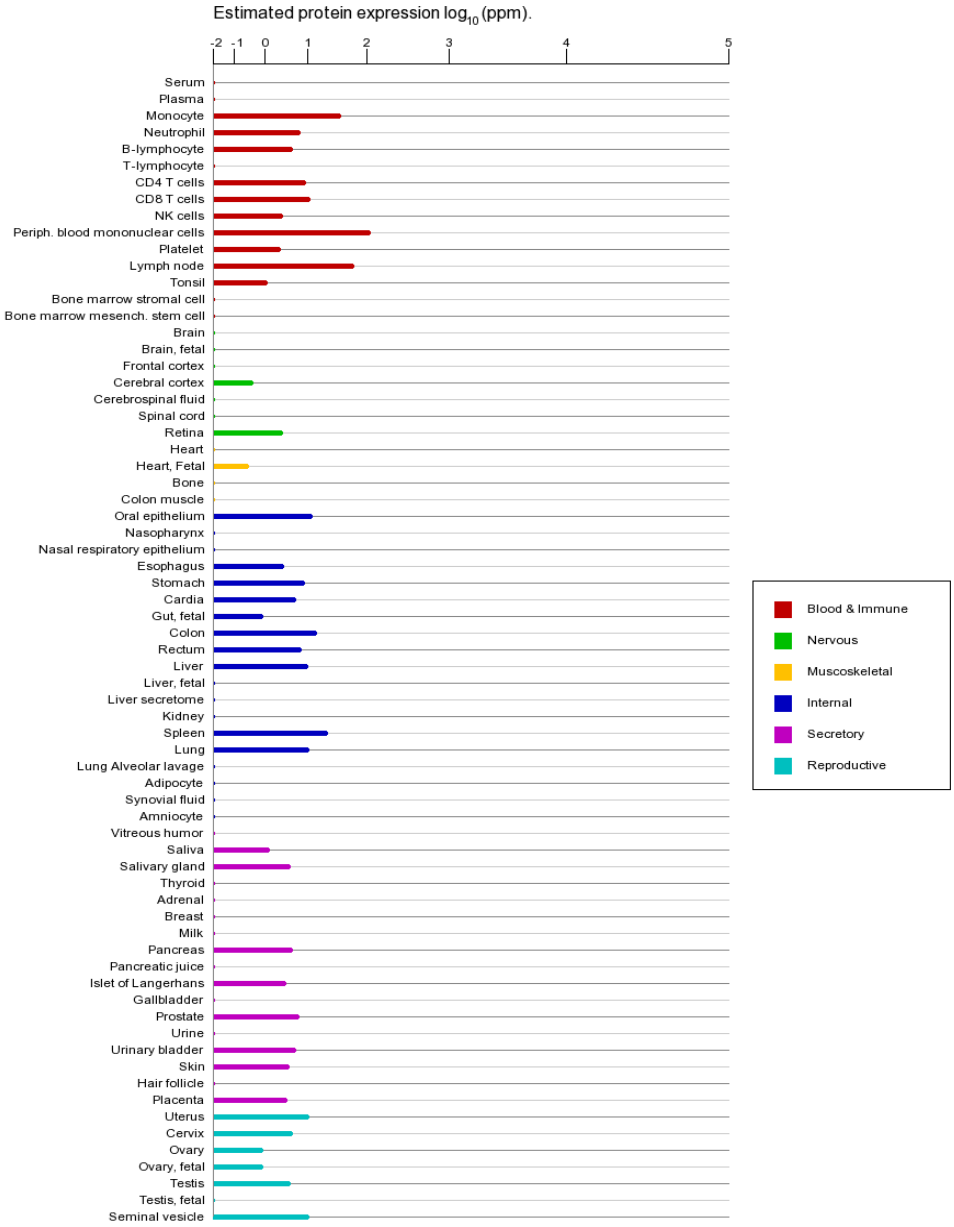


Figure 3.1.28: Image extracted from GeneCards that express the protein expression of *NMI* gene, according to ProtemoixDb, PaxDb, MaxDb and MOPED.

Family Case 13

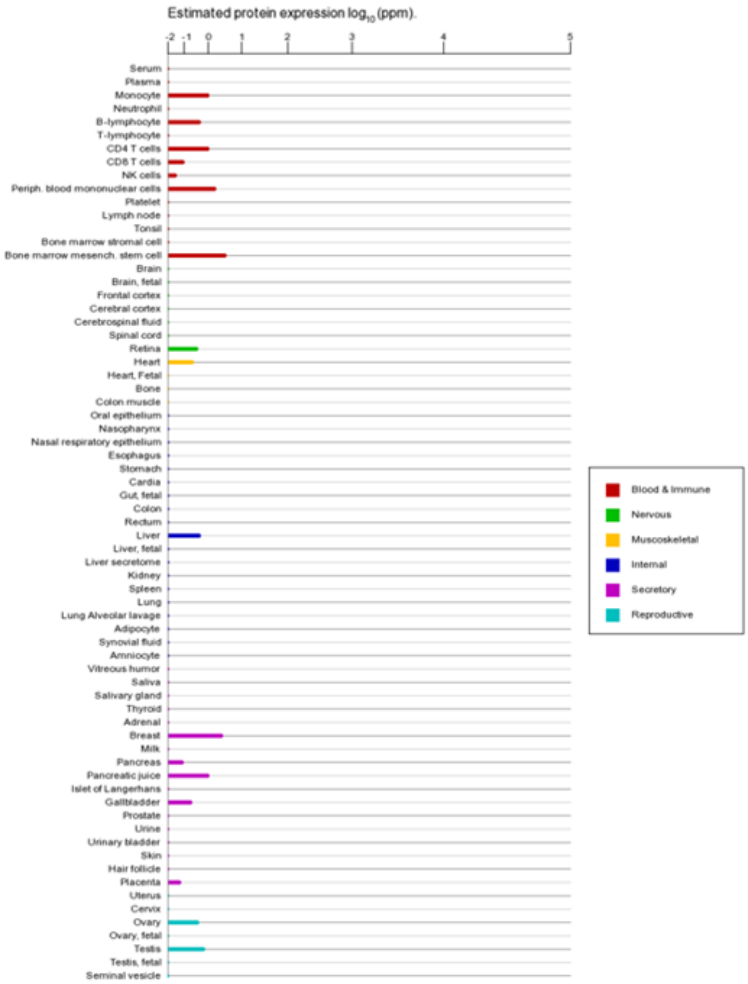
This caucasian family is another case of a trio. The patient is affected with recurrent invasive pyogenic infections. In this family we found 24 de novo, three compound heterozygous and four homozygous mutations (**Table 3.1.13**). A very strong candidate gene is *LYST* gene, which presents a compound heterozygous mutation.

Table 3.1.13: Different inheritance patterns and respective possible mutated genes in family case 13 patient (obtained after filtering process).

Inheritance pattern	**Patient**
de novo	*FGFRL1; IL17D; PCP2; PNMAL2; ZNF880; CCDC22; AQPEP; AQPEP; SOX18; CTRC; PRKRA; GTPBP2; PRIM2; TPM2; CDK20; SWAP70; MTCH2; DRD2; CCNT1; TYRO3; SLC27A2; PATZ1; MKL1;*
Compound Heterozygous	***LYST****; ASPN; DGKH*
Homozygous	*TRAK1; SELPLG; IL17D; SKA3*

The mutations in *LYST* gene are in exon 6 (C3083G:p.S1028C) and 7 (A3544G:p.M1182V). The first mutation is inherited from the mother and the second from the father. The mutation in exon 7 (rs749553632) is private in 1000G_EUR, Exac_ALL and ESP65000siv2_ALL databases. In the patient, the genotype quality is 99 and, allele depth is 24 for reference allele and 15 for alternative allele. The mutation in exon 6 (rs150636017) is not present in the 1000g_EUR database and is present in 0.05% of Exac_ALL and ESP65000siv2_ALL database. The genotype quality is the same as the first mutation but the allele depth was better: 44 for reference and 45 for alternative allele. The *LYST* gene is a known PID gene and the mRNA expression in normal tissues is overexpressed in whole blood, in agreement with GTEx (**Figure 3.1.29a**). This gene encodes a protein that is overexpressed in bone marrow mesenchymal stem cell, breast and peripheral blood mononuclear cells according to HIPED database (**Figure 3.1.29a**). The MSC prediction tool gives a high probability of the mutation in exon 7 to be not damaging, but the mutation in exon 6 is predicted to be highly damaging by CADD, MSC-CADD, SIFT and MSC-SIFT (**Figure 3.1.29b**). The protein encoded by *LYST* gene has WD repeated regions (Trp-Asp repeat) and 2 regulatory domains.

a)



b)

Chromome	1	1
Position	235967815	235969353
ID	-	-
Reference_Allele	T	G
Alternative_Allele	C	C
Gene	LYST	LYST
CADD_Score	0.001	6.739
MSC-CADD_Score	0.456	0.456
MSC-CADD_Impact_Pred	low	high
PolyPhen2_Score	0.000	0.001
PolyPhen2_Pred	benign	benign
MSC-PolyPhen2_Score	0.435	0.435
MSC-PolyPhen2_Impact_Pred	low	low
SIFT_Score	0.910	0.010
SIFT_Pred	T	D
MSC-SIFT_Score	0.080	0.080
MSC-SIFT_Impact_Pred	low	high

Figure 3.1.29: Impact prediction of compound heterozygous mutation and protein expression of *LYST*. a) Image extracted from GeneCards regarding the protein expression in normal tissues and cell line of *LYST* proteomicDB, PaxDB, MaxQB and MOPED databases. b) Data extracted from MSC impact prediction of the compound heterozygous mutations.

The *LYST* gene has been associated with the autosomic recessive disorder that leads to a decrease in the phagocytosis, resulting in recurrent pyogenic infections, albinism and peripheral neuropathy [191–193]. The autosomal recessive inheritance model of the mutation agrees with the Chediak-higashi syndrome inheritance pattern, a syndrome that has been reported to be caused by mutations in *LYST* gene [191,192]. This syndrome is extremely complex and the phenotype goes from severe immunologic defects, recurrent infections, neurologic dysfunction to oculocutaneous albinism and prolonged bleeding [194]. Besides that, the knockout of this gene in mice showed a reduced bacterial protection and dysregulated phagosomal maturation through a regulation of TLR4 and TLR3-induced endosomal TRIF (TIR domain-containing adapter-inducing interferon β) signaling pathways [195]. Mutations in this gene seem to affect several tissues and systems

but the exact pathways and function of the sequence domains have not yet be unraveled, as we can see in the family and domains of the protein, according to UniProt. We suggest that this loss of function mutation affect a specific protein domain responsible for regulation of TLR, affecting the immune receptor signaling pathways and inflammatory response, which leads to an increased susceptibility to bacterial infections. The further experiments could test the involvement of *LYST* with TLR signaling and test if the mutated protein has the same involvement as the wild type protein. This can be done performing a Western Blot.

The SANGER sequence of LYTS gene confirmed that the patient is a compound heterozygote (**Figure 3.1.30**). To have more confidence that these variants are disease causing we could test other patients with suspected *LYST* deficiency for this two mutations since we did not found in the literature the mutations observed.

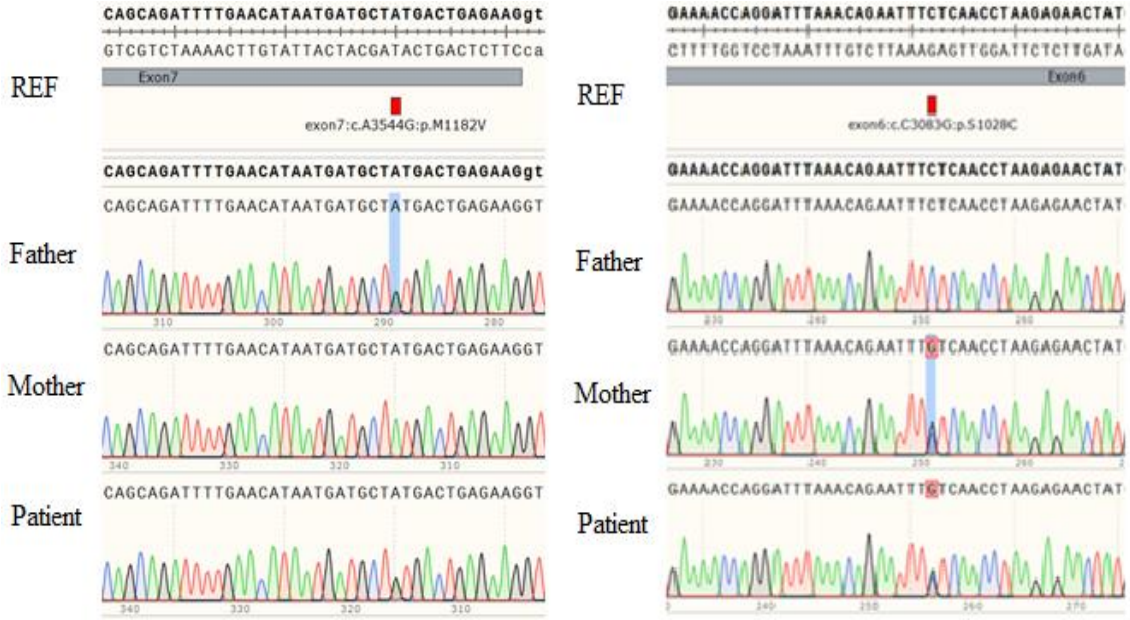


Figure 3.1.30: SANGER sequence of gene *LYST* in two different exon regions. Confirmation of the presence of a compound heterozygous mutation in the patient.

Family Case 14

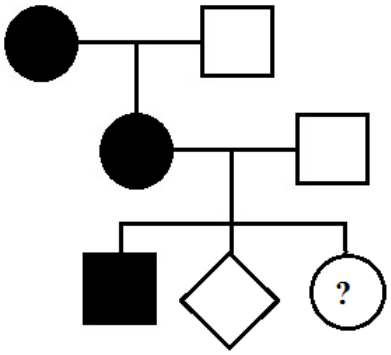


Figure 3.1.31: Pedigree of family 14, composed by two grandparents and parents and three children. Affected family members are shown in black. We are not sure that the sister is affected (?).

In this family, the mother is affected with severe allergies, such as hay fever (rhinitis) and food allergy and the grandmother has psoriasis. The family constitutes five individuals and one son is affected with atopic dermatitis from birth, severe food allergy, Immunoglobulin E > 5000 kU/L, neutropenia and auto-immune hepatitis (**Figure 3.1.31**). The way the disease is transmitted is not clear because the symptoms of the mother and grandmother are very different to those described in the son and the sister had an episode of hyper IgE. The mother might only be expressing a severe allergy or be a case of clinical heterogeneity. However, after whole exome sequencing analysis and filtering, we subdivided the variants in *de novo*, inherited by the mother and X-linked, all absent in the siblings (**Table 3.1.14**). We could find some candidate genes, such as *SH2D3C*, *SP1* and *UNC5CL*, that presented mutations inherited from the mother and a compound heterozygous mutation in *TMED3*.

Table 3.1.14: Different inheritance patterns and respective possible mutated genes in family case 13 patient (obtained after filtering process).

Inherited Patterns	Patient
de novo	*CFAP57; ATF6; ESRRG; PLCD4; RNF180; ZFYVE16; ANKRD34B; SSBP2; RGMB; MUC21; LOC100130705; PPP2R1B; MYO9A; NFE2L1; ABI3; NETO1; RASAL3; LILRB4; MECP2;*
Inherited by mother	*IGSF3; DNAH6; DNAH6; ACOX3; SRP72; CCDC109B;* ***UNC5CL****; ASZ1; AGBL3;* ***SH2D3C****; HMCN2; APBB1; C11orf24; KRT7;* ***SP1****; ALDH1L2; ZNF598; FHOD1; KCTD19; DHX38; NCOR1; TMEM106A; ADGRE2; ZNF486; ADGRG2;*
X- linked mutation	*ADGRG2; ATP2B3; MECP2*
Compound Heterozygous	*FOXD4L1; FOXD4; ANKRD20A1; ADAM21;* ***TMED3****;*

The *SH2D3C* mutation is present in exon 7, of the transcript NM_170600, at allele position 1529 and is a substitution of a cytosine by a thymine. The amino acid change from an alanine to a valine. The mutation is not present in 1000g_EUR and ESP6500 databases but has a frequency of 0,00329% in Exac database. In the patient the genotype quality is 99 and allele depth 19 and 21 for reference and alternative allele, respectively (Supplementary data: **Table 6.3**). According to GTEx, the mRNA of this gene is overexpressed in whole blood and, the protein expression, according to HIPED, is overexpressed in the lymph node, peripheral blood mononuclear cells, CD8 T cells and plasma. This gene plays a role in the maturation and function of marginal zone B cells and is thought to also be involved in T cell activation, what might lead to dysregulation of the levels of immunoglobulin E [196–199]. The mutation was confirmed to be present in the mother and patient by SANGER sequencing (**Figure 3.1.32**). Although it is known that this gene is involved in B cell maturation and function and in T cell activation, we did not find any evidence that mutations in *SH2D3C* could cause the phenotype observed in the patients. The location of the high expression of *SH2D3C*, such as in T cells lymph node and plasma support that this gene has a role in immune responses. It is possible that our mutation is gain of function, modulating B cell function, increasing T cell activation or increasing levels of IgE in plasma cells, which can lead to the hyper IgE observed in the patient phenotype [196,199,200]. This association might be a little far-fetched and not specific but the next candidate genes are better disease causing candidates.

