Exome Sequencing in Gastrointestinal Food Allergies Induced by Multiple Food Proteins

Alba Sanchis Juan

Department of Biotechnology
Universitat Politècnica de València

Supervisors: Dr. Javier Chaves Martínez
Dr. Ana Bárbara García García
Dr. Pablo Marín García

This dissertation is submitted for the degree of
Doctor of Philosophy

September 2019

If you know you are on the right track, if you have this inner knowledge,
then nobody can turn you off... no matter what they say.

Barbara McClintock

INCLIVA | VLC
Instituto de Investigación Sanitaria

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Declaration

FELIPE JAVIER CHAVES MARTÍNEZ, PhD in Biological Sciences, ANA BÁRBARA GARCÍA GARCÍA, PhD in Pharmacy, and PABLO MARÍN GARCÍA, PhD in Biological Sciences,

CERTIFY:

That the work "Exome Sequencing in Gastrointestinal Food Allergies Induced by Multiple Food Proteins" has been developed by Alba Sanchis Juan under their supervision in the INCLIVA Biomedical Research Institute, as a Thesis Project in order to obtain the degree of PhD in Biotechnology at the Universitat Politècnica de València.

Dr. F. Javier Chaves Martínez

Dr. Ana Bárbara García García

Dr. Pablo Marín García

Acknowledgements

Reaching the end of this journey, after so many ups and downs, I cannot but express my gratitude to all those who supported me and helped me through this challenging but rewarding experience.

First and foremost, I would like to thank my supervisors, Dr. Javier Chaves Martinez, Dr. Ana Barbara García García and Dr. Pablo Marín García. They gave me the opportunity to work with them when I was still an undergraduate student, and provided me the opportunity to contribute to several exciting projects since then. They also encouraged me into the clinical genomics field and provided me guidance and advice throughout the last four years. I am also grateful to my mentor Prof. José Gadea Vacas for his support during this time.

Thank you to the additional members of the Genotyping and Genetic Diagnostics Unit from the INCLIVA research foundation. I very much enjoyed being part of this big team, and learned a lot from it. Thank you to the Garmitxa organisation for generously funding this work and the first year of my PhD. I am also grateful to all patients who participated in the study, without whose generosity none of this work would have been possible.

I am extremely grateful to my superiors at University of Cambridge, Prof. Lucy Raymond and Dr. Keren Carss. They provided a huge

amount of support during my project, and were very understanding with me combining my PhD with my job at University of Cambridge. They encouraged, mentored and inspired me throughout my career. I have been very fortunate to work with them. Thank you to Dr. Courtney French for her most valuable comments to this dissertation. I am also grateful to all my colleagues from University of Cambridge and Universitat Politècnica de València, specially Dr. Detelina Grozeva, Dr. Beatriz Ballester Llobell and Dr. Marcia Hasenahuer, for their motivation and support especially in the most difficult moments.

I thank Prof. Willem Ouwehand and the NIHR BioResource for providing me the opportunity to work in this stimulating group. I am grateful for all the opportunities that had arisen during this time.

Last, but not least, I would like to thank my parents and my sister for their unconditional support, for their encouragement and for believing in me. I also thank my friends, but especially I am grateful to my endlessly supportive partner Mateus Patricio, who motivated me to go every weekend to the library, rain or shine, to make this thesis possible.

Agradecimientos

Llegando al final de este viaje, después de tantos altibajos, no puedo dejar de expresar mi gratitud a todos aquellos que han estado conmigo a lo largo esta dura pero gratificante experiencia.

En primer lugar, me gustaría agradecer a mis supervisores, el Dr. Javier Chaves Martínez, la Dra. Ana Barbara García García y el Dr. Pablo Marín García, por abrirme las puertas a su grupo cuando aún era estudiante. Gracias por introducirme en el campo de la genómica clínica y por brindarme orientación y asesoramiento durante los últimos cuatro años. Asimismo, me gustaria agradecer a mi tutor, el Profesor José Gadea Vacas, por todo su apoyo durante este tiempo.

Gracias a los compañeros de la Unidad de Genotipado y Diagnóstico Genético de la fundación de investigación INCLIVA. He disfrutado de ser parte de este gran grupo y he aprendido mucho de todos vosotros. Gracias a la organización Garmitxa por financiar generosamente este trabajo y el primer año de mi doctorado. También quiero agradecer a todos los pacientes que han participado en el estudio, sin cuya generosidad no hubiera sido posible.

Gracias a mis supervisoras en la Universidad de Cambridge, la Profesora Lucy Raymond y la Dr. Keren Carss, por su enorme apoyo durante mi proyecto y por ser tan comprensivas conmigo compaginando mi doc-

torado con mi trabajo en la Universidad de Cambridge. Me animaron, me orientaron y me inspiraron a lo largo de mi carrera. He sido muy afortunada de trabajar con ellas.

Gracias a la Dr. Courtney French por los valiosos comentarios en el desarrollo de esta tesis. También estoy agradecida a todos mis compañeros de la Universidad de Cambridge y de la Universitat Politècnica de València, especialmente a la Dra. Detelina Grozeva, a la Dra. Beatriz Ballester Llobell y a la Dra. Marcia Hasenahuer, por su motivación y apoyo, especialmente en los momentos más difíciles.

Gracias al Profesor Willem Ouwehand y al NIHR BioResource por brindarme la oportunidad de trabajar en este estimulante grupo. Estoy agradecida por todas las oportunidades que han surgido durante este tiempo.

Por último, y más importante, me gustaría agradecer a mis padres y a mi hermana por su apoyo incondicional y por creer en mí. También agradezco a mis amigos, pero especialmente a Mateus Patricio, que me apoyó incesantemente y me motivó a ir todos los fines de semana a la biblioteca, llueva o truene, para hacer posible esta tesis.

Abstract

The study of genetics has been making significant progress towards understanding the causes of rare and common disease during the past decades. Across a wide range of disorders, there have been hundreds of associated loci identified and associated with multiple disorders. Now, with the advent of next-generation sequencing technologies, we are able to interrogate the contribution of high and low frequency variation to disease in a high throughput manner. This provides an opportunity to investigate the role of rare variation in complex disease risk, potentially offering insights into disease pathogenesis and biological mechanisms.

In this thesis, it has been assessed the use of whole-exome sequencing technology to investigate the role of rare variation in a complex disease, gastrointestinal food allergy induced by multiple food proteins. For that, a cohort of 31 individuals (eight affected and 23 non-affected) from seven different families was whole exome sequenced. Data obtained from multiple sequencing systems and libraries were analysed, and a workflow was developed, focusing on a comprehensive quality control to maximise the number of real positive calls. Different types of genome variations were investigated, including single nucleotide variants, insertions/deletions, copy number variants and HLA haplotypes. By approaching different methods of variant filtering, a set of rare variants

that could be associated with the disease was identified. The possible role of these candidate variants in the pathogenesis of gastrointestinal food allergies was also discussed.

These results reveal important insights into the genetic architecture of gastrointestinal food allergies and lead to additional lines of investigation that will be required in order to fully understand the genetic basis of this disease.

Resumen

Durante las últimas décadas, se han realizado importantes avances en el estudio de las causas genéticas de enfermedades raras y comunes, donde un gran número de variantes han sido identificadas y asociadas a múltiples enfermedades. Con las tecnologías de secuenciación de nueva generación, hoy en día somos capaces de investigar, con un alto rendimiento, la contribución de variantes de alta y baja frecuencia a distintos tipos de enfermedades, permitiéndonos así estudiar su importancia en el desarrollo de las mismas.

En ésta tesis se ha utilizado la secuenciación del exoma como tecnología para el estudio de variantes raras en una enfermedad compleja, la alergia gastrointestinal inducida por múltiples alimentos. Para ello, se realizó la secuenciación del exoma completo de una cohorte de 31 individuos (ocho afectados y 23 no afectados) provenientes de siete familias diferentes. Se desarrolló un flujo de trabajo para procesar los datos generados a partir de diferentes librerías e instrumentos de secuenciación, así como un control de calidad exhaustivo con el fin de maximizar el número de variantes de alta calidad. Diferentes tipos de mutaciones fueron investigadas, incluyendo polimorfismos de nucleótido único, inserciones/deleciones, variantes del número de copia y haplotipos HLA, y se realizaron diferentes métodos de filtrado para su interpretación.

Finalmente, se encontraron una serie de mutaciones que podrían estar asociadas con la enfermedad y se describe su posible papel en la patogénesis de las alergias gastrointestinales. Los resultados de esta tesis suponen importantes avances en el estudio de la compleja arquitectura genética de las alergias gastrointestinales y abren las puertas a futuras líneas de investigación, que serán necesarias para entender completamente las bases genéticas de esta enfermedad.

Resum

Durant les últimes dècades, s'han realitzat importants avanços en l'estudi de les causes genètiques de malalties rares i comunes, on un gran nombre de variants han sigut identificades i associades a múltiples malalties. Amb les tecnologies de seqüenciació de nova generació, avui en dia som capaços d'investigar, amb un alt rendiment, la contribució de variants d'alta i baixa freqüència a diferents tipus de malalties, permetent-nos així estudiar la seva importància en el desenvolupament de les mateixes.

En aquesta tesis s'ha utilitzat la seqüenciació del exoma com a tecnologia per a l'estudi de variants rares en una malaltia complexa, l'al·lèrgia gastrointestinal induïda per múltiples aliments. Per això, es va realitzar la seqüenciació del exoma complet d'una cohort de 31 individus (vuit afectats i 23 no afectats) provinents de set famílies diferents. Es va desenvolupar un flux de treball per a processar les dades generades a partir de diferents llibreries e instruments de seqüenciació, així com un control de qualitat exhaustiu amb la fi de maximitzar el nombre de variants d'alta qualitat. Diferents tipus de mutacions foren investigades, incloïent polimorfismes de nucleòtid únic, insercions/delecions, variants del nombre de còpia i haplotips HLA, i es realitzaren diferent mètodes de filtrat per a la seva interpretació.

Finalment, es trobaren una sèrie de mutacions que podrien estar associades amb la malaltia i es descriu el seu possible paper en la patogènesis de les al·lèrgies gastrointestinals. Els resultats d'aquesta tesis suposen importants avanços en l'estudi de la complexa arquitectura genètica de les al·lèrgies gastrointestinals i obrin les portes a futures línies d'investigació, que seran necessàries per entendre completament les bases genètiques d'aquesta malaltia.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AD Autosomal dominant

AR Autosomal recessive

BAM Binary alignment map

CNV Copy number variant

DNA Deoxyribonucleic acid

dNTP Deoxynucleotide triphosphates

EoE Eosinophilic esophagitis

ER Endoplasmic reticulum

FMT Faecal microbiota transplant

FPE Food protein-induced enteropathy

FPIAP Food protein-induced allergic proctocolitis

FPIES Food protein-induced enterocolitis syndrome

GATK Genome analysis toolkit

GI Gastrointestinal

GWAS Genome-wide association study

HLA Human leukocyte antigen

HMM Hidden markov model

HPO Human Phenotype Ontology

IBD Inflammatory bowel disease

IFNγ Interferon γ

IgA Immunoglobulin A

IgE Immunoglobulin E

IGV Integrative genomics viewer

IL Interleukin

IOIBD Infantile onset inflammatory bowel disease

LOF Loss-of-function

LPS Lipopolysaccharide

MAF Minor allele frequency

MAPK Mitogen-activated protein kinases

MOI Mode of inheritance

NGS Next generation sequencing

OAS Oral allergy syndrome

OIT Oral immunotherapy

SAM Sequence alignment map

SBS Sequencing by synthesis

SNP Single nucleotide polymorphism

SNV Single nucleotide variant

STR Short tandem repeat

SV Structural variant

T2D Type-2 Diabetes

TGF Transforming growth factor

TNF Tumour necrosis factor

TPM Transcripts per million

VCF Variant call format

VEP Variant effect predictor

WES Whole-exome sequencing

WGS Whole-genome sequencing

XHMM eXome-hidden markov model

XL X-linked

Chapter 1

Introduction

Connecting phenotype with genotype is a fundamental aim of genetics. The knowledge of mutant alleles responsible for a disease aids in predicting the prognosis of an affected individual and provides a better selection of the therapeutic strategies.

Next generation sequencing has been invaluable in the elucidation of the genetic aetiology of many human disorders in recent years, providing researchers with the opportunity to interrogate large numbers of candidate genes in order to establish key components of disease. In particular, exome sequencing offers an efficient method to investigate disease, as the exome only constitutes 1-2% of the whole genome, and contains the majority of known disease-causing variants. This study explores the potential of whole exome sequencing to elucidate the genetic basis of gastrointestinal (GI) food allergies induced by multiple food proteins.

A GI food allergy is a type of adverse immune response where exposure to certain food(s) induces allergy rather than tolerance, mainly affecting the GI system. Although this is a complex disorder, genetic

factors play an important role and are one of the major risk factors of its development [1–3]. The study of common variants across the genome by genome-wide association studies found an association with allergies and genetic variations in genes that play crucial roles in immune responses, such as interleukins, genes from the JAK-STAT signalling pathway, or genes that play an important role in skin barrier, such as *FLG*, which encodes for the filaggrin precursor [1]. They also found that the Human Leukocyte Antigen (HLA) locus plays a major role in immune regulation [4], and it has been significantly associated with multiple immune disorders, including allergic diseases [5, 6].

However, very few well-grounded associations have been established for GI food allergies. The fact that candidate gene studies have been carried out for decades without consistent findings supports a possible role for rare variation. Still, there have been no reports of rare variants associated with GI food allergy via next generation sequencing.

In this chapter, it is explained the value of whole-exome sequencing for the discovery of genetic rare variants and its utility for the study of GI food allergies. Finally, the classification, pathogenesis and genetics of this disorder is described and a case study of seven families with individuals affected with severe GI food allergy is presented.

1.1 The genetics of disease

In an oversimplified categorisation, human diseases can be separated into Mendelian or complex disorders, depending on the underlying genetic cause.

A disease is termed to be Mendelian if it segregates according to Mendel's laws of inheritance: dominant, recessive or X-linked. These are usually caused by highly penetrant mutations, meaning that almost all individuals who carry a disease-causing mutation express the phenotype. Mendelian diseases are usually caused by very rare mutations in one or very few genes, and that is why they are often referred as monogenic or oligogenic diseases, respectively (Figure 1.1). The frequency of these mutations tends to be very low because they undergo negative selection due to the highly deleterious effects. Although many specific disorders are very rare, altogether, Mendelian disorders affect between 5-10% of the population which encompasses millions of people in the world. There are at least 6,000 disorders in the Online Mendelian Inheritance in Man database (http://www.omim.org) and 4,000 genes with disease-causing mutations.

In contrast, diseases which do not follow a classic Mendelian pattern of inheritance are complex diseases (also called polygenic or multifactorial). These do not have a single cause, but several of them have been shown to have a genetic component from twin and family studies [7, 8]. These disorders can be the result of incomplete penetrance, poligenic risks or mutations in multiple genes that can be present at higher population frequencies (Minor Allele Frequency (MAF) >5%). The variants associated with complex disorders do not directly cause disease individually, but influence disease risk [9, 10].

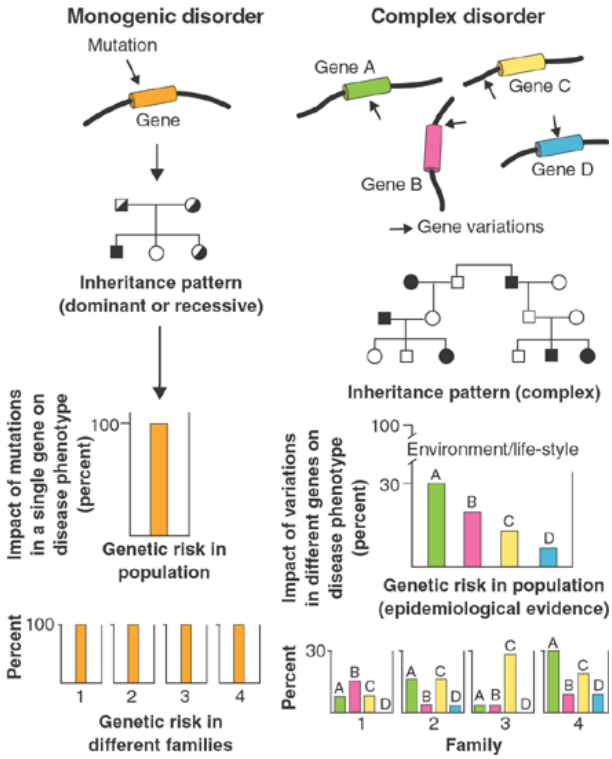


Fig. 1.1 Inheritance of monogenic and complex disorders. In monogenic diseases, mutations in a single gene, often highly penetrant, are sufficient to produce the clinical phenotype and to cause the disease. In complex disorders, variations in a number of genes encoding different proteins result in a genetic predisposition to a clinical phenotype. Pedigrees reveal no Mendelian inheritance pattern, and gene mutations are often neither sufficient nor necessary to explain the disease phenotype. Incomplete penetrance, environment and life-style are major contributors to the pathogenesis of complex diseases. Adapted from [10].

Over the past decade, Genome Wide Association Studies (GWAS) have been developed upon the common disease-common variant hypothesis. This argues that "genetic variations with appreciable frequency in the population at large, but relatively low penetrance are the major contributors to genetic susceptibility to common diseases" [11]. These studies

have played a critical role using an advanced high-density genotyping approach to characterise the contribution of single-nucleotide polymorphisms (SNPs) scattered across the genome to the genetic susceptibility of individuals. GWAS have identified hundreds of common risk alleles for complex human diseases, such as osteoporosis, autoimmune diseases and diabetes [12–14].

Even though these studies have provided several biological insights, most common variants have only subtle functional consequences and therefore only explain a low percentage of the genetic risk component of disease. For example, GWAS in type-2 diabetes (T2D) have identified more than 70 loci at genome-wide significance, but that only explains about 11% of T2D heritability [15]. Similarly, around 70 loci have been associated with Crohn's disease but these only explain 23% of heritability. This problem is referred to as "missing heritability" (Figure 1.2) [16].

In order to solve the question of the missing heritability, the "common disease-rare variant" hypothesis was raised, suggesting that multiple rare DNA sequence variations, each with relatively high penetrance, are the major contributors to genetic susceptibility to common diseases [11]. Since then, the focus on the discovery of rare variants with an important effect was inevitable.

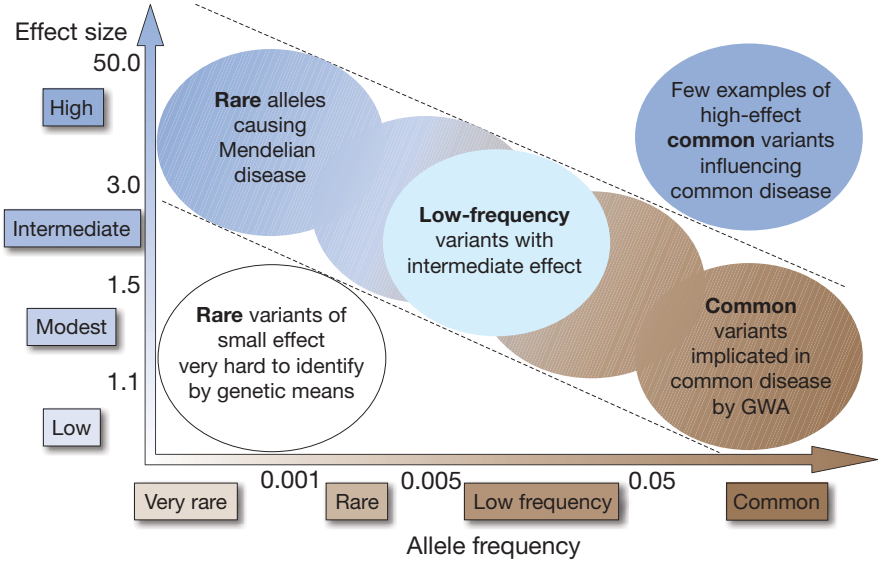


Fig. 1.2 Genetic variants frequency and disease susceptibility. Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from [17].

Hence, the field of human genetics typically separates rare and complex depending on whether a phenotype is caused by mutations in one gene or many genes, with the ambiguous term "oligogenic" being used as an intermediate. Nevertheless, there are several cases where this rule does not apply. There are Mendelian diseases where the "single gene" mutation does not correlate absolutely with the clinical phenotype because of the effects of additional independently inherited genetic variations and/or environmental influences. Those are Mendelian diseases where the phenotypes are in fact complex traits [18]. For example, Fuchs corneal dystrophy is caused by autosomal dominant mutations in *TCF4* [MIM: 602272] (which encodes transcription factor 4) but is defined as

a non-penetrant Mendelian disorder or a complex trait, since modifier genes and/or environmental factors influence the observed phenotype [19]. There are also monogenic disorders that violate the "one gene, one phenotype" assumption. For example, recessive loss-of-function (LOF) mutations in *CEP290* gene [MIM: 610142] (which encodes for centrosomal protein 290) can cause a range of conditions, from relatively mild disorders (such as Leber congenital amaurosis or nephronophthisis) to the perinatally lethal Meckel-Gruber syndrome [20].

Therefore, it may be appropriate to consider human diseases as a continuum of causality from Mendelian to complex, where some disorders do not fit neatly into either one of these categories. As such, genetic diseases would present with diminishing influence from a single primary gene, then a single primary gene influenced by modifier genes, to increasingly shared influence by multiple genes.

1.2 Next generation sequencing

Since Sanger sequencing was introduced in 1977, it has been used as a gold standard for the study of disease-causing genes. During this time, the technology has been enhanced to sequence longer DNA fragments and for a higher level of parallelism. However, this method achieves only a limited level of parallelization that does not allow the analysis of the DNA in a high-throughput manner [21, 22].

Encouraged by the Human Genome Project in 2004, Next Generation Sequencing (NGS) technologies emerged [23]. These are based on new sequencing instruments, which are capable of producing millions of DNA sequence reads in a single run. Since then, the advent of NGS has revolutionised the genomics field by enabling the fast and inexpensive

sequencing of entire genomes. This has led to very successful large-scale sequencing projects, such as the 1000 Genomes [24], UK10K [25], and Genome Aggregation Consortium (gnomAD) projects [26], among others. In the clinical field, this technology has also been used for identifying the causes of disorders with the ultimate goal of establishing therapeutic treatments and finding cures.

Some different technologies have been developed. The Illumina/-Solexa platforms (Illumina Inc., San Diego, CA, USA) are most common and offer diverse systems, from relatively small machines such as MiSeq to population-scale machines (HiSeq X Ten). These are based on the sequencing of short reads (100-150 bp) of fragmented DNA (Figure 1.3).

In 2007, "targeted capture" was created by Nimblegen. This method is able to select specific DNA sequences by microarray hybridisation for further sequencing [27]. Targeted capture allowed the sequencing of only a subset of the genome, e.g. specific genes or the whole exome, increasing speed of analysis and reducing cost. An alternative platform is Oxford Nanopore Technologies (ONT, Oxford, UK), which performs sequencing of long-reads, with a median size of 10Kb, although reads longer that 100Kb have been sequenced (Figure 1.3). This approach is especially useful for phasing variants that are farther away than the short-read sequencing read length and for identifying structural variants (SVs), which usually happen in repetitive regions where short-read sequencing has lower sensitivity [28]. However, the number of base-pairs sequenced per run is lower compared to other technologies [29].

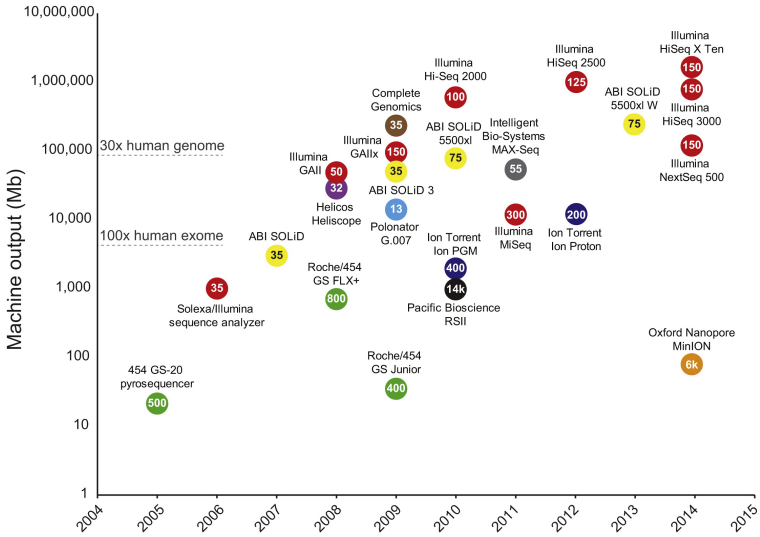


Fig. 1.3 High throughput sequencing technologies. Timeline of commercial release dates versus machine output per run. Numbers inside data points denote current read lengths. Sequencing platforms are colour coded. [29].

Nowadays, different NGS strategies are used for discovering genetic variations contributing to rare or common diseases. The simplest one is gene panel sequencing, based on high-throughput sequencing (HTS) for only specific genes. Since the introduction of NGS into clinical practice, the number and variety of disorders for which gene panel tests are being offered have increased dramatically [30, 31]. This is a good strategy in genetically heterogeneous diseases (where different genes can be responsible) given the reduced cost and the easy interpretation. However, only known genes that have been previously associated with the disease can be analysed.

Alternatively, whole-exome sequencing (WES) is based on the sequencing of the entire exome, which constitutes only about 1-2% of the human genome, and requires sequencing of just 30-65 mega bases (Mb)

of coding regions [32]. This is considered a suitable approach because it allows gene discovery and less time of analysis and cost compared to the sequencing of the whole genome, as well as a relatively simple final interpretation of the results [33–35, 30]. The main limitation of WES is that it does not detect non-coding variants and is limited to identify CNVs in coding regions, missing structural variants that don't present a copy number change, or that extend beyond the exome.

Whole-genome sequencing (WGS) can instead identify SNVs, indels and all types of structural variants in coding and non-coding regions with high confidence. WGS also performs better to detect exome variants than WES (where the proportion of variants missed by WES is higher) since the distribution of coverage depth and genotype quality are more uniform [36, 37]. For SNVs, the proportion of false-positive variants has been seen to be higher for WES (reported 78%) than for WGS (17%), making more difficult the analysis of true variants in the former.

However, despite the benefits, WGS is still more expensive than WES and the analysis, storage and interpretation of full genomes in a large number of individuals remains a challenge. Furthermore, it has been reported that the difference in the diagnostic utility of WGS over WES is not significant yet [38, 39], since most of WGS studies are limited to coding variants (or non-coding but previously reported) due to the challenges of analysing non-coding regions. Additionally, due to the large amount of data produced by WGS experiments, turnaround times take longer than for exome and panel experiments (although recent studies have demonstrated the possibility of rapid turnaround of WGS of $\sim$2 weeks [40, 41]). This is especially relevant in a clinical context rather in a research setting, where performing fast clinical diagnoses in as many individuals as possible is a priority. Therefore, the sequencing strategy

Table 1.1 Comparison of NGS strategies. [42–45]

	Targeted sequencing	WES	WGS
Targeted region	Variable (∼5Mb)	64Mb	3Gb
Number of variants	Variable (∼1500)	∼20,000	∼4,000,000
Cost	£200 - 400	£382 - £3,592	£1,312 - £17,243
Clinical coverage	80x	120x	30x
Advantages	(1) Can be customised (2) The cheapest and easiest to analyse	(1) High coverage of exons (2) Less expensive and easier to analyse than WGS	(1) Uniform coverage (2) Can detect SNVs/indels and all types of SVs in coding and non-coding variants
Disadvantages	(1) No gene discovery (2) Cannot detect SVs	(1) Cannot detect non-coding variants (2) Limited to detect CNVs in coding regions	(1) Highest cost (2) Largest volume of data and the most complex analysis

needs to be chosen accordingly to the aims the study. A comparison between panel sequencing, WES and WGS is represented in Table 1.1.

The availability of sequencing technologies led to the characterisation of many forms of variation in the human population, including SNPs/indels, SVs and more complex rearrangements [24, 28], providing the first insights into the scale of variation within the human genome. It

was found that: 1) there are at least 3.5 million positions and approximately 1,000 large Copy Number Variants (CNV) in each individual that differ from the reference genome [46], 2) most of these variants are common in the population [47], and 3) individuals from older ancestral origins (such as African) present higher variation with respect to the human genome of reference [48].

One of the largest datasets from NGS data is gnomAD, which provides 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals, sequenced as part of various disease-specific and population genetics studies. It provides allelic frequencies for specific ancestries and is largely used for the efficient filtering of candidate disease-causing variants [26].

gnomAD provides not only allelic frequencies, but also information about mutational recurrence, metrics of pathogenicity for sequence variants, and information about which genes are subject to strong selection against various classes of mutations, since deleterious variants are expected to have lower allele frequencies than neutral ones, due to negative selection. This information can be used for the discovery of human 'knockout' variants in protein-coding genes. For example, the probability of being a loss-of-function (LOF) intolerant gene is measured by the pLI score, which is calculated by comparing the observed and expected protein truncating variant population counts within each gene. This provides a probability of being intolerant (pLI > 0.9) or tolerant (pLI < 0.1) to LOF variation, for a range between 0 and 1. Using a similar approach, the tolerance to missense or synonymous variants is measured by a Z score, which for synonymous variants is centred at zero, but is significantly shifted towards higher values (greater constraint) for missense variants. Positive Z scores indicate increased constraint

(intolerance to variation) and negative values are given to genes that had more variants than expected [26].

This information allows us to identify those genes, or regions within specific genes, where variation is more likely to present with deleterious consequences, therefore to be associated with a disease. An example for the *KMT2B* gene [MIM: 606834] is represented in (Figure 1.4). *De novo* mutations in this gene have been previously associated with dystonia. *KMT2B* is a highly constrained gene for LOF (pLI = 1) and missense variants (Z score = 4), meaning there is less LOF and missense variation than expected in a control population. In (Figure 1.4), the gnomAD missense distribution across the entire gene is lined up with the encoded protein coordinates, showing the specific regions that are more constrained for variation. Interestingly, this matches to regions where pathogenic missense mutations have been reported. Thus, having this information is a good approach for predicting whether other variants are likely to be pathogenic or not.

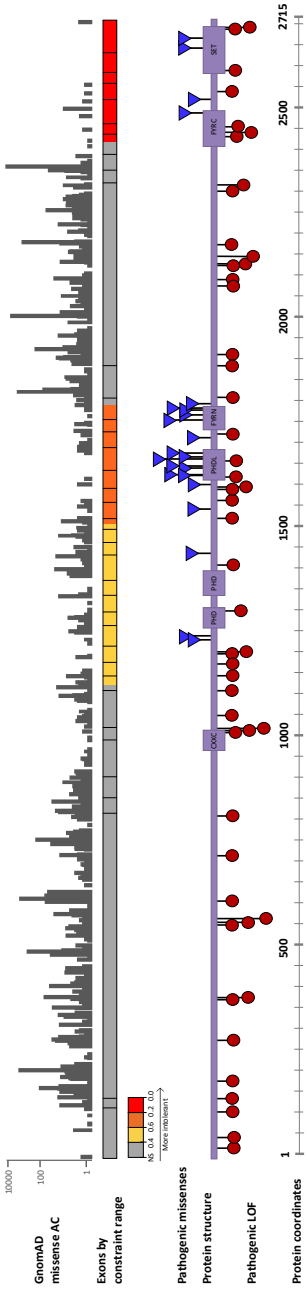


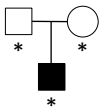
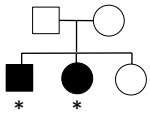
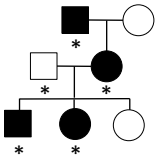
Fig. 1.4 Schematic diagram of *KMT2B* protein structure. Allele count (AC) of missense variants obtained from gnomAD is represented, followed by the constrained coding regions [49]. Pathogenic missenses and LOF were obtained from ClinVar (http://www.ncbi.nlm.nih.gov/clinvar). Protein structure information obtained from Pfam (http://pfam.xfam.org, using *KMT2B*'s UniProt ID: Q9UMN6)

1.2.1 Exome sequencing in Mendelian diseases

The first monogenic disorder to be resolved by exome sequencing was the multiple malformation disorder Miller syndrome in 2009. By doing WES in four affected individuals from three unrelated families, they found pathogenic compound heterozygous mutations in *DHODH* [MIM: 126064]. This demonstrated the value of this technology even without pedigree information or any biological information related to the mechanism of the disease [50]. Since then, more than 800 novel monogenic disease genes have been identified by similar approaches [50, 51]. Notable studies that have used NGS strategies for large-scale sequencing of patients with Mendelian disorders are the ESP project [52], the UK10K project [25] and the NIHR BioResource [53].

Currently, high coverage (60-120x) WES is one of the most popular approaches for discovering genes underlying Mendelian diseases, especially because the vast majority of disease-associated mutations that have been previously identified by result in the disruption of protein-coding sequences [54]. Genetic studies of Mendelian diseases are usually performed on family-based designs (Table 1.2). Different pedigree structures can be used: trios (where the proband and both parents are sequenced), duos (where the proband and a family member, both usually affected, are sequenced) or even larger pedigrees including multiple relatives. This design depends on the suspected inheritance. For example, for highly penetrant autosomal dominant inheritance, where the main mechanism of disease is usually sporadic mutations, trio analysis is especially useful because it allows identification of *de novo* variants (those that are not present in either of the parents) very efficiently. On the other hand, a duo study can also be very powerful, for example, when a recessive or X-linked inheritance is suspected. However, if the sequenc-

Table 1.2 Pedigree structures for NGS family-based studies. Asterisks represent sequenced individuals. Males are in squares and females are in circles. White colour are unaffected and black colour are affected individuals. AD=autosomal dominant, AR=autosomal recessive, XL=X-linked disorders, aff=affected.

Family structure	Trio	Duo	Multiple
Pedigree			
Suited for	AD,AR,XL	AR,XL (if aff+aff)	AD,AR,XL
Advantage	*De novo* and recessive variants characterisation	Identification of cosegregating biallelic variants	Combines the advantage of trios with multiple affected relatives
Disadvantage	If budget is limited, fewer patients can be sequenced	It needs posterior segregation analysis	Difficult to collect. If pedigree is very large, it can be more expensive

ing of affected individual/s alone (singleton) is performed, due to the large number of candidate variants that would be identified, usually only those which are in genes associated with the phenotype are considered. Current diagnostic rates for singleton analysis is 22–25%, whereas for a trio it can reach up to 33% [55–57]. This is probably because inheritance pattern or *de novo* status of variants can be considered by this approach, providing an extra evidence for a variant to be deemed as pathogenic.

NGS technologies have revolutionised the field of genetics by allowing fast and accurate identification of disease-causing mutations.

However, the identification of variants in genes of uncertain significance is also dramatically increasing. Even if a variant segregates within the family, if the gene has not been previously described as disease-causing, a single family on its own is not sufficient evidence that the mutation is causative. Therefore, observations in the same gene in additional families or individuals with a similar phenotype provide an important statistical support. Current guidelines for investigating causality of variants in new candidate genes suggest that more than three unrelated individuals with mutations in the same gene and consistent phenotypes are required to demonstrate that a gene is disease-causing [58]. Additional supporting evidence, such as functional assays and animal models, are often considered, as well as *in silico* evidence (eg. how tolerant is the gene to the observed class of variation) although the last one with a minor impact on decision.

Other than family-based design, a strategy that is increasingly being used in disease studies is case-control enrichment. In this approach, rare variants identified in a cohort of cases and a large cohort of controls are used. A statistical test is then applied to identify if there are a set of variants enriched in cases and not present in controls. Importantly, this approach considers non-classical contributors to disease, such as incomplete penetrance and variants that contribute to the phenotype in combination with others.

There are different types of statistical methods that can be used to perform case-control enrichment of rare variants [16]. Some of the most common ones are CAST [59], Sum [60], SKAT/SKAT-O [61] and other Bayesian methods such as BeviMed [62]. A summary of them is shown in Table 1.3.

Table 1.3 Rare variant association analysis methods. MAF=Minor allele frequency. SNP=Single nucleotide polymorphism

Category	Description	Method	Assumptions
Burden tests	Weighted average of rare allele counts	CAST, CMC	The mode of inheritance is jointly dominant [59, 63]
		SUM	All variants in the set have the same effect size [60]
		VT	All variants with MAF $\leq \zeta$ have the same effect size [64]
		WSC	The effect size is inversely proportional to MAF [65]
		RWAS	All SNPs have the same population attributable risk [66]
Variance component tests	Test of the variance of variant effect sizes	C-ALPHA	Variants are both protective and at risk [67]
		SKAT	The variance is $w_j\tau^2$ with beta w_j weights [61]
Combination tests	Combination of burden and variance component tests	SKAT-O	The test is based on an optimal combination of burden and variance statistics [68]
		MIST	The effect sizes are explicitly modelled using a mixed effect model [69]
		EMMPAT	The effect sizes are explicitly modelled using a mixed effect model that incorporates SNP annotation [70]
Other tests	Tests that enforce sparsity	EC, LASSO	Only a few of the variants are associated [71, 72]
	Replication-based test	RBT	Inference is based on separate statistics for protective and at-risk SNPs [73]
	Bayesian methods	BeviMed	Information on variant effect size and sparsity is incorporated in priors [62]

Although this strategy was designed for complex diseases, it is often used in Mendelian studies, and several works have demonstrated its utility. For example, heterozygous LOF variants in *NFKB1* [MIM: 164011] were observed to be the most common cause of primary immunodeficiency using BeviMed [74]; a BURDEN test was used in 2,536 schizophrenia cases and 2,542 controls identifying an enrichment for rare disruptive mutations in particular gene sets, including the voltage-gated calcium ion channel and the signalling complex formed by the scaffold protein ARC of the postsynaptic density [75], among others [76–78]. Nevertheless, case-control studies require a significant number of cases and controls. Additionally, any baseline differences, for example technical artefacts from the sequencing, can yield to false-positive signals, so results from these kinds of studies need particular attention and careful review of all significant results.

1.2.2 Exome sequencing in complex diseases

GWAS has been broadly used for the study of common variants in complex diseases. However, it presents two main limitations. First, it cannot detect rare variants since only SNPs with allele frequencies greater than 5% in the population can be analysed. Second, it is based on a genotyping array of known SNPs, therefore, the detection of novel variants or genes is not achievable directly, but feasible by imputation and haplotype analysis [79, 79, 80].

For this reason, several studies have used WES as technique for the discovery of rare variants involved in complex diseases [81–83]. For example, the NHLBI ESP has sequenced 6,500 individuals to study phenotypes such as heart attack, stroke and blood lipid levels. Like-

wise, T2D-GENES Consortium has sequenced the exomes of ∼10,000 individuals to identify variants associated with T2D, and the UK10K Project has sequenced the exomes of 6,000 individuals with multiple phenotypes [16]. Smaller projects have also used WES for the discovery of disease-causing genes in familial cases with complex traits [84], and several methods for performing rare-variant association test in families have been developed [85–90].

The main limitation of WES is that it does not consider non-coding regions, while GWAS has previously demonstrated several variants associated with disease in non-coding regions [80]. An alternative to this is to perform low coverage (∼10x) WGS to maximise cost and statistical power when budget is limited, so more individuals can be sequenced but at a lower depth [91].

1.2.3 Other applications of exome sequencing

Copy number variants

Copy Number Variants (CNVs) are a major source of variation in the human genome, contributing to many human diseases including neuropsychiatric disorders and cancer [92–95]. Microarrays have been typically used to identify copy number changes with great accuracy, but with the inconvenience of a minimum probe resolution of 10Kb. Therefore, using WES to identify CNVs is advantageous, since it provides not only SNP/indel information at the exonic regions, but also exonic copy number changes smaller than the microarray minimum resolution.

There are different approaches to detect CNVs from NGS data: 1) split reads, based on split mapping of reads that span a CNV breakpoint [96, 97]; 2) read pairs, based on an improbable distance of mapped

read pairs[96, 97] and 3) read depth, based on drops or increases in read depth [98]. Approaches one and two are of limited utility in WES, since they will only detect breakpoints that fall within an exon. For that reason, multiple software algorithms for WES are based on approach three, including ExomeDepth [99] and XHMM [100], among others [101, 102].

Read depth can be affected by other factors besides copy number, such as alignability, exome capture efficiency and GC content - especially in exome data, where PCR amplification is performed on the enriched reads in most of the protocols. To minimise those, different strategies can be used. For example, ExomeDepth considers read depth as relative to a reference sample (an average of many other exomes). A reference file is first created with as many unrelated samples as possible, and with minimum technical variability (samples need to be prepared in the same way or using the same library prep kits).

An alternative is used by XHMM, which is also based on read depth but uses principal components to handle normalisation. Basically, it creates a matrix of the depth of all exons in all samples, and the principal components of this matrix are expected to capture many of the artefacts. Once normalisation is done, XHMM calls CNVs using a Hidden Markov Model (HMM). This is based on the fact that if an exon is deleted, then the prior probability for the adjacent exons to be deleted is considerably higher than if no CNV had yet been detected in the gene.

However, CNV detection from WES data is still limited due to the variable coverage distribution across the genome that negatively affects the variant detection, and exons that are not well covered (especially GC-rich content regions and capturing limitation) [103]. It also relies on read depth as the sole source of information, ignoring split read and

read pair information. Another limitation of CNV detection in WES is that the breakpoints might not be exact, since there is only access to the coding regions.

Nevertheless, some works have successfully found pathogenic CNV by the analysis of WES data, demonstrating that an 'exome-first' approach for clinical genetic investigations may be considered for the analysis of CNV as well [104, 105]. For example, Spataro *et al.* identified ten patients with Parkinson's disease and a gene dosage alteration in *PARK2, GBA,* and *PARK7* [106]. An example of a deletion they identified in *PARK7* (also known as *DJ1*) is shown in (Figure 1.5). Throughout this thesis, when a gene symbol is not followed by the MIM number to avoid confusion due to the presence of multiple genes, these can be found in the Appendix (Section 7.2, Gene information).

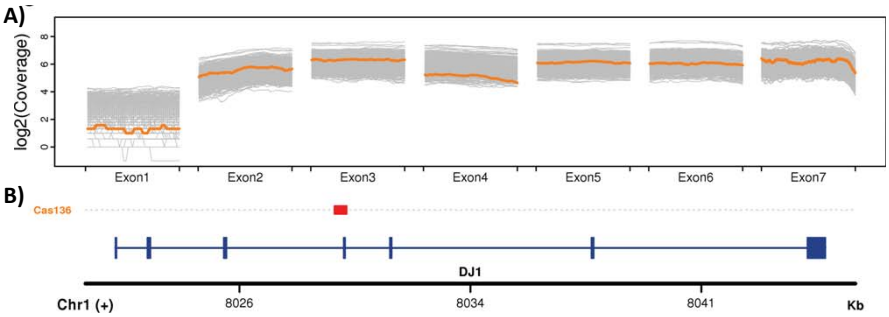


Fig. 1.5 Deletion in *PARK7* gene detected by exome sequencing. A) Sequencing depth of coverage for all samples in the study was used to infer copy number variants by the XHMM software [107, 100]. The orange line is the coverage for a patient with a deletion of exon 4. The grey background represents sequencing depth of coverage. B) Schematic representation of the corresponding validated copy number variant and the bottom track represents a schematic representation of the gene structure. *PARK7* is also known as *DJ1* [106].

HLA haplotyping

Since human Major Histocompatibility Complex (MHC) variation was first linked to disease via association to Hodgkin lymphoma [108] it has been intensively studied. Today, MHC, also called Human Leukocyte Antigen (HLA) in Humans, has been established as the region of the genome that is associated with the greatest number of human diseases. HLA genes are crucial to the immune system function and they play important roles in allergies, pathogenesis of autoimmune diseases, immune responses to infection and transplant rejection among others [109].

HLA is divided into three subclasses: class I region, which includes classical (*HLA-A*, *HLA-B*, *HLA-C*) and non-classical (*HLA-E*, *HLA-F*, *HLA-G*) genes; class II region, which includes *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQA2*, *HLA-DQB1*, *HLA-DQB2*, *HLA-DRA*, *HLA-DRB1*, *HLA-DRB2*, *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5*; and the class III region, which contains genes that are involved in leukocyte maturation, inflammatory responses and the complement cascade. The organisation of the HLA gene region is represented in Figure 1.6.

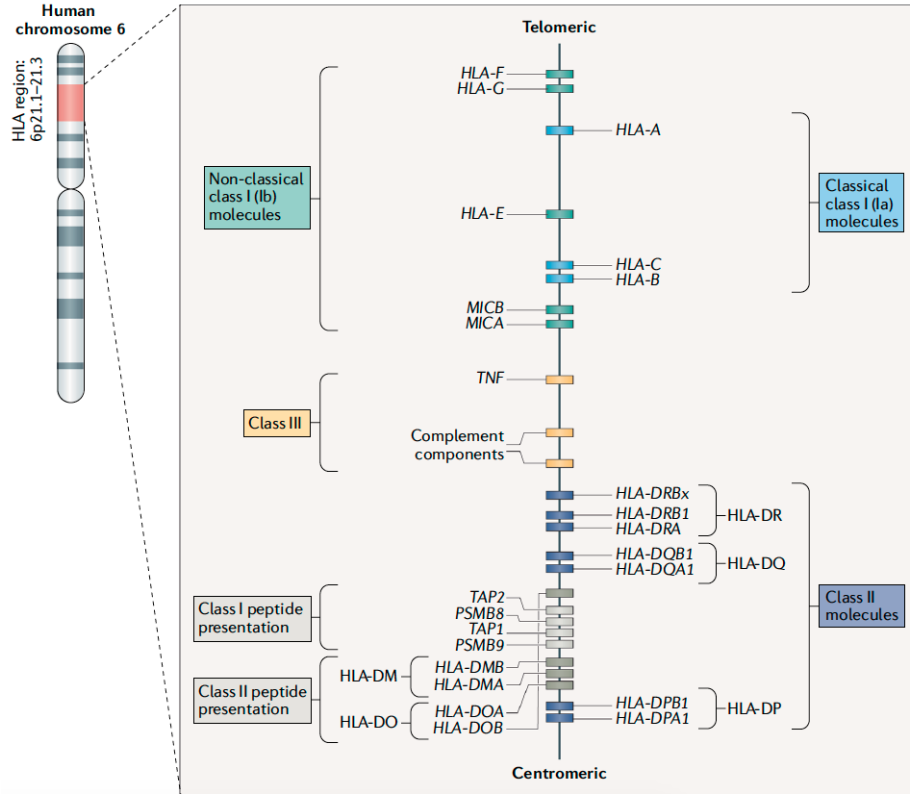


Fig. 1.6 Organisation of the HLA gene region. The HLA gene region is shown, with each bar representing a gene. Class II molecules (HLA-DP, HLA-DQ, HLA-DR) are towards the centromeric end, while class I molecules are located in the telomeric end region. Classical class I and II genes display extraordinary allelic variation, except for HLA-DRA1. Interspersed among the class II loci are genes that regulate antigen presentation (in grey colour). The class III region encodes non-polymorphic immune molecules that are not directly involved in antigen presentation (such as complement components and TNF). From [110].

The HLA locus is extremely polymorphic and is in strong linkage disequilibrium (the non-random association of alleles at different loci), complicating the determination of the exact genes and their association to disease. This variation may arise due to point mutations, but also often

by mechanisms such as gene conversion (when one allele is converted to another by mismatch repair mechanisms) [109]. More than 15,000 classical HLA alleles have been identified. This diversity likely exists to maximise the probability of some individuals successfully mounting an immune attack against a possible infection and survive.

HLA alleles have been associated with several disorders, mostly by conferring risk to disease. For example, *HLA-DR15* and *HLA-DR4* have been associated with Multiple sclerosis, *HLA-A*02:01* increases risk to type-1 diabetes (T1D), and *HLA-DQ2.5* and *HLA-DQ8* have been seen in several individuals with Coeliac disease [109]. These, and other examples, are shown in Table 1.4.

The current gold standard for high resolution typing of HLA alleles is sequence-based typing, that uses Sanger sequencing or targeted amplification of the HLA genes followed by HTS. Previous studies have already used NGS data for HLA typing. For example, 1000 genomes and exomes were typed by Major *et al.* [111], demonstrating that *HLA-A*, *HLA-B* and *HLA-C* can be typed from exome data with an accuracy higher than 90%. Moreover, with the growth of NGS technologies, methods for inferring HLA types have been developed. One example that uses a Population Reference Graph (PRG) is HLA*PRG [112], that can use data from both WGS and WES.

1.2.4 NGS summary

WES has been proven to be of great potential value as a diagnostic tool in clinical practice. It allows analysis of SNPs/indels in the coding regions of genome, and also investigation of CNVs and HLA alleles. It has been

Table 1.4 HLA haplotypes associated with disease. Autoimmune disease HLA associations for which molecular mechanisms of action have been identified. Adapted from [109].

Autoimmune disease	HLA allomorph (effect on disease)
Type-1 Diabetes	*HLA-A**02:01 (risk)
	HLA-DQ2 (risk)
	HLA-DR4 (risk)
	HLA-DQ8 (risk)
	HLA-DQ6 (protection)
	HLA-DQ2 and HLA-DQ8 (risk)
Coeliac disease	HLA-DQ2(.5) and HLA-DQ8 (risk)
Goodpasture disease	HLA-DR15 (risk)
	HLA-DR1 (dominant protection)
Systemic lupus erythematosus	MHC risk variants in distal intergenic XL9 regulatory element
Crohn's disease	Highly expressed *HLA-C* allotypes (risk)
Autoimmune polyglandular syndrome, IgA deficiency	HLA-DQ6 (protection)
Multiple sclerosis	HLA-DR15 (risk)
	HLA-DR4 (risk)
Rheumatoid arthritis	HLA-DR4 (risk)

used for the study of Mendelian and complex disorders, demonstrating its value for both in numerous cases. This technology is currently the gold standard for diagnostic and clinical research in many centres. It is used for the identification of known and novel variants in disease-associated genes, as well as discovery of novel genes. In this dissertation, this technology was used for the study of genetic variants in individuals with a severe form of gastrointestinal food allergy to multiple food proteins. Therefore, the classification, pathogenesis and genetics of the disorder is next explained.

1.3 Gastrointestinal food allergies

1.3.1 Introduction to food allergies

Certain foods or components of food may cause adverse reactions ranging from a slight rash to a severe allergic response. Adverse reactions to foods can be classified into non-toxic (immune and non-immune mediated reactions) and toxic reactions produced by, for example, bacterial toxins (Figure 1.7). The symptoms range from slight inconveniences to life-threatening shock reactions. Some reactions are difficult to recognise, diagnose and treat, while other dermatological, respiratory and systemic manifestations are readily recognisable.

The most common adverse reaction to foods is food allergy, which is an immune-mediated response that occurs after the ingestion of a specific type of food protein and is absent during avoidance. The current definition of food allergies is "adverse health effect arising from a specific immune response that occurs reproducibly on exposure to a given food" [113]. Food allergies can be classified into IgE (Immunoglobulin E)-

mediated, non-IgE-mediated and mixed responses. Other non-toxic adverse responses to foods are not immune mediated, and these can be classified into food intolerance due to toxicity, pharmacological reactions and even psychological food intolerance (Figure 1.7).

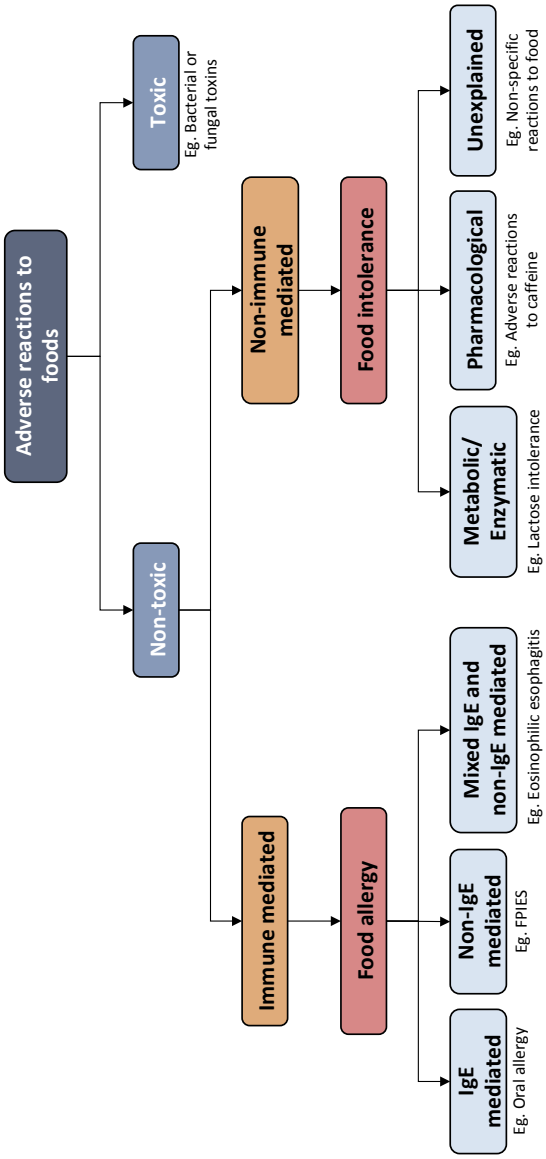


Fig. 1.7 Classification of adverse reactions of foods.

Symptoms of food allergies usually start within minutes of exposure to the trigger food and always occur within a few hours after the ingestion. Clinical presentation varies depending on the type of food allergy. For example, IgE-mediated food allergy responses are most commonly localised and affect the lips, mouth and throat. Additionally, they can also be associated with more systemic reactions, involving gastrointestinal (GI) manifestations, respiratory effects and skin manifestations [114]. Sometimes they can even entail a severe and life-threatening systemic hypersensitivity reaction that involves multiple organ systems and is called anaphylaxis [115]. Augmenting factors (such as alcohol, nonsteroidal anti-inflammatory drugs, concomitant infections or physical exercise) can increase the severity of the reaction in up to 30% of cases [116].

Non-IgE mediated food allergy primarily affect the GI tract, and are often classified as dietary protein enteropathies. Common examples are allergy to cow's milk or soy protein, and can cause variable small and/or large bowel injury associated with nonspecific villous atrophy and inflammation [117]. Symptoms may include repetitive emesis and diarrhoea after one or two hours of ingestion of offending foods. On the other hand, mixed responses can present manifestations characteristic of both IgE-mediated responses (like atopic dermatitis) and non-IgE-mediated responses (such as GI disorders).

Approximately 20% of the population in industrialised nations has been reported to experience adverse reaction to foods, which varies in clinical presentation, severity and underlying aetiology [114]. However, when placebo-controlled food challenges studies were performed, the prevalence of true reactions dropped to between 2% and 4% [115]. This difference highlights the difficulty of measuring the prevalence of true adverse reaction to food.

Table 1.5 Classification of food allergies. GI=Gastrointestinal; OAS=Oral allergy syndrome; Eo=Eosinophilic; EoE=Eosinophilic oesophaguitis; FPIES=Food protein–induced enterocolitis syndrome; FPIAP=Food protein–induced allergic proctocolitis; FPE=Food protein–induced enteropathy. Adapted from [119].

	GI	Cutaneous	Respiratory	Generalised
IgE mediated	OAS, GI anaphylaxis	Urticaria, angioedema, morbilliform rashes and flushing	Acute rhinoconjunctivitis, bronchospasm	Anaphylactic shock
Mixed	EoE, Eo, gastroenteritis	Atopic dermatitis	Asthma	-
Non-IgE mediated	FPIES, FPIAP, FPE	Contact dermatitis, dermatitis herpetiformis	Heiner syndrome	-

Food allergy is more common in children than adults, and the prevalence is increasing in many countries. This disorder is often developed in early childhood, affecting up to 6%-8% of children younger than ten years and between 1%-4% of the adult population. Accurate determinations are elusive because different factors influence the estimation, such as sex, ancestry, geographic location, ages and dietary exposures [118, 113]. Moreover, there are different ways to classify food allergies: by affected system or by immune-type response. Allergies can have generalised responses or can affect specific systems such as GI, cutaneous or respiratory system. They can also be classified into IgE mediated, non-IgE mediated, and those that are mediated by both commonly referred to as mixed food allergies (Table 1.5).

This dissertation focuses on GI food allergies, hence other types of food allergies will not be discussed. The different subtypes, pathogenesis and management of GI food allergies are next explained.

1.3.2 Pathophysiology

Regulation of the Intestinal Immune Response

Appropriate regulation of the intestinal immune response is essential to maintain balance and avoid potentially deleterious immune responses to foods [120]. This is achieved by down-regulating the normal immune response to bacteria and food antigens (also termed "oral tolerance"). This hyporesponsiveness, that seems to be impaired in GI food allergy, is regulated by two major pillars: the innate, general defence and the adaptive, specialised defence, both working closely together and taking on different tasks.

First, the innate immune mechanisms in the GI system include gastric acid, bicarbonate, intact epithelial layer with tight junctions, mucus secretion, digestive enzymes and peristaltic movement among others [118]. These mechanisms are involved in the control of invasion and prevention of infection of pathogens, so a dysregulation could lead to GI problems. For example, it has been seen that Humans and animal models treated with proton pump inhibitors and with other anti-secretory drugs presented increased sensitisation to food antigens, probably due to less effective gastric proteolysis [121]. The permeability of the intestinal barrier also plays an important role. Infants with an incompletely matured intestinal mucosa or individuals with an impaired barrier have an increased uptake of molecules. Increased intestinal permeability and subsequent uptake of food antigens has also been observed in patients

with food allergy [120, 122]. This may be secondary to the intestinal inflammation. Additionally, the properties of the triggering antigen influence the type of immune response (where more soluble proteins are more tolerogenic than particulate of globular antigens) [123].

Second, the balance of adaptive immune response in the gut is also important for its maintenance, since uncontrolled inflammation could drive an inappropriate immune response. In response to specific food antigens, T-cells produce cytokines to induce B-cells to produce specific antibodies. There are two different types of T-cell responses, Th1 and Th2. Th1 cytokines, like IFN-γ, tend to produce the proinflammatory responses responsible for killing intracellular parasites. In contrast, Th2 cytokines include *IL-10* [MIM: 124092], which has more of an anti-inflammatory response, and *IL-4* [MIM: 147780], *IL-5* [MIM: 147850] and *IL-13* [MIM: 147683], which are associated with the promotion of IgE and eosinophilic responses in atopy. The interplay between Th1 and Th2 responses can be regulated by multiple factors, including the expression of costimulatory molecules, different type of dendritic cells and the cytoplasmic milieu. A dysregulation on the balance between Th1 and Th2 responses could lead to an uncontrolled inflammatory response, that could drive to GI disorders such as atopy and food allergic reactions.