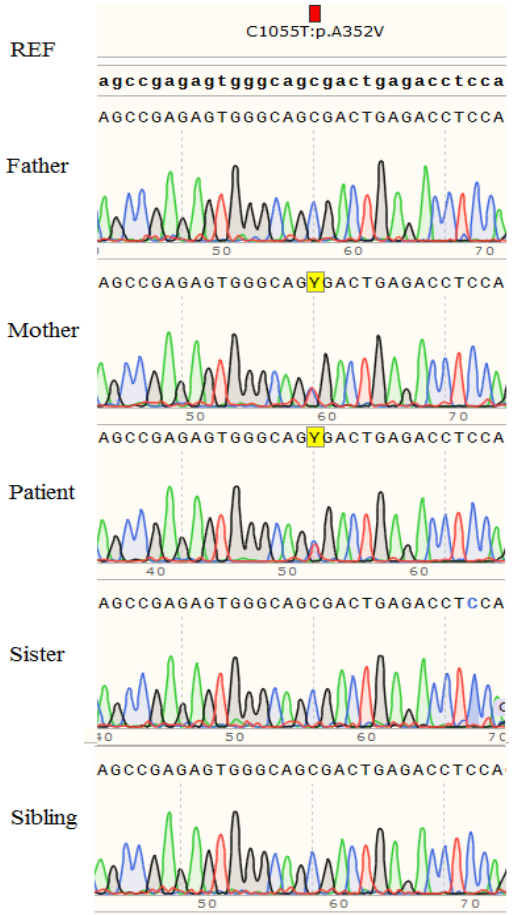


Figure 3.1.32: SANGER sequence of a region of *SH2D3C* gene. Heterozygous mutation present in the Mother and the patient.

The *UNC5CL* mutation is present in exon 4 and is a substitution of a cytosine to a thymine at position 796 that consequently encodes a cysteine instead of an arginine, in NM_173561 transcript. This mutation is also not present in 1000g_EUR and ESP6500 databases. In ExAC database is present in 0,0011% of the individuals. The genotype quality of this call in the patient is 99 and the allele depth 13 and 8 for reference and alternative allele, respectively (Supplementary data: **Table 6.3**). The mRNA of this gene is overexpressed in the cortex kidney, liver and pancreas according to GTEx. The protein differential expression in normal tissues is raised in plasma, Islet of Langerhans and kidney according to HIPED. The protein has two regulatory domains and a region of interaction with RELA and NFKB1, according to UniProt. The mutation we found is positioned in the region of interaction with the NFKB1 protein. This gene translates a protein that works at the cytoplasmic level and associates itself with subunits of nuclear factor-kappa-B (NFKB), causing the inhibition of its activation and binding to target sequences [201]. The NFKB pathway affects a wide range of processes. It is known to play

a role in immune and inflammatory processes and also differentiation of monocytes and other cell types, brain function, cancer, etc. One of the NFKB subunits, the NFKB1, is crucial to the class switch to IgE [202]. The mutation was confirmed by SANGER sequencing and is present in the mother, maternal grandmother and patient (**Figure 3.1.33**). The mutation in *UNC5CL* gene might be changing the way this protein interacts with NFKB1, since is present in the domain responsible for this interaction. *UNC5CL* cytoplasmic protein inhibits the NFKB-dependent transcription by both TNF and IL-1 and their downstream signaling protein due to interaction with NFkappaB subunit p105 and p65, which prevents target sequence binding [201]. According to Geldmeyer-Hilt et al. (2011), activation of vitamin D receptor reduced the expression of p105/p50 protein and mRNA in peripheral B cells, which is mediated by the impairment of nuclear translocation of p65, and caused the inhibition of IgE production. *UNC5CL* gene also prevents binding of NFkappaB subunits p105 and p65 to the target sequences but the heterozygous mutation we found might not let this binding occur reducing the levels of IgE inhibition. This supposed loss of function mutation, not common in heterozygous mutations, must be tested by Western Blot to check if the impact of the mutation is enough to prevent binding of the *UNC5CL* protein to NFkappaB subunits p105 and p65. We also found a mechanism involving the dead domain of *UNC5CL* and the activation of IL-1R-associated kinases (IRAKs) independently of MyD88, that ends in the activation of the transcription factor NFKB and c-Jun N-terminal kinase. This study suggests that *UNC5CL* plays a role in epithelial inflammation and immunity and is involved in mucosal diseases [203]. The mutation in *UNC5CL* is not present in the death domain but the loss of function of the region that binds to NFKB can cause less interaction of *UNC5CL* with p65 and p105 and subsequently increase the dead domain availability, increasing its activation and leading to the skin inflammation observed in the patient (atopic dermatitis). We could not find an association of this gene with autoimmune hepatitis but the liver presents a high expression of this gene and so, some not known *UNC5CL* mechanism might be causing the autoimmune phenotype.

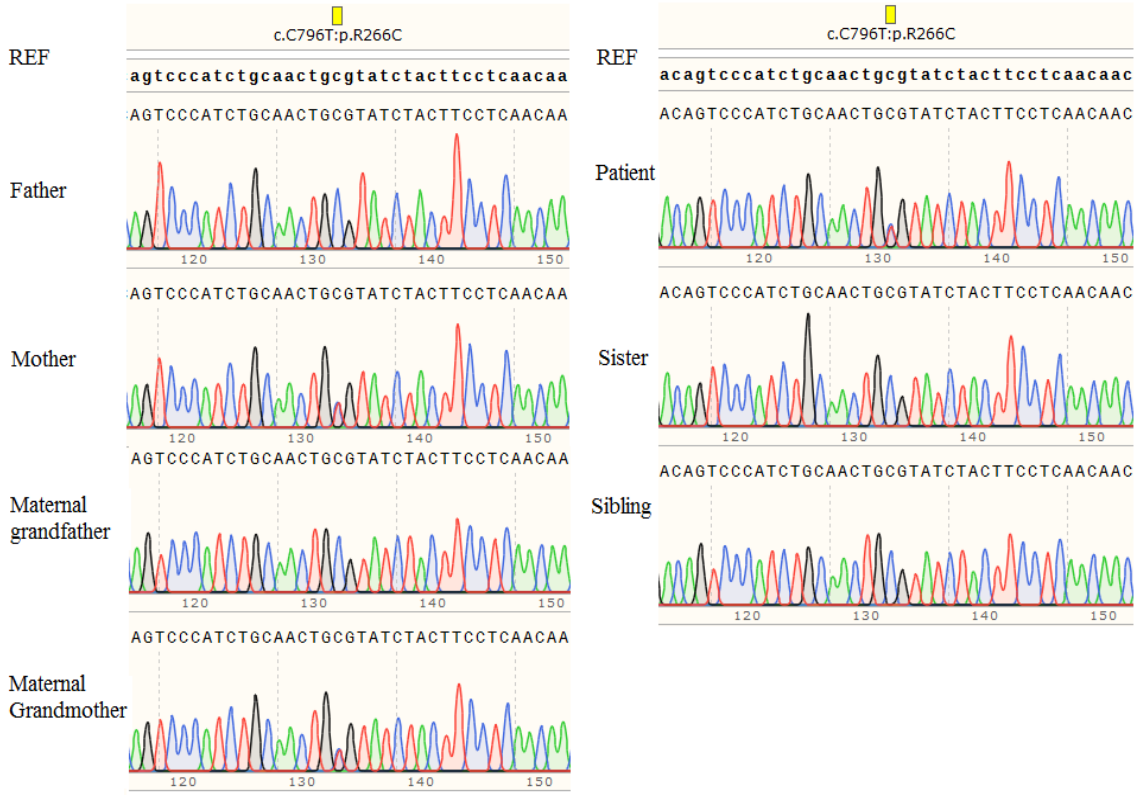


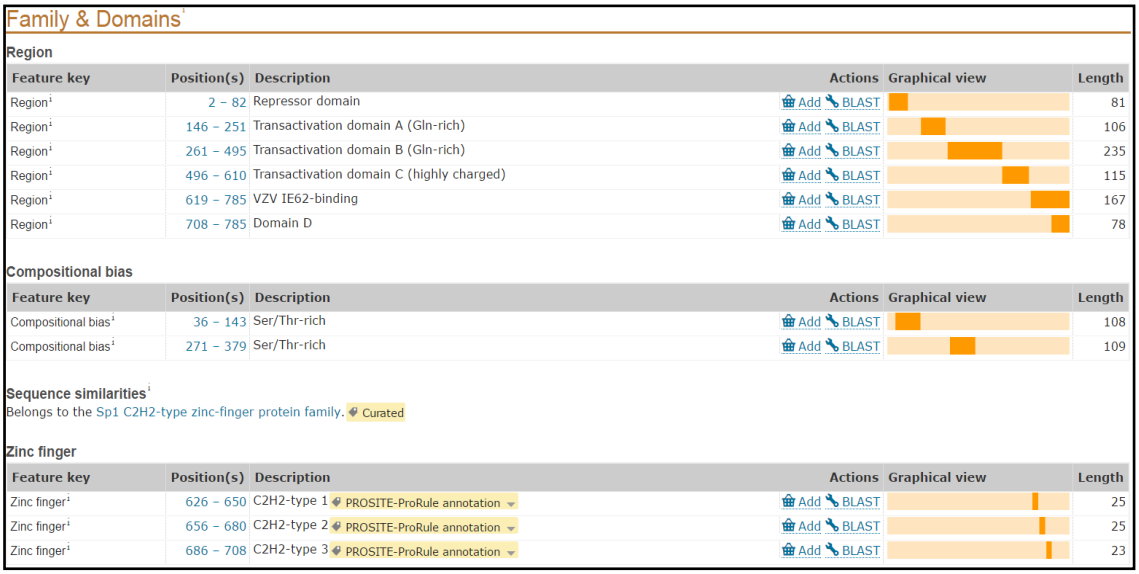
Figure 3.1.33: SANGER sequence of the region of the mutation in gene *UNC5CL*. We also had access to the DNA of the maternal grandmother and maternal grandfather. Confirmation of a mutation present in the mother, the maternal grandmother and the patient.

The mutation in *SP1* is present in the beginning of the gene and is a substitution of a guanine to an adenine in position 5, changing the second amino acid of the transcript NM_138473 from a serine to asparagine. This mutation is completely private in 1000g_EUR, ESP6500 and ExAC database. The genotype quality in the patient is 99 while the allele depth is 17 for reference and 16 for the alternative allele (Supplementary data: **Table 6.3**). The mRNA transcribed from this gene in normal tissues is differentially overexpressed in whole blood, as provided by GTEx. The protein expression, according to HIPED database, is overexpressed in CD8 T cells, lung and fetal gut but also shows high expression in other cells of the immune and blood system (**Figure 3.1.34**). All aforementioned variants are predicted to be damaging and the *SP1* variant is no different (**Figure 3.1.35**). The *SP1* protein has very different functions with several domains and regions (**Figure 3.1.36**). Our mutation is present in the repressor domain that has NFKB1 as a target. According to UniProt, the second amino acid, Serine, can be a site for post-transcriptional phosphorylation and/ or acetylation that is involved in the cleavage of the initiator methionine [204,205].

Figure 3.1.34: Protein expression in normal tissues and cell lines of *SP1* by proteomicDB, PaxDB, MaxQB and MOPED databases.

Chromome	Position	ID	Reference _Allele	Alternative _Allele	Gene	CADD Score	MSC-CADD_ Score	MSC-CADD Impact _Pred	PolyPhen2 _Score	PolyPhen2_Pred	MSC-PolyPhen2 _Score	MSC-PolyPhen2 Impact Pred	SIFT Score	SIFT_Pred	MSC-SIFT Score	MSC-SIFT Impact Pred
9	130507114	-	G	A	SH2D3C	12.370	3.313	high	0.003	benign	0.239	low	0.120	T	0.156	high
12	53774080	-	G	A	SP1	23.000	3.313	high	0.213	benign	0.450	low	0.010	D	0.271	high
6	41000776	-	G	A	UNC5CL	34.000	3.313	high	0.635	possibly damaging	0.239	high	0.000	D	0.243	high

Figure 3.1.35: Prediction of the damage caused by the mutations using the MSC server.



Family & Domains

Region

Feature key	Position(s)	Description	Actions	Graphical view	Length
Region	2 – 82	Repressor domain	Add BLAST		81
Region	146 – 251	Transactivation domain A (Gln-rich)	Add BLAST		106
Region	261 – 495	Transactivation domain B (Gln-rich)	Add BLAST		235
Region	496 – 610	Transactivation domain C (highly charged)	Add BLAST		115
Region	619 – 785	VZV IE62-binding	Add BLAST		167
Region	708 – 785	Domain D	Add BLAST		78

Compositional bias

Feature key	Position(s)	Description	Actions	Graphical view	Length
Compositional bias	36 – 143	Ser/Thr-rich	Add BLAST		108
Compositional bias	271 – 379	Ser/Thr-rich	Add BLAST		109

Sequence similarities

Belongs to the Sp1 C2H2-type zinc-finger protein family. Curated

Zinc finger

Feature key	Position(s)	Description	Actions	Graphical view	Length
Zinc finger	626 – 650	C2H2-type 1 PROSITE-ProRule annotation	Add BLAST		25
Zinc finger	656 – 680	C2H2-type 2 PROSITE-ProRule annotation	Add BLAST		25
Zinc finger	686 – 708	C2H2-type 3 PROSITE-ProRule annotation	Add BLAST		23

Figure 3.1.36: Extracted image from UniProt link regards the domains regions and their functions of the protein encoded by *SP1* gene.

SP1 is thought to regulate the expression of several genes involved in cell growth, apoptosis and immune responses [206–208]. This gene has also been indirectly associated with atopic dermatitis and hyper IgE by promoting the production of IL31 in association with EPAS1, in CD4 T cells with Dock8 knockout mice [209]. In humans, *DOCK8* deleterious mutations have been linked to atopic dermatitis and hyper IgE [210,211]. *DOCK8* is a negative regulator of EPAS1 nuclear translocation and, this last one, is thought to induce the production of IL-31 in $CD4^+$ T cells. Atopic dermatitis normally arises due to a high level of IL-31 and the transcription factor EPAS1 induces this production in association with *SP1* [209]. Furthermore, with knockdown of *SP1*, the EPAS1-mediated IL31 promoter activation was significantly reduced. We suggest that the heterozygous mutation in *SP1* causes a gain of function that leads to the upregulation of EPAS which subsequently increases IL31 production developing atopic dermatitis and producing high IgE levels. This gain of function also makes sense when analyzing a different mechanism where *SP1* is involved. Kim et al. (2012) showed that HDAC3 (histone deacetylase 3) mediates allergic skin inflammation by induction of monocyte chemoattractant protein 1. This induction occurs when *SP1* and c-Jun are associated with HDAC3 and so, if *SP1* had a gain of function mutation, the induction of monocyte chemoattractant protein 1 would increase, leading to skin inflammation.

The mutation in *SP1* was confirmed in the mother, grandmother and patient by SANGER sequencing (**Figure 3.1.37**). It is important to perform an extensive immune phenotyping so we can, for example, see if IL31 levels are high, supporting that the association of *SP1* and EPAS1 is increased. Besides that, RNA-seq could be done to evaluate which genes are upregulated and if they are *SP1* molecular targets. We should also synthetize cDNA and perform the SANGER sequence of this region because the mutation is very close to the beginning of the transcript and it is likely to disturb the normal splicing, according to MutationTaster. For all candidate genes described above (*SH2D3C*, *UNC5CL* and *SP1*) we assume that the mother and/or grandmother are asymptomatic for most of the disease phenotype characteristics.

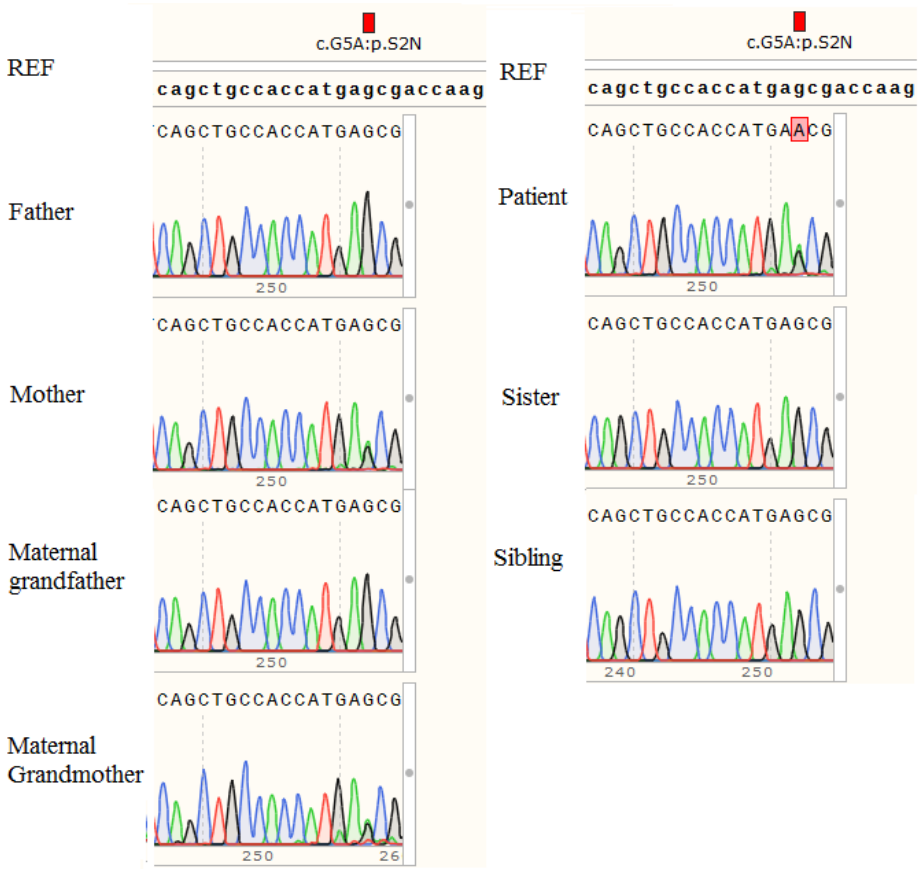


Figure 3.1.37: SANGER sequence of a region of *SP1* gene aligned with a reference sequence using SnapGene program. Confirmation of the mutation. Present in the mother, maternal grandmother and patient.