Allergic inflammation

Pathogenic mechanisms of GI food allergies differ depending on if they are IgE-mediated or non-IgE mediated (Figure 1.8). On one hand, IgE-mediated food allergies require an initial food allergen sensitisation. This occurs when Th2 cytokines such as *IL-4* and *IL-13* are produced by T cells in response to specific food antigens, and induce B cells to produce food-specific IgE antibodies. These antibodies then bind to the

surface of mast cells and basophils. Upon re-exposure to the offending foods, the food antigens bind to the food-specific IgE antibodies, causing their activation and degranulation. Released mediators such as histamine and leukotrienes cause inflammation, the allergic response and the development of signs and symptoms [124, 114, 125, 116].

On the other hand, non-IgE mediated food allergies are independent of IgE-mediation mechanisms. These are less understood than the IgE-mediated ones and are usually confined to childhood, being less recognised in adults. In non-IgE mediated mechanisms, inflammatory cytokines (such as TNF-α) are produced antigen-specifically by T-cells in response to specific food antigens. Inflammatory cytokines increase the intestinal permeability, which facilitates the uptake of undigested food antigens. Other Th2 cytokines such as IL-4, IL-5 and IL-13 are also produced by T cells. Here, IL-4 and IL-13 don't induce production of food antigen-specific IgE antibodies by B cells, but induce intestinal epithelial damage, while IL-5 accumulates and activates eosinophils in GI tissues [124]. Mixed responses can present both IgE and non-IgE mediated pathogenic mechanisms.

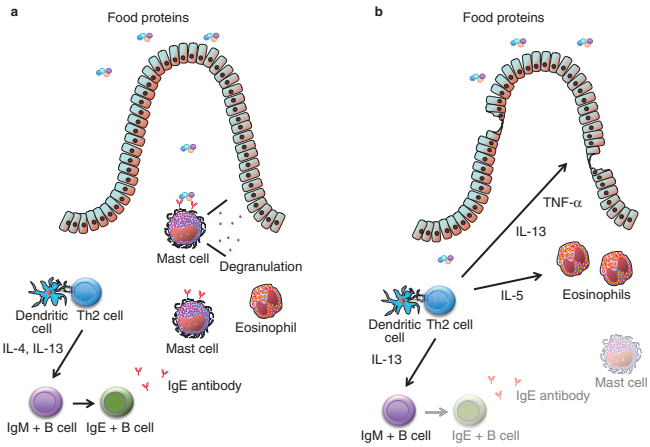


Fig. 1.8 Pathogenic mechanisms of food allergy. (a) IgE-mediated food allergy. (b) Non-IgE mediated food allergy. Adapted from [124].

Therefore, regulation of T and B cells responses play an important role in the development of GI food allergy. Patients with food allergy present with the allergen-specific Th2 cell releasing cytokines in blood, skin and mucosal sites [126], which play an important role in the induction of allergic responses by either regulating IgE synthesis (*IL-4* and *IL-13*) or chemoattraction of proinflammatory cells (*IL-4* and *IL-5*).

It has also been seen that the risk of presenting allergic disease is increased by a delayed development on the IgA system or the enhanced switch to IgE producing B cells. The major inducer of IgA synthesis is TGF-β, whereas the switch to IgE depends on *CD40L* [MIM: 300386], *IL-4* and *IL-13*, derived from Th2 and inflammatory cells [118].

1.3.3 Classification

In general, allergic reactions to foods affecting the GI tract are characterised by symptoms such as vomiting, diarrhoea and bloody stool after the ingestion of offending foods. However, depending on their type of response, specific characteristics and pathogenesis, they can be classified into specific subtypes (as previously shown in Table 1.5).

IgE mediated

IgE mediated GI food allergies are Oral Allergy Syndrome (OAS) and GI anaphylaxis. OAS is the most common manifestation of food allergy in adults. Exposure to certain types of allergens (such as plant proteins) may lead to itchy skin, or even more systemic reactions occurring a few minutes after the ingestion of the allergen [118]. Differently, in GI anaphylaxis the phenotypes (vomiting, nausea, abdominal pain and diarrhoea) typically occur in conjunction with allergic manifestation in other organs. The responsible foods usually are cow's milk, eggs, peanuts, seafood and fish. It can also be confirmed by measurement of specific IgE levels or skin prick test.

Mixed IgE and non-IgE mediated

Mixed GI food allergies can be classified into eosinophilic oesophagitis (EoE) and eosinophilic gastroenteritis (EG). EoE is an increasingly recognised chronic inflammation of the oesophagus, that usually affects children and adults. This is characterised by the infiltration of eosinophils in the oesophageal mucosa, presenting symptoms like vomiting, pain, reflux and dysphagia. Some patients have concomitant asthma or other

chronic respiratory disease. Individuals with EoE often have positive skin prick tests and specific IgE (sIgE) to foods, although these are weaker in adults [118]. Diagnosis in children usually occurs within the first three years of life. In EoE, diagnosis can be supported by endoscopic findings and histological features of eosinophilic inflammation (with >15 eosinophils per high power field) [115]. Three types of EG include eosinophilic gastritis, eosinophilic enteropathy and eosinophilic colitis. They are characterised by eosinophilic inflammation of the GI tissues, and can manifest at any age (with male predominance) [115].

Eosinophilic infiltration location and depth determine the manifestations of this condition, which is characterised by abdominal pain, nausea, vomiting and diarrhoea. Because these symptoms are also characteristics of Inflammatory Bowel Disease (IBD), diagnosis is not always straightforward – the current gold standard for diagnosis is demonstrated by characteristic endoscopic and histopathological features. EG is rare and managed with corticosteroids in most cases. Successful resolution of symptoms has been reported in a series of children [115].

Non-IgE mediated

Non-IgE mediated GI food allergies encompass three main types: Food Protein-Induced Enterocolitis Syndrome (FPIES), Food Protein-Induced Allergic Proctocolitis (FPIAP) and Food Protein-Induced Enteropathy (FPE). A comparison is shown in Table 1.6.

The first one, FPIES, is an uncommon food allergy that causes GI symptoms (vomiting with or without diarrhoea) as a reaction to the ingestion of specific food proteins. The underlying pathophysiology is not well defined, but it is suspected that the resulting inflammation from

the stimulation of mucosal T-cells and TNF-α could explain the clinical effects. In a prospective study on 13,019 infants, 0.34% (44/13,019) presented FPIES [127]. It is present in babies and young children, and most become tolerant by three years of age. Symptom onset could be within weeks of birth, but if the babies are breast-fed, it could be up to months, with the introduction of solid foods. There are three common foods that lead to FPIES: cow's milk, soy and rice, however, other aliments such as vegetables, egg white, legumes and meat can also trigger the symptoms. Symptoms can manifest two hours after the exposure to the offender aliment/s, presenting with vomiting with or without diarrhoea. And although they usually resolve in 6-12h, these children can present as acutely unwell. Affected individuals are frequently mistreated for sepsis, pyloric stenosis or inherited metabolic disease, especially since blood test results may present metabolic acidosis, neutrophilia and thrombocytosis [115]. The diagnosis has to be clinical, although the presence of blood, eosinophils and lymphocytes in stools is supportive. Due to overlapping features with the previously mentioned disorders, it can take up to five episodes to establish diagnosis [128]. One difference is that individuals with FPIES recover more rapid than those with sepsis or surgical conditions.

The second type is FPIAP. Patients with FPIAP present blood and mucus in the stool. Here, inflammation of the colon and rectum is due to eosinophilic and lymphocytic inflammation. Diagnosis is based on the presence of fresh rectal bleeding, and the absence of other symptoms, as well as eosinophilic infiltration performed on mucosal biopsies from colonoscopy. Symptoms resolve when eliminating the offending proteins (cow's milk and soy protein) from the diet. This is important to perform

in order to differentiate infants with FPIAP than infants with transient colitis, whom can resolve even without a change in diet.

Lastly, FPE is a disease of infants, characterised by malabsorption mainly caused by cow's milk. Affected infants develop chronic diarrhoea, steatorrhoea and poor weight gain. It is often seen with anaemia and hypoalbuminemia. The most common offending foods are cow's milk, but also soy, rice, chicken and fish. The symptoms are observed in the first few months of life, and resolution generally occurs in 1-2 years [118]. The underlying mechanisms involve T-cell immune responses within the small intestine, with villous atrophy and lymphocytic infiltration. It is similar to Coeliac disease, but the main difference is that in FPE symptoms may appear before the introduction of dietary gluten. Diagnosis is based on elimination diets and endoscopy/biopsy to identify an increased intraepithelial lymphocytes and eosinophils and villous injury.

Table 1.6 Comparison of non-IgE mediated GI food allergies. FPIES=food protein-induced enterocolitis syndrome; FPIAP=food protein-induced allergic proctocolitis; FPE=food protein-induced enteropathy, FTT=failure to thrive, LNH=lymphonodular hyperplasia, OFC=oral food allergy. Adapted from [129].

	FPIES	FPIAP	FPE
Age of onset	Usually one day to one year	Days to six months	Dependent on age of exposure to antigen
Common food proteins	CM, soy, rice, multiple	CM	CM
React to ≥2 foods	Up to 35%	Up to 20%	Rare
IgE positive	4% to 30%	Negative	Negative
Transition to IgE positive	Up to 35%	None reported	None reported
Family history of atopy	40% to 70%	Up to 25%	Unknown
Symptoms	Emesis, severe diarrhoea, severe bloody stools, severe oedema, shock (15%)	Mild diarrhoea, prominent bloody stools, mild/infrequent oedema	Intermittent emesis, moderate diarrhoea, rare bloody stools, moderate oedema, moderate FTT
Laboratory findings	Moderate anaemia, acute hypoalbuminemia, possible methemoglobinemia, possible acidaemia, prominent leukocytosis with neutrophilia, moderate thrombocytosis	Mild/infrequent anaemia, mild/infrequent hypoalbuminemia, mild thrombocytosis, occasional peripheral blood eosinophilia	Moderate anaemia, moderate hypoalbuminemia, mild thrombocytosis, malabsorption, steatorrhoea

Continued on next page

Table 1.6 – continued from previous page

	FPIES	FPIAP	FPE
Treatment	Food elimination; symptoms clear within hours in patients with acute FPIES and in 3-10 days in patients with chronic FPIES	Food elimination from the maternal diet or hypoallergenic formula. Food reintroduction after 12 months	Food elimination, symptoms clear in 1-3 weeks, re-challenge and biopsy in 1-2 years
Resolution	Varies by population, CM tends to resolve by age 3-5 years; rice-induced FPIES, 50% outgrow by age five years	Majority resolve by age 12 months	Most cases resolve in 24-36 months
T-cell response	Inconclusive, TH2 skewing	Unknown	Increased intestinal intraepithelial suppressor/cytotoxic CD8+ T cells
B-cell response	Absent IgE, IgG4, IgA responses	Unknown	Absent
Cytokine imbalance	Decreased TGF-β, increased TNF-α and IFN-γ	Unknown	Increased IFN-γ and *IL-4* level in jejunal biopsy specimens

1.3.4 Offending foods

The most common food allergen in patients with GI food allergy are cow's milk, soy and cereals, including rice and oats. FPIES is usually caused by a single food (60-80% cases), but there are cases with reaction to two foods (30-50%) or even more, though these are very rare. It's also been seen to vary with geographic differences (for example, high frequency of fish allergy in infants from Italy and Spain), but feeding routines, age of induction and genetic predisposition might also underpin this.

Interestingly, the study of allergens that may cause allergic reactions in the GI tract revealed that inhalant allergens such as pollens can also be swallowed and detected in faecal samples of affected individuals. Pollen shares morphological features with certain parasite eggs [130]. Major epitopes (the part of an antigen that is recognised by the immune system) in pollen are Bet v1 and Bet v2. Specific IgE in patients with allergy to pollen are directed to Bet v1, emphasising the importance of this protein as a major epitope [131]. This opens up the opportunity for genetically modified and recombinant food antigens, offering new possibilities for both diagnosis and treatment of patients with food allergies. For example, a cloned peanut allergen (Ara h3) has already been developed which binds less efficiently to IgE but keeps the ability to stimulate T-cell activation [132].

1.3.5 Diagnostic approach

It is very important for the proper management of the patient to diagnose and properly differentiate between GI food allergies and other types of GI pathologies with different aetiology, such as food intolerance, inflam-

mation (IBD, Crohn's disease, ulcerative colitis), anatomic problems (pyloric stenosis, which is a narrowing of the opening from the stomach to the first part of the small intestine), malignancy, and infections or metabolic disorders.

Food intolerances are different than immune-mediated allergies, where patients may experience anaphylactic reactions and must avoid all foods containing the specific allergen. Unlike a food allergy, for intolerance there is a delay in symptom onset (several hours), a prolonged symptomatic phase (can last for hours or days) and negative IgE serology [115]. Therefore, one main difference is that most GI food allergies exhibit severe symptoms within one hour after ingestion of the offending food, while other disorders present delayed manifestation of symptoms (up to several hours after the ingestion) [118]. Nevertheless, the overlapping phenotypes and the poorly understood pathophysiologic mechanisms makes very challenging proper diagnosis of GI food allergies. GI food allergy diagnosis highly depends on the clinical history of the patient, the exclusion of other conditions and the observation of the patient after the ingestion of offending foods. The diagnostic algorithm for food allergy, developed by the American Gastroenterological Association [133], is shown in (Figure 1.9).

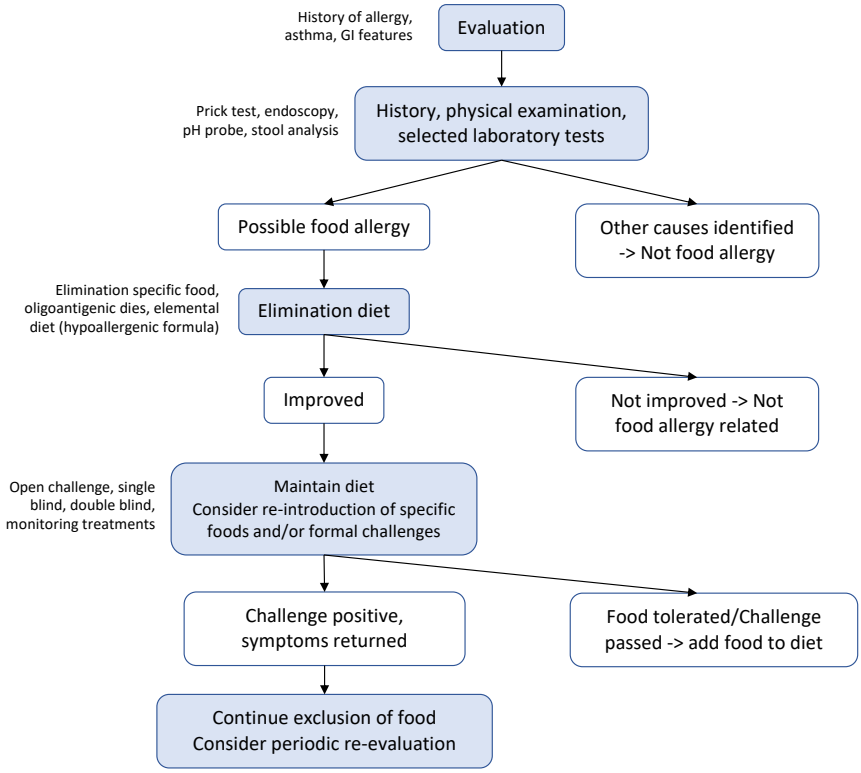


Fig. 1.9 Diagnosis evaluation approach in GI disorders. Adapted from American Gastroenterological Association [133].

Diagnosis of subtypes of GI food allergy mediated by IgE can be performed with the combination of the skin prick test along with the measurement of food-specific IgE antibody levels. Measurement of specific IgE by a radioallergosorbent test or a newer nonradioactive test is also possible. These have higher specificity and reliability than the skin prick test. However, these tests have some disadvantages: first, false positive results are fairly common, and cannot be distinguished between a sensitised individual to the allergen and one who is clinically allergic. Second, IgE is also produced locally in the GI mucosa, so serum IgE

measurements do not correlate well with mucosal allergic responses in the intestine [130]. Consequently, in those cases where a food challenge has not been performed, the classic elimination of foods followed by the observation of the patient well-being is considered to be a good approach [115, 116].

Due to the absence of IgE in non-IgE mediated food allergies, its exact diagnosis is more challenging. Some approaches include T-cell cytokine assays and serum measurements of markers of eosinophil activation (for example, eosinophil cationic protein). Measurement of IgE, TNF-α and eosinophil mediators in stool samples are also interesting tools, but they are not yet established for use in clinical practice [118]. Colonoscopy allergen provocation is a technique equivalent to skin testing in which a panel of antigens are injected in the GI mucosa and then the response is observed by endoscopy [134]. However, although it has been reported to be an advance in the field, its incorporation into routine clinical practice has been limited.

Furthermore, the fact that precise pathogenic mechanisms of GI food allergies remain poorly understood, makes difficult the identification of more specific types. For example, different subtypes of non-IgE mediated food allergy (FPIES, FPIAP and FPE) exhibit similar symptoms such as vomiting and diarrhoea. It is not well described if these disorders have similar pathogenesis and just differ in severity, or whether the pathogenesis of each is distinct, meaning they should be classified as separate entities. Previous works have tried to approach this issue by performing cluster analysis on the clinical and laboratory findings in order to characterise between these types of allergies and determine whether the pathogenesis is different [135].

1.3.6 Treatment

Since there is no curative treatment for food allergy, the main strategy for management is the avoidance of the allergen and the preparation for future accidental exposures. Patients should also learn to read and understand labels for food allergens. Sometimes it is also important for those affected individuals and family members to join local foundations and support groups that can provide information and support. In instances where an elimination diet cannot be followed (e.g. multiple food allergies), antiallergic medications should be tried. For example, mast cell-stabilising agent disodium cromoglycate can act locally in the GI tract and can be tried in such cases, although its supporting evidence is limited [118].

Patients and their families should also be prepared for accidental exposures. For individuals with IgE-mediated response, the first-line of defence is the emergency medication during the anaphylaxis, through the intramuscular auto-injection of an epinephrine-containing syringe. Antihistamines and corticosteroids have a supportive role in treating anaphylaxis, but they should not replace the adrenaline injection [113, 116, 114]. Furthermore, patients with non-IgE mediated exposures will require intravenous or oral rehydration. Steroids can also be provided, although there is no evidence that they hasten recovery [118].

Nevertheless, the rise in prevalence of food allergy has led to significant interest in developing better therapeutic strategies for its management. Treatment approaches for food allergies, including the most promising advances, are shown in Table 1.7. One example is Oral Immunotherapy Treatment (OIT), which offers the best efficacy as compared to other routes of immunotherapy but also the highest probability

for adverse effects. The use of *Omalizumab* in conjunction with OIT may improve the safety profile.

Table 1.7 Food allergy treatments. OIT=Oral Immunotherapy Treatment; IT=Immunotherapy.

Type	Strategy	Description	Ref
First line treatment	Dietary strategies	Targeted elimination diet or elemental diet	[115, 136]
	Cortico-steroids	First-line treatment for induction of remission	[115, 117, 136]
	Steroid sparing agents	Include selective leukotriene inhibitors (e.g. *Montelukast*), mast cell stabilizers (e.g. *Sodium cromoglycate*) or 2nd generation H1-antihistamine agents (e.g. *Ketotifen*)	[115, 117, 136]
Immuno-therapies	OIT	Involves the administration of increasing doses of the offending food over months and then a maintenance dose for years	[137–143]
	Sublingual IT	Delivers the antigen under the tongue in a liquid form. Patients receive gradually escalating doses until a maintenance dose period is achieved	[144–147]
	Epi-cutaneous IT	Delivers the offending antigen via a patch through the skin	[148–150]

Continued on next page

Table 1.7 – continued from previous page

Type	Strategy	Description	Ref
Future therapeutic strategies	Hypo-allergenic antigens	Reduce the allergic potential of foods by genetically or chemically modifying their structure (e.g. substitutions in the IgE binding site of a peanut allergen)	[132, 151]
	Anti-monoclonal antibodies IT	Use of anti-IgE antibodies for the specific region that binds to receptors on mast cells and basophils. E.g.: *Omalizumab*	[152, 153]
	Antagonist of Th2 response	Strategies to antagonize Th2 response, such as Th1-type cytokines (including *IL-12* and IFN-γ). *IL-12* provides benefit in a murine model with peanut hypersensitivity	[154–158]
	Serotonin 5-HT3 receptor antagonist	Individuals with FPIES demonstrated to resolve symptoms with *Ondansetron*	[159–161]
	Anti-*IL-33* [MIM: 608678]	Knocking out the *IL-33* receptor, ST2, in a mouse model showed this pathway is necessary for driving the Th2 cellmediated allergic response	[162, 163]
	GSK3 inhibitors	GSK3 promotes inflammation, and it has been associated with diseases that involve inflammation, including Alzheimer's disease, diabetes, and cancer	[164, 165]
	Apoptosis of T and B cells	*Azathioprine* is a corticosteroid sparing agent and has been used for the treatment of asthma and eosinophilic enteritis	[166]

Continued on next page

Table 1.7 – continued from previous page

Type	Strategy	Description	Ref
	Toll like receptors antagonists	R848, a TLR7 agonist, was found to decrease airway inflammation. TLR4 agonist has been effective in treating pollen allergy	[167–169]
Others	Peanut vaccine	Demonstrated in mice using oral delivery of a DNA plasmid encoding the Ara h 2 protein on a nanoparticle carrier. Subsequent Ara h 2 expression in the gut epithelium resulted in partial protection from anaphylaxis. A clinical trial is currently underway to test a DNA vaccine for peanut allergy	[167]

1.3.7 Animal models

Multiple animal models have been used to investigate the pathogenesis of allergic diseases *in vivo* [120, 170, 171]. Animal models vary in terms of animal used (rat, mouse, pig, guinea pig, dog), methods used (measurement of inflammatory mediators, morphologic studies, functional assays of gut function) or sensitization protocols (type of food allergen, route of administration, dose).

However, despite the benefits and the advances on the study of food allergy that these studies provided, there is no animal model that can mimic the human food-allergic sensitization and allergic responses. Therefore, it is still a challenge to extrapolate results observed in animal models to human.

1.3.8 Prevention

Common recommendations in infants with GI food allergy have been made. These include the exclusive use of breast-feeding and delayed introduction of solid foods up to 4-6 months, avoidance of all cow's milk protein and, if formula is needed, the use of extensively hydrolysed or amino acid-based formula [118]. Probiotics have been suggested to be beneficial in food allergies. For example, *Lactobacillus rhamnosus* was given to pregnant woman during the last 4 weeks of pregnancy and subsequent breast-feeding until infants were three months of age resulted in only 15% of offspring presenting allergic eczema, compared to the 47% that received placebo [172]. However, beneficial results were not observed in a different study of young adults and teenagers with oral allergy syndrome [173], and it was suggested that the use of probiotics in allergic diseases is especially beneficial shortly after birth, when the normal enteric flora has just been established.

1.3.9 Heritability of food allergy

The association between genetic variants of nearly a dozen candidate genes and food allergies were first identified via positional cloning and candidate gene approaches. Mutations in proteins that play a role in the gut motility, inflammation, microflora, visceral hypersensitivity, and dietary factors were identified to be relevant. For example, LOF mutations in *FLG* gene [MIM: 135940] (which encodes for filaggrin protein) play a role in food allergy, since it is involved in the maintenance of an effective skin barrier including allergens. Children with LOF variants in *FLG* were 1.5 times more likely to react during food challenge to at least one food as compared to carriers of the wild-type alleles [116].

From GWAS studies, the search for common variants across the genome showed that HLA-DR and HLA–DQ regions at locus 6p21.32 were significantly associated with peanut allergy in a cohort of 2,197 US subjects of European ancestry [174]. Another study identified that copy number variants in *CTNNA3* [MIM:607667] and *RBFOX1* [MIM: 605104] were associated with food allergy [175] and that knockdown of *CTNNA3* resulted in up-regulation of CD63 and CD203c in mononuclear cells, suggesting a role in sensitisation to allergen.

After the introduction of NGS technologies, the role of rare coding variants was also considered for food allergy and other atopic phenotypes such as asthma, eczema and atopic dermatitis. Consequently, rare coding variants in the genes *PDE4DIP* [MIM: 608117], *CBLB* [MIM: 251110], *KALRN* [MIM: 604605], *DPP10* [MIM: 608209], *IL12RB1* [MIM: 601604], *IKBKAP* [MIM: 603722] and *AGT* [MIM: 106150] were reported in patients with asthma [176, 177].

1.3.10 Environmental factors

During the past years the prevalence of allergic diseases has been rising more rapidly than changes to the genome sequence would indicate, suggesting an important role of environmental factors. Numerous hypotheses have been postulated to lead to an increased prevalence of allergic diseases. For example, the hygiene hypothesis, postulated in 1989 by Strachan, proposed that increased prevalence of allergic diseases could be affected by an increased cleanliness, decreased family size and decreased childhood infections [116]. Since then, other life environmental and style characteristics have also been considered. A summary of the

environmental factors that have been proposed to influence food allergy or sensitisation are described in the following Table 1.8.

Table 1.8 Environmental factors of food allergy.

Factor	Evidences	References
Hygiene hypothesis	Proposes that the lack of early childhood exposure to infectious agents, gut flora, and parasites increases susceptibility to allergic diseases by modulating immune system development, although limited data for the hygiene hypothesis exist with respect to FA	[178]
Microbiota	Gut microbial composition and colonisation early in life influence the development of atopic diseases. Differential composition of the microbiome could be explained by the fact that specific intestinal microorganisms can downregulate inflammation by counterbalancing type-2 T-helper cell responses, enhancing then allergen exclusion through an immunological response	[179–184]
Skin	Skin damage, such as eczema, is frequently associated with food allergy, and approximately one in five infants with infantile eczema will go on to develop a food allergy. This occurs because in damaged skin, depending on the nature of the allergen, epithelial cells can produce cytokines that instruct dendritic cells on the skin	[185, 178]
Exposure to foods	Late introduction of allergenic foods into the diet has been associated with higher risks of food allergy, compared with an early introduction	[186]

Continued on next page

Table 1.8 – continued from previous page

Factor	Evidences	References
Genetic sex	The male/female ratio of children with food allergy is 1.8, whereas for adults, it is 0.53. Studies identified even higher disparity for specific food allergens, such as peanut, where the male/female ratio in children was almost five, whereas for adults it was less than one. This disparity has been usually ascribed to sex hormones, since these are one of the most obvious physiological differences between adult males and females, and their impact on immune system function is well recognised	[187–189]
Dietary factors	Exposure to an increased diversity of allergenic foods in early life is inversely associated with allergic diseases including food allergy. It's been proposed that the increased consumption of fatty acids from margarine and vegetable oils, and through reduced consumption of animal fats, led to an increase in allergies. Also, breast milk modulates microbiota and confers immunological protection when the infant's immune system is immature (it contains, hormones, growth factors and cytokines among many others)	[190, 187, 191]
Dietary antioxidants	Increased beta-carotene intake was associated with a reduced risk of allergic sensitisation and lower IgE levels in 5- and 8-year-old children	[192]
Obesity	It induces an inflammatory state associated with an increased risk of atopy and theoretically could lead to an increased risk of FA	[187]

Continued on next page

Table 1.8 – continued from previous page

Factor	Evidences	References
Vitamin D	Epidemiological and immunologic data that suggest that either excessive vitamin D or, conversely, vitamin D deficiency (predominantly caused by low sunlight exposure) results in increased allergies	[187]
Contamination	Chemical contamination affecting plant foods have been suggested to influence on plant food allergens	[193]

Another interesting aspect are the different effects of the environment on individuals with specific variants (also called gene-environment interactions). These reflect the complex interplay between environmental exposures (including lifestyle and diet) and genetic predispositions to modify disease risk, and could explain why food allergies, like many other complex diseases, exhibit a heritable component but do not follow Mendel's laws. Recent studies have shown that gene-environment interactions may explain a proportion of phenotypic variance.

For example, the *GSTP1* [MIM: 134660] NP_000843:p.Ile105Val polymorphism modifies the effect of air pollution on allergic sensitisation to inhalant and/or food allergens [194], and the NM_000591:c. 159 CC>TT polymorphism in the *CD14* gene [MIM: 158120], which has an increased protection from eczema with dog exposure [195], could depend on the microbial stimulation from the environment [195–197].

On the other hand, gene-gene interactions are also likely to contribute to the complexity of food allergies, where genetic variants in genes involved (e.g. in the Th2-cell differentiation and signalling pathways) can

also contribute to the allergic phenotype. A study performed in Germany with 1,120 children aged from nine to eleven years old genotyped several polymorphisms in the respective genes of the IL-4/IL-13 pathway. They observed that combining polymorphisms leads to an increased risk for asthma and high serum IgE levels, compared with the maximum effect of any single polymorphism [198].

1.3.11 Epigenetics

Epigenetics mechanisms such as methylation, acetylation, phosphorylation, ubiquitylation, and sumoylation play an important role in gene expression patterns and can be inherited independently of changes in DNA sequence. An increasing number of studies suggest that allergic disorders can also be affected by epigenetic regulations. Syed, *et al.* [199] found that CpG sites in *FOXP3* [MIM: 300292] were differentially demethylated in children with immune tolerance of peanut allergy compared to children without tolerance. At the same time, Martino *et al.* [200, 201] examined DNA methylation profiles in CD4+ T-cells in 24 infants with and without IgE-mediated FA diagnosed at 12 months. The authors suggested that the allergic phenotype may be affected by dysregulated DNA methylation in genes involved in the mitogen-activated protein kinase (MAPK) cascade during early CD4+ T-cell development. Therefore, DNA methylation in the regions of genes related to T-cell differentiation and balance between Th1 and Th2 during the critical period of early life may be potential mechanisms of allergic disease development.