The compound heterozygous mutation in *TMED3* gene includes a deletion in exon 2 that cause a frameshift, R79fs and a substitution, R91Q. The deletion is present in the mother, patient and sister and the substitution in the father, patient and in the other sibling. The sequencing quality of both mutations are very good in all individuals with a genotype quality of 99, coverage depth supporting the mutation of at least 14 reads. Besides those good sequencing quality scores, both mutations also pass the confidence filters. The deletion is only present in 0.009% of Exac_all database and the substitution has 0.5% and 0,45% frequency in ESP6500siv2_all and Exac_all databases, respectively. The substitution is predicted to be damaging by the mutational significance cut-off (MSC)-CADD tool. This gene mRNA is overexpressed in bone, cartilage, pancreas, stomach and esophagus, according to GTEx Portal. The protein has a high expression in the skin, hair follicle and breast but also shows high expression in several cells from the immune system and the blood, specifically in peripheral mononuclear cells, according to GeneCards. This gene is not yet well characterized but in thought to be involved in IL-11/STAT3 signaling

to regulate STAT3 activity [213]. In turn, STAT3 is known to be involved in the regulation of IgE production and has been associated with gastrointestinal manifestations [214]. *TMED3* knockdown causes a decrease in *STAT3* activation [213]. Deficiencies in *STAT3* are strongly associated with hyper IgE syndrome and also with gastrointestinal manifestations [214,215].

The deletion was confirmed to be present in the mother and patient (**Figure 3.1.38**) and the substitution confirmed to be present in the father, patient and sibling (**Figure 3.1.39**). The compound heterozygous mutation in *TMED3* suggest a recessive inheritance pattern of the disease, assuming that the sister, mother and grandmother do not have the same disease. The patient presents both frameshift deletion and substitution and, the sister and mother only the deletion, which explain why they have a much lower degree of disease severity. The confirmation by SANGER sequence in the sister must be repeated in order to confirm the presence of the deletion. This inheritance pattern deduces a loss of function mutation and both frameshift mutation and inheritance pattern gives confidence to this assumption.

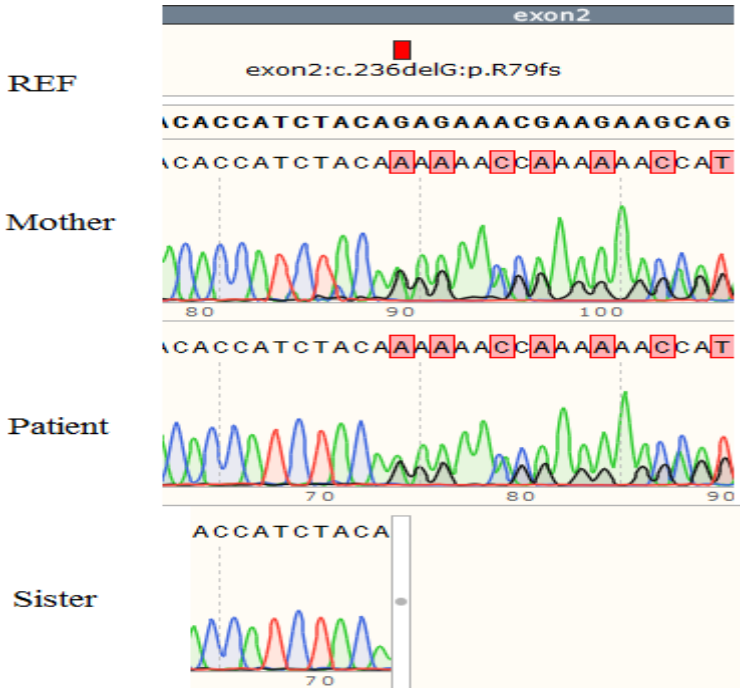


Figure 3.1.38: SANGER sequence of *TMED3* gene on the Guanine deletion in position 236 of the mother patient and sister. SANGER sequence aligned with reference sequence using SnapGene program. Confirmation of the mutation in the mother and patient.

Figure 3.1.39: SANGER sequence of *TMED3* gene on the Guanine substitution to Adenine in position 272 of the father patient and sibling, aligned with reference sequence using SnapGene program. Confirmation of the mutation in the father, patient and sibling.

We can hypothesize that our predicted loss of function mutation in *TMED3* might strongly reduce STAT3 activation, which will cause an increase in IgE levels. This can be linked to atopic dermatitis and the gastrointestinal manifestation that in our patient is demonstrated by food allergy. To unravel this hypothesis qPCR could performed to analyze the expression of *STAT3*. This mutation is an alternative to the autossomal dominant variants previously mentioned.

Family Case 15

In this family trio the parents are self-reported healthy and the patient has eczema, streptococcus and staphylococcus skin infections, incredible hyper IgE, candida (fungal infections) and herpes. The mother doesn't have allergies but has dry skin. The father has pollen allergy and chronic sinusitis, perleche and normal IgE values.

In this family, using our standard filtering we could not find any disease causing variant and because of that, we included addicional variants by increasing the frequency level of the variants and also allowed segmental duplications. The clinician that was treating this patient, with all her experience, highlighted that the seen eczema could be caused by a defect in filaggrin (*FLG)* protein. We checked the *FLG* gene and founded two homozygous candidate mutations: a substitution of a guanine to a thymine at position 10691 changing the amino acid from arginine to a leucine and a substitution of cytosine to a thymine at position 1501 that causes a gain of a stop codon also in exon 3 and in the transcript NM_002016. The first substitution (exon3:c.G10691T:p.R3564L) is present in 4,43%, 1,63% and 3,39% in ESP6500 database, 1000g2014act_all and Exac03, respectively. Besides that, the quality of the genotype is 99 and coverage depth 55 in the patient. The second substitution has frequency of 1.41%, 0.33% and 0.87% in ESP6500, 1000g2014act_all and Exac03 databases, respectively. The genotype quality is also 99 and the coverage 48. The genomicSuperDups have a score of 0.929016 and the GDI score of 27,35. These mutations are defined as being probably damaging to the protein but we can be sure that the stop gain variant is damaging (**Figure 3.1.40**). According to HIPED the differential protein expression in normal tissues and cell this lines is observed to be raised in urinary bladder and skin. This gene is essential for skin differentiation and the formation of skin barrier. Mutations or disruption of the function of this gene seem to predispose or lead to atopic dermatitis, allergy and asthma [216,217]. The loss stop codon mutation was confirmed to be homozygous in the patient and heterozygous in both parents by SANGER sequence and the other mutation was found to also be heterozygous in the parents (**Figure 3.1.41**). Loss of function mutations in *FLG*, both homozygous and compound heterozygous, have been found in patients that presented atopic dermatitis or ichthyosis vulgaris [217,218]. The two authors reported the same mutation that we found in our patient, R501X, and other deletion. These mutations were strongly associated with extrinsic atopic dermatitis, high level of IgE, asthma and palmar hyperlinearity [219]. Another author studied the same mutations reported by Palmer et al. (2006) and Smith et al. (2006), showed that they are ancestral European variants carried on conserved haplotypes, which explain the high frequency of these mutations [220]. All these studies support that our mutation might be the cause of, at least, eczema and hyper IgE in the patient. Moreover, using mice with a homozygous deletion analogous to common *FLG* human mutations, Fallon et al. (2009) demonstrated that the application of allergens resulted in skin inflammation and enhanced cutaneous allergen that activated and

developed allergen-specific antibody response, suggesting that deficiencies in the epidermal barrier play a role in elevated IgE levels and cutaneous inflammation. These gene mutations are strongly linked to the observed phenotype and so, *FLG* might be an important target for new therapeutic approaches.

Chromome	Position	ID	Reference Allele	Alternative Allele	Gene	CADD Score	MSC-CADD Score	MSC-CADD Impact Pred	PolyPhen 2 Score	PolyPhen 2 Pred	MSC-PolyPhen 2 Score	MSC-PolyPhen2 Impact Pred	SIFT Score	SIFT Pred	MSC-SIFT Score	MSC-SIFT Impact Pred
1	152276671	-	C	A	FLG	7.927	0.011	high	0.532	possibly damaging	0.239	high	0.590	T	0.261	low
1	152285861	-	G	A	FLG	25.700	0.011	high	NA	NA	0.239	NA	1.000	T	0.261	low

Figure 3.1.40: Table of Impact prediction tool (MSC) obtained after input the specific mutation in the MSC server.

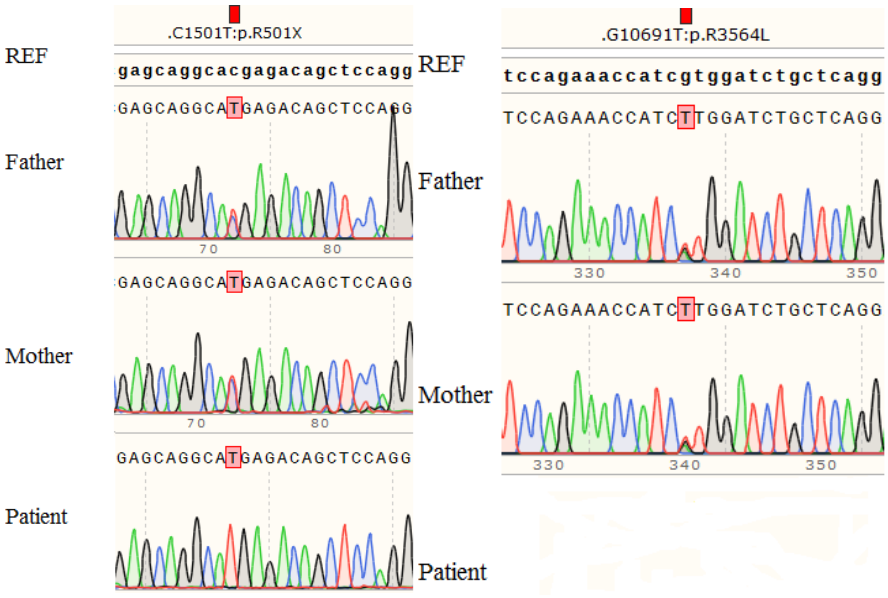


Figure 3.1.41: SANGER sequence of *FLG* gene on both substitution in the father, patient and sibling. SANGER sequence aligned with reference sequence using SnapGene program. This figure confirms the presence of both heterozygous mutations in the father and mother, and the presence of a homozygous mutation in the patient (R501X).

Analysis of a single patient

After clinicians diagnosed a patient through targeted Sanger sequencing we reanalyzed our output file to see how the causal variant was missed by our pipeline. After changing several filtering steps, to see if we could detect the mutation, we found the reason for this missed diagnosis. We were excluding the variant by excluding the segmental duplication regions. The genomicSuperDups had a score of 0.897187. The gene is a known PID gene, *PIK3CD*. This gene has a substitution mutation of a guanine to an adenine at position 3061 causing the amino acid change from glutamic acid to lysine in the transcript NM_005026. The genotype quality is 99 and the allele depth is 7 for reference and 12 for alternative allele. This mutation is completely private in pop_freq_all_20150413.

3.2. Targeted genetic analysis based on clinical phenotype

Family Cases 16, 17, 18

These three family cases presented with different phenotypes. Family case 16 is a patient with a Castleman disease'-like phenotype which includes generalized polyclonal lymphoproliferation, systemic inflammation and multiple organ system failure resulting from hypercytokinemia, especially of interleukin 6 (IL6). Full clinical remission was achieved with IL-6 blockade. Whole exome sequencing was performed in this patient and a mutation in the gene that encodes for adenosine deaminase (ADA2) (*CECR1*) was found. We confirmed that the mutation was present in the patient, mother, father, two sisters and two brothers but was exclusively homozygous in the patient and brother 2 (Figure 3.2.1). Furthermore, the activity of ADA2 protein was found to be reduced in both, confirming the deficiency of ADA2 (DADA2).

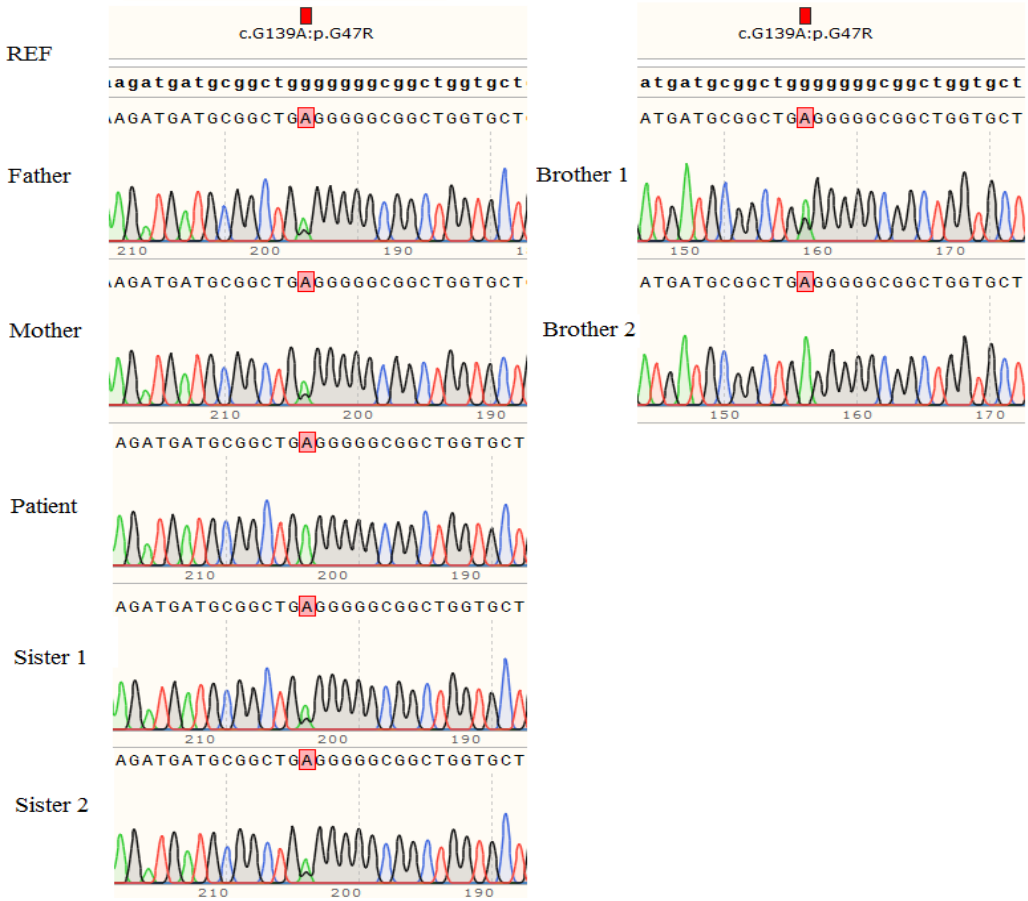


Figure 3.2.1: Results of the SANGER sequencing of *CECR1*, aligned with a Reference sequencing using SnapGene. We could confirm the existence of this mutation that was present in exon 2 of the *CECR1* gene of all individuals (NM_001282225:exon2:c.G139A:p.G47R). The patient and brother 2 were the only homozygous for this mutation.

Mutations in *CECR1* gene have been reported to induce an inflammatory condition that is characterized by a heterogeneous clinical manifestation that includes polyarteritis nodosa, cerebral stroke and immune deficiency and also reported to cause ADA2 deficiency [222–226]. Furthermore, the homozygous mutation found in family 16 (p.G47R) has been reported to cause ADA2 deficiency in consanguineous and not consanguineous families and in kindred [227]. The presence of the homozygous mutation in an unaffected brother displays an incomplete penetrance or demonstrates an wide phenotypic variability, which has been reported before [227]. However, testing the ADA2 activity in the brother is necessary to define if this homozygous mutation can display incomplete penetrance or if the brother has somehow overcome the ADA2 deficiency. Anyhow, our results in this family provide one more evidence that this mutation is disease causing and that incomplete penetrance is possible. Besides that our results show that this mutation is enough to be causing the ADA2 deficiency since we did not find any other mutation in *CECR1*. The mutation causing ADA2 deficiency, in this patient, is probably causing IL6 stimulation mediating lymphoproliferation and systemic inflammation [228,229]. This knowledge that hyper IL6 is a downstream cause of ADA2 deficiency make ADA2 and/or its signaling pathways potential therapeutic targets for the phenotypes observed in this patient [228].

The patient of family case 17 presented with a neonatal hemorrhage in the brain which can be the first presentation of DADA2. As an ADA2 deficiency was suspected, we designed primers to amplify all exons and flanking regions of the *CECR1* gene. Then, by SANGER sequencing, checked if the patient had any mutations in this gene. After a careful analysis, we could not find any relevant mutations. Whole genome sequence should be performed to infer about mutations in intronic regions or in other genes that might regulate ADA2 expression or function [230].

Family case 18 had a child affected with refractory polyarteritis nodosa lesions in the legs. The doctors suspected a DADA2 based on this classical presentation and ADA2 enzyme activity proved to be low. Refractory polyarteritis nodosa has been linked with ADA2 deficiency and subsequently to mutations in *CECR1* gene[222,224–226,231]. Whole exome sequencing was not performed in this patient. We used the *CECR1* primers designed for previous family to perform a full molecular analysis of the gene in this family. We identified several mutations and annotated these using Annovar (**Table 3.2.1**).

Table 3.2.1: Annotation of identified variants: to make this table, we had to identify the chromosome, position and the reference and alternative allele using the SANGER sequence results aligned with a reference sequence.

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	AAChange.refGene	genomicSuperDups	popfreq_all_2015041
22	17702737	17702737	A	AT	UTR5	CECR1	NA	NA	NA	NA
22	17702662	17702662	A	G	UTR5	CECR1	NA	NA	NA	NA
22	17669339	17669339	A	G	splicing	CECR1(NM_001282229:exon7:c.613-2T>C	NA	NA	NA	NA
22	17662917	17662917	C	G	intronic	CECR1	NA	NA	NA	NA
22	17662793	17662793	T	C	exonic	CECR1	stopgain SNV	CECR1:NM_001282225	NA	NA
22	17662575	17662575	T	C	intronic	CECR1	NA	NA	NA	NA
22	17662555	17662555	G	A	intronic	CECR1	NA	NA	NA	NA
22	17662039	17662039	G	C	UTR3	CECR1	NA	NA	NA	NA
22	17661811	17661811	T	A	UTR3	CECR1	NA	NA	NA	NA
22	17661798	17661798	G	C	UTR3	CECR1	NA	NA	NA	NA
22	17661667	17661667	G	A	UTR3	CECR1	NA	NA	NA	NA
22	17661640	17661640	A	G	UTR3	CECR1	NA	NA	NA	NA
22	17661223	17661223	G	C	UTR3	CECR1	NA	NA	NA	NA
22	17661054	17661054	T	C	UTR3	CECR1	NA	NA	NA	NA
22	17660522	17660522	T	C	UTR3	CECR1	NA	NA	NA	NA
22	17660377	17660377	G	C	UTR3	CECR1	NA	NA	NA	NA
22	17660179	17660179	T	A	UTR3	CECR1	NA	NA	NA	NA
22	17659932	17659932	C	T	UTR3	CECR1	NA	NA	NA	NA
22	17690360	17690360	G	A	exonic	CECR1	stopgain SNV	CECR1:NM_001282225	NA	NA
22	17670849	17670849	G	T	exonic	CECR1	nonsynonymous SNV	CECR1:CECR1:NM_001	NA	NA

We found one homozygous mutation that could cause an aberrant splicing. We confirmed that this mutation is homozygous in the patient and heterozygous in the mother and father (**Figure 3.2.2**). To demonstrate altered splicing, we extracted RNA from PBMCs (peripheral blood mononuclear cells) and executed a reverse transcription process to obtain cDNA. After that, we performed a PCR run using primers positioned outside exon 7 and submitted the PCR product to an electrophoresis run (**Figure 3.2.3**). The 2nd and 3rd wells showed two bands, one of the same size as control (4th and 5th wells) and another with approximately 100 bp less. The smaller band is very likely to represent the loss of exon 7 given that this exon has 109bp. The results we expected was only the presence of the small band due to the homozygous nature of the mutation. The top band present in the 2nd and 3rd wells of the electrophoresis run might represent the transcripts that don´t have and open reading frame which means that they are not translated and show that this trasncripts are not sensible to this intronic mutation.

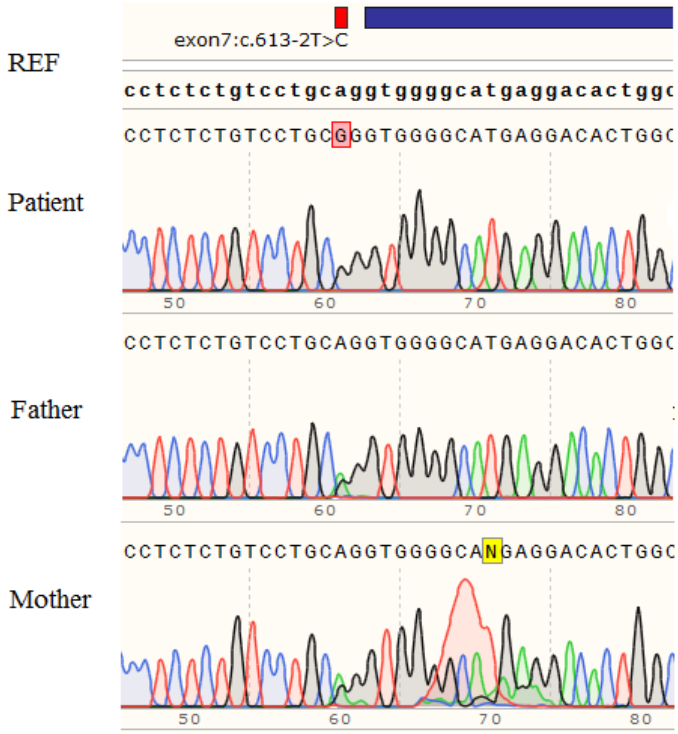


Figure 3.2.2: SANGER sequence of the homozygous substitution of an Adenine to Guanine in gene *CECR1* of family 18. Confirmation that the mutation is heterozygous in the parents and homozygous in the patient.

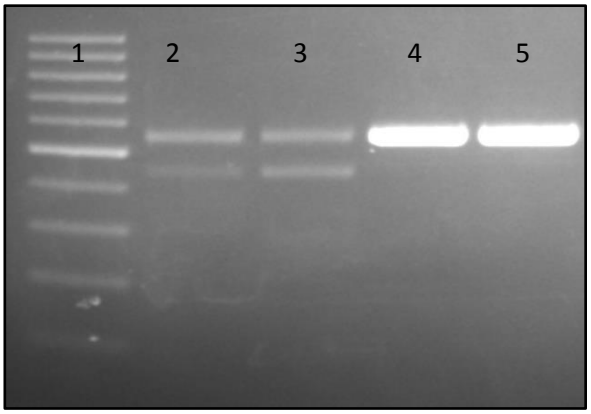


Figure 3.2.3: Photograph of the end of electrophoresis run. **The first well is the 100bp DNA ladder, the second and third is the cDNA of the patient and the last two bands are cDNA control.**

These results suggested that the mutation can cause the exon 7 skipping although the normal splicing still occurs. The exon 7 skipping was confirmed by comparing SANGER sequencing on each of the two bands separately generated by the patient cDNA with the band attained from control cDNA (**Figure 3.2.4**). We also demonstrated reduced expression of *CECR1* mRNA likely due to the exon7 skipping by executing a qPCR (**Figure 3.2.5**). However, the expression in the father is higher than the control instead of having the same expression of the mother. Perhaps some problem in the mRNA quantification of the father happened and thus, redoing this qPCR using at least one more constitutively transcripted gene as a control is necessary to have a better relative quantification. Our results showed evidence that the mutation causes an alternative splicing in *CECR1* gene. A compound heterozygous mutation with a substitution and a deletion of exon 7 and a has been described in two children but here we show a newly homozygous mutation in an intronic region that leads to this deletion [232]. Our results show a new homozygous mutation in the *CECR1* introgenic region, not yet reported that causes inflammatory lesions in the legs of the patient.

Figure 3.2.4: Sager sequence of cDNA control and the two bands present in the patient' electrophoresis run, aligned with a transcript reference sequence (NM_001282225).

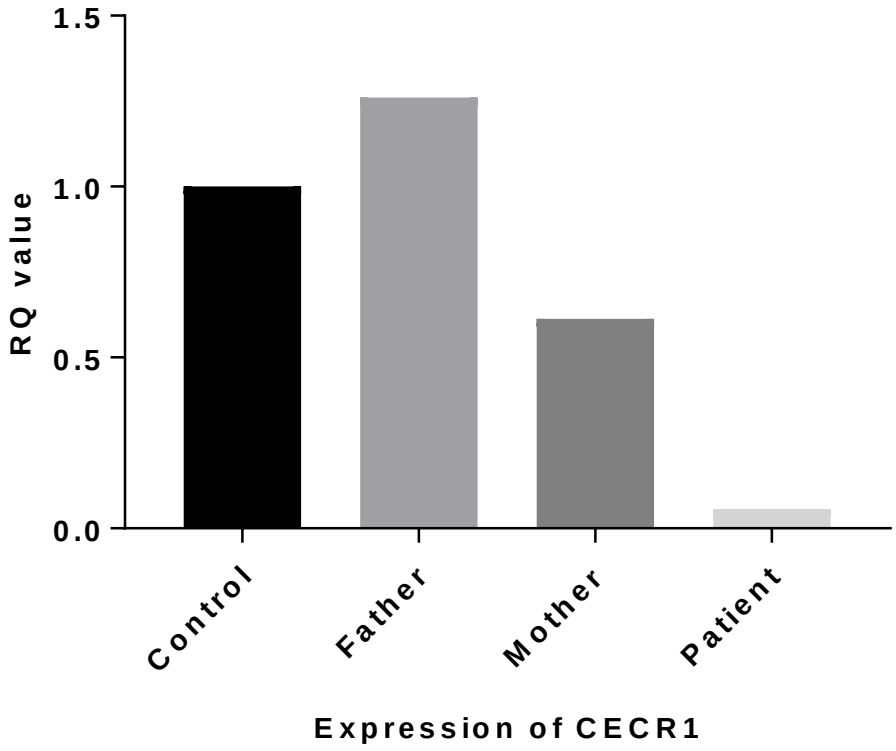


Figure 3.2.5: Relative quantification (RQ value) of *CECR1* mRNA expression of blood cells in different individuals (Father, mother, patient and control), confirming a very low rate of expression in the patient.

3.3. Investigation of rare variants in polygenic disease

Family cases 19, 20, 21

Through WES we found two *IL6R* mutations present in three unrelated patients with Juvenile Idiopathic Arthritis (JIA) as part of 57 JIA patients and 81 unrelated healthy control samples. Three polyarticular JIA patients, but no controls, were heterozygous for mutations in the *IL6R* gene. Specifically, the P65L variant found in one JIA patient has been previously reported in the 1000 Genomes project (freq=0.0009) and exac3 (freq=0.0005), but was predicted to be possibly damaging based on the mutation significance cutoff (MSC)-CADD score. The second variant, L379F, has been previously reported in exac3 (freq=0.001), was predicted to be benign based on the MSC-CADD score, but was shared by two polyarticular patients. These mutations were confirmed by SANGER sequence in the cases' families (**Figure 3.3.1** and **3.3.2**).

Figure 3.3.1: Results of SANGER sequence of *IL6R* aligned with reference sequence using SnapGene. These results confirm the presence of the mutation in family 20 (c.C194:p.P65L).

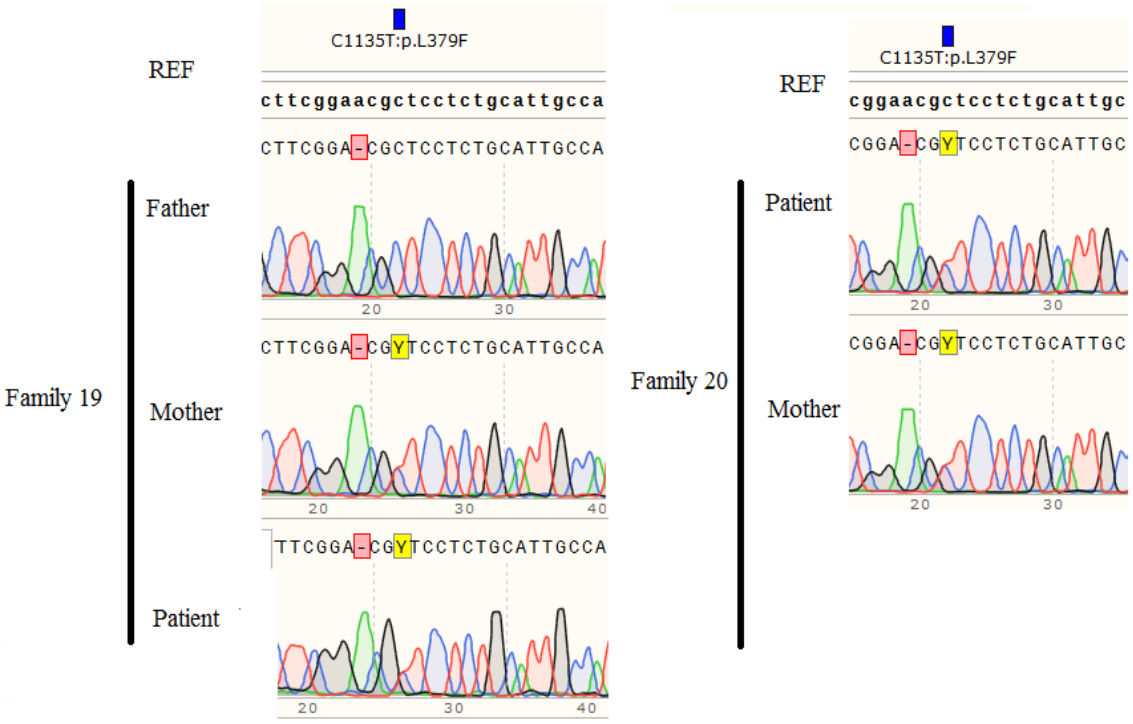


Figure 3.3.2: Results of the SANGER sequence aligned with ref sequence using SnapGene in families 19 and 20. Confirmation of the mutation (c.C1135T:p.L379F).

After confirming these mutations in our patients, we were interested in checking whether these mutations were present in an independent set of patients with the same disease. With this objective, we received 48 DNA samples of patients with Juvenile Idiopathic Arthritis from our collaborators from the United States. We performed SANGER sequencing of the two mutations in all of the patient samples but none of the mutations were detected in these additional samples (some of Sanger results are shown in Supplementary data, **Figure 6.2**).

4. Discussion

4. Discussion

In the results, we already discussed the functional aspects of each one of the gene variant candidates found in each family and therefore, in this section, I will focus on the discussion of the WES results and the pipeline, as well as the filtering options we used. I chose this approach in order to be more elucidative, understandable and clear about the work performed to the readers of this thesis.

To choose variant candidate genes we took into account several aspects, such as, gene function, mutation frequency, expression profiles, functional impact predictions and genotype qualities. SANGER sequencing was performed in several candidate variants demonstrating either the veracity of the mutation or the presence of a sequencing artifact. Not all of the candidates we found were tested by Sanger sequencing, some due to the lack of functional association, and others to the low quality of the genotype. The sequencing quality was given by several parameters in the VCF file. The candidate genes in family case 1 are examples of variants that were not SANGER sequenced due to lack of a strong functional association between the genes and the phenotype presented by the patients. Besides that, the sequencing quality of the variant genes *POLE* and *SHANK1* were low. The mutation in *POLE* has a genotype quality (GQ) of 61 and the mutation *SHANK1* has a filtering depth (DP) of only 6, which means that this supposedly alternative allele was only supported by 6 reads. This sequencing parameters scores demonstrate a low confidence that the alternative alleles are indeed present in the patients. On the other hand, the recessive mutations in *DIDO1* has good sequencing quality indicating with more confidence that this variant is present in the patient and the SANGER sequencing confirmation of this variant must be done.

On family case 2, the inheritance pattern of *TRPM5* mutation we were looking for was not confirmed by the SANGER mutation, as we could suspect by the bad sequencing quality of the father, demonstrated by the genotype quality (0), allele and coverage depth (2). The father does not show the patient phenotypes and so, this gene mutation was excluded. The deletion in *PSMB11*, other potential candidate gene, was not confirmed by SANGER sequence results, although it had a very good genotype quality (99) and allele depth (15). Is important to note that something rather than the nucleotide deletion is happening in the patient and father, as we can see in SANGER sequence results, that the bioinformatic process was not able to identify. However, the genotype quality of the father was very low (0) and that is a warning that this region was badly sequenced or that

it was difficult to align. Some kind of mutation is happening in this gene region and both father and son have it, which does not fit into our suspected inheritance because the father does not show any pathogenic phenotype. This gene was also excluded from the set of candidate genes. The other potential candidate was *CST7*. Although the variant in *CST7* had a good general sequencing quality, it was not confirmed by SANGER sequence and proved to be an artifact. The only parameters that could alert us for this results were the relatively low QUAL score (52,27), which expresses the level of confidence of a mutation being real, and furthermore, the information that the mutation did not pass the FILTER VQSR Tranche SNP 99.90 to 100.00, considering the mutation a false positive (Supplementary data: Table 6.3). The mutation was obviously excluded. Unfortunately we could not find any variant that could be correlated with the phenotype in this family. However we could proceed to WGS in order to analyze mutations outside exon regions. These results demonstrate that it is important to notice the QUAL score, Genotype quality (GQ) and also the FILTER tranches to choose a good candidate mutation.

The WES results for family case 5 were uncommon because after executing our filtering process, no compound heterozygous or homozygous mutation was found. This result make us question if we are using the appropriate frequency cutoff for recessive variants. Perhaps this cutoff is too restrictive when analyzing variants with this type of inheritance. We could also inquire about the quality of the DNA used for sequencing or the quality of the sequencing itself but, when comparing the number of variants obtained at the beginning and at the end of the filtering of this family with other families analyzed we conclude that there is no serious difference between them (Supplementary data: Table 6.4). However, we found a potential candidate gene in this family, *SERTAD1* which had a variant with an average genotype quality and a low coverage depth of the alternative allele, suggesting that this variant might not be confirmed by SANGER sequence. The quality of the same region in the parents was good with a genotype quality of 81 and a good coverage depth (30). The good quality of the parents in this region tell us that this region is not difficult to sequence and suggest that the low quality of the variant in the patient might be due to its nature (small nonframeshift deletion of 3 nucleotides) or to bad genome reference alignment, considering that this variant is present in the patient. SANGER sequencing should be executed for this variant.

In family case 6, one of the possible candidate genes is *TYRO3*, which has a deletion of one nucleotide that will cause a shift in the reading frame. However, by

SANGER sequence, we demonstrated that this mutation is not present in the individuals. Once again, a deletion is not confirmed and although the exome quality was good, the mutation did not pass all FILTER, specifically the VQSR (variant quality score recalibration) Tranche INDEL 99.00 to 99.90. This VQSR is a newly calculated score that is supposedly super well calibrated. This new score allows us to balance sensitivity (find real variants) and specificity (limiting the false positives) and is an important information provided by the VFC file. This type of quality TRANCHEs represents the estimated probability of each variant to be true. The tranches are used so we can establish thresholds easier. This mutation is in the confidence interval between 99 and 99.90 and as we saw, the mutation was not real. If the mutation does not pass the FILTER field it means that the mutation is probably not real.