1.3.12 GI food allergies summary

Food allergy is a complex disorder presenting with a wide variety of phenotypes that make the proper diagnosis and management difficult. The molecular mechanisms of this disease still remain poorly characterised, and the absence of a suitable treatment also reveals the need for understanding the molecular mechanism of the disease. This disorder is likely to be a result of a complex interplay between epigenetics, environmental factors and genetics. In order to elucidate the genetics part, numerous studies have been focused in the study of common variants in food allergies by GWAS. However, results have been modest so far and the understanding of the complex biological pathways and mediators involved remains unknown.

Recent advances in NGS have increased the analysis throughput while reducing costs, turning it into a candidate technology to pursue other types of genetic variation of interest to food allergy. Therefore, when eight affected individuals from seven families with severe GI food allergy to multiple food proteins where gathered by INCLIVA research institute (Valencia, Spain), exome sequencing was selected as technology to investigate the effect of rare variants in the phenotype of these individuals.

1.4 Clinical case

Eight children from seven families affected with severe GI food allergy to multiple food proteins were identified. Affected individuals presented non-IgE mediated allergic responses after the ingestion of most solid foods since their first year of life. These individuals presented with vom-

iting, diarrhoea, abdominal weakness and severe pain after the ingestion of multiple solid foods. Most of the patients had abnormal breastfeeding and manifested the phenotypes in the first month of life. Due to severity of the phenotypes, food intolerance was promptly discarded. EoE, gastritis and Coeliac disease were also discarded for some patients by endoscopic biopsies. Affected individuals were under examination for many years without a clear diagnosis. After a long diagnostic odyssey, the majority of them were diagnosed with severe FPIES (Food Protein-Induced Enterocolitis Syndrome). These individuals could only be fed by a Percutaneous Endoscopic Gastrostomy (PEG) or with *Neocate*, a hypoallergenic amino acid-based infant formula for the dietary management of different kinds of allergies. In early adolescence, specific cases started tolerating some types of aliments. All this together, and the presence of blood in stools and transition to IgE positivity in some cases, was consistent with the FPIES diagnosis. However, the phenotypic presentation of these individuals was somewhat different.

Whereas FPIES is triggered by specific offending foods (e.g. cow's milk, soy and rice), these children were symptomatic after the exposure to multiple types of solid food, triggering similar symptoms of FPIES. Extreme presentations of suspected FPIES have also been reported, where individuals were symptomatic with the introduction of most solid foods [202]. In this work, authors argued that this could be a presumed severe form of non-IgE mediated food allergy, but it could also represent a new syndrome. These individuals fulfilled three main criteria: 1) non-IgE mediated cow's milk and soy allergy commencing in infancy, 2) asymptomatic on amino acid-based formula, and 3) GI symptoms (diarrhoea, vomiting, abdominal distension and severe irritability) with the introduction of a broad range of foods. This criteria was consistent

with the one our patients presented. Most of them were also males (87%).

Interestingly, a strong family history of allergic phenotypes was observed in almost all the families, and relatives often presented lactose intolerance, pollen and food allergies or other diarrhoea issues. Due to these correlations and the role that genetic factors play in food allergies, the demand for discovering new genes that may be involved in the pathogenesis of this disease was raised. Furthermore, because the affected individuals were very severely affected, they were suspected to harbour more deleterious variants in candidate genes than individuals mildly affected. Identification of these genes could help us to understand the molecular basis of the disease, which is important to perform adequate diagnosis, and to discover new therapeutic targets.

Given the capacity for discovering genetic variations contributing to rare diseases and the availability of resources, WES was chosen as first approach for the study of rare variation in these eight patients with severe GI food allergies induced by multiple food proteins and their relatives. This work is the first study of individuals with this phenotype, and presents potentially interesting results that could allow us to understand the pathogenesis of this complex disease.

Chapter 2

Hypothesis and Aims

2.1 Hypothesis

The hypothesis of this work are that:

- Rare genetic variants contribute to the development of GI food allergies induced by multiple food proteins, and these are likely to be present in coding regions of one or multiple genes.
- Whole-exome sequencing is a powerful technology to investigate multiple types of genomic variation in these affected individuals.
- Identification and interpretation of these variants in eight severely affected individuals could facilitate the understanding of its pathogenesis, hence providing a better diagnosis and management of other cases affected with this disorder.

2.2 Aims

The main aim of this work is:

- To characterise the mutational spectrum of seven families affected with gastrointestinal food allergy induced by multiple food proteins, in order to investigate the role that rare genetic variants may play in the development of the disease.

The detailed aims of this work are:

- To develop a workflow to process the exome sequencing data from raw signal to genetic variants, including SNV/indels, CNVs and HLA haplotypes.
- To assemble a list of candidate genes associated with immunological disorders.
- To perform a comprehensive quality control analysis of the data obtained.
- To identify rare genetic variants and pathways associated with the disease, and to assess the possible contribution they may have in the development of gastrointestinal food allergy.

Chapter 3

Methods

In this chapter the recruitment criteria for the affected individuals, as well as the methods for the WES analysis are described. Because sequencing was performed in multiples batches, data had to be merged and filtered in order to remove errors from the sequencing. Therefore, a number of recommendations that can be used in order to maximise calling of true sites of variation are suggested. The workflow for the automated analysis of rare SNVs/indels and CNVs, as well as HLA typing is presented. Finally, the quality control analysis of the data is also included in the workflow.

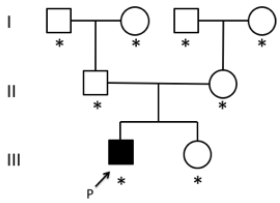
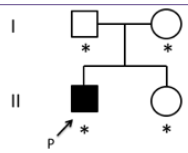
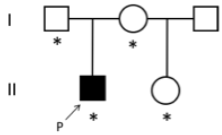
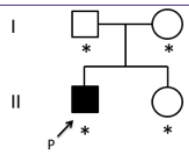
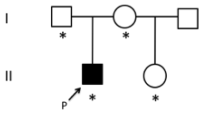
3.1 Patient recruitment

Recruitment was performed by the collaboration between the Genotyping and Genetic Diagnosis unit of the INCLIVA research institute (Hospital Clínico de Valencia, Valencia, Spain), Garmitxa association (Basque Country, Spain), Euskal BioBankoa (Basque Country, Spain) and the

Institute of Medical and Molecular Genetics (INGEMM, Hospital Universitario la Paz, Madrid, Spain). The criteria for selecting individuals for sequence analysis were i) affected individuals had to present with gastrointestinal food allergy after the ingestion of most solid foods, ii) no known genetic cause of disease previously identified and iii) family pedigree had to be available for further study and sequencing.

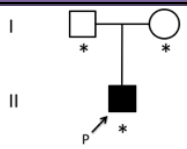
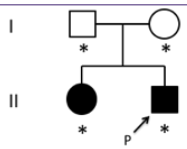
The cohort consisted of DNA samples from 31 individuals from seven families, eight of which were affected. Within research ethical framework (IRAS 03/0/014 and 13/EE/0325) participants, parents or guardians provided written informed consent to participate in the study. Family pedigrees are presented in Table 3.1. Individual identifiers were constituted by the family number followed by the individual identifier based on the family relationship to the proband. Therefore, affected probands have the extension 01, then mothers have 02, fathers 03 and siblings, if present, 04. For larger pedigrees, IDs were given by proximity of relationship to the proband.

Table 3.1 Familial pedigree structures. Affected individuals are indicated with a P (of proband). Sequenced individuals are shown with an asterisk. Individual IDs are only provided for sequenced individuals.

Family	Pedigree structure	Relationship	Individual ID
F01		I-1 Paternal grandfather	F01_05
		I-2 Paternal grandmother	F01_06
		I-3 Maternal grandfather	F01_07
		I-4 Maternal grandmother	F01_08
		II-1 Father	F01_03
		II-2 Mother	F01_02
		III-1 Proband	F01_01
		III-2 Sister	F01_04
F02		I-1 Father	F02_03
		I-2 Mother	F02_02
		II-1 Proband	F02_01
		II-2 Sister	F02_04
F03		I-1 Father	F03_03
		I-2 Mother	F03_02
		II-1 Proband	F03_01
		II-3 Half-sister	F03_04
F04		I-1 Father	F04_03
		I-2 Mother	F04_02
		II-1 Proband	F04_01
		II-2 Sister	F04_04
F05		I-1 Father	F05_03
		I-2 Mother	F05_02
		II-1 Proband	F05_01
		II-2 Half-sister	F05_04

Continued on next page

Table 3.1 – continued from previous page

Family	Pedigree structure	Relationship	Individual ID
F06		I-1 Father I-2 Mother II-1 Proband	F06_03 F06_02 F06_01
F07		I-1 Father I-2 Mother II-1 Sister II-2 Proband	F07_03 F07_02 F07_01 F07_04

3.2 Exome Sequencing

Exome sequencing is based on the sequencing of millions of short length reads of DNA, which are enriched for the exome sequence. WES sequencing workflow is based on three main steps: library preparation (from nucleic acid sample), amplification (to produce clonal clusters) and sequencing (using massively parallel synthesis).

In this study, sample preparation was done using two different protocols: SureSelectXT Human All Exon V5 + UTRs kit (Agilent Technologies, Santa Clara, CxA, USA) [203], and Nextera Rapid Capture Exome kit (Illumina, San Diego, CA, USA) [204], termed later for short SureSelect and Nextera. Sequencing was performed in three different platforms (HiScanSQ and HiSeq1500 for Nextera, and HiSeq2000 for SureSelect), and in order to reach high coverage, seven different batches

of sequencing for different samples were done, across three different centres: INCLIVA (Valencia, Spain), Health in Code (HIC, La Coruña, Spain) and Centre for Genomic Regulation (CRG, Barcelona, Spain).

3.2.1 Sample preparation

The genomic library is formed by genomic fragments of DNA (gDNA) with the adapters added at the ends of the fragments, ready for further amplification and sequencing. In order to obtain the libraries, DNA needs to be fragmented into smaller fragment size (ranged from 200 to 800 bp), since the platforms that were used here can read sequences until 100-150 bp of length from both ends of the fragment. Then the sequencing adapters with the barcodes are added, constituting the genomic library. Finally, this is enriched for the exome by using probes marked with biotin, that will hybridise to the complementary DNA, and will then be captured back using streptavidin beads.

Genomic libraries

gDNA from 31 individuals was obtained from blood extraction using Chemagen o Maxwell systems following the corresponding protocols, and quantified with Quant-iTTM PicoGreen® dsDNAAssay Kit (InvitrogenTM). Measures were done by spectrofluorometer GLOMAX® Multi Detection System (Promega) following the specifications. All samples were diluted to start thereby with the DNA recommended by Illumina (1μg of gDNA for SureSelect and 50ng for Nextera).

Of the seven batches of sequencing performed, five had libraries constructed with Nextera and two with SureSelect. The main difference between the protocols is that in Nextera, fragmentation and adapter

ligation occurs simultaneously since this is mediated by tagmentation (which involves transposons cleaving and tagging the double-stranded DNA, with a minimum distance of 300 bp) (Figure 3.1 A).

Instead, SureSelect protocol needs the gDNA to be fragmented, ends repaired and adapters ligated in different steps (Figure 3.1 B). Here, Covaris S220 technology, a focalized utrasonicator, was used to fragment the DNA. Settings were as recommended by Illumina, and fragmentation was done making 200-300 bp length fragments of DNA (Figure 3.1 B-A). After fragmentation, an End-Repair Mix with a 3' to 5' exonuclease was used to remove the 3' overhangs and the polymerase activity filled in the 5' overhangs (Figure 3.1 B-B). Then, a single 'A' nucleotide was added to the 3' ends of the blunt fragments to prevent them from ligating among themselves during the adapter ligation reaction. A corresponding single 'T' nucleotide on the 3' end of the adapter provides a complementary overhang for ligating the adapter to the fragment. This strategy ensures a low rate of chimera (concatenated template) formation (Figure 3.1 B-C).

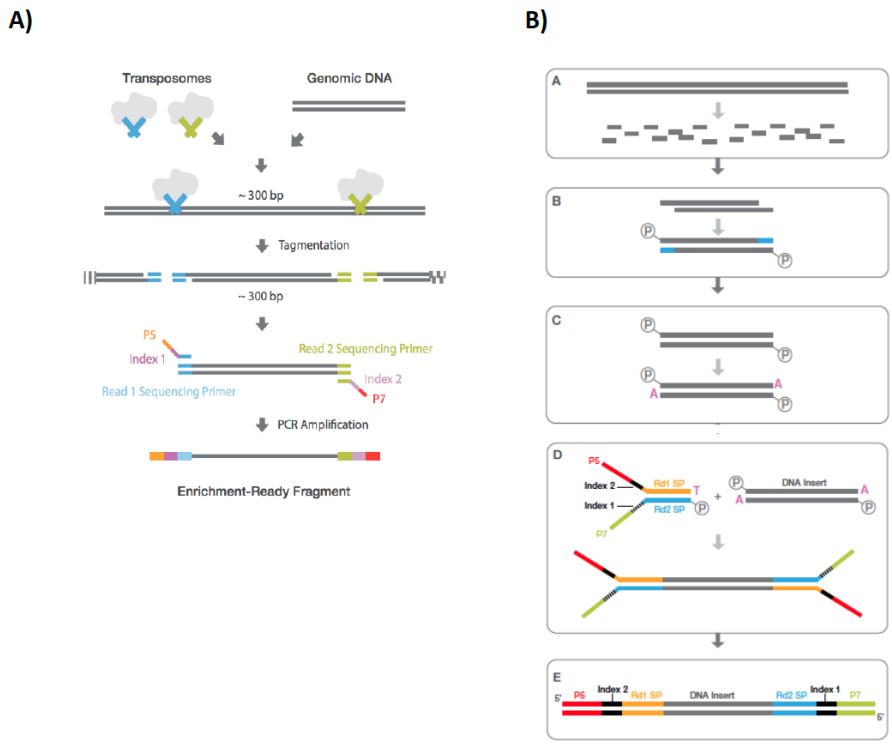


Fig. 3.1 Library preparation steps. (A) Schema of Nextera one-step protocol. (B) Schema of SureSelect four-step protocol.

The adapters, that are required for both SureSelect and Nextera protocols, are formed by different sequences, important for posterior steps during the sequencing:

- P5 and P7 are primers that contain an attachment site to the flow cell.
- Rd1 SP and Rd2 SP are complementary to the primers that start the sequencing of the fragment.

- Index: is a unique identifier of 6 bp for each sample. It allows multiplexed sequencing, running multiple individual samples in one lane. After the sequencing, all the reads are mixed together, and they will be separated (demultiplexed) by sample using this unique identifier.

Different combinations of indexes were used for each sample, following the Illumina recommendations. The genomic library was finally enriched by PCR for those fragments that have adapter molecules on both ends. The PCR was performed with a PCR primer cocktail that anneals to the ends of the adapters, following the instructions from the manufacturer. This final mixture contained the genomic library, amplified and ready for enrichment.

Exome enrichment

Target enrichment was performed with SureSelect and Nextera. Specifications for targeted regions are shown in Table 3.2.

Nextera and SureSelect systems use different types of baits for enrichment. SureSelect uses biotinylated cRNA baits, and Nextera uses biotinylated DNA baits to capture known coding DNA sequences (CDS) from the NCBI Consensus CDS Database, as well as other major RNA coding sequence from databases like miRbase (microRNA database from Sanger institute). Genomic libraries were hybridised with these biotinylated baits, complementary to CDS. The captured sequences were then enriched with streptavidin-conjugated paramagnetic beads and further amplified before being subjected to Illumina sequencing (Figure 3.2).

Table 3.2 Enrichment set characteristics

	SureSelect	Nextera
Target size	75 Mb	62 Mb
Number of exons	359,555	201,121
Overall workflow	1.5 days	1.5 days
Genomic DNA input	1 μg	50 ng
Adapter ligation	Ligation	Transposase
Baits	Biotinylated cRNA	Biotinylated DNA
Expected on-target reads	>80%	>70%

The size of the DNA fragments was checked throughout the protocol procedure, using a capillary electrophoresis gel technology (QIAxcel DNA Screening Kit from QIAxcel (Qiagen)), since it is more sensitive than traditional agarose gel method.

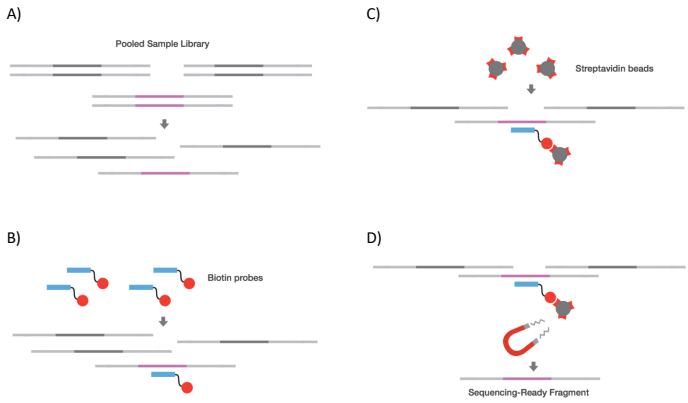


Fig. 3.2 Exome enrichment steps. (A) Denaturalization of double-stranded DNA library (for simplicity, adapters and indexes are not shown); (B) Hybridisation of biotinylated probes to targeted regions; (C) Enrichment using streptavidin beads; (D) Elution from beads.

3.2.2 Clonal amplification

Prior to sequencing, single-molecule DNA templates were bridge amplified to form clonal clusters inside the flow cell. Clonal amplification for each single-molecule DNA was performed with the cBOT system from Illumina (San Diego, CA, USA). Essentially, clonal amplification has three steps:

1. Immobilisation of single-molecule DNA templates: hundreds of millions of templates are hybridised to the flow cell surface and copied using a DNA polymerase. The original templates are denatured, leaving the copies immobilised on the flow cell surface (Figure 3.3-A).

2. Isothermal bridge amplification: immobilised DNA template copies are then amplified by isothermal bridge amplification to create millions of individual, dense clonal clusters containing ∼2,000 molecules (Figure 3.2-B).

3. Linearization, blocking, and primer hybridisation: each cluster of double strand DNA bridges is denatured, and the reverse strand is removed, leaving only the forward DNA strand. The sequencing primer is hybridised to the complementary sequence on the Illumina adapter, and this is ready to be sequenced. At this point the flow cell contains >200 million clusters with ∼1,000 molecules/-cluster (Figure 3.2-C).

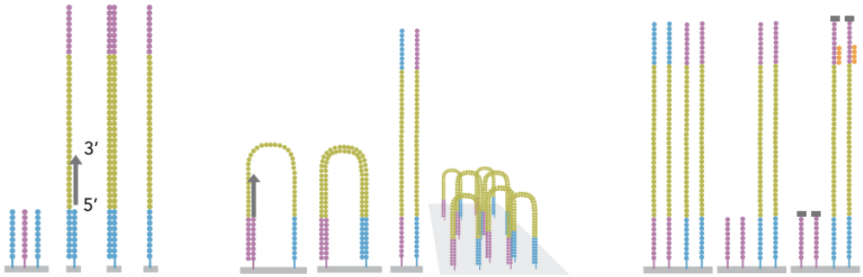


Fig. 3.3 Cluster generation. Cluster generation from single-molecule DNA templates occurs within the sealed Illumina flow cell on the cBOT instrument, and involves immobilisation and 3' extension, bridge amplification, linearization, and hybridisation.

3.2.3 Sequencing

Posterior to the clonal amplification, sequencing of 100bp paired-end reads was carried out on different Illumina HiSeq systems. Illumina sequencers are based on Sequencing By Synthesis (SBS) technology, that uses four fluorescently labelled nucleotides with reversible terminators

[205] to sequence the tens of millions of clusters on the flow cell surface in parallel. During each of the 100-150 sequencing cycles, a single labelled deoxyribonucleoside triphosphate (dNTP) with reversible terminator is added to the nucleic acid chain. If the nucleotide is incorporated, it acts as a terminator for polymerisation, and the fluorescent dye is imaged to identify the base and then the terminator and the fluorescent tag are cleaved enzymatically to allow incorporation of the next nucleotide. Base calls are made directly from signal intensity measurements during each cycle. (Figure 3.4).

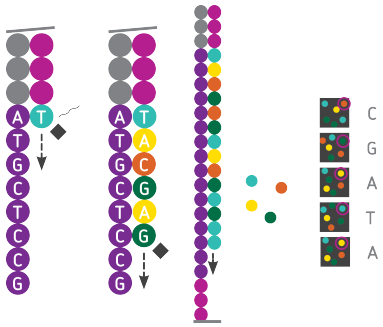


Fig. 3.4 Sequencing by synthesis. Each dNTP has a corresponding fluorophore attached to it. When the DNA polymerase elongates the strand with a fluorescently-labelled dNTP, the clusters are then excited by a light source and the colour is recorded by an optical detector. After incorporation occurs, the fluorophore is cleaved, unblocking for the next nucleotide to be incorporated in the next cycle. Since each cycle one permits the elongation of a single dNTP at a time, homopolymers are determined precisely.

3.3 Data processing

The first computational step entails the conversion of the raw data (fluorescent signal) into nucleotide bases. This process is termed "base calling" and, as mentioned above, it occurs in the sequencing machine.

The output are the sequenced reads in a text file. A general workflow for variant discovery is based on the alignment of these reads to the genome of reference and the subsequent identification of those positions that differ from the reference which will be called as variants. Variants are then annotated with additional information and filtered by different criteria for further investigation. In order to carry out this analysis, a customised workflow was developed to perform an automated analysis of the data, using a specific selection of the most suitable algorithms. In this pipeline, the Genome Analysis Toolkit (GATK, Broad Institute) Best Practices recommendations [206] were followed, using multiple programs and custom scripts. All the programs and commands used are publicly available in GitHub (http://github.com/alsanju/wes-pipeline). A schema with more detailed information of the workflow, that will be explained in this section, is shown in (Figure 3.5).

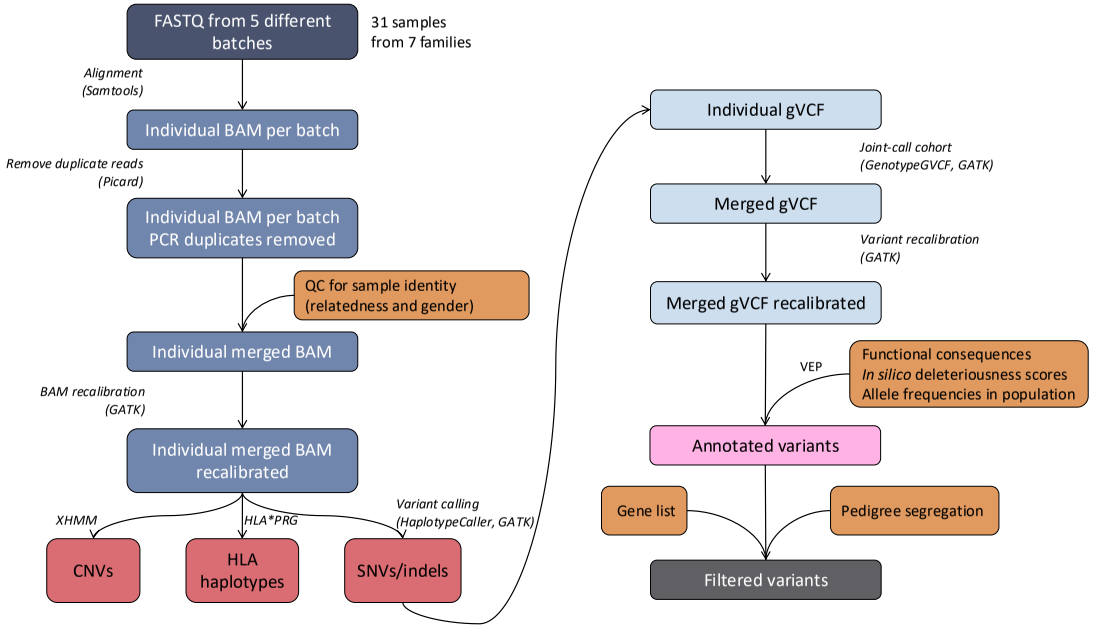


Fig. 3.5 Schema of the WES analysis workflow. Sequencing data from 31 individuals was obtained from seven different batches. Alignment was performed for each individual per batch. Quality control analysis was carried out to confirm sample identity, and individual BAM files were merged. Three different types of analyses were done: Copy Number Variant (CNV) analysis, HLA haplotyping and SNV/indel calling. SNVs and indels were processed using GATK pipeline, then annotated with information for functional consequences, deleteriousness scores and allele frequencies using Variant Effect Predictor (VEP) and other sources. Finally, candidate variants were filtered by mode of inheritance using pedigree segregation information, or by gene list.

3.3.1 Image analysis and demultiplexing

Illumina sequencing instruments generated per-cycle BCL basecalled files as primary sequencing output, which were converted to FASTQ files by the software bcl2fastq (Illumina, San Diego, CA, USA). This also performed the demultiplexing, where samples were separated into individual ones by their specific indexes (the 6 bp sequences that were in the adapter, and were unique for each sample).

FASTQ files store the sequences and their corresponding quality scores, encoded as a single ASCII character for pairing the array of letters with the array of its qualities. These files use four lines per sequence, as shown in Figure 3.6.

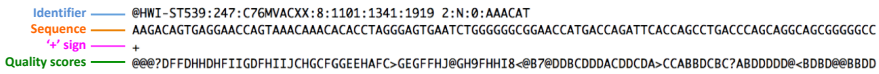


Fig. 3.6 FASTQ file format. Shows the information for one read in a FASTQ file: first line is the read identifier, second line is the read sequence, third line is the '+' sign and fourth line are the quality scores for each of the bases in line two.

3.3.2 Alignment

The sequencing reads were aligned to the human genome reference sequence (with decoy, hs37d5), based on the GRCh37 assembly, using Burrows-Wheeler Alignment (BWA) tool [207]. The decoy human genome integrates the reference sequence of the GRCh37 primary assembly (chromosomal plus unplaced contigs), the revised Cambridge Refer-

ence Sequence (rCRS) mitochondrial sequence (AC:NC_012920), Human herpesvirus 4 type-1 (AC:NC_007605) and the concatenated decoy sequences (concatenated sequences with 20 "n" bases filled between adjacent sequences) (ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz). Therefore, when doing the alignment against the decoy genome as a reference, some reads will quickly find a very confident alignment in the decoy, avoiding countless compute cycles spent trying to Smith-Waterman align it to someplace it doesn't belong. This results in a significantly higher speed of the alignment step.

After the alignment step, the output are BAM files, the compressed binary version of the Sequence Alignment Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments. They had information for the coordinates of the mapped read, as well as for the read quality, length, read group, flow cell and library information among others (Figure 3.7).

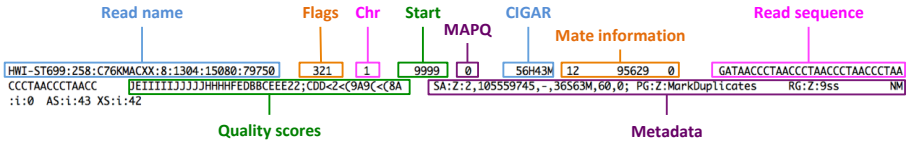


Fig. 3.7 BAM file format. For one read, the following information appears in the BAM file: the read name, flags, chromosome, start position of the alignment, mapping quality, CIGAR string, information about the mate, the actual read sequence, quality scores and metadata that contains additional information about the read.

The resulting BAM files needed to be processed. First, BAM files were sorted and merged by sample using Samtools (a suite of programs

for interacting with HTS data), options *sort* and *merge* respectively. However, because samples had been sequenced in different lanes/runs/centres and by different staff, they were merged after a relatedness analysis corroborated the identity of each sample, avoiding possible labelling mistakes or sample swaps while processing the sequencing libraries. The methods for the relatedness analysis are next described in Section 3.4.3. Duplicates can arise during PCR amplification steps, incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument. These duplication artefacts were flagged and taken into account for future steps using Picard tool (option *MarkDuplicates*) which locates and tags duplicate reads in a BAM file.

Afterwards, *IndelRealignment* was also used to perform a local realignment of specific reads to minimise the number of mismatching bases. This two-step indel realignment process first identifies such regions where alignments may potentially be improved (which are those with indels or repetitive regions), then realigns the reads in these regions using a consensus model that takes all reads in the alignment context together.

Lastly, the quality base score was recalibrated by adjusting the phred quality scores (quality of each base that has been read by DNA sequencer machine) to be more accurate, using GATK recommendations, as specified in the workflow. All commands used are publicly available in GitHub (http://github.com/alsanju/wes-pipeline).

3.3.3 Variant calling and annotation

SNVs and indels

Variant calling was performed to identify the sites where there was variation respect to the reference genome, then presented in VCF format (Figure 3.8). The calling depends heavily on accurate mapping to the reference genome, and is accomplished by statistical modelling methods that are optimised to distinguish genuine variation from sequencing errors [208]. One such improvement was the incorporation of a level of uncertainty for calling a genotype at a specific position, rather than just simply determining the genotype based on read counts.

The average error rate of NGS per single read is reported to be 0.1% per nucleotide, most of which are single nucleotide substitutions [209]. This is higher than the error rate of Sanger sequencing, that can read lengths of up to ∼1,000 bp at a per-base accuracy of 99.999%. As these errors are mainly random, the problem is usually attenuated by sequencing at a high depth. This was approached by the design of this study, which aimed for a high coverage (>= 50x/sample), and by downstream QC of variants and samples. Additionally, joint variant calling was performed using GATK *HaplotypeCaller* [206], which calculates the likelihoods of each possible genotype, and selects the most likely by applying a Bayesian model.

HaplotypeCaller is one of the best-established tools for calling SNVs and indels, and was the one used in this workflow [206]. This has two separate steps: per-sample calling and genotyping across samples. *HaplotypeCaller* runs first on each sample separately in gVCF mode, to produce an intermediate file format called gVCF (for genomic VCF). A gVCF is similar to the VCF format, so that the basic format specification

is the same, but a genomic VCF contains not only the sites with variation but also extra information with all sites with no variation, allowing to differentiate homozygous reference positions from no calls. A gVCF therefore has records for all sites, whether there is a variant call there or not. It contains information for the coordinates of the variant, the reference and alternative alleles, and genotype quality scores (Figure 3.8).

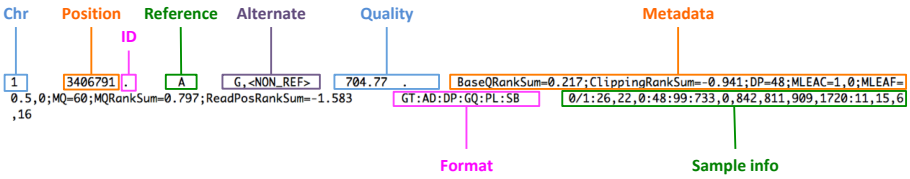


Fig. 3.8 VCF file format. In the variant call format file, there is an entry per variant called. For each variant, there is information for: chromosome, position, identifier (if the variant already has been reported), reference allele, alternate allele, quality information, metadata, format and sample information, which includes, among others, genotype and PL values (probabilities of the variant for being homozygous for the reference allele, heterozygous, or homozygous for the alternate allele).

The gVCFs of multiple samples are then run through a joint genotyping step using *GenotypeGVCFs*, to produce a multi-sample VCF callset, which can then be filtered to balance the sensitivity and specificity as desired. The multi-sample joint calling merges the records at each position of the input gVCF, producing correct genotype likelihoods. It also resolves the so-called N+1 problem. The N+1 problem occurs when a large number of samples sequenced in different batches is obtained. When new sample/s sequence are included, if a true joint analysis is desired, the re-call of all samples from scratch would need to be performed every time. Running *HaplotypeCaller* on each sample separately and

then performing a joint genotyping by family scales better and resolves the problem.

After the variant calling, the GATK Best Practices suggest performing a variant quality score recalibration to filter the variants and identify annotation profiles of variants that are likely to be real. However, this step was not performed in this workflow since this method requires a large callset (and there were only 31 samples included in this study). The number of variants identified at this point depends on many factors, but it can range from 10,000-50,000 variants in exome sequences. While these numbers represent a challenge in interpretation, there are several biological annotations that are normally added at this stage to facilitate downstream genetic analyses and extract meaningful biological information from the data itself.

Functional-based annotations determine the effect of a variant on the transcript/s and encoded protein/s, based on the resulting amino acid change. For this, Variant Effect Predictor (VEP) version 88 [210] was used, providing well-defined terms for each variant (Figure 3.9).

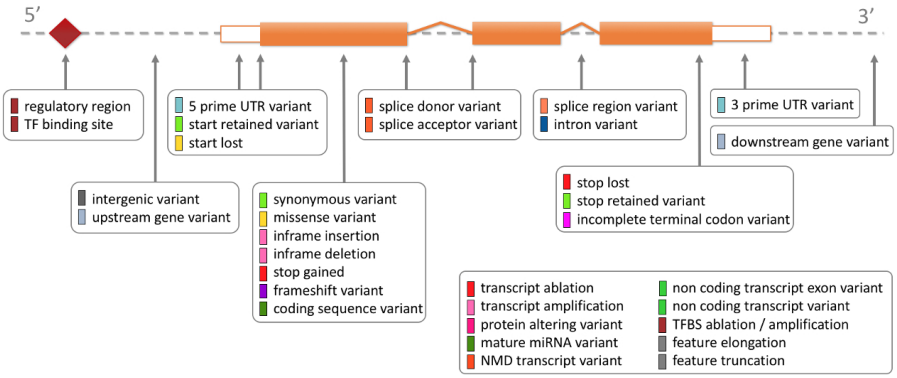


Fig. 3.9 Functional consequences at the protein level. The diagram illustrates the functional terms given by VEP tool. Detailed descriptions of each term are represented in: http://www.ensembl.org/info/genome/variation/predicted_data.html.

Annotation of deleteriousness of changes on the resulting protein can also be done, taking into account sequence conservation in homologous sequences (eg. SIFT, CADD) or structural properties, such as the impact in the tri-dimensional protein structure (e.g. PolyPhen). Finally, annotation with allele frequency information from population databases is a crucial step to differentiate between common and rare/ultra-rare variation. A list of the sources for variant annotation used in this study is represented in Table 3.3.

Copy Number Variants

Copy number changes (deletions and duplications) were detected based on the read depth using the eXome Hidden Markov Model (XHMM) program [100]. Because CNV detection from WES data is challenging due the variable coverage across the genome, only samples with an

Table 3.3 Annotation sources

Source	Description
SIFT	Predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. Substitutions with a score < 0.05 are called 'deleterious' and all others are called 'tolerated'. Version: sift5.2.2
PolyPhen	Predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. Values nearer one are more confidently predicted to be deleterious. Version: 2.2.2
CADD	Tool for scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome [211]. It uses many different annotations for its combined score. A scaled C-score of greater of equal 10 indicates that this variant is predicted to be the 10% most deleterious substitutions that you can do to the human genome, a score of greater or equal 20 indicates the 1% most deleterious and so on. Version: v1.4
Minor Allele Frequencies	Data for existing variants from major genotyping projects: 1000 Genomes Project: contains variation and genotype data from 1000 individuals from different ancestries (Version: phase3). NHLBI-ESP: well phenotyped populations from the United States of more than 200,000 individuals with different disorders (Version: 20141103). gnomAD: resources of sequencing data from 123,136 exomes and 15,496 genomes from unrelated individuals sequenced as part of various disease-specific and population genetic studies (Version: r2.0.2) [26]
ClinVar	Archive of human variations and phenotypes, with supporting evidence. It allows identification of variants previously that have been reported as associated with disease. Version: 20170530
GTEx	Resource of tissue-specific gene expression and regulation data from 53 non-diseased tissue sites across nearly 1000 individuals. Version used: GTEx Analysis Release V7 (dbGaP Accession phs000424.v7.p2)

average coverage higher than 80x were considered for this analysis (23 individuals). The key steps in running XHMM include 1) running coverage calculations from alignment files, 2) data normalisation, 3) CNV calling and 4) statistical genotyping.

XHMM relies on read depth as the sole source of information on CNV events, ignoring split read and read pair information. To handle normalisation, it creates a matrix of the depth of all exons in all samples, and the principal components of this matrix are expected to capture many of the non-CNV factors that affect an exon's read depth. XHMM performs better in detecting rare CNVs, whereas common CNVs may go undetected since they are present in the reference samples used for PCA.

After normalisation, XHMM calls CNVs using a Hidden Markov Model (HMM). HMM is based on the fact that (sufficiently large) CNVs will affect a whole contiguous swath of exons, so the probability of an exon to be deleted/duplicated would be considerably higher if the neighbour exon is (Figure 3.10). M. Fromer *et al.* previously described how to run XHMM [107], and this script was implemented in the pipeline. This sofware was selected because its has been largely implemented to study CNVs in 60,642 individuals [212], which data was used to annotate the variants obtained in this study.

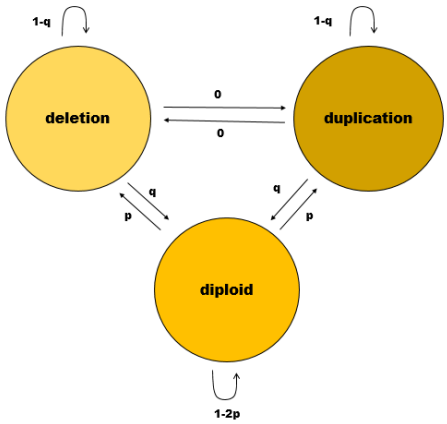


Fig. 3.10 XHMM strategy. Hidden Markov models rely on probabilities of transitions between states, and the XHMM needs just two quantities from which to base all of its probabilities. p is the rate of exonic CNVs, and q is the reciprocal of the average CNV length (number of exons). From http://www.cureffi.org/2014/01/17/comparison-of-tools-for-calling-cnvs-from-sequence-data.

HLA typing

HLA haplotypes were inferred using HLA*PRG [112]. HLA*PRG addresses the unique challenges of calling HLA haplotypes by aligning reads from the HLA genes to a Population Reference Graph (PRG) of the HLA genes and then evaluating the graph-aligned reads in a likelihood framework. A PRG is a graphical model for genetic variation, where alternative alleles, insertions and deletions are represented as alternative paths through the graph [213]. The reads from the HTS that are likely to arise from the HLA region are mapped directly to the graph structure, thus enabling the identification of the greatest continuity along a path (Figure 3.11). This step is very expensive computationally and needs 70-80GB of memory per sample.

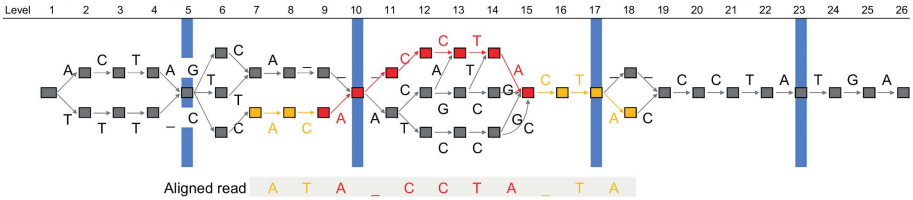


Fig. 3.11 Schematic HLA type inference. The aligned sequence of the read is displayed below the PRG, and the alignment path is highlighted. The red component of the alignment path corresponds to the exact-match component of the alignment, whereas the yellow components correspond to those components of the alignment where mismatches are allowed. From [112]

Each HLA allele name has an HLA prefix followed by the gene, a separator and a unique number corresponding to up to four sets of digits separated by colons (Figure 3.12). The digits between the separator and the first colon describe the type, which often corresponds to the serological antigen carried by an allotype (allele of the antibody). The next set of digits provide information about the subtypes, synonymous nucleotide substitutions and non-coding substitutions in the third and fourth set of digits respectively.

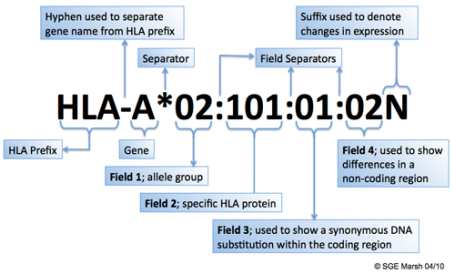


Fig. 3.12 Nomenclature for factors of the HLA system. Each allele name has a unique corresponding set of numbers and letters. HLA prefix is followed the by HLA gene before the separator. Then, the different fields are comma separated. First appears the allele group, followed by the specific protein, synonymous substitutions within the coding regions, and differences in a non-coding region. Source: http://hla.alleles.org/nomenclature/naming.html

Here, Sequence Based Typing (SBT) was carried out at 6-digit "G" resolution (three sets of digits). Only sequences of the exons encoding the peptide binding groove - exons two and three of the class I genes (*HLA-A, -B, -C*), and the exon two of the class II genes (*HLA-DQA1, -DQB1, -DRB1, -DRA1, -DPB1*) - were considered. In order to get good quality HLA types, HLA*PRG was run on samples with an average coverage greater than 80x (23 individuals). In order to perform later an association test, HLA typing was also done on 120 internal controls with no reported food allergy.

3.4 Quality control

Before starting with the variant interpretation, a series of quality control (QC) assessments were performed at different stages of the analysis, to make sure the sequencing data were of high quality. Because in this study samples were recruited at different times by different centres, involving multiple associated staff performing independent data collection, there was a need to perform exhaustive quality control of the genomic data.

3.4.1 Assessing sequencing quality

Quality of the sequenced samples was assessed by detecting: 1) the per base quality, 2) exome coverage, 3) the ratio of transitions (interchanges of two-ring purines (A G) or pyrimidines (C T)) to transversions (interchanges of purine for pyrimidine bases) (Ts/Tv ratio), and 3) the number of variants called.

The per base quality was obtained running FastQC software on the FASTQ files. Additional information such as read length distribution

and GC content of the sequences was also obtained by this software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc).

Alignment performance was checked using different mapping statistics (such as percentage of mapped reads, or percentage of properly paired reads) obtained from the BAM files using Samtools *stats* option. Samtools was also used to calculate coverage in the exome, with the depth option. Variant evaluation metrics were obtained using *CollectVariantCallingMetrics* tool from Picard, which calculates general statistics, such as the number of SNPs and indels, and the Ts/Tv ratio.

3.4.2 Computation of genomic sex

Genomic sex was estimated from the BAM files. For each sample and chromosome, the number of aligned reads (obtained running Samtools *idxstats* option) was normalised by dividing them by the number of bases which are non-N in the reference genome. The X/Auto and Y/Auto ratios were defined as the normalised read counts on X and Y divided by the median of the normalised read counts on the autosomes (Auto).

Females should have higher X/Auto ratio (theoretical 1) than males (theoretical 0.5), and males should present with higher Y/Auto (theoretical 0.5) than females (theoretical 0). Here, it was established that if the X/Auto – Y/Auto was higher than 0.5, the sample was deemed to be female; if smaller, it was deemed to be male.

3.4.3 Inferring relatedness status

Genetically inferring the relatedness status is important for multiple reasons. First, it is used as a QC before merging data from different

Table 3.4 Kinship coefficients. Theoretical value and observed range of kinship coefficients per relationship type. MZ=monozygotic.

Relationship	Theoretical Value	Range
MZ twins / Self	0.5	>0.354
1st Degree	0.25	[0.177, 0.354]
2nd Degree	0.125	[0.0884, 0.177]
3rd Degree	0.0625	[0.0442, 0.0884]

lanes/runs/centres, to confirm sample identity. Second, checking family relationships facilitates the identification of any discrepancies. And third, the presence of consanguinity needs to be determined, since offspring of related parents will present a higher number of homozygous variants [214].

To obtain kinship coefficients and relationships, the method of Manichaikul *et al.*, [215] was used. This implements the same algorithm used in KING (a toolset to explore genotype data from a genome-wide association study (GWAS)), and works in a fast and robust manner for pedigrees with WES data. The input was the merged VCF file, and the output was a *relatedness2* file with the kinship coefficient (relatedness phi) each sample comparison. This coefficient value changes for the different relations between individuals as follows (Table 3.4).

3.4.4 Inferring ancestry origin

Ancestry origin of a sample can lead to different genomic metrics. For example, individuals with African ancestry have higher number of variants compared to individuals with European genetic background, due to the higher genetic diversity across African genomes [216].

The genetic background of the individuals was inferred to check if the samples were genetically homogeneous and to asses to which ancestries they were more similar. This information was used to interpret variants using specific population allele frequencies. For that, assessment of the ancestry origin of each individual was done using the R package EthSEQ [217]. EthSEQ categorises each individual in a VCF file into European, African, East Asian or South Asian ancestries. As input the tool requires a merged VCF file of individuals with unknown ethnicity and a reference model (genotype data at SNPs positions for a set of individuals with known ethnicity, obtained from 1000 Genome Project).

EthSEQ first builds a reference model from 1,000 Genome Project individual's genotype data for which ethnicity is known at 4,561 SNPs positions for the Exome dataset. Then, a target model is similarly created for the individuals with unknown ethnicity. Principal component analysis (PCA) is next performed using SNPRelate R package on aggregated target and reference models genotype data. The space defined by the first two PCA components is then inspected to generate the smallest convex sets, identifying the ethnic groups described in the reference model and next to annotate individuals with unknown ancestry origin.

3.5 Variant interpretation

Variant interpretation is one of the most challenging steps, where pathogenic mutations have to be identified among thousands of non-pathogenic. Here, different strategies were applied depending on the type of variants.

3.5.1 SNVs and indels

The merged VCF file was uploaded to Genome MINIng (GEMINI) framework, version 0.19.1 [218], along with a pedigree file (tabular file describing meta-data about the samples and their relationship). GEMINI stores all the information in a portable SQLite database, allowing easy exploration of the data.

Single nucleotide variants (SNVs) and indels variants were filtered by rare frequency, MAF <= 0.01 in control datasets (gnomAD). Next, a filter by consequence in the protein was applied. The consequences considered to have a functional effect in the protein were defined as any that fell in the following consequence classes: transcript ablation, splice acceptor variant, splice donor variant, stop gained, frameshift variant, stop lost, start lost, transcript amplification, inframe insertion, inframe deletion, missense, variant, and splice region variant. Candidate disease causing mutations were then identified using two different strategies (Figure 3.13).

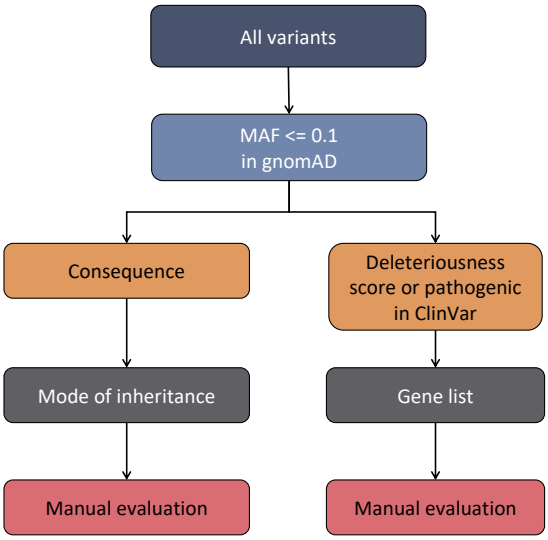


Fig. 3.13 Filtering strategy used to identify candidate variants.

1) Variants in genes that followed a Mendelian mode of inheritance (MOI). This focused on identification of variants that were in biallelic status (either autosomal recessive or compound heterozygous variants), X-linked recessive (XLR, where the mother was heterozygous and the affected male individual was hemizygous) or *de novo* (present in the child but not in the parents).

2) In order to consider variants in genes that did not follow a Mendelian model (due to eg. incomplete penetrance, polygenic traits), those present in a list of candidate genes previously associated with immune system disorders were considered. The gene list was assembled from literature searches for allergy and immunodeficiency, as well as associated Human Phenotype Ontology (HPO) terms (accessed March 2018), comprising a total number of 1,346 genes. The distribution of HPO terms is shown in Figure 3.14. The gene list is listed in Appendix.

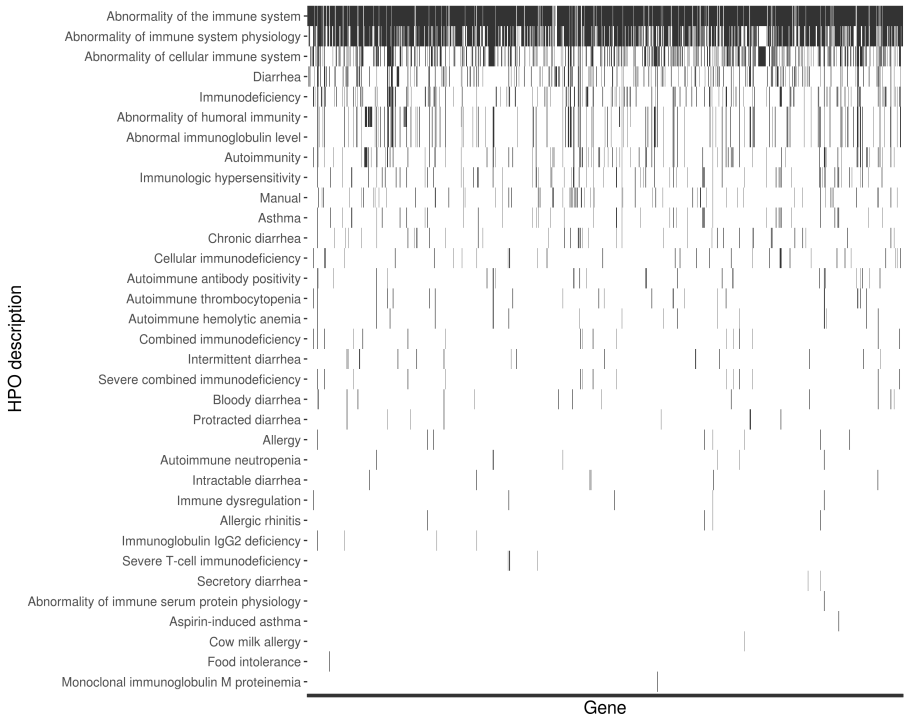


Fig. 3.14 Gene list. Heatmap of the HPO terms distribution of genes included in the gene list. Data accessed on March 2018. Manual=genes included from literature searches.

Due to a large number of candidate variants, stricter filters were applied in this case. Only mutations with high predicted deleterious score (CADD phred >= 20) or that had been previously reported as pathogenic in ClinVar were kept.

3.5.2 Copy Number Variants

The large number of CNVs were filtered by quality to get those that have a high probability to be real, as previously recommended [107]. CNVs

were also filtered by internal overlap removing those that occur in more than 10% of all samples in our cohort (a relatively liberal frequency threshold to remove only common CNVs and artefacts). *IntersectBed* function from Bedtools toolset was used to get the number of overlapped samples, requiring a reciprocal 50% of overlapping.

Lastly, CNVs were annotated with gene information from Ensembl (http://www.ensembl.org), in order to identify which genes were present within each structural variant. Due to the high number of false positive CNVs obtained from WES analysis, three situations were considered: CNVs only present in the proband (*de novo*), CNVs in genes from the gene list, and CNVs overlapping genes that had a candidate variant from the SNV/indel analysis. All of these were carefully evaluated and inspected using Integrative Genomics Viewer (IGV) [219].

3.5.3 HLA typing

Results from the HLA typing were analysed with PyHLA [220]. PyHLA is a tool for the association analysis between diseases and HLA types inferred from NGS data. It detects HLA association in antigen (two-digit allele level), protein (four-digit allele level) and amino acid levels. Zygosity tests examine monoallelic and biallelic zygosity associations.

Chapter 4

Results

4.1 Patients and phenotypes

A total number of 31 individuals from seven families with eight affected children were enrolled in this study. All the affected individuals presented severe gastrointestinal (GI) food allergies to multiple food proteins, and the majority of them had been diagnosed with severe FPIES. Allergic responses after the ingestion of most solid foods included vomiting, diarrhoea, abdominal weakness and severe pain since their first year of life. Seven of the eight affected individuals were males (87%). All members were part of the Garmitxa association (http://garmitxa.org/es), founded by the parents of these children. Phenotypic information was collected and is presented in Table 4.1.

Table 4.1 Clinical features of affected individuals.

	Family 1 F01_01	Family 2 F02_01	Family 3 F03_01	Family 4 F04_01	Family 5 F05_01	Family 6 F06_01	Family 7 F07_01	Family 7 F07_04	Observed
A. Patient information									
Gender - Male	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	7/8
City of birth (in Spain)	Bilbao	Bilbao	Bilbao	Bilbao	Bilbao	Castellón	Xàtiva	Xàtiva	-
YOB	2004	2009	2005	2005	2009	2013	2012	2014	-
Manifested symptoms	3 months	Few months	3 months	7 months	-	7 months	2 years	1 month	-
Diagnosis	FPIES	FPIES	FPIES	FPIES	FPIES	Likely FPIES	Likely FPIES	Likely FPIES	-
Family history of allergies	Father with GI problems, sister lactose intolerant, mother with allergies. Grandmother had 10 daughters that died from diarrhoea.	Father with allergy to beef and nuts.	Father with GI problems, mother allergic to legumes and with articular and muscular pain, brother allergic to mites, uncle with milk intolerance.	Mother allergic to mites, shrimps, dogs and cats.	Mother with tolerance problems to legumes and artichoke, father with psoriasis.	Father with intolerance to milk during the first months of life (in Hospital for 6 months), maternal grandfather with antibiotic allergy.	Brother with same phenotype, mother with dermatitis and allergy to pollen and nickel.	Sister with same phenotype, mother with dermatitis and allergy to pollen and nickel.	-

Continued from previous page

	Family 1 F01_01	Family 2 F02_01	Family 3 F03_01	Family 4 F04_01	Family 5 F05_01	Family 6 F06_01	Family 7 F07_01	Family 7 F07_04	Observed
B. Symptoms and presentation									
Intestinal problems during breast-feeding	Yes	-	Yes	Yes	Yes	-	Yes	Yes	6/6
Breast-feeding duration	5 months	-	-	0 months	0 months	-	15 days	1 month	-
Aliments tested	Multiple	-	Multiple	-	Multiple	Multiple	Multiple	Multiple	-
Offending foods	Multiple	Multiple	Multiple	Multiple	Multiple	Multiple	Multiple	Multiple	-
Diarrhoea	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	8/8
Blood in sediments	-	-	Yes	Yes	No	Yes	-	No	3/5
Vomiting	-	-	Yes	Yes	Yes	Yes	Yes	Yes	6/6
Muscular pain	Yes	Yes	Yes	Yes	-	-	No	No	4/6
Articular pain	Yes	Yes	Yes	Yes	-	-	No	No	4/6
Fatigue	Yes	Yes	Yes	Yes	-	Yes	No	No	5/7

Continued from previous page

	Family 1 F01_01	Family 2 F02_01	Family 3 F03_01	Family 4 F04_01	Family 5 F05_01	Family 6 F06_01	Family 7 F07_01	Family 7 F07_04	Observed
Abdominal distension	-	-	Yes	Yes	-	-	-	-	2/2
Gastroesophageal reflux	-	-	-	-	-	-	Yes	Yes	2/2
Esophagitis	-	-	-	-	-	-	Yes - until 5 months	Yes - until 5 months	2/2
Gastroscopy abnormal	-	-	-	-	-	-	No	Yes	1/2
Eczema	Yes	-	-	-	-	-	-	-	1/1
IgE-mediated sensitization	No	Yes	No	No	Yes	No	No	No	2/8
Loss of consciousness	No	Yes	No	No	No	-	-	-	1/5
Laboratory findings	No significant findings in biopsies from duodenum, stomach, oesophagi and colon.	-	-	-	No significant findings from biopsy of duodenum and colon. Normal stools.	-	No significant findings from biopsy of stomach and duodenum. EoE discarded.	No significant findings from biopsy of oesophagi and colon.	

Continued from previous page

	Family 1 F01_01	Family 2 F02_01	Family 3 F03_01	Family 4 F04_01	Family 5 F05_01	Family 6 F06_01	Family 7 F07_01	Family 7 F07_04	Observed
C. Resolution and follow-up									
Elemental formula	Yes	-	-	Yes	Yes	-	Yes	Yes	5/5
Steroids	Yes	-	-	Yes	Yes	-	-	-	3/3
PEG	Yes	-	Yes	Yes	No	-	-	-	3/4
Phenotype after alternative feeding/PEG/ steroids	Diarrhoea or constipation, blood in sediments, vomiting, muscular and articular pain.	-	Muscular and articular pain.	Muscular and articular pain.	Diarrhoea, vomiting, muscular and articular pain.	-	-	-	-
Resolution	Tolerates specific food, but he still has articular and muscular pain.	Food introduction at 4 years old - he still doesn't tolerate meat.	Tolerates specific food.	2015, although he still has muscular and articular pain.	-	-	Tolerates specific food.	Tolerates specific food.	-

4.2 Quality control

Overall, the data generated were of high quality. Sequencing was done by three different centres, in seven batches, using two different platforms and pulldown arrays. Therefore, a thorough quality control analysis was performed.

4.2.1 Per base quality

Median quality score by position in the read sequenced was analysed for the seven different batches of sequencing. The sequencing quality score of a given base (Q), is defined by the phred quality score [221, 222] in the following equation:

$$Q = -10log_{10}(e)$$

Where e is the estimated probability of the base call being wrong. A higher Q score means a smaller probability of error. For example, a quality score of 20 represents an error rate of 1 in 100, with a corresponding call accuracy of 99%.

FastQC was used to obtain per base quality scores for each batch of sequencing (Andrews S. (2010), available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc). In Figure 4.1 it is shown that all batches of sequencing had good median quality scores (over 25, as recommended by FastQC). However, quality of the INCLIVA batch was not only lower than the others, but also had higher dispersion. This is because the platform used for this batch was HiScanSQ while the others were HiSeq1500 and HiSeq2000, that have higher throughput and sequencing quality.

Additionally, the relative lower quality of the first eight bases was due to technical reasons, since the first cycles of sequencing are used for cluster calling and for establishing metrics (that are used to correct subsequent calls), as well as by possible artefacts due to non-random fragmentation performed during the sample preparation. Otherwise, quality scores behaved as expected.