Other candidate for this family was a variant in *NFKBIZ* gene which appear to fit very well with the phenotype, but we discovered that the actual mutation was quite frequent among all databases, even when we calculated the frequency for the presence of two alternative alleles. The sequencing quality provided by the bioinformatic process in the VCF file only warned us in the FILTER field. The mutation did not pass the Tranche for INDELs, with a confidence of 99.00 to 99.90. This erroneous frequency must have happened on the alignment with the reference genome or a misleading in the annotation made by ANNOVAR. Due to the new frequency of the mutation, we excluded this mutation from being disease causing.

In family case 7 we had a very low number of mutations which is quite abnormal. This low number of variants might indicate that the quality of the DNA that was used for whole exome sequencing was not good [233]. New whole exome sequencing can be done in this family. Everything in terms of sequencing quality and DNA quality might have gone correctly, and this disease is maybe driven by mutations in intronic sequences which can regulate the expression of genes [52,234]. Instead of redoing the WES, it would be better to perform the whole genome sequence so we could sequence intronic regions. Comparing this family with family case 5 we can see the difference between a family with low number of variants, which is probably a consequence of low quality DNA, and a family with normal number of variants but with no variants following the recessive inheritance (Supplementary data: Table 6.4).

Another wrong call from the WES was in family case 8. In this family the SANGER sequence demonstrate that the expected variation *ILF3* was not a substitution

but an adenine insertion. The variant did not pass the confident filter tranche of SNP 90 to 99.90 (Suplemmentary data: **Table 6.3)**. This is an example of a probable miss alignment of the reads with the reference genome leading to the wrongly assumption that this mutation was a substitution. This gene was put aside due to the confirmation that the mutation was inherited from the father and he does not show any signals of the disease that affects the son. Besides *ILF3* variants we consider *SEC61A1* as one good candidate disease-causing gene. The information about the sequencing quality of *SEC61A1* shows a very good sequence quality in terms of genotype quality and coverage in all individuals, and also passed the FILTER field. Although we can be confident that this mutation is present in the patient, the SANGER sequence of this mutation must be executed.

In family case 9 the deletion in *SELPLG* was not present in the patient. The bioinformatic analysis did not point a low confidence on this variant call (Supplementary dada: **Table 6.3**), but the number of variants calls obtained after filtering process was 65, which is very low compared to the mean number of total variants found in other families, indicating a low quality or quantity of the DNA at the time of sequencing (Supplementary data: **Table 6.4**). The WES should be repeated again in high DNA concentration and quality.

The sequencing quality of *TBK1* variant was not good in the family case 12. The genotype quality and coverage depth of the deletion was poor, and the mutation did not pass all confidence filters (Supplementary data: **Table 6.3**). Once again, SANGER sequence showed that the bioinformatics process is not very accurate when it comes to identifying deletions. However, we can see background noise in the SANGER sequence of the patient and so it would be better to repeat it.

The other candidate mutation in family case 12, *NMI,* does not have a good genotype quality in the patient. The mutation also did not pass all confidence filters, precisely the VQSR tranche SNP 99.00 to 99.90. We have been learning from the other families analysis and saw that all the mutations that did not pass the filter tranches were proven to be artifacts by SANGER sequence. This mutation will probably not be confirmed by SANGER sequence either.

The family case 15 allow us to realize one limiting aspect of our filtering. In this family, the specialized dermatology clinician that analyzed this patient phenotype had suggested a filaggrin protein deficiency. Applying our initial filtering we did not find any

candidate gene variants that could adapt to the phenotype or inheritance mode but, when we included recessive variants with a higher frequency cutoff in the databases we found two homozygous mutations in *FLG*, one of them confirmed by SANGER sequence (R501X) and the other one need to be executed again in the patient. Mutations in this gene have been reported to cause skin inflammation and the mutation R501X has also been previously reported in a similar phenotypic case, as referred in the results [235,236].

To test our filtering process, we analyzed a single patient that was treated in the clinic and had already been molecularly diagnosed. Our standard filtering process was excluding the reported mutation by filtering out all segmental duplications. We performed this filter because segmental duplications are still very difficult to resolve accurately due to the small sequencing reads generated by WES, which makes impossible to differentiate a duplicated gene and its parent gene during the alignment of the reads with the reference genome [237]. This problem induced researcher to exclude segmental duplications of the down-stream analysis in order to reduce artifacts. We must be careful in the analysis of variants in segmental duplication, because the large segmental duplications can only display 3% of inter-variability and because mutations in this duplications can be highly damaging [238]. In this thesis we described two cases where damaging mutations were excluded by the filtering process. The first are the mutations in family case 15, which had homozygous mutations in *FLG* missed by the frequency filter and also by the exclusion of segmental duplications. The sequencing of *FLG* was problematic due to its molecular structure. *FLG* has a large exon 3 comprising 10 to 12 identical full tandem repeats [220]. Mutations in C-terminal repeats can be pathogenic and result in ichthyosis vulgaris or predisposition to eczema and allergic diseases or atopic dermatitis [220,239,240]. The second, is the mutation in *PIK3CD*, the known pathogenic E1021K substitution, which was also missed because of a duplication in exon 24 [162].

We state that variants in segmental duplications cannot be excluded blindly and should be analyzed when no variant to explain the phenotype can be found using the standard filtering process of WES data. Furthermore, when no candidate variants are found, we can also increase the mutation frequency for the recessive mutations, as shown in family case 15. In this way we can be sure that we analyze all possible gene variants present in the patient and then, we can think of new methods to unravel the cause of the disease outside the limits of WES. There were several families, such the case of case 10 and 11, were no gene variants that fitted the suspected inheritance and could be disease

causing were found. On those, whole genome sequencing should be done in all family individuals to search for a disease causing mutation on intronic regions or in exonic regions with better coverage and also analyze variants excluded by the filtering process.

The last families analyzed had patients affected with polyarticular juvenile idiophatic arthritis. This disease is characterized by arthritis in 5 or more joints in the first moths of the disease, and is thought to be driven by several genes and environmental factors [241–243]. Several patients that suffer from JIA have been treated with *IL6R* blockade, tocilizumab, and had successful clinical responses [244–246]. Besides that, variants in *IL6R* and/or *IL6* genes have been proposed to induce susceptibility to develop JIA and other forms of arthritis mediating different expression of IL6 levels [247–251].

Our results show that the rare mutations we found in *IL6R* were not replicable, decreasing the confidence that these mutations cause an increase in susceptibility to develop JIA, but increasing their rare frequency state in the population. However, the mutations were not absent in some databases possibly indicating that other individuals have this polypmorphisms and might not present the disease. Although, we must take into account that the ethnic background can strongly influence the genetic associations and that several identified polymorphisms also vary between ethnic groups [243,252]. The mutations described have never been associated with JIA before. Both mutations can influence the level of expression of IL6 and be defined as a gain of function [249]. The first mutation, P65L, does not have such a strong credibility to be disease-causing as the second one, L379F, because the second was found to be present in two individuals from unrelated families.

We report two novel mutations in *IL6R* gene that are probably increasing the susceptibility to develop Juvenile Idiopathic arthritis, whereas the L379F substitution has a greater reliability. Although, they might only have a modest effect on risk and only explain part of the variance disease risk. Nonetheless, we must investigate these mutations in a larger cohort and, if possible, have the ethnic groups clearly identified. Besides that, whole exome sequencing should be performed on the samples received from our US collaboration to see if they present any other mutations in *IL6R* and, perhaps, provide one more evidence that mutations in this gene cause a susceptibility to develop JIA.

5. Conclusions and future perspectives

This thesis work identified strong biological candidate gene variants and indicated new possible pathways involved in the immune system regulation. However, further studies are needed to confirm the function of these mutations. The functional confirmation is very complicated but, some widely used techniques include Western blot, qPCR and cloning assays, for example associated with CRISPR-Cas9 methodology and the use of mouse models of diseases. Besides that, immunophenotyping should also be done in all affected or mildly affected individuals due to its power to exploit human immune phenotype and to identify immunologic changes. This technique is capable of analyzed multiple parameters on different individual cells and characterize many subset of cells in a complex mixture such as blood, allowing the measure of, for example, B and T cell subsets and cytokine expression patterns. Through this analysis we can make associations between genes and the immune phenotype meaning that a certain gene variant can be causing the overexpression of some cell subset or cytokine and with this information dander on the molecular mechanisms involved.

Whole-exome sequencing and the available bioinformatics tools are very useful for the identification of disease causing variants in families with Mendelian forms of diseases, which are crucial for better understanding of the disease and for choosing the correct medical treatment. However, there are several data analysis challenges and can lead to false positives results. For example, it is still difficult to detect accurately indels with short sequence reads generated by next-generation sequencing technologies. The criteria used for filtering the long list of mutations generated should be decided carefully and thresholds should not be followed so strictly. Our standard filtering pipeline would exclude mutations present in segmental duplication, mutations with high gene damage index score and mutations with a frequency of more than 0.5% presence in public database. By applying these filtering criteria to several family trios, we concluded that: 1) GDI should be used only for prioritization because, large genes have a higher mutational rate and subsequently a higher GDI score 2) mutations present in segmental duplication regions should also be investigated after failing to find any good candidate genes with the first analysis and 3) increase the threshold of allele frequency when looking for recessive variants (homozygous and compound heterozygous) because the frequency of having one alternative allele can be considered common (>1% in public databases) in the population but the presence of a second allele at the same position is probably a very rare episode and probably has a completely different biological impact.

Furthermore, we observed that some genes were frequently showing up at the end of the filtering process across different families and sometimes represented false positives variants. Finally, sharing the list of filtered variants with clinicians is also important as they can select the good candidates based on the clinical presentation of the patient.

Despite the fact that whole-exome sequencing is used in the clinical setting for the identification of mutations contributing to rare genetic diseases, there is still much to be improved as many cases remain unresolved. Computation and bioinformatics can provide better tools for analysis and interpretation of the sequencing data. Furthermore, whole-genome sequencing that examines other areas of the genome can also facilitate to solve difficult cases. Finally, a more targeted analysis approach looking for mutations in genes that are already known to have a role in the disease pathogenesis can also be the key to the discovery of the disease-causing variant.

6. Supplementary data

Table 6.1: Families and individuals analyzed and their clinical data, disease status, ethnic origin and presence of consanguinity.

Family/individual	Clinical data	Disease Status	Origin	Consanguinity
Case 1_Father	Patient 1: Schwachman-like. Patient 2: Autism, delayed development. Patient 3: Autism, pancreas insufficiency	Unaffected	Caucasian	No
Case 1_ Mother		Unaffected	Caucasian	
Case 1_Patient1		Affected	Caucasian	
Case 1_Patient2		Affected	Caucasian	
Case 1_Patient3		Affected	Caucasian	
Case 2_Father	Colitis and pyogenic infections	Unaffected	Caucasian	No
Case 2_Mother		Unaffected	Caucasian	
Case 2_Patient		Affected	Caucasian	
Case 3_Father	PID + tumor (medulloblastoom) -	Unaffected	Caucasian	No
Case 3_Mother		Unaffected	Caucasian	
Case 3_Patient		Affected	Caucasian	
Case 4_Father	Syndromic - retardation - PID – aneurysm	Unaffected	Caucasian	No
Case 4_Mother		Unaffected	Caucasian	
Case 4_Patient		Affected	Caucasian	
Case 5_Father	EBV-related PID and lymphoma - SAP and XIAP (-) combined PID	Unaffected	Caucasian	No
Case 5_Mother		Unaffected	Caucasian	
Case 5_Patient		Affected	Caucasian	
Case 6_Father	Invasive pyogenic infections	Unaffected	Caucasian	No
Case 6_Mother		Affected	Caucasian	
Case 6_Patient1		Affected	Caucasian	
Case 6_Patient2		Affected	Caucasian	
Case 7_Father	Patient: hypogammaglobinemia, Burkitt lymphoma, sepsis strep (bacterial infection) and severe varicella. Sister: granuloma annulare (inflammatory dermatosis) and low level of naive T cells.	Unaffected	Caucasian	No
Case 7_Mother		Unaffected	Caucasian	
Case 7_ Patient		Affected	Caucasian	
Case 7_ Sister		Mildly Affected	Caucasian	
Case 8_Father	Neutropenia, WHIM like syndrome, severe congenital neutropenia. Warts, hyper-IgG, NK cells "lowish" (but warts indicative of NK defect), pneumonia, VZV infection, mollusca.	Unaffected	Caucasian	No
Case 8_Mother		Unaffected	Caucasian	
Case 8_Patient		Affected	Caucasian	
Case 9_ Father	Hyper Immunoglobulin E and familial systemic sclerosis	Unaffected	Caucasian	No
Case 9_ Mother		Unaffected	Caucasian	
Case 9_ Patient		Affected	Caucasian	
Case 10_Father	Combined Immune Deficiency	Unaffected	Caucasian	No
Case 10_Mother		Unaffected	Caucasian	
Case 10_Patient		Affected	Caucasian	

Case 11_Father	Graves' disease, general autoimmunity disease and several skin manifestations (vitiligo, verrucosis/papillomatosis and keloid formation)	Unaffected	Caucasian	No
Case 11_Mother		Unaffected	Caucasian	
Case 11_Patient		Affected	Caucasian	
Case 12_Father	Autoimmune neutropenia and eczema	Unaffected	Caucasian	No
Case 12_Mother		Unaffected	Caucasian	
Case 12_Patient		Affected	Caucasian	
Case 12_Sister		Unaffected	Caucasian	
Case 13_Father	Invasive pyogenic infections	Unaffected	Caucasian	
Case 13_Mother		Unaffected	Caucasian	
Case 13_Patient		Affected	Caucasian	
Case 14_Father	Patient: Atopic dermatitis from birth, severe food allergy, Hyper IgE, neutropenia and auto-immune hepatitis. Sister: Hyper IgE episode Mother: bad allergy	Unaffected	Caucasian	No
Case 14_Mother		Mildly Affected	Caucasian	
Case 14_Patient		Affected	Caucasian	
Case 14_Sister		Mildly affected	Caucasian	
Case 14_Sibling		Unaffected	Caucasian	
Case 15_Father	Eczema, streptococcus and staphylococcus skin infections, incredible Hyper IgE, candida (fungal infections) and herpes	Unaffected	Not known	
Case 15_Mother		Unaffected	Not known	
Case 15_Patient		Affected	Not known	
Single patient	Treated in clinic	Affected	N/A	N/A
Case 16_Mother	Castleman disease'-like phenotype	Unaffected	N/A	No
Case 16_Father		Unaffected	N/A	
Case 16_Patient		Affected	N/A	
Case 17_Patient	Neonatal hemorrhage in the brain (low ADA2 activity)	Affected	N/A	N/A
Case 18_Patient	Refractory polyarteritis nodes lesion (low ADA2 activity)	Affected	N/A	N/A
Case 19, 20 and 21_Patients	Juvenile idiophatic artritis	Affected	Caucasian	No

TABLE 6.2: Sequence of primers forward (F) and reverse (R), in SANGER sequence confirmation of several gene variants. The name of the primers coincide with the gene that supposedly contains the mutation.

Primer Name		Sequence (5' -> 3')
TRPM5	F	CTGGTCAGCAACAAGCCCG
	R	CCTTGAGTACGCGGGAGACC
PSMB11	F	GTTGTGACCCTCAAACCTTCC
	R	CAGAGATCCAGTCCCCGGTA
CST7	F	AGAGAACGGGAACACAGCAA
	R	GTGCTACCATGCTGCCATTC
KMT2D_1	F	CCCAACAGCCCATGCTAGAG

	R	TGTGGGTTTTTGGCCAGGAC
KMT2D_2	F	GGAAGGTAGTGGGGGTCTCA
	R	GATGAGAGTGGGTGGTGTGG
TYRO3	F	GGGACCTCATGGTGTCATCCTA
	R	TTTGTCCTTCCCACGATCAC
NFKBIZ	F	TGCCTTTGACAAGGATCGCA
	R	CGAGTACTTGGGTCTGCTCC
ILF3	F	GTTCTACAGCAACGGAGGGC
	R	TCGAGTAGGAGTAGCTGCCTT
SELPLG	F	TCCACGGATTCAGCAGCTATG
	R	TGGGAATGCCCTTGTGAGTA
TBK1	F	ATGTGCCTACATTCAGTTCCAC
	R	GCAAAAGCACAAACTGCAATGAAA
LYST_1	F	GCCAGCGAGCCTTTAAGTCA
	R	AATAGCTTTGCTTCCTCGGG
LYST_2	F	GAGGGACCATCTTTCCCAGT
	R	TCAGCGTCCTAGTGTCATCC
SH2D3C	F	CACTCAGCAGCTACAGTGACC
	R	GTGACTTCCACGATGGGGAC
UNC5CL	F	TGGCAGTGTGTTGGTGTCTT
	R	TCTCCAGATGTCCAGGGGAG
SP1	F	CCAGGCACGCAACTTAGTCT
	R	AAGCTTGGAGTGGACTCATCC
TMED3	F	AGTGGTGGTTGCTCCTTTGAT
	R	CAGCCCCGAACTGCTAATGT
FLG_1	F	CTGGACGTTCAGGGTCTTCC
	R	TGATGACGTGACCCTGAGTG
FLG_2	F	AAGCAGAAGAGGAAGGCAGG
	R	TGCAGATGAAGCTTGTCCGT
CECR1_exon1	F	CAACGCTCAGAGTCACACCT
	R	CTCCAAGTCACACCTTGGCA
CECR1_exon2	F	GTTTTCACTCCCCCACTCTCC
	R	TAGCCCCAAGATGAGTCCCT
CECR1_exon3	F	GGGACTTCACCCCCTCCTTT
	R	AAGGGAGACACCTACCCACTG
CECR1_exon4	F	TGGCCAAGCATACAGAGGGA
	R	GTCAGGCCAGAGCAAAGGAG
CECR1_exon5	F	CAGGTGTATGAGCTCAGTGG
	R	GCACTGGTGCCAAGGAGCT
CECR1_exon6	F	CCACCTGGCCTCCTCTAAAC
	R	CATGCCCCCTTAACAGGCAG
CECR1_exon7	F	GCGCGCCATCAGCAAGTG
	R	GCTCTCCCATTGACCACCTC
CECR1_exon8	F	CATCCATCCCCCATGGAAGAC
	R	GAGCAGAGGTTGTGGTTAGGGG

CECR1_exon9	F	TGATGGGGCTCAAGGTCTCA
	R	AGGAACCATCGAGGCATCTG
CECR1_exon10A	F	GTGCTCTGCAAGGCTCTAATG
CECR1_exon10A	R	TCTGGTCTCTTTAACTCTCTCCT
CECR1_exon10B	F	CCAACTCTTGACCTCAGGTGAT
	R	CTTCCCTTTGCCTCCTCCAG
CECR1_exon10C	F	CCTCCTCACTGATCTCCCCT
	R	CCAAGGCAGAGGGGAGTAGA
CECR1_exon10D	F	GCAGGCGCTCATGATTGTTT
	R	TGCCTTTGTGCTTTGTGTCC
***CECR1*_cDNA_splicing of exon7**	F	GCTGCTGCCGGTGTATGA
	R	GAGTAAGTCCTGACTGCGGG
IL6R_1	F	GTACCACTGCCCACATTCCT
	R	GCACCTAAAACACGGCTTGG
IL6R_2	F	CTGCCCTAATCCAGGCAGAG
	R	GGGAACTATCTCCGGGACCT

Table 6.3: Information about the gene, mutation, and sequencing quality of the assigned genotype. Columns: Gene name; NM_number: representing the RefSeq accession number of the mRNA sequence; Mutation: position and nature of the mutation in mRNA sequence (ins= insertion; del=deletion; the first two letters represent the change of the main bases and the last two letters the protein change); Confirmation by SANGER sequence: "YES" presence of the mutation and expected inheritance pattern, "NO" mutation not present or different inheritance pattern than expected; Inheritance: extracted by the analysis of the final VCF file generated; Exonic function: provided by the VCF file and represents the consequence of the mutation in terms of mRNA and protein sequence; QUAL: Confidence through a Phred Scale probability that the Reference and Alternative allele are real in a specific position. A score of 10 represents a change of errors of 1 in 100, while 100 represent a change of error of 1 in 10 ^10. FILTER: contains the name of the filters that the variant failed to pass but, if Filter=PASS, the variant passed all filters applied; F_DP: father filtered depth that represents the number of filtered reads that support the reported alleles but the uninformative reads are also counted. F_GQ: Father Genotype Quality, that represents the Phred-scaled confidence that the allele assigned is correct. The maximum score is 99 meaning that the confidence of that allele call over another allele is very high; M_DP, M_GQ, P_DP and P_GQ is the same but for the called alleles of the Mother (M) and the Patient (P); P_AD is the patient Allele depth that is similar to the DP but in this case the uninformative reads are not included, so this number provides the number of reads that support each one of the called alleles. (Next page)

Gene	NM_number	Mutation	Confirmed?	Inheritance	Exonic function	QUAL	FILTER	F_DP	F_GQ	M_DP	M_GQ	P_AD	P_DP	P_GQ
CARD11	NM_001324281	G3325A:p.E1109K	NO	Compound HET	nonsynonymous SNV	1.69	.	.	25	.	.	8,3	.	.
CARD11	NM_001324281	A1508U. K503M	NO	Compound HET	nonsynonymous SNV	1.69	.	.	.	.	.	6,2	.	25
GUCY1B3	NM_001291953	C415A:p.R139S	YES	from Father	nonsynonymous SNV	1097.19	PASS			35	99	61,44	105	99
ITPR3	NM_002224	C7570T:p.R2524C	YES	de novo	nonsynonymous SNV	1307.13	PASS	34	84	28	75	78,49	127	99
KMT2D	NM_003482	C12542G:p.S4181C	YES	compount HET	nonsynonymous SNV	406,16	PASS	18	51	20	99	15,5	20	99
KMT2D	NM_003482	C3392T:p.P1131L	YES	compount HET	nonsynonymous SNV	491,36	PASS	11	99	5	15	3,9	12	43
NR1H2	NM_007121	G1169A:p.R390H	NO	de novo	nonsynonymous SNV	34.9	PASS	0	.	0	.	3,2	5	61
PHLPP2	NM_015020	C3349T:p.R1117C	YES	Compound HET	nonsynonymous SNV	938.16	PASS	32	99	38	99	16,19	35	99
PHLPP2	NM_015020	A820G:p.I274V	YES	Compound HET	nonsynonymous SNV	2539.16	PASS	41	99	94	99	35,45	80	99
SH2D3C	NM_170600	C1280T:p.A427V	NO	de novo	nonsynonymous SNV	42.05	PASS	2	6	6	15	5,4	9	72
SLC8A3	NM_183002	G1636A:p.G546S	YES	Compound HET	nonsynonymous SNV	1390.16	PASS	66	99	33	99	25,24	49	99
SLC8A3	NM_183002	C1462T:p.R488C	YES	Compound HET	nonsynonymous SNV	1738.16	PASS	38	99	60	99	11,27	38	99
SH2D3C	NM_001142533	C1055T:p.A352V	YES	from mother	nonsynonymous SNV	1802,17	PASS	43	99	87	99	19,21	40	99
TRAF6	NM_004620	T608C:p.I203T	YES	from father	nonsynonymous SNV	1625,9	PASS	53	99	36	99	21,24	45	99
CD93	NM_012072	362_363insA:p.E121fs	No	de novo	frameshift insertion	395.09	VQSRTrancheINDEL99.90to100.00	36	99	63	99	90,32	122	99
AVEN	NM_020371	A638G:p.Q213R	YES	N/A	nonsynonymous SNV	1439,92	PASS	35	99	49	99	25,15	40	99
SUPT6H	NM_003170	C901T:p.R301C	YES	N/A	nonsynonymous SNV	1896,92	PASS	50	99	48	99	31,25	56	99
TRAK1	NM_001265608	2064_2066del:p.688_689del	NO	de novo	nonframeshift deletion	650.88	PASS	21	99	16	99	8,0,10	18	99
ZNF337	NM_015655	C1411T:p.Q471X	Yes	de novo	stopgain SNV	361.93	PASS	33	90	2	6	9,15	24	99
CEP295NL	NM_001243541	G286C:p.G96R	Yes	de novo	nonsynonymous SNV	71.89	PASS	18	51	0	.	4,4	8	99
TYRO3	NM_006293	1382_1382del:p.461_461del	NO	de novo	frameshift deletion	32,67	VQSRTrancheSNP99.00to99.90	25	72	113	99	97,21	118	99
PSMB11	NM_0010099780	del293	NO	de novo	frameshift deletion	426,64	PASS	11	0	26	66	17,15	32	99
NFKBIZ	NM_001005474	exon11:c.1635+2->ACTTTTAGAAAG(	YES	homozygous	splicing	15195.3	VQSRTrancheINDEL99.00to99.90	58	99	70	99	0,78	78	99
MAPKBP1	NM_014994	C3925G:p.P1309A	YES	de novo?	nonsynonymous SNV	976.19	PASS	/	/	36	99	22,29	51	99
MBD2	NM_003927	.345_346insGGC:p.S116delinsGS		de novo	nonframeshift insertion	217.5	PASS	/	/	6	12	4,7	11	99
PTGDR2	NM_004778	C1096T:p.P366S	Yes	de novo?	nonsynonymous SNV	47.86	PASS	/	/	3	9	1,2	3	34
VEZF1	NM_007146	1046_1047insGCA:p.Q349delinsQQ	YES	de novo	nonframeshift insertion	447.15	PASS	/	/	42	99	20,15	35	99
UNC5CL	NM_173561	C796T:p.R266C	YES	from mother	nonsynonymous SNV	437,17	PASS	28	81	42	99	13,8	21	99
SP1	NM_001251825	G5A:p.S2N	YES	from mother	nonsynonymous SNV	1011,17	PASS	32	87	36	99	17,16	33	99
TRPM5	NM_014555	exon9:c.G1241A:p.R414Q	NO	de novo	nonsynonymous SNV	34,12	PASS	2	0	2	6	4,2	6	61
CNTN5	NM_001243270	exon3:c.264_265ins	NO	de novo	stopgain	224,17	PASS	10	24	13	12	10,9	19	99
LYST	NM_000081	A3544G:p.M1182V,	YES	from father	nonsynonymous SNV	1301.16	PASS	56	99	39	99	24,15	39	99
LYST	NM_000081	C3083G:p.S1028C	YES	from mother	nonsynonymous SNV	2512.16	PASS	37	99	78	99	44,45	89	99
SELPLG	NM_001206609	exon2:c.452_481del:p.151_161del	NO	de novo	nonframeshift deletion	188,09	PASS	12	36	24	69	10,6	16	99
TYRO3	NM_006293	.1382_1382del:p.461_461del	NO	from mother	frameshift deletion	1133.86	VQSRTrancheINDEL99.00to99.90	25	72	113	99	97,21	118	99
IGFLR1	NM_024660	T883C:p.W295R	??	de novo?	nonsynonymous SNV	100.19	PASS	/	/	10	30	3,4	7	69
TBK1	NM_013254	1159+2_del	NO	de novo	splicing delection	40,29	VQSRTrancheINDEL99.00to99.90	22	54	26	72	19,3	22	81

Table 6.4: Number of variants after Whole exome sequencing and each filtering step in all family cases.

Filtering steps / Family Cases		Number of variants before filtering	Filter in Exonic and splicing variants	Filter out synonimous variants	Filter out duplications	Filter out common variants			Filter out higly mutated genes	Variants present in the patient
						1000Genomes_EUR	Exac_ALL	ESP6500siv2_ALL		
Case 1	Nº of variants	78183	30605	15758	12760	1885	1029	1005	758	/
Case 2	Nº of variants	75112	29709	15175	12277	1644	855	839	658	360
Case 3	Nº of variants	74177	29366	15003	12157	/	924	529	679	388
Case 4	Nº of variants	75912	29997	15551	12698	1393	553	528	400	194
Case 5	Nº of variants	75862	29890	15406	12470	1362	541	532	393	207
Case 6	Nº of variants	76369	29929	15445	12599	1809	943	930	679	/
Case 7	Nº of variants	39453	14722	7643	6030	763	379	375	308	159
Case 8	Nº of variants	75128	29564	15167	12295	1703	873	861	633	351
Case 9	Nº of variants	74775	29598	15246	12337	1525	698	686	78	65
Case 10	Nº of variants	97957	31948	16681	13811	1423	622	613	542	345
Case 11	Nº of variants	99040	31804	16442	13571	/	791	759	629	340
Case 12	Nº of variants	108636	34313	18256	14541	2060	1071	1049	819	456
Case 13	Nº of variants	74211	29361	15144	12387	1707	950	933	647	354
Case 14	Nº of variants	110364	34442	18145	14533	12649	1075	1015	782	334

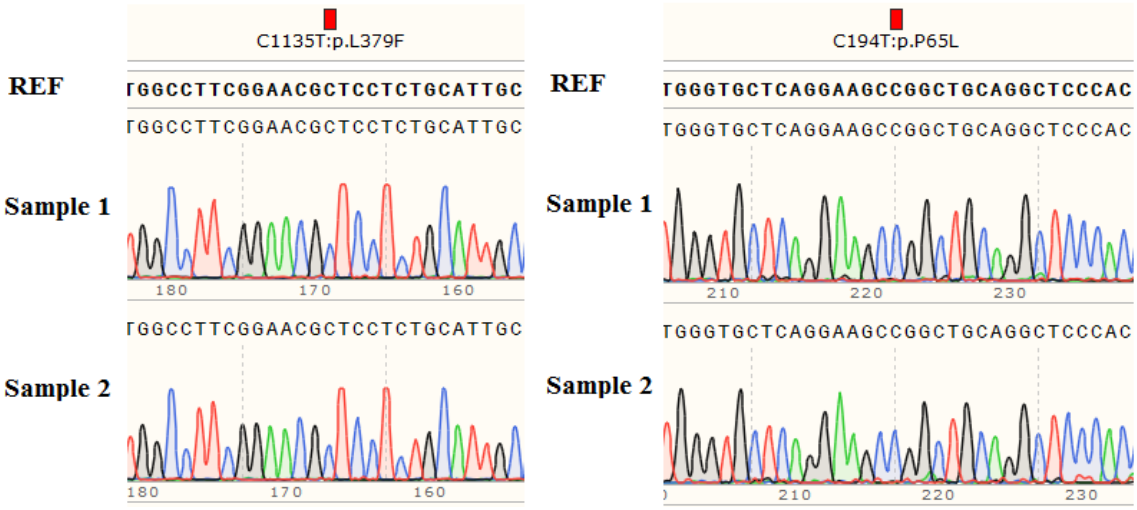


Figure 6.1. Sanger sequence of Juvenile Idiopathic arthritis patients for the two mutations we found. Due to the amount of samples we choose to only present two examples.

7. References

1. Sanabria, N., Goring, D., Nürnberger, T. & Dubery, I. Self/nonself perception and recognition mechanisms in plants: A comparison of self-incompatibility and innate immunity. *New Phytol.* **178,** 503–514 (2008).

2. Martinon, F., Mayor, A. & Tschopp, J. The inflammasomes: guardians of the body. *Annu. Rev. Immunol.* **27,** 229–265 (2009).

3. Doria, A. *et al.* Autoinflammation and autoimmunity: Bridging the divide. *Autoimmun. Rev.* **12,** 22–30 (2012).

4. Kawasaki, T., Kawai, T. & Akira, S. Recognition of nucleic acids by pattern-recognition receptors and its relevance in autoimmunity. *Immunol. Rev.* **243,** 61–73 (2011).

5. Mosser, D. M. & Edwards, J. P. Exploring the full spectrum of macrophage activation. *Nat. Rev. Immunol.* **8,** 958–969 (2008).

6. Ohta, Y. *et al.* Primitive synteny of vertebrate major histocompatibility complex class I and class II genes. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 4712–4717 (2000).

7. Martinon, F., Burns, K. & Tschopp, J. The Inflammasome: A molecular platform triggering activation of inflammatory caspases and processing of proIL-β. *Mol. Cell* **10,** 417–426 (2002).

8. Faustin, B. & Reed, J. C. Sunburned skin activates inflammasomes. *Trends Cell Biol.* **18,** 4–8 (2008).

9. Lamkanfi, M., Walle, L. Vande & Kanneganti, T. Deregulated inflammasome signaling in disease.pdf. **243,** 163–173 (2011).

10. Sims, J. E. & Smith, D. E. The IL-1 family: regulators of immunity. *Nat. Rev. Immunol.* **10,** 89–102 (2010).

11. Chung, Y. *et al.* Critical Regulation of Early Th17 Cell Differentiation by Interleukin-1 Signaling. *Immunity* **30,** 576–587 (2009).

12. McGonagle, D., Savic, S. & McDermott, M. F. The NLR network and the immunological disease continuum of adaptive and innate immune-mediated inflammation against self. *Semin. Immunopathol.* **29,** 303–313 (2007).

13. Shinkai, K., McCalmont, T. H. & Leslie, K. S. Cryopyrin-associated periodic syndromes and autoinflammation. *Clin. Exp. Dermatol.* **33,** 1–9 (2008).

14. Aksentijevich, I. *et al.* The clinical continuum of cryopyrinopathies: Novel CIAS1 mutations in North American patients and a new cryopyrin model. *Arthritis Rheum.* **56,** 1273–1285 (2007).

15. Masters, S. L. *et al.* Familial autoinflammation with neutrophilic dermatosis reveals a regulatory mechanism of pyrin activation. *Sci. Transl. Med.* **8,** 332ra45-332ra45 (2016).

16. Leadbetter, E. A. *et al.* Chromatin–IgG complexes activate B cells by dual engagement of IgM and Toll-like receptors. *Nature* **416,** 603–607 (2002).

17. Moghaddas, F. *et al.* Whole exome sequencing in systemic juvenile idiopathic arthritis. *Pediatr. Rheumatol.* **13,** O2 (2015).

18. Stock, C. J. W. *et al.* Comprehensive association study of genetic variants in the IL-1 gene family in systemic juvenile idiopathic arthritis. *Genes Immun* **9,** 349–357 (2008).

19. McGonagle, D. & McDermott, M. F. A proposed classification of the immunological diseases. *PLoS Med.* **3,** 1242–1248 (2006).

20. Pontillo, A. *et al.* Polimorphisms in inflammasome genes are involved in the predisposition to systemic lupus erythematosus. *Autoimmunity* **45,** 271–8 (2012).

21. Masters, S. L., Simon, A., Aksentijevich, I. & Kastner, D. L. Horror autoinflammaticus: the molecular pathophysiology of autoinflammatory disease (*). *Annu Rev Immunol* **27,** 621–668 (2009).

22. Touitou, I. The spectrum of Familial Mediterranean Fever (FMF) mutations. *Eur. J. Hum. Genet.* **9,** 473–83 (2001).

23. Peterson, P. & Peltonen, L. Autoimmune polyendocrinopathy syndrome type 1 (APS1) and AIRE gene: New views on molecular basis of autoimmunity. *J. Autoimmun.* **25,** 49–55 (2005).

24. Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. *Nat Rev Genet* **7,** 277–282 (2006).

25. Meyts, I. *et al.* Exome and genome sequencing for inborn errors of immunity. *J. Allergy Clin. Immunol.* **138,** 957–969 (2016).

26. Zen, M. *et al.* Clinical guidelines and definitions of autoinflammatory diseases: Contrasts and comparisons with autoimmunity- A comprehensive review. *Clin. Rev. Allergy Immunol.* **45,** 227–235 (2013).