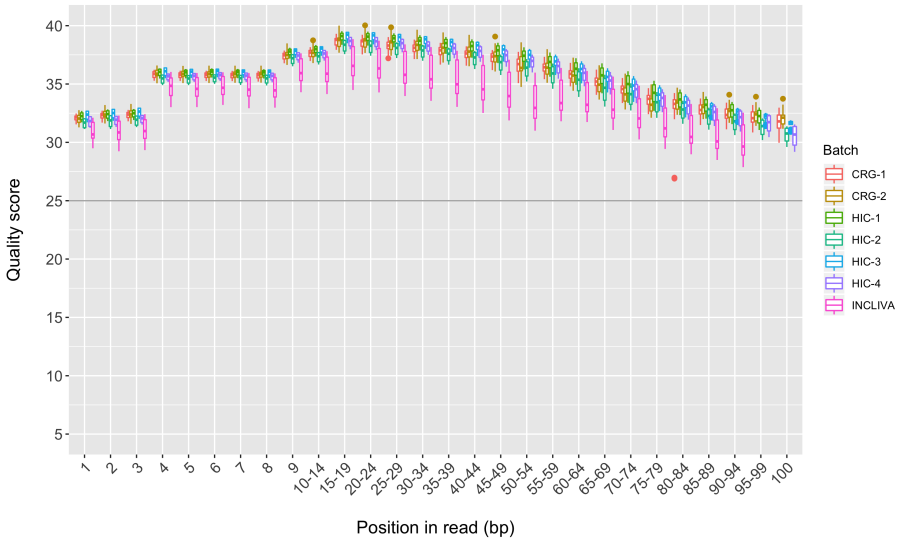


Fig. 4.1 Quality score results. Data obtained for each batch of sequencing using FastQC. A threshold of 25 (horizontal grey line) was set to determine good quality scores.

4.2.2 Coverage

The coverage distribution of the exome was compared across the different samples. There were three major groups: individuals that had been sequenced only with Nextera at INCLIVA and HIC centres (IN-

CLIVA_HIC), individuals sequenced only with SureSelect at CRG, and individuals sequenced at both centres with both sets.

As expected, those that were only sequenced with Nextera (INCLIVA_HIC) had lower coverage since the amount of Giga bases (GB) sequenced by sample was lower due to technical reasons (sequencing with HiScanSQ or HiSeq 1500, which have lower throughput) and experimental limitations (lower coverage in general aimed by sample). Additionally, the HiScanSQ machine was at the end of its life span, also explaining the lower amount and poorer quality of data produced. This is shown in Figure 4.2. INCLIVA_HIC sequenced individuals had a minimum coverage of 20x for 50% of the exome, while those that were sequenced at CRG or both had 90% of the exome at a minimum coverage of 20x.

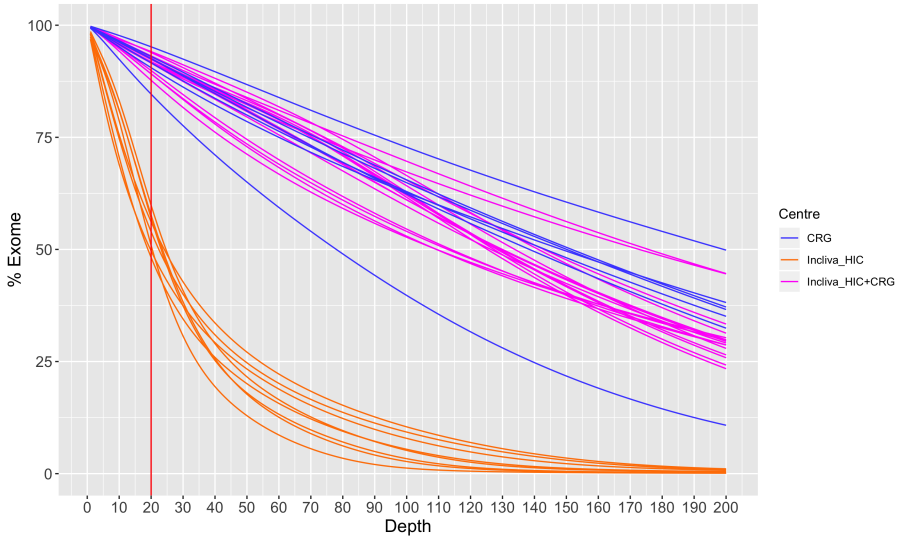


Fig. 4.2 Coverage results. Percentage of exome covered at a minimum depth. Each line is a sample. Colours are shown by centre of sequencing.

While coverage for clinical WES has to be higher than 80-120x, research WES is endorsed to be performed at a minimum coverage of 20-30x for accurate detection of variants [223, 224]. In this project, eight individuals had lower coverage, with only the 50% of the exome at a minimum coverage of 20x: F01_04, F01_05, F01_06, F01_08, F02_01, F02_02, F02_03 and F02_04, although only one (F02_01) was an affected individual. The consequence of lower coverage is an increased number of false negatives, as well as false positives due to bad mapping and wrong calling, that difficult variant filtering and interpretation. This was taken into account when performing analysis of these individuals.

4.2.3 Variant metrics

The number of variants that are identified in exome sequencing studies varies greatly, depending on the exome enrichment set used, the coverage reached, the sequencing platform and the algorithms used for mapping and variant calling. Here, the number of total variants detected per sample was compared by enrichment set used: Nextera and SureSelect. The median number of variants called per sample and enrichment set were 102,630 SNVs and 14,241 indels with Nextera, and 136,073 SNVs and 23,589 indels with SureSelect. The number of variants were similar, although slightly higher with SureSelect (Figure 4.3 A-B).

When only considering high quality variants (defined by depth and mapping quality higher than 20), the median number of SNVs (42,444) and indels (4,171) identified with Nextera were much lower than the median number of SNVs (103,021) and indels (15,181) identified with SureSelect (Figure 4.3 C-D). These numbers were within the expected range seen in other exome studies [225–227], and were also consistent

with the fact that i) SureSelect enrichment kit contains more regions than Nextera, including UTR regions and miRNAs, and ii) SureSelect variant calls were more reliable and had better quality due to a higher coverage.

Additionally, it has been seen that SureSelect outperforms Nextera in coverage uniformity, quality of the mapping and variant calls, exome capture rates and low PCR duplicate rates [228, 229]. The results shown in Figure 4.3 supported this, where the the number of raw SNVs/indels was comparable between both capture methods, but was higher for SureSelect calls when considering high quality variants only.

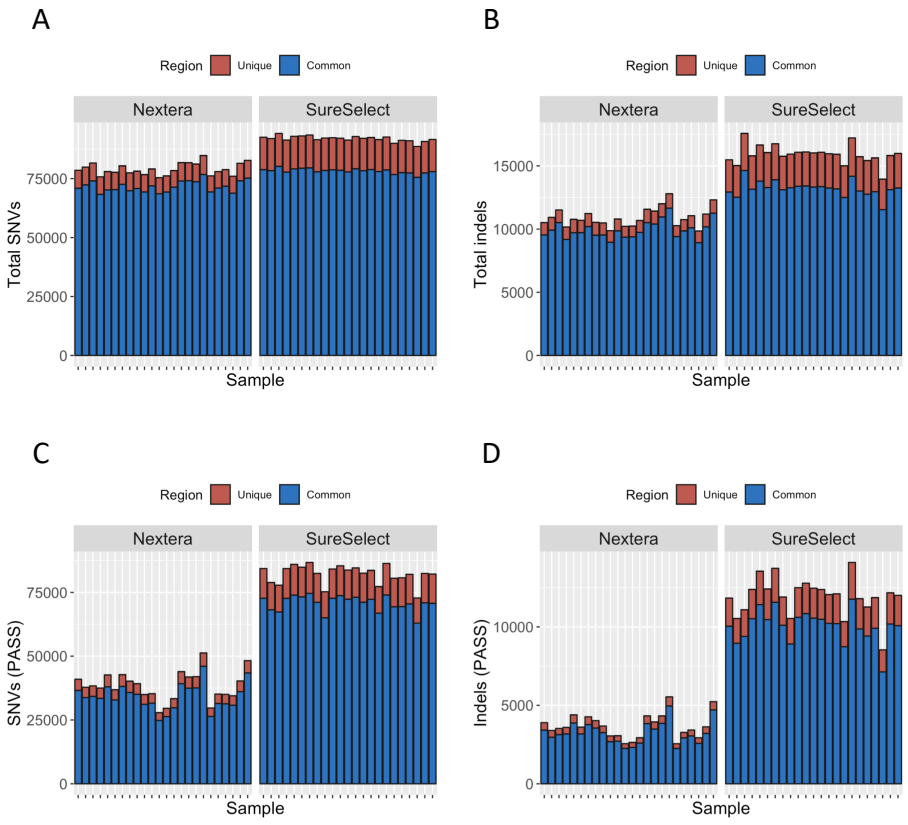


Fig. 4.3 Number of variants per sample and enrichment set. Number of variants are coloured by region, depending on if they are located in the regions present in both enrichment sets (blue) or in the unique ones (red). A) Total number of SNVs called. B) Total number of indels called. C) Number of SNVs passing QC called. D) Number of indels passing QC called. PASS=variant with depth and mapping quality higher than 20

The transition/transversion (Ts/Tv) ratio is also a useful metric because, in nature, transitions (A<-> G and C <-> T) occur much more often than transversions (A <-> C, A <-> T, G <-> C or G <-> T). For

exome datasets, the ratio should be a little above 2.0 [225]. Here, the ratio obtained in average were 2.32 for SureSelect samples and 2.37 for Nextera, as expected (Figure 4.4-A).

The heterozygosity to non-reference homozygosity ratio (Het/Alt) is another quality control parameter for DNA sequencing. For genome sequencing data, this ratio should be around 2.0 for variants in Hardy–Weinberg equilibrium, and little below for exome sequencing. In this case, the average Het/Alt obtained was 1.8, close to the expected value [230] (Figure 4.4-B).

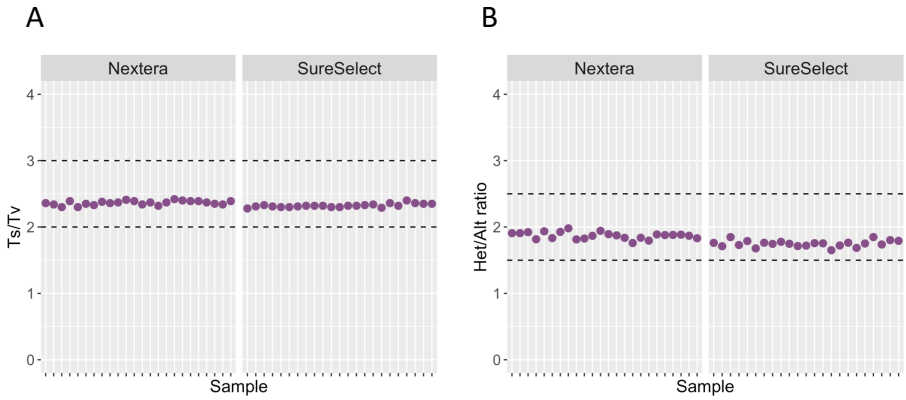


Fig. 4.4 Ts/Tv and Het/Alt ratios. A) Transitions to transversions ratio (Ts/Tv) per sample and enrichment set. B) Heterozygous to homozygous (alternative allele) ratio per sample and enrichment set.

Overall, a total number of 293,092 SNVs and indels, with an average coverage and mapping quality across the 31 individuals higher than 20, were called for all samples. Of these, 70,443 were rare (MAF <= 0.01 in gnomAD).

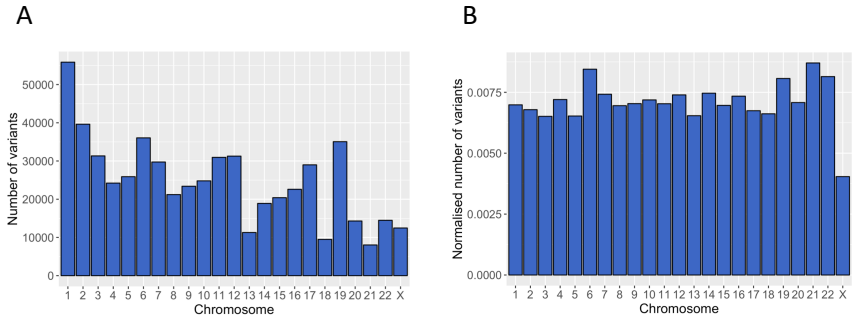


Fig. 4.5 Number of variants per chromosome. A) Number of variants per chromosome. B) Number of variants normalised by coding base pairs in each chromosome.

The number of variants per chromosome is represented in Figure 4.5-A. This was normalised by the number of exonic base pairs by chromosome (Figure 4.5-B). A uniform distribution of the number of variants was observed for the autosomal chromosomes, but not in the chromosome X. This is consistent with previous results [231], where it has been observed that the number of genes constrained for LOF variants is higher on chromosome X, so rare variants, which are more likely to have a moderate or high effect, are less likely to be found on that chromosome.

4.2.4 Ancestry origin

The ancestry origin of each individual was determined using the R package 'EthSEQ' [217]. This performed a principal component analysis (PCA) on the 31 individuals, and placed them into a reference PCs, space constructed from the reference model (individuals from the 1000G project, with known ancestries). PCA analysis revealed that all the

samples were of European ancestry (Figure 4.6). Therefore, European MAF was used later for filtering rare variants.

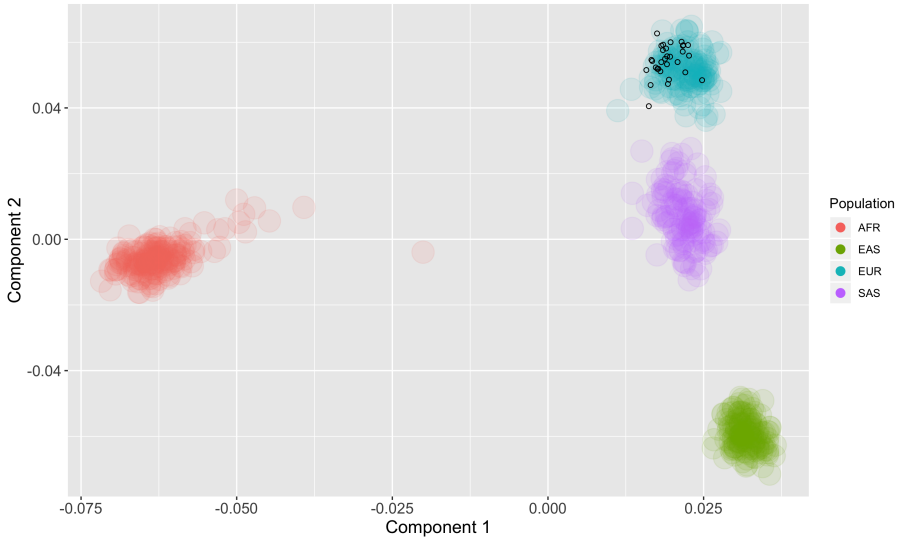


Fig. 4.6 Ancestry origins. PCA results from the ancestry analysis performed on the 31 individuals. The black dots represent individuals in this study.

4.2.5 Relatedness status

Relatedness between individuals was estimated using KING: Kinship-based INference for Gwas [215]. This was performed at a library level, and samples prepared with Nextera and SureSelect enrichment sets were compared amongst themselves, to not only confirm relatedness between individuals, but also to confirm self-identity (defined by PHI score of 0.5). This analysis is of especial importance when data has been prepared in different laboratories, using different methodologies and platforms. Therefore, the KING analysis was performed on the VCFs obtained from individual BAM files, before being merged by sample.

The kinship coefficient obtained was compared to the expected kinship for all individuals, and observed to correspond as expected (Figure 4.7). No consanguinity was identified in any of the families, and all them confirmed self-identity.

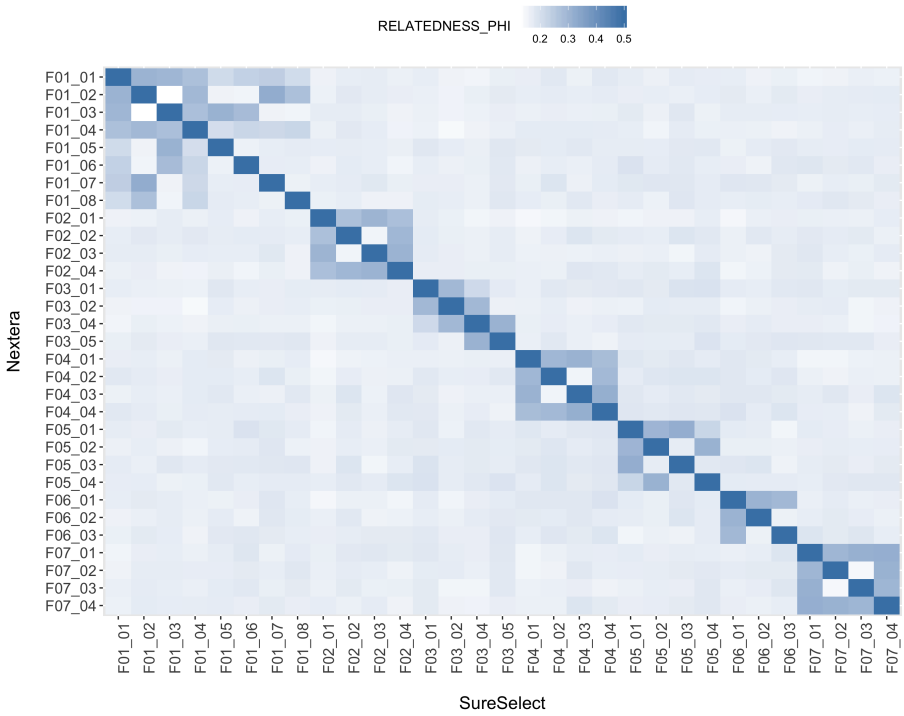


Fig. 4.7 Kinship coefficient results. Heatmap representation of the phi scores obtained from the relatedness analysis.

4.2.6 Genomic sex

After the relatedness analysis, genomic sex was compared to declared gender for all individuals. For that, normalised read counts on chromosomes X and Y divided by the median of the normalised read counts on

the autosomes was obtained and represented in Figure 4.8. All individuals' ratio clustered into their declared gender and no discrepancy was identified.

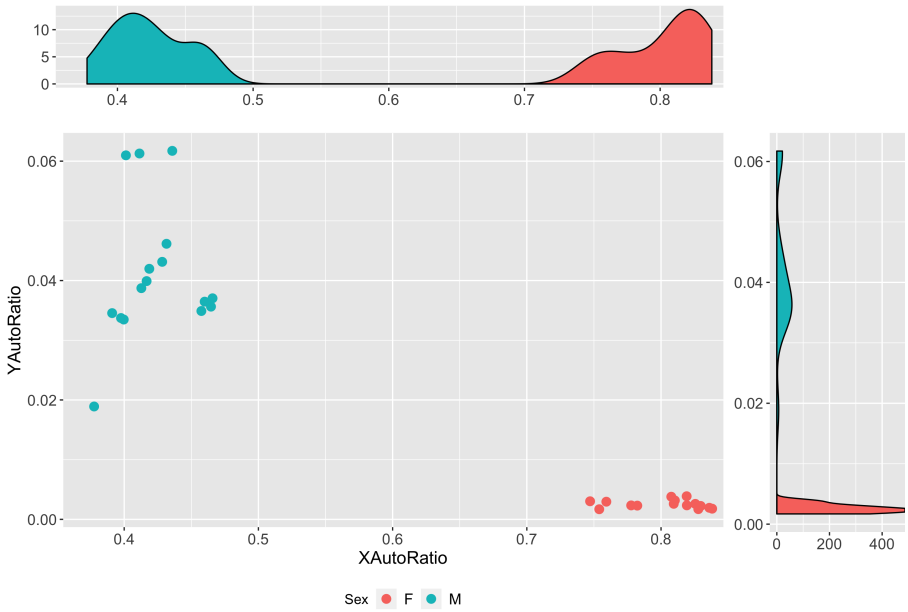


Fig. 4.8 Genomic sex. Representation of normalised read counts on X and Y divided by the median of the normalised read counts on the autosomes, showed as X/Auto and Y/Auto respectively.

4.3 Variant filtering and prioritisation

Following the QC analysis, the identification of candidate rare variants in the seven families was performed for high quality variants, following the two different strategies previously described (Methods, Figure 3.13). The MIM (Mendelian Inheritance in Man) numbers for all the genes in this chapter can be found in Appendix, Section 7.2 (Gene information).

Strategy one, based on filtering by mode of inheritance (MOI), identified a total number of nine candidate variants in eight different genes (Table 4.2), of which four were X-linked recessive, one *de novo* and four biallelic compound heterozygous. Regarding the consequences, one was predicted to be a splice donor variant and one was a splice region variant. The rest of the variants were missense. None of the them had previously been reported as pathogenic in ClinVar database.

Seven of these genes had been previously associated with the immune system, and observed to be involved in a variety of processes, including cytokine signalling, antigen processing and presentation, innate immune system and cell cycle control. Only *MAP3K15* had not been reported as linked to the immune system, but was also a candidate because two unrelated probands had hemizygous variants in this same gene.

The results of strategy two, based on the analysis of rare variants in genes associated with immune-related disorders, are shown in Table 4.3. Single variants in recessive genes (or suspected to be recessive due to pRec >= 0.9, which is the pLI equivalent for falling into the recessive category [231]), were not considered, as well as those where the phenotype was clearly not consistent. A total number of seven variants in six different genes were identified (variants in *NLRP12* were found in two unrelated families). Two of them were frameshift and the others were missense. All these genes play a role in the regulation of the immune system, and two of them have already been seen in genes with incomplete penetrance (*NLRP12* and *ANKZF1*).

Other variants, apart from the ones reported in Table 4.2 and Table 4.3 were also identified. However, they were in genes previously associated with different phenotypes, therefore their consideration was not pertinent in this study.

Table 4.2 Candidate variants identified by MOI filtering. AC, Hom and Hemi were obtained from gnomAD [231]. Csq = consequence. AC = allele count. Hom = number of homozygous. Hemi = number of hemizygous. RS = Reference SNP.

Sample	Gene	pLI	pRec	Variant (RS)	MOI	Csq	CADD	HGVSc	HGVSp	AC	Hom	Hemi
F02_01	*IL13RA2*	0	0.75	X:114250253 T>A	De novo	Missense	4.28	ENST00000243213.1:c.226A>T	ENSP00000243213.1:p.Ile76Phe	0	0	0
	ZNF645	0	0.89	X:22291936 A>G *(rs750306802)*	XLR	Missense	0.00	ENST00000323684.1:c.828A>G	ENSP00000323348.1:p.Ile276Met	11	0	4
F03_05	*LAMA5*	0.94	0	20:60887583 C>T *(rs144323773)*	Comp. het	Missense	11.5	ENST00000252999.3:c.9233G>A	ENSP00000252999.3:p.Arg3078Gln	87	0	0
				Multiple: 20:60904044 C>T 20:60911398 G>A *(rs145721906)*	Comp. het	Missense	25.2, 23.2	ENST00000252999.3:c.4303G>A ENST00000252999.3:c.2321C>T	ENSP00000252999.3:p.Ala1435Thr ENSP00000252999.3:p.Thr774Ile	0, 393	0, 2	0, 0
	MAP3K15	0	0	X:19391684 G>A *(rs138433947)*	XLR	Missense	5.04	ENST00000338883.4:c.2903C>T	ENSP00000345629.4:p.Ala968Val	28	0	13
	TNFRSF1A	0.99	0	12:6438989 G>A *(rs1251500082)*	Comp. het	Missense	11.3	ENST00000162749.2:c.1012C>T	ENSP00000162749.2:p.Leu338Phe	1	0	0
				12:6443001 G>A *(rs4149637)*	Comp. het	Missense	22.2	ENST00000162749.2:c.224C>T	ENSP00000162749.2:p.Pro75Leu	1914	41	0

Continued from previous page

Sample	Gene	pLI	pRec	Variant (RS)	MOI	Csq	CADD	HGVSc	HGVSp	AC	Hom	Hemi
F04_01	*PPL*	0	0	16:4935504 TT>AC *(rs148151950)*	Comp. het	Missense	9.37	ENST00000345988.2:c.3151_3152delinsGT	ENSP00000340510.2:p.Lys1051Val	163	0	0
				16:4960955 G>A *(rs189380553)*	Comp. het	Splice region	11	ENST00000345988.2:c.63-5C>T	-	441	1	0
F06_01	*GPR50*	0.45	0.53	X:150348444 T>A	XLR	Missense	28	ENST00000218316.3:c.389T>A	ENSP00000218316.3:p.Ile130Asn	0	0	0
	MAP3K15	0	0	X:19389462 C>A *(rs147323806)*	XLR	Splice donor	27.3	ENST00000338883.4:c.3294+1G>T	-	4	0	1
	STAB1	0	1	3:52538857 G>A *(rs779364897)*	Comp. het	Missense	26.7	ENST00000321725.6:c.1342G>A	ENSP00000312946.6:p.Gly448Arg	6	0	0
				3:52558162 C>T *(rs766033396)*	Comp. het	Missense	22.7	ENST00000321725.6:c.7589C>T	ENSP00000312946.6:p.Thr2530Ile	2	0	0

Table 4.3 Candidate variants identified by gene list filtering. GT = genotype for all individuals. Csq = consequence. AC = allele count. Hom = homozygous. Hemi = hemizygous. Pheno = phenotype. RS = Reference SNP.

Fam	Individuals	Gene	pLI	pRec	GT	Variant (RS)	Csq	CADD	HGVSc	HGVSp	AC	Hom	Hemi	PMID	Pheno
F01	F01_01 F01_04	*ANKZF1*	0	0.3	0/1	2:220100258 G>A *(rs189875478)*	Missense	34	ENST00000323348.5:c.1754G>A	ENSP00000321617.5:p.Arg585Gln	1713	12	-	28302725	IO IBD
					0/1	2:220100476 G>A *(rs201069890)*	Missense	29.3	ENST00000323348.5:c.1850G>A	ENSP00000321617.5:p.Arg617Gln	664	4	-	-	-
	F01_01 F01_03 F01_04 F01_05 F01_06	*NLRP12*	0	0	0/1	19:54314003 G>A *(rs141245482)*	Missense	24	ENST00000324134.6:c.910C>T	ENSP00000319377.6:p.His304Tyr	1254	4	-	25064839	Immunodeficiency common variable
F03	F03_01 F03_02 F03_04	*SKIV2L*	0	0.9	0/1	6:31931897 C>A	Missense	28.9	ENST00000375394.2:c.1855C>A	ENSP00000364543.2:p.Pro619Thr	-	-	-	-	-
					0/1	6:31929737 C>T *(rs36038685)*	Missense	33	ENST00000375394.2:c.970C>T	ENSP00000364543.2:p.Arg324Trp	2225	17	-	26122175	Immunodeficiency common variable

Continued from previous page

Fam	Individuals	Gene	pLI	pRec	GT	Variant (RS)	Csq	CADD	HGVSc	HGVSp	AC	Hom	Hemi	PMID	Pheno
F04	F04_01 F04_02 F04_04	*NLRP12*	0	0	0/1	19:54313621 G>A	Missense	26.5	ENST00000324134.6:c.1292C>T	ENSP00000319377.6:p.Thr431Ile	-	-	-	-	-
F06	F06_01 F06_02	*GFI1*	0	0.9	0/1	1:92946548 CA>C	Frameshift	-	ENST00000294702.5:c.395delT	ENSP00000294702.5:p.Leu132ArgfsTer66	-	-	-	-	-
	F06_01 F06_03	*INO80*	1	0	0/1	15:41387810 T>G	Missense	23.9	ENST00000401393.3:c.370A>C	ENSP00000384686.3:p.Lys124Gln	-	-	-	-	-
F07	F07_01 F07_03 F07_04	*CAPN14*	0	0.14	0/1	2:31422363 AT>A *(rs1169440343)*	Frameshift	-	ENST00000403897.3:c.761delA	ENSP00000385247.3:p.His254LeufsTer11	1	0	-	-	-

As additional quality control, all affected individuals were observed to have a mean coverage of at least 20x across all candidate genes identified by gene list or MOI filtering (Table 4.2 and Table 4.3). This was important to exclude the presence of another individual with a rare variant in one of the candidate genes that had not been called because of low coverage.

4.3.1 Pathway analysis

Analysis of the pathways in which these genes were involved was performed using Reactome [232], a curated and peer-reviewed pathway database (http://reactome.org, date of accession 12/04/2019). Genes whose function had not been previously demonstrated to play a role in a specific pathway were: *NLRP12*, *MAP3K15*, *ANKZF1*, *GFI1* and *GPR50*.

Five of the nine genes that were present in Reactome had been associated with the immune system: *LAMA5*, *ZNF645*, *TNFRSF1A*, *IL13RA2*, and *PPL* (*p-value* of 1.98E-2). Three of them played a role in signalling by interleukins (*LAMA5*, *TNFRSF1A*, *IL13RA2*, with a *p-value* of 5.73E-4) (Figure 4.9), more specifically, the IL-10, IL-4 and IL-13 signalling pathways.

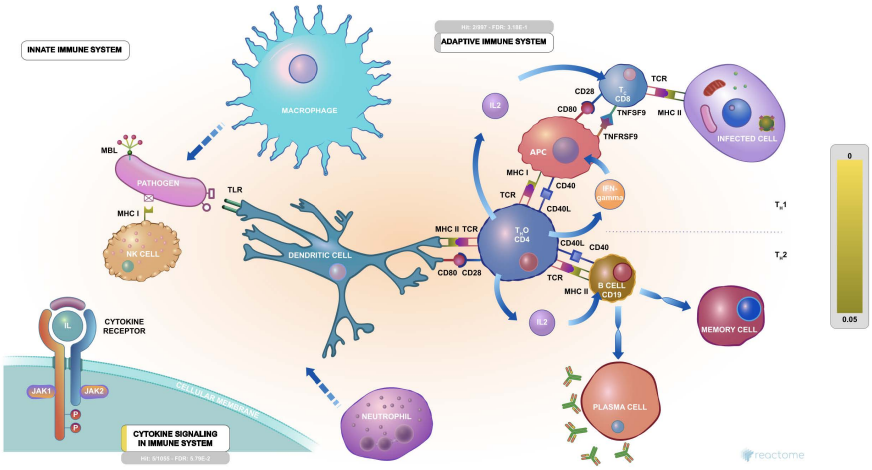


Fig. 4.9 Reactome enrichment analysis. From http://reactome.org [232].

The other four genes had been associated in Reactome to the following pathways: *SKIV2L* to Metabolism of RNA (mRNA decay by 3' to 5' exoribonuclease) and Metabolism of proteins (Association of TriC/CCT with target proteins during biosynthesis); *INO80* to DNA repair (DNA Damage Recognition in GG-NER); *CAPN14* to Extracellular matrix organisation (Degradation of the extracellular matrix); and *STAB1* to Vesicle mediated transport (Scavenging by Class H Receptors).

From manual interpretation of the candidate genes, it was noticed that three of them play a role in the NF-$\kappa\beta$ pathway: *GFI1* (which antagonises NF-$\kappa\beta$ p65 [233]), *NLRP12* (which suppresses non-canonical NF-$\kappa\beta$ pathway [234]), and *TNFRSF1A* (which activates NF-$\kappa\beta$ signalling [235]). Mutations in *TNFRSF1A* have already been reported as associated with autosomal dominant auto-inflammatory disorder by enhanced activation of NF-$\kappa\beta$ and cytokine secretion, constitutive activation of IL-1R pathway and inhibition of apoptosis [236]).

4.3.2 Family 1: *ANKZF1* and *NLRP12*

Family 1, the biggest pedigree family that enrolled this study, consists of one affected individual, both parents, one sister and the four grandparents. Two variants were identified in *ANKZF1* and *NLRP12* genes, in the affected individual and multiple relatives.

ANKZF1: Ankyrin Repeat and Zinc Finger Domain Containing 1

First, compound heterozygous variants in *ANKZF1* were identified in trans in the proband (F01_01) and the unaffected sister (F01_04) (ENSP00000321617.5, p.Arg585Gln and p.Arg617Gln). Both SNVs were missense variants, very rare and predicted to be damaging. *ANKZF1* plays a role in the cellular response to hydrogen peroxide and in the maintenance of mitochondrial integrity under conditions of cellular stress. Although the gene is not constrained for recessive LOF variation in gnomAD, it has been previously reported as associated with infantile-onset inflammatory bowel disease (IO IBD) [237]. Specifically, one of the variants (p.Arg585Gln) has been observed in one individual with IO IBD.

IO IBD is an early onset form of IBD, a chronic inflammatory condition of the gastrointestinal tract. The symptoms include abdominal pain, diarrhoea, and blood in stool being most common [238]. Our patient had diarrhoea, blood in sediments, vomiting and muscular and articular pain, presenting overlapping features with IO IBD.

Upon cellular stress conditions, the protein encoded by *ANKZF1* is located diffusely in the cytoplasm and translocates to the mitochondria. Depletion of *ANKZF1* reduces mitochondrial integrity and mitochondrial respiration under conditions of cellular stress. Mutations in this gene,

including p.Arg585Gln, result in an increased level of apoptosis in patients' lymphocytes, a decrease in mitochondrial respiration in patient fibroblasts, and an inability to rescue the phenotype of yeast deficient in Vms1, the yeast homologous of *ANKZF1* [237].

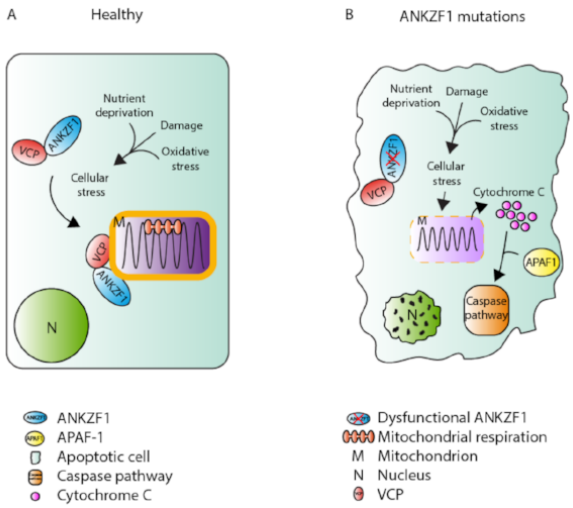


Fig. 4.10 Suggested pathogenesis mechanism of *ANKZF1*. A) Healthy cell. B) Cell with dysfunctional *ANKZF1*. From: http://cofferlab.science/new-blog/2017/8/4/ankyrin-repeat-and-zinc-finger-domain-containing-1-mutations-are-associated-with-infantile-onset-inflammatory-bowel-disease.

Nevertheless, both p.Arg585Gln and p.Arg617Gln were observed to be in homozygous individuals in gnomAD (12 and four individuals, respectively). However, it has been suggested that mutations in this gene could present incomplete penetrance [237], and this could explain the observation of homozygous and healthy individuals in gnomAD and the presence of the variants in this combination in the unaffected sister (F01_04).

To prove if these variants are causal, mRNA and protein expression of ANKZF1 analyses could be performed, since these have been seen to be reduced in patients with IO IBD and the p.Arg585Gln mutation [237]. Additionally, functional studies could also be done to determine if increased level of apoptosis and decreased mitochondrial respiration under conditions of cellular stress in lymphocytes are observed.

NLRP12: NLR Family Pyrin Domain Containing 12

The second variant identified in this family was in *NLRP12* gene. This was a heterozygous mutation in ENSP00000319377.6, p.His304Tyr, predicted to be damaging, and present in the affected individual, the father, the sister and two grandparents. This variant, however, is observed to be in 1250 heterozygous individuals and four homozygous in gnomAD and has conflicting interpretations of pathogenicity (likely benign and VUS) in ClinVar. This exact mutation has also been observed to be in compound heterozygosity with p.Ala629Asp [239] in an affected female, but with more severe phenotype of common variable immunodeficiency.

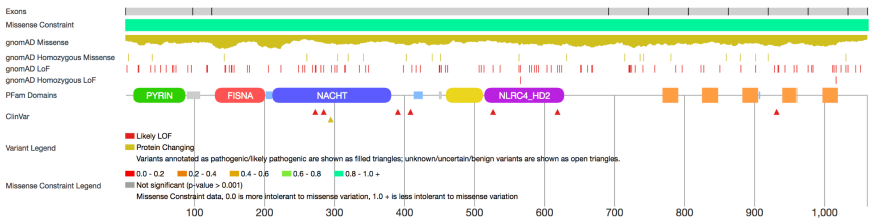


Fig. 4.11 Schematic representation of *NLRP12*. First track represents the exons of this gene followed by the gnomAD missense count. Representation of homozygous missenses and LOF variants in this gene are also shown, and line up with the secondary structure of the protein. Pathogenic variants in ClinVar are shown under the protein in yellow (missense) and red (LOF) triangles. Data obtained from DECIPHER [240].

Heterozygous mutations in *NLRP12* are associated with periodic fever syndromes and atopic dermatitis in humans, by negatively regulating pathogenic T cell responses [241]. Phenotype of mutations in this gene include fever, severe fatigue and musculoskeletal symptoms, which are typically activated or worsened by cold exposure. The protein encoded by *NLRP12* inhibits the transcription factor NF-$\kappa\beta$, and when mutated, an elevated non-canonical NF-$\kappa\beta$ activation and increased expression of target genes has been observed. Reduced *NLRP12* expression increased the activation of NF-$\kappa\beta$ and proinflammatory cytokine expression, leading to subverted pattern of inflammation. Interestingly, this mutation is in the NACHT domain, which is a key region in the clinical molecular diagnosis of Familial Cold Auto-inflammatory syndrome [242], and where the only pathogenic missense variant has been identified [243]. Although the gene is not constrained for dominant or recessive LOF variation in gnomAD, low penetrance has been reported [243]. Therefore, it could be possible that dysregulation of NF-$\kappa\beta$ pathway could accentuate the severe phenotype present in the individual F01_01, and that this variant is acting as risk factor rather than likely Mendelian pathogenic variant.

4.3.3 Family 2: *IL13RA2* and *ZNF645*

Family 2 is composed by the proband, father, mother and sister. Two candidate variants in the *IL13RA2* and *ZNF645* genes were identified in this family by filtering by MOI.

IL13RA2: Interleukin 13 Receptor Subunit Alpha 2

The mutation in *IL13RA2* was *de novo* (Figure 4.12) and missense (ENSP00000243213.1:p.Ile76Phe), and was not observed to be present in any other individual in the cohort or in gnomAD. Although this variant had a relatively low CADD phred score, it caused the change of the hydrophobic side chain Isoleucine to a Phenylalanine, which has a bigger side chain and could have consequences in the protein structure.

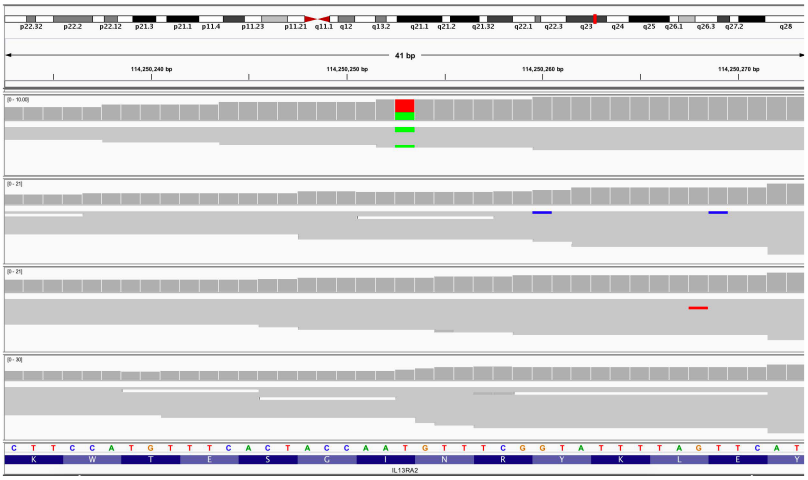


Fig. 4.12 *De novo* variant in *IL13RA2*. Integrative Genomics Viewer snapshot of the reads of the four individuals in this family, showing that the variant in *IL13RA2* is *de novo* in the proband. Alignment tracks correspond to the proband, father, mother and sibling.

IL13RA2 encodes for a membrane bound protein (IL-13Rα2) that binds IL-13 with high affinity. Although it does not appear to function as a signal mediator since it lacks any significant cytoplasmic domain, this protein can act as a decoy receptor regulating the effects of IL-13 and its internalisation.

IL-13 is a cytokine that acts as central regulator in IgE synthesis; it influences isotype class switching to IgE and is a mediator of allergic inflammation and eosinophil chemotaxis. This cytokine is critical to the induction and perpetuation of the T-helper type 2 (Th2)-mediated allergic immune responses (Figure 4.13), and has been implicated in multiple atopic diseases [244].

Variants in IL13 gene have already been associated with IgE-mediated paediatric food allergy [245] and EoE [246]. IL-13 is the chief stimulus for the production of eotaxin-3, an eosinophil-selective chemo-attractant and activating cytokine, along with *CAPN14* from oesophageal epithelial cells.

IL-13 signalling begins through a heterodimer receptor complex consisting of alpha IL-4 receptor (IL-4Rα) and alpha Interleukin-13 receptor (IL-13Rα1). Heterodimerisation activates STAT6 (a transcription factor) signalling, which is important in initiation of the allergic response [247]. The other receptor of IL-13 is IL-13Rα2, encoded by *IL13RA2* gene, which has 50-times greater affinity to IL-13 than IL-13Rα1. However, IL-13Rα2 lacks a signalling motif and has a truncated cytoplasmic domain suggesting that it functions as a decoy receptor for IL-13 (Figure 4.13).

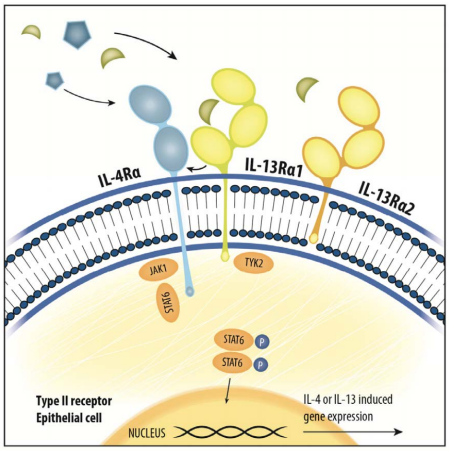


Fig. 4.13 Receptor system for IL-4 and IL-13. Both, IL-4 and IL-13 bind the type II receptor complex. IL-13 also binds IL-13Rα2 with higher affinity than IL-13Rα1. Binding of these cytokines to their respective receptor complexes leads to activation of protein kinases, JAK1 or JAK3, or Tyk2 and subsequently phosphorylation of the transcription factor, STAT6. Phosphorylated STAT6 dimerises and translocates to the nucleus to activate IL-4- and IL-13-induced genes transcription (e.g. periostin) [244].

The association of IL-13 to atopic disease, along with other Type 2 response cytokines (IL-4), has been widely reported. Interestingly, the use of *dupilumab* (Regeneron and Sanofi), a fully human monoclonal antibody that blocks both IL-4 and IL-13 signalling, has demonstrated to have unprecedented efficacy on multiple atopic diseases [244]. Therefore, the fact that this mutation was observed to be *de novo* in the proband, and present in the receptor of IL-13, which has been associated with food allergy, makes it a good candidate to be associated with the FPIES present in this affected individual, though functional analysis would be required to elucidate the exact mechanism of pathogenesis.

ZNF645: Cbl Proto-Oncogene Like 2

A hemizygous missense variant was also identified in *ZNF645* (ENSP000 00323348.1:p.Ile276Met). The mother was a carrier of this variant, which was observed in four hemizygous individuals in gnomAD. The gene was not constrained for missense variation, but was observed to be constrained for hemizygous LOF variants (with a pRec of 0.89) therefore a recessive mechanism was suspected. This gene, also known as *CBLL2*, encodes a member of the zinc finger domain-containing protein family, and it may function as an E3 ubiquitin-protein ligase. Although there is not much known about the gene, protein localisation suggests a role in human sperm production and quality control [248], and gene expression studies showed high gene expression in testis (GTEx). However, it has also been related to Class I MHC mediated antigen processing and presentation (Reactome identifier: R-HSA-8851646), so a possible role in the immune system cannot be ruled out.

4.3.4 Family 3: *LAMA5*, *MAP3K15*, *TNFRSF1A* and *SKIV2L*

Family 3 is formed by the affected individual and the father, mother and half-sister. Four candidate variants were identified in the affected individual of this family. Three of them were identified by the filtering by inheritance (in *LAMA5*, *MAP3K15* and *TNFRSF1A*), and one by the gene list filtering (in *SKIV2L*).

LAMA5: Laminin Subunit Alpha 5

The first gene was *LAMA5*, which encodes for laminin $\alpha 5$, one of the vertebrate laminin alpha chains, an extracellular matrix glycoprotein. A compound heterozygous variant in this gene was identified in the affected individual of this family. This was formed by one missense variant inherited from the father (ENSP00000252999.3, p.Arg3078Gln) and two missense variants inherited from the mother (p.Ala1435Thr, p.Thr774Ile).

LAMA5 is a constrained gene for LOF variation, and has not been previously associated with disease. However, it's been seen that laminin $\alpha 5$ deletion in mice leads to a number of developmental abnormalities, including hyper-proliferation of basal keratinocytes and a delay in hair follicle development [249, 250]. Loss of laminin $\alpha 5$ has resulted in increased numbers of CD45+, CD4+ and CD11b+ immune cells in the skin, indicating that immune cell changes are the consequence of keratinocyte hyper-proliferation.

Furthermore, dominant mutations in this gene have been associated with Ehlers-Danlos syndrome, a complex multi-system syndrome due to dysfunction of the extracellular matrix [251]. Affected individuals presented with kin anomalies, impaired scarring, night blindness, muscle weakness, osteoarthritis, joint and internal organs ligaments laxity, malabsorption syndrome and hypothyroidism.

LAMA5 is largely expressed across multiple human cells, including oesophagus (GTEx). Interestingly, because this gene encodes for a laminin alpha chain, it plays an important role in extracellular matrix organisation. Previous genes involved in matrix organisation have already been associated with GI disorders, such as *CAPN14*, which is

associated with EoE. Nevertheless, the loss of laminin $\alpha 5$ has not been investigated yet as associated with GI disorders, and this finding opens new possibilities of research in the field, which would be required to confirm the role of this gene in FPIES.

MAP3K15: Mitogen-Activated Protein Kinase Kinase Kinase 15

A hemizygous missense variant was also identified in *MAP3K15* (ENSP0 0000345629.4:p.Ala968Val), and the mother of this individuals was seen to be carrier of the variant. The protein encoded by this gene, also known as ASK3, is a member of the mitogen-activated protein kinase (MAPK) family. The gene has not previously been associated with disease and is not constrained for LOF (pLI = 0) or missense variation (Z score = -0.78) in gnomAD. However, *MAP3K15* was considered to be relevant because two unrelated probands in this study (F03_01 and F06_01) presented a rare hemizygous variant in this gene.

Kaji *et al.* demonstrated that knockdown of *MAP3K15* protected HeLa cells against cytotoxicity induced by anti-Fas monoclonal antibody, TNF-α, or oxidative stress [252], suggesting that *MAP3K15* is a member of apoptosis signal-regulating kinases and that it plays a pivotal role in the signal transduction pathway implicated in apoptotic cell death triggered by cellular stresses. The gene is highly expressed in Adrenal gland (GTEx). Furthermore, proteins from the same family of kinases have been previously associated with inflammation [253]. Tartey *et al.* observed that apoptosis signal–regulating kinases 1 and 2 (ASK1 and ASK2) mediated footpad inflammation by controlling proinflammatory signalling in the neutrophils. The possible role of ASK3 in inflammation and how mutations in this gene could be involved in FPIES is yet to be determined.

TNFRSF1A: TNF Receptor Superfamily Member 1A

A compound heterozygous variant was identified in *TNFRSF1A* gene (ENSP00000162749.2, p.Leu338Phe and p.Pro75Leu). This is a constrained gene for LOF and missense (Z score = 2.1) mutations and encodes a member of the TNF receptor super-family of proteins. The ligand of this receptor is tumour necrosis factor alpha (TNF-α), and when it binds its receptor, it induces receptor trimerisation and activation, which plays a role in cell survival, apoptosis, and inflammation. Mutations in this gene may also be associated with multiple sclerosis in human patients.

TNF-α is a principal mediator of the acute inflammatory response, and has been previously associated with IBD and several other immune-driven disorders. Currently, anti-TNF treatments are already in use to treat IBD and other GI disorders [254, 255].

The greatest producers of TNF-α are activated macrophages and monocytes, particularly when stimulated with lipopolysaccharide (LPS), though the gene is widely expressed across different cell types (GTEx). Uncontrolled TNF-α release can cause chronic inflammation, cachexia, septic shock, and many inflammatory diseases, including IBD [256]. IBD is characterised by unregulated inflammation of the intestinal tract, and it's been seen that affected individuals with IBD have higher TNF-α concentrations than controls.

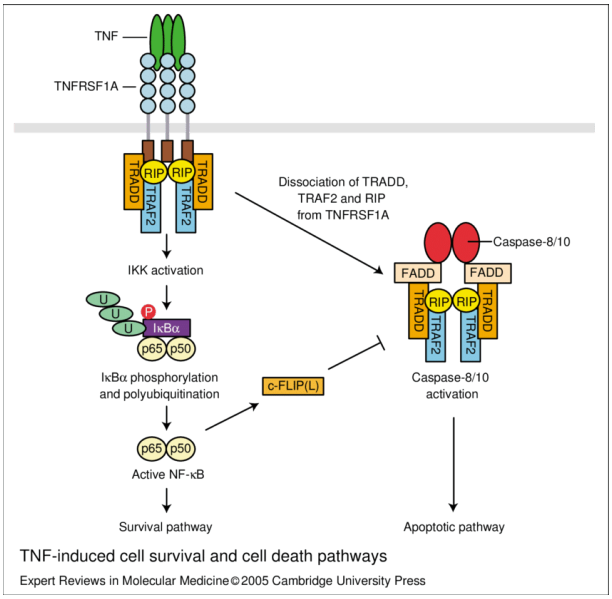


Fig. 4.14 TNF-induced cell survival and cell death pathways. Binding of TNF to its receptor *TNFRSF1A* regulates cycle cell. Activation of IKK leads to I$\kappa\beta$(NF-$\kappa\beta$ inhibitor) phosphorylation and degradation. This process allows translocation of the NF-$\kappa\beta$ p50-p65 heterodimer to the nucleus to bind DNA and induce gene expression for cell survival. However, if NF-$\kappa\beta$ is not activated upon *TNFRSF1A*-mediated signalling, apoptotic pathway is induced leading to cell death [257].

Therefore, dysregulation of a TNF-α receptor could lead to intestinal inflammation, which is consistent with our patient's phenotype. The exact mechanism is yet to be elucidated.

SKIV2L: Ski2 Like RNA Helicase

Two more mutations were identified in this individual in *SKIV2L* gene by gene list filtering. Both of them were missense, predicted to be damaging, and present in the mother and the sister, therefore in the same allele (in cis). Considering the canonical transcript ENSP00000364543.2,

p.Pro619Thr was not present in gnomAD, while p.Arg324Trp was relatively common (AC = 2225) and observed to be in 17 homozygous in gnomAD. The latter was also present in ClinVar as associated with Immunodeficiency common variable, although posterior studies reported it as likely benign, due to the high frequency in the population. This gene is not constrained for LOF variants (pLI = 0) but it is for recessive LOF variants (pRec = 0.96). Autosomal recessive mutations in *SKIV2L* cause trichohepatoenteric syndrome (syndromic diarrhoea) [258], thus phenotype could be relevant for this affected individual since she presents severe diarrhoea.

Although the specific function of *SKIV2L* is not very well understood, it could be possible that these two mutations, in combination with others present in the proband not in the relatives, could contribute to the patient's phenotype, especially since this gene has been seen in a digenic form with *AKR1D1* to cause severe infantile liver disease [259].

4.3.5 Family 4: *PPL* and *NLRP12*

Family 4 is formed by a proband and mother, father and sister. Two candidate variants were identified in this family in *PPL* and *NLRP12* genes.

PPL: Periplakin

Compound heterozygous variants were identified in the *PPL* gene. One of the variants was missense (ENST00000345988.2, c.3151_3152delins GT) and the other was in the splice region, at position -5 (c.63-5C>T). They were observed to be in trans in the affected child, and absent in this combination in the unaffected sibling. Although both mutations

had a CADD phred lower than 20, the missense was absent in homozygous individuals in gnomAD and the splice region variant had only a homozygous count of one, so both were very rare in biallelic state.

The protein encoded by this gene is a component of desmosomes and of the epidermal cornified envelope in keratinocytes. *PPL* acts as a linking protein: its N-terminal domain interacts with the plasma membrane and its C-terminus interacts with intermediate filaments. AKT1/PKB, a protein kinase mediating a variety of cell growth and survival signalling processes, has been seen to interact with this protein, suggesting a possible role as a localisation signal in AKT1-mediated signalling [260].

PPL is highly expressed in oesophagus (Figure 4.15). This is relevant because genes that play a role in the maintenance of the oesophagus mucosa, such as *CAPN14*, have already been associated with GI disorders. Therefore, mutations in genes involved in the pathway could also lead to similar phenotypes.

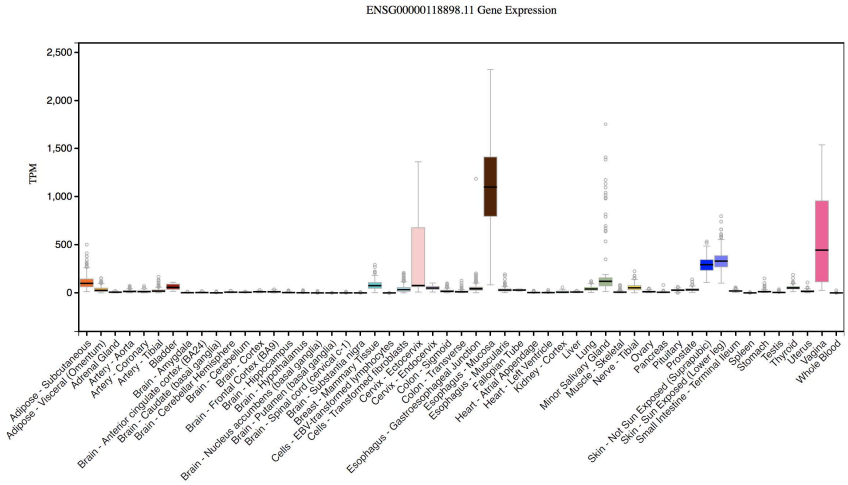


Fig. 4.15 Gene expression of *PPL*. Gene expression of *PPL* across different tissues. From GTEx. TPM = Transcripts Per Million.

NLRP12: NLR Family Pyrin Domain Containing 12

As in Family 1, a missense variant was identified in *NLRP12* in Family 4. This variant was ENSP00000319377.6:p.Thr431Ile in the protein, and it was present in the unaffected mother and sister. The exact mutation was absent in gnomAD and predicted to be damaging.

This gene negatively regulates T cell responses and inhibits the transcription factor NF-$\kappa\beta$. Since low penetrance has been reported, and the phenotype is very variable, it is potentially interesting that two different families present mutations in this gene. Both are rare and damaging, and could be contributing to the regulation of the T cell signalling.

4.3.6 Family 6: *GPR50*, *MAP3K15*, *STAB1*, *GFI1* and *INO80*

Family 6 is a trio, formed by the affected individual and both parents. Three variants identified in the *GPR50*, *MAP3K15* and *STAB1* genes were observed in the affected child of this family after performing the filtering by inheritance. Two other variants were identified in the filtering by gene list: a frameshift mutation in *GFI1* present in the affected child and the mother, and a missense mutation in *INO80* present in the affected child and the father.

GPR50: G Protein-Coupled Receptor 50

A hemizygous missense variant in the transmembrane receptor domain of *GPR50* was observed to be in F06_01 (ENSP00000218316.3: p.Ile130Asn). This was absent in gnomAD and predicted to be damaging, with a CADD phred score of 28. The mother was observed to be a carrier of this variant.

GPR50 gene encodes for a G-protein coupled receptor that inhibits melatonin receptor function through heterodimerisation. Variants in this gene have been previously associated with bipolar affective disorder and depression in women [261, 261–263].

Melatonin, a hormone secreted by the pineal gland, plays a role in regulating sleep and circadian rhythm as well as a possible role in gut-brain signalling [264]. Extrapineal melatonin has been detected in multiple tissues such as the skin, lymphocytes, mast cells, airway epithelium and GI tract among others [265]. This "sleep" hormone has demonstrated to play a role in oesophagitis and chronic inflammation [266] (Figure 4.16).

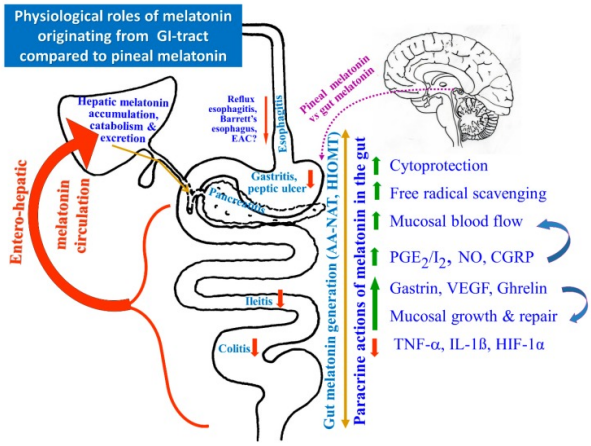


Fig. 4.16 Mechanistic effects of melatonin in the GI tract. From [266].

Recent discoveries suggested that changes in the microbiota modulate the host immune system by modulating Tryptophan (Trp) metabolism. Endogenous Trp metabolites include serotonin and melatonin [267]. Abnormal regulation of serotonin (5-HT) has already been associated with GI disorders, such as IBD and IBS. In fact, administration of *ondansetron*, a serotonin 5-HT3 receptor antagonist, has already been used to treat FPIES reactions, suggesting the potential role for serotonin in the pathophysiology of acute FPIES [268, 129]. Therefore, mutations in *GPR50* could impair the metabolism of melatonin, affecting the maintenance of the homeostasis in the GI tract.

MAP3K15: Mitogen-Activated Protein Kinase Kinase Kinase 15

A splice donor mutation was identified in *MAP3K15* (ENST00000338883 .4:c.3294+1G>T). This was the second affected child with a hemizygous mutation in this gene (also F03_01). As it was previously mentioned, this gene is highly expressed in the adrenal gland, and plays a pivotal role

in the signal transduction pathway implicated in apoptotic cell death triggered by cellular stresses and inflammation. Its role in immune system is yet to be determined.

STAB1: Stabilin 1

A compound heterozygous variant in *STAB1*, formed by two missense mutations (ENSP00000312946.6, p.Gly448Arg and p.Thr2530Ile), was observed to be in trans in F06_01. One of the variants, p.Gly448Arg, was located in the Fasciclin domain, while p.Thr2530Ile was nearly at the end of the protein (full size of 2570aa). Both variants were in conserved positions and predicted to be damaging (CADD phred > 20). This gene is highly conserved for recessive LOF variants. Neither of the variants were in any homozygous individual in gnomAD.