27. Touitou, I. & Koné-Paut, I. Autoinflammatory diseases. *Best Pract. Res. Clin. Rheumatol.* **22,** 811–829 (2008).

28. Pétrilli, V., Dostert, C., Muruve, D. A. & Tschopp, J. The inflammasome: a danger sensing complex triggering innate immunity. *Curr. Opin. Immunol.* **19,** 615–622 (2007).

29. Sfriso, P. *et al.* Infections and autoimmunity: the multifaceted relationship. *J. Leukoc. Biol* **87,** 385–395 (2010).

30. Doria, A., Zampieri, S. & Sarzi-Puttini, P. Exploring the complex relationships between infections and autoimmunity. *Autoimmun. Rev.* **8,** 89–91 (2008).

31. Agmon-Levin, N., Paz, Z., Israeli, E. & Shoenfeld, Y. Vaccines and autoimmunity. *Nat Rev Rheumatol* **5,** 648–652 (2009).

32. Shoenfeld, Y. & Agmon-Levin, N. 'ASIA' - Autoimmune/inflammatory syndrome induced by adjuvants. *J. Autoimmun.* **36,** 4–8 (2011).

33. Kastner, D. L., Aksentijevich, I. & Goldbach-Mansky, R. Autoinflammatory Disease Reloaded: A Clinical Perspective. *Cell* **140,** 784–790 (2010).

34. Doria, A., Sherer, Y., Meroni, P. L. & Shoenfeld, Y. Inflammation and Accelerated Atherosclerosis: Basic Mechanisms. *Rheum. Dis. Clin. North Am.* **31,** 355–362 (2005).

35. Cho, J. H. & Gregersen, P. K. Genomics and the multifactorial nature of human autoimmune disease. *N. Engl. J. Med.* **365,** 1612–23 (2011).

36. Sfriso, P. *et al.* Blau syndrome, clinical and genetic aspects. *Autoimmun. Rev.* **12,** 44–51 (2012).

37. Muscari, I. *et al.* The diagnostic evaluation of patients with potential adult-onset autoinflammatory disorders: Our experience and review of the literature. *Autoimmun. Rev.* **12,** 10–13 (2012).

38. Gale, E. A. M. Type 1 diabetes in the young: the harvest of sorrow goes on. *Diabetologia* **48,** 1435–1438 (2005).

39. Soriano, A. & Manna, R. Familial Mediterranean fever: New phenotypes. *Autoimmun. Rev.* **12,** 31–37 (2012).

40. Bano, S. *et al.* Lupus erythematosus and the skin. *Clin. Exp. Rheumatol.* **24,** S26 (2006).

41. Drenth, J. P. H., Boom, B. W., Toonstra, J. & Van der Meer, J. W. M. Cutaneous manifestations and histologic findings in the hyperimmunoglobulinemia D syndrome. *Arch. Dermatol.* **130,** 59–65 (1994).

42. Punzi, L., Scanu, A., Ramonda, R. & Oliviero, F. Gout as autoinflammatory disease: New mechanisms for more appropriated treatment targets. *Autoimmun. Rev.* **12,** 66–71 (2012).

43. Touitou, I. New genetic interpretation of old diseases. *Autoimmun. Rev.* **12,** 5–9 (2012).

44. Uttenthal, B. J., Layton, D. M., Vyse, T. J. & Schreiber, B. E. The Wolf at the Door. *N. Engl. J. Med.* **366,** 2216–2221 (2012).

45. Farasat, S., Aksentijevich, I. & JR, T. Autoinflammatory diseases: Clinical and genetic advances. *Arch. Dermatol.* **144,** 392–402 (2008).

46. Gattorno, M. *et al.* Diagnosis and management of autoinflammatory diseases in childhood. *J Clin Immunol* **28,** (2008).

47. Nielsen, N. M. *et al.* Type 1 diabetes and multiple sclerosis. *Arch Neurol* **63,** 1001–1004 (2006).

48. Tettey, P., Simpson, S., Taylor, B. V & van der Mei, I. A. F. The co-occurrence of multiple sclerosis and type 1 diabetes: Shared aetiologic features and clinical implication for MS aetiology. *J. Neurol. Sci.* **348,** 126–131 (2015).

49. Morris, A., Voight, B. & Teslovich, T. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44,** 981–990 (2012).

50. Stranger, B. E., Stahl, E. A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187,** 367–383 (2011).

51. Mäkinen, V.-P. *et al.* Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet.* **10,** e1004502 (2014).

52. Hindorff, L. a *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–7 (2009).

53. Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* **39,** 596–604 (2007).

54. Parkes, M. *et al.* Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* **39,** 830–2 (2007).

55. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42,** 1118–25 (2010).

56. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science (80-.).* **291,** 1304 LP-1351 (2001).

57. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell* **155,** 27–38 (2013).

58. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).

59. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30,** 418–426 (2014).

60. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).

61. Valouev, A. *et al.* A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18,** 1051–1063 (2008).

62. Ju, J. *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 19635–19640 (2006).

63. Shendure, J. *et al.* Molecular biology: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (80-.).* **309,** 1728–1732 (2005).

64. Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27,** 847–850 (2009).

65. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).

66. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

67. Price, A. L., N.j.patterson, R.m.plenge, M.e.weinblatt & N.a.shadick. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet* **38,** 904–909 (2006).

68. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).

69. Fang, M., Abolhassani, H., Lim, C. K., Zhang, J. & Hammarstr??m, L. Next Generation Sequencing Data Analysis in Primary Immunodeficiency Disorders ??? Future Directions. *J. Clin. Immunol.* **36,** 68–75 (2016).

70. Conley, M. E. & Casanova, J. L. Discovery of single-gene inborn errors of immunity by next generation sequencing. *Curr. Opin. Immunol.* **30,** 17–23 (2014).

71. Geha, R. S. *et al.* Primary immunodeficiency diseases: An update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee. *Journal of Allergy and Clinical Immunology* **120,** 776–794 (2007).

72. Itan, Y. *et al.* HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. *BMC Genomics* **15,** 256 (2014).

73. Ciancanelli, M. J. *et al.* Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency. *Science (80-.).* **861,** 111–115 (2014).

74. Itan, Y. *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 13615–20 (2015).

75. Deschamps, M. *et al.* Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. *Am. J. Hum. Genet.* **98,** 5–21 (2016).

76. Alcaïs, A. *et al.* Life-threatening infectious diseases of childhood: Single-gene inborn errors of immunity? *Ann. N. Y. Acad. Sci.* **1214,** 18–33 (2010).

77. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91,** 839–848 (2012).

78. Cibulskis, K. *et al.* ContEst: Estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27,** 2601–2602 (2011).

79. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

80. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* **11,** 473–483 (2010).

81. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

82. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

83. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

84. Li, H. *et al.* Mapping short DNA sequencing reads and calling variants using mapping quality scores Mapping short DNA sequencing reads and calling variants using mapping quality scores. 1851–1858 (2008). doi:10.1101/gr.078212.108

85. Habegger, L. *et al.* Vat: A computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* **28,** 2267–2269 (2012).

86. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

87. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res.* **21,** 1529–1542 (2011).

88. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

89. Zhang, Y., Su, H. C. & Lenardo, M. J. Genomics is rapidly advancing precision medicine for immunological disorders. *Nat. Immunol.* **16,** 1001–4 (2015).

90. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12,** 628–40 (2011).

91. Itan, Y. & Casanova, J.-L. Can the impact of human genetic variations be predicted? *Proc. Natl. Acad. Sci.* **112,** 11426–11427 (2015).

92. Belkadi, A. *et al.* Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proc. Natl. Acad. Sci.* 201606460 (2016). doi:10.1073/pnas.1606460113

93. Veltman, J. a & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–75 (2012).

94. Adam, R. *et al.* Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis. *Am. J. Hum. Genet.* **99,** 337–351 (2016).

95. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15,** 121–32 (2014).

96. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

97. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133,** 1–9 (2014).

98. Scott, E. M. *et al.* Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat. Genet.* **48,** 1071–1076 (2016).

99. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–755 (2011).

100. Quintana-Murci, L. & Clark, A. G. Population genetic tools for dissecting innate immunity in humans. *Nat. Rev. Immunol.* **13,** 280–93 (2013).

101. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

102. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).

103. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

104. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46,** 310–315 (2014).

105. Itan, Y. *et al.* The mutation significance cutoff: gene-level thresholds for variant predictions. *Nat. Methods* **13,** 109–10 (2016).

106. Itan, Y. *et al.* The human gene connectome as a map of short cuts for morbid allele discovery. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 5558–63 (2013).

107. Casanova, J.-L., Conley, M. E., Seligman, S. J., Abel, L. & Notarangelo, L. D. Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies. *J. Exp. Med.* **211,** 2137–49 (2014).

108. Miosge, L. A. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E5189-98 (2015).

109. Narasimhan, V. M. *et al.* R es e a rc h | r e po r ts 24. **1320,** 1312–1316 (2014).

110. Holzelova, E. *et al.* Autoimmune lymphoproliferative syndrome with somatic Fas mutations. *N. Engl. J. Med.* **351,** 1409–1418 (2004).

111. Pessach, I. M. *et al.* Induced pluripotent stem cells: A novel frontier in the study of human primary immunodeficiencies. *J Allergy Clin Immunol* **127,** 1400–7 (2011).

112. Raje, N. *et al.* Utility of Next Generation Sequencing in Clinical Primary Immunodeficiencies. *Curr. Allergy Asthma Rep.* **14,** 1–13 (2014).

113. Notarangelo, L. D. & Sorensen, R. Is it necessary to identify molecular defects in primary immunodeficiency disease? *J. Allergy Clin. Immunol.* **122,** 1069–1073 (2008).

114. Fischer, A. Gene therapy: Myth or reality? *C. R. Biol.* **339,** 314–318 (2016).

115. Hubbard, N. *et al.* Targeted gene editing restores regulated CD40L expression and function in X-HIGM T cells. *Blood* **127,** blood-2015-11-683235 (2016).

116. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *NIH Public Access* **44,** 623–630 (2013).

117. Belinky, F. *et al.* PathCards: multi-source consolidation of human biological pathways. *Database* **2015,** bav006 (2015).

118. Rappaport, N. *et al.* in *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc., 2002). doi:10.1002/0471250953.bi0124s47

119. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45,** 580–585 (2013).

120. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

121. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33,** D514–D517 (2005).

122. Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44,** D7–D19 (2016).

123. Bell, J. R. A simple way to treat PCR products prior to sequencing using ExoSAP-IT?? *Biotechniques* **44,** 834 (2008).

124. Sato, D. *et al.* SHANK1 Deletions in Males with Autism Spectrum Disorder. *Am. J. Hum. Genet.* **90,** 879–887 (2017).

125. Wöhr, M., Roullet, F. I., Hung, A. Y., Sheng, M. & Crawley, J. N. Communication impairments in mice lacking shank1: Reduced levels of ultrasonic vocalizations and scent marking behavior. *PLoS One* **6,** (2011).

126. Pachlopnik Schmid, J. *et al.* Polymerase ε1 mutation in a human syndrome with facial dysmorphism, immunodeficiency, livedo, and short stature ('FILS syndrome'). *J. Exp. Med.* **209,** 2323 LP-2330 (2012).

127. Böckers, T. M. *et al.* Differential expression and dendritic transcript localization of Shank family members: identification of a dendritic targeting element in the 3′ untranslated region of Shank1 mRNA. *Mol. Cell. Neurosci.* **26,** 182–190 (2004).

128. Silverman, J. L. *et al.* Sociability and motor functions in Shank1 mutant mice. *Brain Res.* **1380,** 120–137 (2011).

129. Fütterer, A. *et al.* Dido gene expression alterations are implicated in the induction of hematological myeloid neoplasms. *J. Clin. Invest.* **115,** 2351–2362 (2005).

130. García-Domingo, D., Ramírez, D., González de Buitrago, G. & Martínez-A, C. Death inducer-obliterator 1 triggers apoptosis after nuclear translocation and caspase upregulation. *Mol. Cell. Biol.* **23,** 3216–3225 (2003).

131. Gomes, I. *et al.* Novel transcription factors in human CD34 antigen–positive hematopoietic cells. *Blood* **100,** 107 LP-119 (2002).

132. Murata, S. *et al.* Regulation of CD8⁺ T Cell Development by Thymus-Specific Proteasomes. *Science (80-.).* **316,** 1349 LP-1353 (2007).

133. Ohigashi, I. *et al.* A human *PSMB11* variant affects thymoproteasome processing and CD8(+) T cell production. *JCI Insight* **2,** e93664 (2017).

134. Matthews, S. P., McMillan, S. J., Colbert, J. D., Lawrence, R. A. & Watts, C. Cystatin F Ensures Eosinophil Survival by Regulating Granule Biogenesis. *Immunity* **44,** 795–806 (2017).

135. Schüttelkopf, A. W., Hamilton, G., Watts, C. & Van Aalten, D. M. F. Structural basis of reduction-dependent activation of human cystatin F. *J. Biol. Chem.* **281,** 16570–16575 (2006).

136. Froimchuk, E., Jang, Y. & Ge, K. Histone H3 lysine 4 methyltransferase *KMT2D*. *Gene* **627,** 337–342 (2017).

137. Lin, J. *et al.* Immunologic assessment and *KMT2D* mutation detection in Kabuki syndrome. *Clin. Genet.* **88,** 255–260 (2015).

138. Stagi, S., Gulino, A. V., Lapi, E. & Rigante, D. Epigenetic control of the immune system: a lesson from Kabuki syndrome. *Immunol. Res.* **64,** 345–359 (2016).

139. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42,** 790–793 (2010).

140. Jang, Y., Wang, C., Zhuang, L., Liu, C. & Ge, K. H3K4 methyltransferase activity is required for MLL4 protein stability. *J. Mol. Biol.* **429,** 2046–2054 (2017).

141. Ang, S.-Y. *et al. KMT2D* regulates specific programs in heart development via histone H3 lysine 4 di-methylation. *Development* **143,** 810–821 (2016).

142. Zhang, J. *et al.* Disruption of *KMT2D* perturbs germinal center B cell development and promotes lymphomagenesis. *Nat. Med.* **21,** 1190 (2015).

143. Rao, R. C. & Dou, Y. Hijacked in cancer: the MLL/KMT2 family of methyltransferases. *Nat. Rev. Cancer* **15,** 334 (2015).

144. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49,** 825–837 (2013).

145. Lee, J.-H. & Skalnik, D. G. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J. Biol. Chem.* **280,** 41725–41731 (2005).

146. Guo, C. *et al. KMT2D* maintains neoplastic cell proliferation and global histone H3 lysine 4 monomethylation. *Oncotarget* **4,** 2144 (2013).

147. Zhu, J. *et al.* Gain-of-function p53 mutants co-opt chromatin pathways to drive cancer growth. *Nature* **525,** 206–211 (2015).

148. Dahlin, A. M. *et al.* CCND2, CTNNB1, DDX3X, GLI2, SMARCA4, MYC, MYCN, PTCH1, TP53, and MLL2 gene variants and risk of childhood medulloblastoma. *J. Neurooncol.* **125,** 75–78 (2015).

149. Parsons, D. W. *et al.* The genetic landscape of the childhood cancer medulloblastoma. *Science* **331,** 435–439 (2011).

150. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).

151. Fuentes-Duculan, J. *et al.* Autoantigens *ADAMTSL5* and LL37 are significantly upregulated in active Psoriasis and localized with keratinocytes, dendritic cells and other leukocytes. *Exp. Dermatol.* **26,** 1075–1082 (2017).

152. Arakawa, A. *et al.* Melanocyte antigen triggers autoimmunity in human psoriasis. *J. Exp. Med.* **212,** 2203–2212 (2015).

153. Mobbs, J. I. *et al.* The molecular basis for peptide repertoire selection in the Human Leucocyte Antigen (HLA) C*06:02 molecule. *J. Biol. Chem.* jbc.M117.806976 (2017). doi:10.1074/jbc.M117.806976

154. Hafer, A. S. & Conran, R. M. Autosomal Recessive Polycystic Kidney Disease. *Academic Pathology* **4,** (2017).

155. Kumar, V., Abbas, A. K., Fausto, N. & Aster, J. C. *Robbins and Cotran Pathologic Basis of Disease, Professional Edition E-Book.* (Elsevier Health Sciences, 2014).

156. Sharp, A. *et al.* Comprehensive genomic analysis of *PKHD1* mutations in ARPKD

cohorts. *J. Med. Genet.* **42,** 336–349 (2005).

157. Gunay-Aygun, M. *et al. PKHD1* Sequence Variations in 78 Children and Adults with Autosomal Recessive Polycystic Kidney Disease and Congenital Hepatic Fibrosis. *Mol. Genet. Metab.* **99,** 160 (2010).

158. Sweeney, W. E. & Avner, E. D. Diagnosis and management of childhood polycystic kidney disease. *Pediatr. Nephrol.* **26,** 675–692 (2011).

159. Chalhoub, V., Abi-Rafeh, L., Hachem, K., Ayoub, E. & Yazbeck, P. Intracranial aneurysm and recessive polycystic kidney disease: The third reported case. *JAMA Neurol.* **70,** 114–116 (2013).

160. Elchediak, D. S., Cahill, A. M., Furth, E. E., Kaplan, B. S. & Hartung, E. A. Extracranial Aneurysms in 2 Patients with Autosomal Recessive Polycystic Kidney Disease. *Case Reports Nephrol. Dial.* **7,** 34–42 (2017).

161. Lanier, L. L. NKG2D receptor and its ligands in host defense. *Cancer Immunol. Res.* **3,** 575–582 (2015).

162. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

163. Cao, W. *et al.* Four novel ULBP splice variants are ligands for human NKG2D. *Int. Immunol.* **20,** 981–991 (2008).

164. Rigaud, S. *et al.* XIAP deficiency in humans causes an X-linked lymphoproliferative syndrome. *Nature* **444,** 110–114 (2006).

165. Sugimoto, M. *et al.* Regulation of CDK4 activity by a novel CDK4-binding protein, p34SEI-1. *Genes Dev.* **13,** 3027–3033 (1999).

166. Hong, S.-W. *et al.* p34^{SEI-1} Inhibits Apoptosis through the Stabilization of the X-Linked Inhibitor of Apoptosis Protein: p34^{SEI-1} as a Novel Target for Anti–Breast Cancer Strategies. *Cancer Res.* **69,** 741 LP-746 (2009).

167. Rothlin, C. V., Ghosh, S., Zuniga, E. I., Oldstone, M. B. A. & Lemke, G. TAM Receptors Are Pleiotropic Inhibitors of the Innate Immune Response. *Cell* **131,** 1124–1136 (2007).

168. Rothlin, C. V, Carrera-Silva, E. A., Bosurgi, L. & Ghosh, S. TAM Receptor Signaling in Immune Homeostasis. *Annu. Rev. Immunol.* **33,** 355–391 (2015).

169. Chan, P. Y. *et al.* The TAM family receptor tyrosine kinase *TYRO3* is a negative regulator of type 2 immunity. *Science (80-.).* **352,** 99–103 (2016).

170. Sundaram, K. *et al.* IκBζ Regulates Human Monocyte Pro-Inflammatory Responses Induced by Streptococcus pneumoniae. *PLoS One* **11,** e0161931 (2016).

171. Kim, Y. *et al.* The resident pathobiont Staphylococcus xylosus in *NFKBIZ*-deficient skin accelerates spontaneous skin inflammation. *Sci. Rep.* **7,** 6348 (2017).

172. Chapman, S. J. *et al. NFKBIZ* polymorphisms and susceptibility to pneumococcal disease in European and African populations. *Genes Immun.* **11,** 319–325 (2010).

173. Coto-Segura, P. *et al. NFKBIZ* in Psoriasis: Assessing the association with gene polymorphisms and report of a new transcript variant. *Hum. Immunol.* **78,** 435–440 (2017).

174. Lougaris, V. *et al.* NFKB1 regulates human NK cell maturation and effector functions. *Clin. Immunol.* **175,** 99–108 (2017).

175. Hollox, E. J. & Armour, J. A. L. Directional and balancing selection in human beta-defensins. *BMC Evol. Biol.* **8,** 113 (2008).

176. Shim, J., Lim, H., R.Yates III, J. & Karin, M. Nuclear Export of NF90 Is Required for Interleukin-2 mRNA Stabilization. *Mol. Cell* **10,** 1331–1344 (2017).

177. Greenfield, J. J. & High, S. The Sec61 complex is located in both the ER and the ER-Golgi intermediate compartment. *J. Cell Sci.* **112,** 1477 LP-1486 (1999).

178. Baron, L. *et al.* Mycolactone subverts immunity by selectively blocking the Sec61 translocon. *J. Exp. Med.* **213,** 2885 LP-2896 (2016).

179. Bolar, N. A. *et al.* Heterozygous Loss-of-Function *SEC61A1* Mutations Cause Autosomal-Dominant Tubulo-Interstitial and Glomerulocystic Kidney Disease with Anemia. *Am. J. Hum. Genet.* **99,** 174–187 (2016).

180. Fuhlbrigge, R. C., Kieffer, J. D., Armerding, D. & Kupper, T. S. Cutaneous lymphocyte antigen is a specialized form of PSGL-1 expressed on skin-homing T cells. *Nature* **389,** 978–981 (1997).

181. Nishimura, Y. *et al.* Human P-selectin glycoprotein ligand-1 is a functional receptor for enterovirus 71. *Nat Med* **15,** 794–797 (2009).

182. Sreeramkumar, V. *et al.* Neutrophils scan for activated platelets to initiate inflammation. *Science (80-.).* **346,** 1234 LP-1238 (2014).

183. Pérez-Frías, A. *et al.* Development of an autoimmune syndrome affecting the skin and internal organs in P-selectin glycoprotein ligand 1 leukocyte receptor-deficient mice. *Arthritis Rheumatol.* **66,** 3178–3189 (2014).

184. Angiari, S. *et al.* Regulatory T Cells Suppress the Late Phase of the Immune Response in Lymph Nodes through P-Selectin Glycoprotein Ligand-1. *J. Immunol.* **191,** 5489–5500 (2013).

185. Wang, H. *et al.* Psgl-1 Deficiency is Protective against Stroke in a Murine Model of Lupus. *Sci. Rep.* **6,** 28997 (2016).

186. Ishii, K. J. *et al.* TANK-binding kinase-1 delineates innate and adaptive immune responses to DNA vaccines. *Nature* **451,** 725–729 (2008).

187. Xiao, Y. *et al.* The kinase *TBK1* functions in dendritic cells to regulate T cell homeostasis, autoimmunity, and antitumor immunity. *J. Exp. Med.* **214,** 1493 LP-1507 (2017).

188. Zhu, M., John, S., Berg, M. & Leonard, W. J. Functional Association of *NMI* with Stat5 and Stat1 in IL-2- and IFN γ-Mediated Signaling. *Cell* **96,** 121–130 (1999).

189. Welsch, K., Holstein, J., Laurence, A. & Ghoreschi, K. Targeting JAK/STAT signalling in inflammatory skin diseases with small molecule inhibitors. *Eur. J. Immunol.* **47,** 1096–1107 (2017).

190. Milner, J. D. *et al.* Early-onset lymphoproliferation and autoimmunity caused by germline STAT3 gain-of-function mutations. *Blood* **125,** 591–599 (2015).

191. Jin, Y. *et al.* Whole Genome Sequencing Identifies Novel Compound Heterozygous Lysosomal Trafficking Regulator Gene Mutations Associated with Autosomal Recessive Chediak-Higashi Syndrome. *Scientific Reports* **7,** (2017).

192. Gil-Krzewska, A. *et al.* Chediak-Higashi syndrome: *LYST* domains regulate exocytosis of lytic granules, but not cytokine secretion by NK cells. *The Journal of allergy and clinical immunology* **137,** 1165–1177 (2016).

193. Nagle, D. L. *et al.* Identification and mutation analysis of the complete gene for Chediak-Higashi syndrome. *Nat Genet* **14,** 307–311 (1996).

194. Ji, X., Chang, B., Naggert, J. K. & Nishina, P. M. in *Retinal Degenerative Diseases: Mechanisms and Experimental Therapy* (eds. Bowes Rickman, C. et al.) 745–750 (Springer International Publishing, 2016). doi:10.1007/978-3-319-17121-0_99

195. Westphal, A. *et al.* Lysosomal trafficking regulator *LYST* links membrane trafficking to toll-like receptor–mediated inflammatory responses. *J. Exp. Med.* **214,** 227–244 (2017).

196. Al-Shami, A. *et al.* The Adaptor Protein *SH2D3C* Is Critical for Marginal Zone B Cell Development and Function. *J. Immunol.* **185,** 327 LP-334 (2010).

197. Park, E. *et al.* MST1 deficiency promotes B cell responses by CD4+ T cell-derived IL-4, resulting in hypergammaglobulinemia. *Biochem. Biophys. Res. Commun.* **489,** 56–62 (2017).

198. Browne, C. D. *et al.* SHEP1 partners with CasL to promote marginal zone B-cell maturation. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 18944–18949 (2010).

199. Sakakibara, A., Hattori, S., Nakamura, S. & Katagiri, T. A novel hematopoietic adaptor protein, Chat-H, positively regulates T cell receptor-mediated interleukin-2 production by Jurkat cells. *J. Biol. Chem.* **278,** 6012–6017 (2003).

200. Wu, L. C. & Zarrin, A. A. The production and regulation of IgE by the immune system. *Nat Rev Immunol* **14,** 247–259 (2014).

201. Zhang, J., Xu, L.-G., Han, K.-J. & Shu, H.-B. Identification of a ZU5 and death domain-containing inhibitor of NF-κB. *J. Biol. Chem.* **279,** 17819–17825 (2004).

202. Geldmeyer-Hilt, K. *et al.* 1,25-dihydroxyvitamin D3 impairs NF-κB activation in human naïve B cells. *Biochem. Biophys. Res. Commun.* **407,** 699–702 (2011).

203. Heinz, L. X. *et al.* The death domain-containing protein *UNC5CL* is a novel MyD88-independent activator of the pro-inflammatory IRAK signaling cascade. *Cell Death Differ.* **19,** 722–731 (2012).

204. Rigbolt, K. T. G. *et al.* System-Wide Temporal Characterization of the Proteome and Phosphoproteome of Human Embryonic Stem Cell Differentiation. *Sci. Signal.* **4,** rs3 LP-rs3 (2011).

205. Olsen, J. V *et al.* Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis. *Sci. Signal.* **3,** ra3 LP-ra3 (2010).

206. Qian, Y. *et al.* aPKC-ι/P-*SP1*/Snail signaling induces epithelial-mesenchymal transition and immunosuppression in cholangiocarcinoma. *Hepatology* n/a-n/a doi:10.1002/hep.29296

207. Deniaud, E. *et al.* Overexpression of *SP1* transcription factor induces apoptosis. *Oncogene* **25,** 7096–7105 (2006).

208. Islam, M. A. *et al.* PBMC transcriptome profiles identifies potential candidate genes and functional networks controlling the innate and the adaptive immune response to PRRSV vaccine in Pietrain pig. *PLoS ONE* **12,** (2017).

209. Yamamura, K. *et al.* The transcription factor EPAS1 links DOCK8 deficiency to atopic skin inflammation via IL-31 induction. *Nature Communications* **8,** (2017).

210. Zhang, Q. *et al.* Combined Immunodeficiency Associated with DOCK8 Mutations. *N. Engl. J. Med.* **361,** 2046–2055 (2009).

211. Engelhardt, K. R. *et al.* The extended clinical phenotype of 64 patients with DOCK8 deficiency. *J. Allergy Clin. Immunol.* **136,** 402–412 (2015).

212. Kim, Y. *et al.* Histone deacetylase 3 mediates allergic skin inflammation by regulating expression of MCP1 protein. *J. Biol. Chem.* **287,** 25844–25859 (2012).

213. Zheng, H. *et al. TMED3* promotes hepatocellular carcinoma progression via IL-11/STAT3 signaling. **6,** 37070 (2016).

214. Arora, M. *et al.* Gastrointestinal Manifestations of STAT3-Deficient Hyper-IgE Syndrome. *J. Clin. Immunol.* **37,** 695–700 (2017).

215. Deverrière, G. *et al.* Life-Threatening Pneumopathy and U urealyticum in a STAT3-Deficient Hyper-IgE Syndrome Patient. *Pediatrics* **139,** (2017).

216. Cookson, W. O. C. M. & Moffatt, M. F. The genetics of atopic dermatitis. *Curr. Opin. Allergy Clin. Immunol.* **2,** 383–387 (2002).

217. Palmer, C. N. A. *et al.* Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat Genet* **38,** 441–446 (2006).

218. Smith, F. J. D. *et al.* Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. *Nat Genet* **38,** 337–342 (2006).

219. Weidinger, S. *et al.* Loss-of-function variations within the filaggrin gene predispose for atopic dermatitis with allergic sensitizations. *J. Allergy Clin. Immunol.* **118,** 214–219 (2006).

220. Sandilands, A. *et al.* Comprehensive analysis of the gene encoding filaggrin uncovers prevalent and rare mutations in ichthyosis vulgaris and atopic eczema. *Nat Genet* **39,** 650–654 (2007).

221. Fallon, P. G. *et al.* A homozygous frameshift mutation in the mouse Flg gene facilitates enhanced percutaneous allergen priming. *Nat Genet* **41,** 602–608 (2009).

222. Stoffels, M. *et al.* Update on *CECR1* molecular diagnostics: new mutations in the deficiency of ADA2 (DADA2) and the North American polyarteritis nodosa (PAN) cohort. *Pediatr. Rheumatol.* **13,** O20 (2015).

223. BATU, E. D. *et al.* A Case Series of Adenosine Deaminase 2-deficient Patients Emphasizing Treatment and Genotype-phenotype Correlations. *J. Rheumatol.* **42,** 1532 LP-1534 (2015).

224. Navon Elkan, P. *et al.* Mutant Adenosine Deaminase 2 in a Polyarteritis Nodosa Vasculopathy. *N. Engl. J. Med.* **370,** 921–931 (2014).

225. Garg, N. *et al.* Novel adenosine deaminase 2 mutations in a child with a fatal vasculopathy. *Eur. J. Pediatr.* **173,** 827–830 (2014).

226. Caorsi, R. *et al.* Prevalence of *CECR1* mutations in pediatric patients with polyarteritis nodosa, livedo reticularis and/or stroke. *Pediatr. Rheumatol.* **13,** 81DUMMY (2015).

227. Belot, A. *et al.* Mutations in *CECR1* associated with a neutrophil signature in peripheral blood. *Pediatr. Rheumatol.* **12,** 44 (2014).

228. Dai, Y. *et al.* A2B Adenosine Receptor-Mediated Induction of IL-6 Promotes CKD. *J. Am. Soc. Nephrol.* **22,** 890–901 (2011).

229. Mutant ADA2 in Vasculopathies. *N. Engl. J. Med.* **371,** 478–481 (2014).

230. Diederichs, S. *et al.* The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol. Med.* (2016).

231. Caorsi, R. *et al.* ADA2 deficiency (DADA2) as an unrecognised cause of early onset polyarteritis nodosa and stroke: a multicentre national study. *Ann. Rheum. Dis.* (2017).

232. Uettwiller, F. *et al.* ADA2 deficiency: case report of a new phenotype and novel mutation in two sisters. *RMD Open* **2,** (2016).

233. Simbolo, M. *et al.* DNA Qualification Workflow for Next Generation Sequencing of Histopathological Samples. *PLoS One* **8,** e62692 (2013).

234. Blattler, A. *et al.* Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biol.* **15,** 469 (2014).

235. Thyssen, J. *et al. Filaggrin loss-of-function mutation R501X and 2282del4 carrier status is associated with fissured skin on the hands: Results from a cross-sectional population study*. *The British journal of dermatology* **166,** (2011).

236. Gruber, R. *et al.* Filaggrin mutations p.R501X and c.2282del4 in ichthyosis vulgaris. *Eur. J. Hum. Genet.* **15,** 179–184 (2007).

237. Wang, Q., Shashikant, C. S., Jensen, M., Altman, N. S. & Girirajan, S. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci. Rep.* **7,** 885 (2017).

238. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41,** 1061–1067 (2009).

239. Ekelund, E. *et al.* Loss-of-function Variants of the Filaggrin Gene are Associated with Atopic Eczema and Associated Phenotypes in Swedish Families. *Acta Dermato-Venereologica* **88,** 15–19

240. O'Regan, G. M. & Irvine, A. D. The role of filaggrin loss-of-function mutations in atopic dermatitis. *Curr. Opin. Allergy Clin. Immunol.* **8,** (2008).

241. Ogilvie, E. M. *et al.* The −174G allele of the interleukin-6 gene confers susceptibility to systemic arthritis in children: A multicenter study using simplex and multiplex juvenile idiopathic arthritis families. *Arthritis Rheum.* **48,** 3202–3206 (2003).

242. Berkun, Y. & Padeh, S. Environmental factors and the geoepidemiology of juvenile idiopathic arthritis. *Autoimmun. Rev.* **9,** A319–A324 (2010).

243. Chistiakov, D. a, Savost'anov, K. V & Baranov, A. a. Genetic background of juvenile idiopathic arthritis. *Autoimmunity* **47,** 351–60 (2014).

244. Tozawa, Y., Fujita, S., Abe, S., Kitamura, K. & Kobayashi, I. Radiological improvement by tocilizumab in polyarticular juvenile idiopathic arthritis. *Pediatr. Int.* **57,** 307–310 (2015).

245. Okuda, Y. & Takasugi, K. Successful use of a humanized anti–interleukin-6 receptor antibody, tocilizumab, to treat amyloid A amyloidosis complicating juvenile idiopathic arthritis. *Arthritis Rheum.* **54,** 2997–3000 (2006).

246. Prakken, B., Albani, S. & Martini, A. Juvenile idiopathic arthritis. *Lancet* **377,** 2138–2149 (2011).

247. Fishman, D. *et al.* The effect of novel polymorphisms in the interleukin-6 (IL-6) gene on IL-6 transcription and plasma IL-6 levels, and an association with systemic-onset juvenile chronic arthritis. *J. Clin. Invest.* **102,** 1369–1376 (1998).

248. Herlin, M., Petersen, M. B. & Herlin, T. Update on Genetic Susceptibility and Pathogenesis in Juvenile Idiopathic Arthritis. *Eur. Med. J.* **1,** 73–83 (2014).

249. Galicia, J. C. *et al.* Polymorphisms in the IL-6 receptor (IL-6R) gene: strong evidence that serum levels of soluble IL-6R are genetically influenced. *Genes Immun.* **5,** 513–6 (2004).

250. Lamas, J. R. *et al.* Influence of *IL6R* rs8192284 polymorphism status in disease activity in rheumatoid arthritis. *J. Rheumatol.* **37,** 1579–1581 (2010).

251. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat. Genet.* **45,** 664–669 (2013).

252. Hersh, A. O. & Prahalad, S. Immunogenetics of juvenile idiopathic arthritis: A comprehensive review. *J. Autoimmun.* **64,** 113–124 (2015).

253. Roach, J. C. *et al.* Genetic Mapping at 3-Kilobase Resolution Reveals Inositol 1,4,5-Triphosphate Receptor 3 as a Risk Factor for Type 1 Diabetes in Sweden. *Am. J. Hum. Genet.* **79,** 614–627 (2017).

254. Oishi, T. *et al.* A functional SNP in the NKX2.5-binding site of ITPR3 promoter is associated with susceptibility to systemic lupus erythematosus in Japanese population. *J Hum Genet* **53,** 151–162 (2008).

255. Beishline, K. & Azizkhan-Clifford, J. *SP1* and the 'hallmarks of cancer'. *FEBS J.* **282,** 224–258 (2015).