The protein encoded by this gene (Satb1) is a genome organiser expressed by T cells. Satb1 plays an essential role in the establishment of immune tolerance, and in the null mice, T cell development is severely impaired.

Because *STAB1* null mice die by week three of age, Kondo *et al.* studied *STAB1* conditional knockout (cKO) mice, in which the *STAB1* gene was deleted from all hematopoietic cells [269]. They observed that i) *STAB1* cKO mice developed autoimmune diseases within 16 weeds after birth, ii) suppressive functions of T regulatory cells, which play a major role in establishment of peripheral tolerance, were affected in the absence of *STAB1*, and iii) negative selection during T cell development in the thymus was severely impaired in *STAB1* deficient mice.

Therefore, although the role of *STAB1* in GI food allergy remains unknown, previous results suggest this protein plays an important role in

T cell development and peripheral tolerance, and this could be related to the phenotype presented in this individual. Further investigation would be required to confirm this association.

GFI1: Growth Factor Independent 1 Transcriptional Repressor

A frameshift mutation was identified in *GFI1* in F06_01 (proband) and F06_02 (mother), at the position 132 of the protein (of 422 amino acids) (ENSP00000294702.5:p.Leu132ArgfsTer66). This was absent in gnomAD and in any other population databases. Heterozygous mutations in *GFI1* have been associated with severe congenital Neutropenia. The protein encoded by this gene, Gfi1, is a transcriptional repressor that promotes T helper type 2 (Th2) cell development and inhibits Th17 and inducible regulatory T-cell differentiation [270]. This happens because Gfi1 inhibits the induction of the Th1 programme in activated CD4 T cells. It has been suggested that it regulates the Th1-type immune response by binding to the gene loci of *TBX21*, *EOMES* and *RUNX2*, and reducing the histone H3K4 methylation levels in part by modulating Lsd1 recruitment (a Lysine-specific histone demethylase). Though the gene was not constrained for LOF variation in gnomAD, manual investigation of the LOF variants in gnomAD revealed that actually only nine mutations were present in this gene, all of them with an allele count of one, therefore being very rare.

Dysregulation of T helper cell response could impair the immune system and response to food exposure. Noval Rivas M *et al.* previously reported that regulatory T cell reprogramming toward a Th2-cell-like impairs oral tolerance and promotes food allergy [271], therefore highlighting the possible association of this gene with food allergy. The fact that the mutation is also present in the mother and that nine individu-

als in gnomAD carry a LOF variant could be explained by a possible incomplete penetrance.

INO80: *INO80* Complex Subunit

A missense mutation in *INO80* was also identified in this family (ENSP0 0000384686.3:p.Lys124Gln). This was absent in gnomAD and predicted to be damaging. The variant was present in the affected child and the father.

INO80 encodes the catalytic ATPase subunit of the chromatin remodelling complex *INO80*, which is suspected to be required for turnover of RNA Polymerase II [272]. Mutations in this gene have previously been associated with immunoglobulin class-switch recombination defects (rare primary immunodeficiencies characterized by impaired production of immunoglobulin isotypes and normal or elevated IgM levels) [273].

This gene is constrained for LOF variation in gnomAD (pLI = 1) and is also constrained for missense variation (Z score = 3.13, which is in the top 10% constraint genes for missenses in the genome) [231]. Therefore, it could be possible that incomplete penetrance of these variants would be contributing to the phenotype of this affected individual. However, the role of *INO80* in the pathogenesis of FPIES is still unknown.

4.3.7 Family 7: *CAPN14*

The last family was formed by two affected siblings and the unaffected parents. A variant was identified in *CAPN14* genes, present in both affected individuals but also in the father.

CAPN14: Calpain 14

A frameshift variant was identified in *CAPN14* (ENSP00000385247.3: p.His254LeufsTer11). This variant was very rare, present in one heterozygous in gnomAD, and was at the position 254 of the protein (of 684 amino acids), with expected activation of the NMD pathway and gene haploinsufficiency.

CAPN14 is a cytosolic calcium-activated cysteine protease, that belongs to the calpain large subunit family, which are involved in a variety of cellular processes including apoptosis, cell division, modulation of integrin-cytoskeletal interactions, and synaptic plasticity [274].

This gene has previously been associated with a specific type of GI food allergy, Eosinophilic Esophagitis (EoE), a chronic inflammatory disorder triggered by allergic hypersensitivity to food [275]. Symptoms of EoE include dysphagia, vomiting, and severe chest pain, which is highly consistent with the phenotype of both affected siblings. The affected male had been diagnosed with chronic oesophagitis grade I, although the sister, who presented the same response to food ingestion, did not present any oesophagitis, and was less severely affected.

CAPN14 is expressed at the highest level in the oesophagus and has been identified as a tissue identity marker (Figure 4.17). It has been hypothesised the protein encoded by *CAPN14* might be a protective protein of the integrity of oesophageal tissue, because oesophageal epithelium is prone to damage because of food consumption.

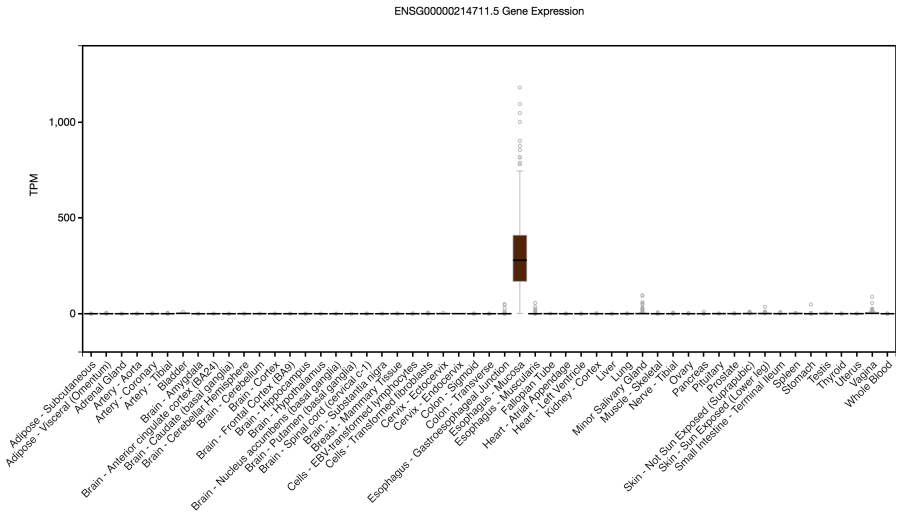


Fig. 4.17 Gene expression of *CAPN14*. Gene expression of *CAPN14* across different tissues. From GTEx. TPM = Transcripts Per Million.

In patients with active EoE, there is a 2- to 6-fold increase in *CAPN14* mRNA levels, and this correlates with disease activity. EoE is driven in part by increased levels of IL-13 in the oesophagus, which leads to disruption of epithelial cell architecture and impaired barrier function. Nevertheless, the lack of *CAPN14* also disturbs IL-13–induced epithelial cell changes, as demonstrated by increased dilated intercellular spaces and basal cell disorder [276]. Even though this gene is generally over-expressed in EoE compared with controls, an EoE risk allele is also associated with reduced oesophageal *CAPN14* expression. In this case, the LOF mutation identified in *CAPN14* gene is a candidate to be associated with FPIES, and incomplete penetrance would be suggested since the father also has the mutation and LOF variants in gnomAD have been observed.

4.4 Copy Number Variants

CNVs were analysed using XHMM software. A total number of 1,506 variants were called for all individuals with at least a median coverage of 80x. Due to the large number of false CNVs that are usually identified from WES data, these were filtered by high quality (as previously recommended, [107]). Next, only variants that were suspected to be unique in the probands (*de novo*), that were overlapping genes present in the gene list or that were overlapping genes with a candidate SNV/indel from Section 4.3 were considered. A total number of 21 CNVs were obtained, and all of them were manually reviewed with Integrative Genomics Viewer (IGV) [219].

Manual review was based on observation of the SNVs present within the CNV boundaries. An example is further explained in Figure 4.18. In this case, a duplication was called at Chr1:161,487,614-161,518,973 in F05_01. The duplication was overlapping the *FCGR3A* gene, which has been previously associated with Immunodeficiency, is present in the gene list. By looking at the SNVs in the highlighted region, it was possible to discern that the proband and the father had a coverage ratio of 2:1 for SNVs in these region, while the mother and the sibling had a ratio 1:1, suggesting that the duplication in the proband was likely to be real and inherited from the father.

FCGR3A mutations cause Immunodeficiency in an autosomal recessive way. Thus, the possibility that the affected F05_01 had a second SNV/indel in trans was considered. However, no second SNV/indel was identified in this gene for this individual, and the variant was deemed to be likely benign.

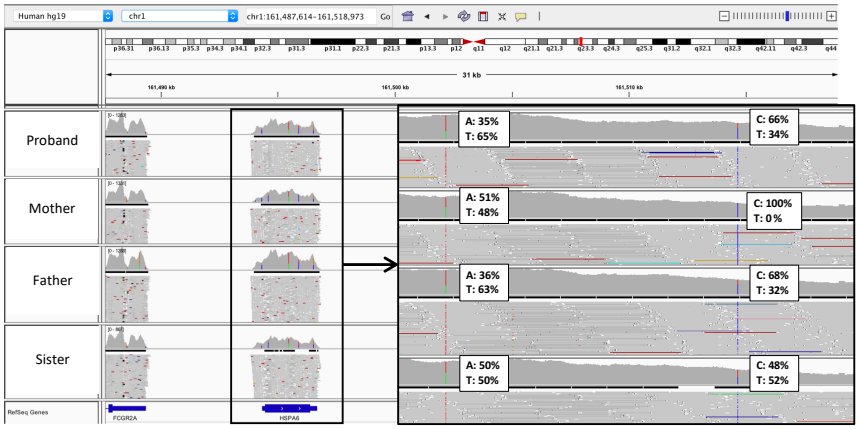


Fig. 4.18 Copy number variant overlapping *FCGR3A* in Family 5. IGV plot of the alignments where the duplication was called in F05_01 and F05_03. The ratio of reads supporting the alternate allele was 65% approximately, consistent with a duplication event.

No pathogenic or candidate CNVs were identified from this analysis. However, WES technology is limited for calling CNVs primarily due to non-uniform coverage. The combination of WES with microarray, or WGS, would be a powerful approach to better identify CNVs.

4.5 HLA typing

HLA typing was performed using HLA*PRG [112] on all individuals with a minimum median coverage of 80x. Because the aim was to identify a possible HLA locus that could be associated or contributing to the disease, the analysis was ran on these participants (cases and relatives), and also on 120 internal controls for which the lab at INCLIVA had previously performed WES.

Stricter filters were applied before further analysis, including a minimum average coverage by locus of 15x and high quality. DRB3 and DRB4 loci were excluded of the analysis due to low coverage. For the remaining loci in the 31 individuals from the seven families, HLA haplotypes were observed to be segregating as expected within the family. PyHLA was used to perform an association test between HLA alleles [220]. The 120 controls were compared to six unrelated cases (Family 2 did not have enough coverage to perform the analysis). For Family 7, which had two affected individuals, the most severely affected individual was selected (the male, F07_04).

First, the data summary function was executed and allele level summary of the frequency was produced in the case and control populations (Figure 4.19).

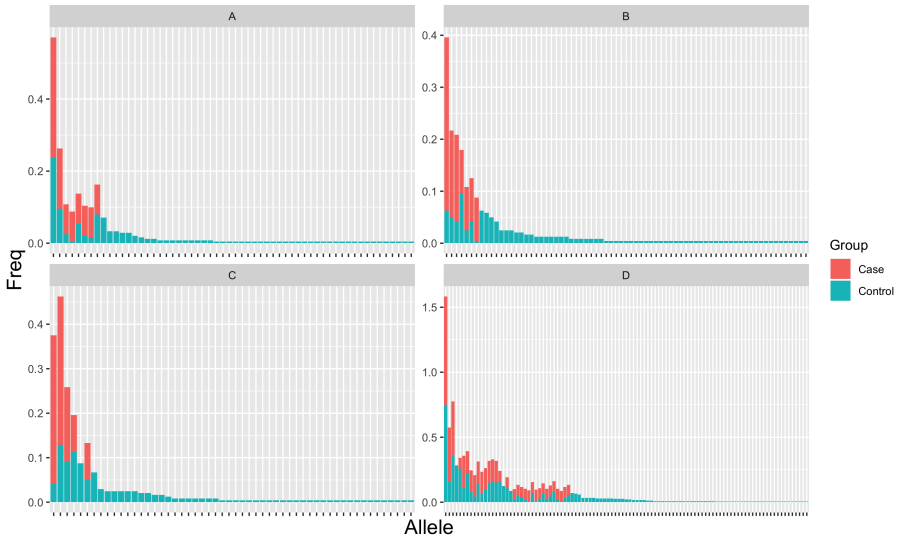


Fig. 4.19 Frequency of HLA alleles by group. Stacked bar plots show frequencies for the different HLA class I alleles, and are coloured by case or control status.

Table 4.4 Output for the Fisher's exact test with adjusted *p-value* > 0.05. Allele: Allele name; A_case: Count of this allele in cases; B_case: Count of other alleles in cases; A_ctrl: Count of this allele in controls; B_ctrl: Count of other allele in controls; F_case: Frequency of this allele in cases; F_ctrl: Frequency of this allele in controls; Freq: Frequency of this allele in cases and controls; P_FET: P-value for Fisher's exact test; OR: Odds ratio; P_adj: Multiple testing adjusted *p-value*.

Allele	A_case	B_case	A_ctrl	B_ctrl	F_case
C*07:02	4	8	10	230	0.3333
B*07:02	4	8	15	225	0.3333

Allele	F_ctrl	Freq	P_FET	OR	P_adj
C*07:02	0.0417	0.0556	0.0023	11.5	0.0116
B*07:02	0.0625	0.0754	0.0079	7.5	0.0236

Then, PyHLA Fisher's exact test was performed. Fisher's exact test first calculates the exact probability of the 2x2 contingency table of the observed values. *p-values* were adjusted by using the false discovery rate (FDR) correction. Two significant *p-values*, which results are in Table 4.4, were identified, for C*07:02 and B*07:02 alleles.

These two alleles were present in five affected individuals, the four included in the Fisher's exact test and also in F07_01, the affected sibling of F07_04 (Figure 4.20). The individual F06_03 (father of F01_01) was also a carrier of these alleles. Interestingly, he presented severe intolerance to milk during the first months of life, and was in hospital for six months. These alleles have not been previously reported alone or in combination as associated with any disease, or to any kind of allergic response to food, hence their role in GI food allergy remains to be confirmed.

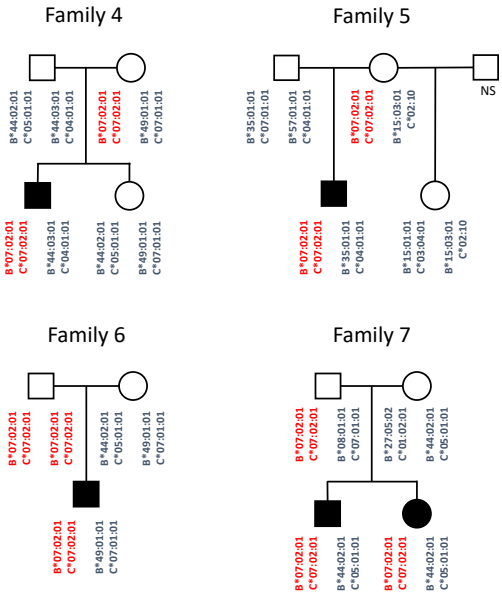


Fig. 4.20 HLA haplotypes in locus B and C in families 4, 5, 6 and 7. Affected individuals are represented in black. NS=not sequenced.

In conclusion, HLA typing was performed to all individuals and 120 controls. Association tests showed a significant association of C*07:02 and B*07:02 alleles, present in combination in six affected individuals (Figure 4.20). Further investigations and HLA typing of larger cohorts of individuals with food allergy are required to interpret the contribution of these alleles to the gastrointestinal food allergy in the affected participants.

Chapter 5

Discussion

5.1 Summary of findings

During the past few years, exome sequencing has successfully been used to identify common and rare variants that confer substantial risk for multiple disorders. In this thesis, WES has been used to study the genetic basis of gastrointestinal food allergy induced by multiple food protein.

Seven families with affected children with severe manifestations of this disorder, diagnosed with FPIES, were whole-exome sequenced, accounting for a total number of 31 individuals. A pipeline was developed to study the possible contribution of rare variants in those individuals. First, an exhaustive quality control of the data was performed. Then, due to the uncertain aetiology of the disorder, candidate variants were identified by mode of inheritance and gene list filtering. A list of candidate variants was obtained. While many of them were in genes that had not been previously described as associated with GI food allergy,

potentially interesting genes were observed to have rare variants in the affected individuals.

In this thesis, WES was also used to analyse CNVs and HLAs in these families. Although no candidate CNV was identified, study of the HLAs uncovered the presence of two alleles that were significantly associated with affected individuals. This work presents the utility of exome sequencing to study, in a single pass, rare SNVs/indels, CNVs and HLAs haplotypes, which could be associated with severe FPIES. This is also the first systematic study of individuals with FPIES by NGS.

5.2 Utility of exome sequencing

Identification of rare variants by exome sequencing

Family-based exome sequencing is an effective strategy that reduces analytic cost and allows the identification of candidate variants in the entire exome, permitting gene discovery. Here, family-based exome sequencing was used to identify candidate variants with and without the use of a gene list. This was crucial because, although 11 candidate variants were identified in genes associated with immune system, additional 17 were observed in genes that had not previously been associated with any disorder.

Because CNVs have been previously associated with paediatric food allergy [175], in this work the identification of CNVs was also performed. However, exome sequencing is a limited technology to detect copy number changes due to the non-uniform coverage distribution and biased amplification of specific regions. For that reason, identification of CNVs in this study was limited to i) *de novo* mutations, ii) genes

from the gene list and iii) genes with a candidate SNV/indel identified in this SNV/indels analysis. However, no candidate CNVs was identified. Nevertheless, the absence of CNVs does not necessarily mean that there are not any CNVs in coding regions contributing to the phenotype. It could be possible that the variant calling of the CNVs in those regions was not possible due to biased amplification and/or low sensitivity of the variant caller algorithm. Or it could also be possible that CNVs are present in a gene not included in the gene list.

Lastly, exome sequencing data was used to perform HLA typing. Because previous HLA haplotypes have been associated with food allergy [277, 174, 278], there was an interest to identify HLA types that could be associated with the phenotype in these affected children. Therefore, a specific algorithm to type HLAs from WES data was used. Association test revealed the presence of two alleles, B*07:02 and C*07:02, that were present in this combination in five of six affected individuals that were considered for this analysis (1 affected individual did not have enough coverage), with an adjusted *p-value* of 0.0236 and 0.0116 respectively. Since the typing for all individuals was available, it was possible to perform phasing of the alleles and check inheritance, which was consistent in all cases. These results showed that WES can indeed be used for HLA typing, at least in a research context.

Altogether, results arising from this thesis show how WES is an good approach to perform a comprehensive analysis of genomic variation in a single pass, including the study of SNVs/indels, CNVs and HLA haplotyping, that can be used to investigate the genetic basis of severe FPIES in affected individuals.

Importance of quality control analysis

In order to identify potential disease-causing mutations with high sensitivity and specificity, multiple quality control analyses needed to be performed at different stages of the pipeline. In this work, quality control was performed on the raw reads, aligned reads and variants called. Overall, all samples passed the quality controls.

Because samples were sequenced at different centres and by different methods, coverage differed between those who had been sequenced with Nextera kit only (at INCLIVA and HIC centres, with lower coverage), those who had been sequenced with SureSelect kit only (at CRG, with higher coverage), and those who had been sequenced by both. A total number of eight individuals in the project did not accomplish the goal of 80x median coverage, therefore they were not included in the CNV and HLA analyses.

Relatedness and gender analyses were also relevant quality control steps since sample extraction and library preparation had been done at different centres. These were used to demonstrate the identity of a sample and its relatedness with the expected relatives. Results showed that genomic data was consistent with expected pedigrees and genders. If this had not been performed and it had been a sampling problem, this would have been identified at the end of the workflow, when a high number of false *de novo* variants would have been observed. However, these analyses provide identification of these kind of problems in early stages of the pipeline, allowing an improved performance of the analysis.

Limitations of Whole-Exome Sequencing

Although short-read WES is a powerful approach, it also presents some limitations that need to be taken into account.

First, variants in regions not covered by design are missed. This is the case of i) non-coding variants, which are particularly relevant in the context of this study since mutations in these areas have previously been associated with food allergy [279, 280], and ii) variants in exons that are not considered in the pull-down array, such as genes from the mitochondrial genome. Mitochondrial variants could play a role in the development of FPIES since two candidate variants in this study are in genes involved in mitochondrial regulation and function (*ANKZF1* in Family 1 and *CAPN14* in Family 7). Although there are techniques to analyse off target reads [281], these are still experimental and unreliable at current standard WES mean coverage.

Second, amplification steps lead to biased amplification of specific regions, where some are over-amplified above others. For example, GC-rich regions tend to be poorly covered due to their high stability and consequent resistance to standard denaturation protocols [37]. Although there are free-PCR WES protocols, these are still less commonly used due to the higher amount of input DNA that is required.

Third non-uniform coverage distribution affects performance of i) variant calling, where variants in regions of low coverage will present higher error rates or will not be called at all, and ii) CNV detection, which is only limited to large copy number gains and losses of exonic regions due to challenges when performing data normalisation.

Additionally, WES is also limited to identify short tandem repeat (STR) expansions and SVs that do not produce any copy number change,

like inversions, large insertions and translocations, due to the very low probability of the breakpoints to be covered by WES reads. Similarly, it also fails to detect other types of variation such as more complex structural rearrangements [28].

Lastly, highly repetitive regions, genes with corresponding pseudogenes or other highly homologous sequences are generally poorly covered in short-read sequencing data, due to the difficulty of uniquely mapping the reads in these regions to the genome.

Most of these limitations can be addressed with the use of WGS, which performs the sequencing of all coding and non-coding regions of the genome, as well as mitochondrial DNA. With the absence of pulldown arrays and the PCR-free sequencing protocols, the coverage achieved is much more uniform, facilitating detection of variant in GC-rich regions and the detection of all types of SVs [37, 282], with high precision, often to single base pair resolution. Additionally, regions of bad mappability due to highly repetitive sequences may be overcome with the use of long-read sequencing technologies such as Nanopore, either in combination with another technology or as a first line approach. These have the advantage of reads of 10–100 Kb allowing for more accurate mapping particularly over repetitive regions and facilitating phasing [28].

Another approach to overcome the limitations of WES could be the combination of technologies, for example, WES with low coverage WGS. However, although in the future WGS may replace WES, the fact that assessing pathogenicity is still mainly linked to coding regions, as well as the substantial extra cost and bioinformatics challenges faced with handling the larger WGS data, makes WES currently the standard NGS technology.

5.3 Variant discovery in FPIES

From this study, at least one SNV/indel or HLA allele has been prioritised and selected for discussion for all affected individuals. Candidate mutations in different genes were identified, highlighting the possible genetic heterogeneity of food allergies. Additionally, variants were identified in genes involved in different pathways and presenting with mutational mechanisms.

The importance of interleukins (IL) signalling was highlighted, as well as the possible role of proteins in the NF-$\kappa\beta$ pathway and extracellular matrix organisation. Interestingly, variants in genes involved in mechanisms that have been recently associated with GI food allergies were also identified, including mitochondrial stress and neuroimmune regulation and homeostasis. A summary for the candidate SNVs/indels identified is shown in 5.1. These results altogether highlight the complex spectrum of GI disorders, and could be the reason why the genetic study of this disease has been hindered during all this time. They also display and reveal important insights into the complex genetic architecture of FPIES, that are thereafter described.

Table 5.1 Summary of candidate variants. Genes are grouped by function/pathway involved. Those with multiple functions/pathways have the number of entries between parenthesis. MOI = mode of inheritance.

Individual	Gene	Function / pathway	Inheritance	Consequence
Interleukins signalling pathway				
F02_01	*IL13RA2*	Cytokine Signalling in Immune system	*De novo*	Missense
F03_01	*TNFRSF1A (2)*	Cell survival, apoptosis, and inflammation; NFKB pathway; Immune system	Comp. het	Missense
			Comp. het	Missense
	LAMA5	Cytokine Signalling in Immune system	Comp. het	Missense
			Comp. het	Missense
NF-κβ pathway				
F06_01 (F06_02)	*GFI1 (2)*	Transcriptional repressor	Inherited	Frameshift
F01_01 (F01_03, F01_04, F01_05, F01_06)	*NLRP12*	Attenuating factor of inflammation	Inherited	Missense
F04_01 (F04_02, F04_04)	*NLRP12*	Attenuating factor of inflammation	Inherited	Missense
F03_01	*TNFRSF1A (2)*	Cell survival, apoptosis, and inflammation; NFKB pathway; Immune system	Comp. het	Missense
			Comp. het	Missense
F06_01	*MAP3K15*	Apoptotic cell death	XLR	Splice donor
T-cell development				
F06_01	*STAB1*	Angiogenesis, cell adhesion, or receptor scavenging	Comp. het	Missense
			Comp. het	Missense
Extracellular matrix organization				
F07_01, F07_04, (F07_03)	*CAPN14*	Cell proliferation	Inherited	Frameshift

Continued from previous page

Individual	Gene	Function / pathway	Inheritance	Consequence
F04_01	*PPL*	Linking protein	Comp. het	Missense
			Comp. het	Splice region
Mitochondrial stress				
F01_01 (F01_04)	*ANKZF1*	Maintenance of mitochondrial integrity	Inherited	Missense
			Inherited	Missense
Neuroimmune regulation and homeostasis				
F06_01	*GPR50*	Inhibition of melatonin receptor	XLR	Missense
Gene expression and chromatin remodelling				
F06_01 (F06_02)	*GFI1 (2)*	Transcriptional repressor	Inherited	Frameshift
F06_01 (F06_03)	*INO80*	Chromatin remodelling complex	Inherited	Missense
Others				
F03_01 (F03_02, F03_04)	*SKIV2L*	Cell proliferation	Inherited	Missense
			Inherited	Missense
F02_01	*ZNF645*	Class I MHC mediated antigen processing & presentation	XLR	Missense

5.3.1 Interleukins signalling pathway

The important role of IL signalling pathways was underscored by the identification of variants in the genes *IL13RA2* (receptor of IL-13), *TNFRSF1A* (receptor of TNF-α) and *LAMA5* (up-regulated by IL-4 and IL-13 signalling [283, 284]). IL are secreted proteins that bind to specific receptors and play a role in intercellular communication among

leukocytes. Several IL have been associated with atopic responses, such as IL-4 and IL-13, which have been used in clinical trials for the treatment of asthma and atopic dermatitis [285]. Furthermore, recent studies showed positive effect of anti-IL-13 treatment on oesophageal eosinophilia in patients with eosinophilic oesophagitis [286, 287]. This, and also the fact that IL-4 and IL-13 are cytokines of type-2 immune response [288], highlight the possible role of IL in the pathogenesis of FPIES, which has been suggested to be part of a type-2 mechanism response.

Furthermore, TNF-α is not a type-2 immune response specifically, but is an important pleiotropic cytokine involved in host defence, inflammation, and apoptosis, and has also been associated with allergic diseases such as asthma and atopic dermatitis [289, 290]. TNF-α blockers have already been used for the treatment of inflammatory bowel disease, thus the role of TNF-α in GI maintenance is of potential interest. These results suggest that IL pathways could be involved in the pathogenesis of FPIES.

5.3.2 NF-$\kappa\beta$ pathway

Another relevant signalling pathway highlighted in this study was the NF-$\kappa\beta$ pathway. Activation of NF-$\kappa\beta$ has already been observed in allergic responses [291, 292], but never demonstrated to play a role in pathogenesis of FPIES. Here, mutations in three genes that directly regulate the NF-$\kappa\beta$ pathway were reported: *GFI1*, *NLRP12* and *TNFRSF1A*. Additionally, two affected individuals from two different families were identified to have a XLR mutation in the *MAP3K15* gene, previously associated with apoptosis. Although no previous associations to NF-

$\kappa\beta$ activation have been reported, this gene belongs to the family of MAP3Ks, some of them notable activators of NF-$\kappa\beta$ pathway, such as *NRK* [293]. This could suggest a possible role of *MAP3K15* in NF-$\kappa\beta$ regulation, and altogether these variants emphasised the interplay of NF-$\kappa\beta$ in allergic diseases, and open a new discussion for its role in the pathogenesis of FPIES.

5.3.3 Mitochondrial dysfunction

Mitochondrial dysfunction has been associated with GI disorders. In general, mitochondrial pathology (as for example, electron transport chain complex dysfunction, diminished mitochondrial membrane potential and changed mitochondrial morphology), have been observed in patients with IBD and EoE [237]. Therefore, and underlined by the findings in *ANKZF1*, these results suggest a role for mitochondrial dysfunction in FPIES, highlighting the phenotypic overlap between different GI disorders (IBD and FPIES).

5.3.4 T cell development

One affected individual presented a compound heterozygous variant in *STAB1*. This gene plays an important role during T cell development and negative selection in the thymus. Negative selection of the T-cell antigen receptors occurs in the thymic cortex, after being generated by recombination. The negative selection shapes the T-cell repertoire to avoid self-reactivity, which powerfully contributes to the avoidance of autoimmunity. This negative selection in the thymus functions as the major mechanism of central immune tolerance. Therefore, the fact that one candidate variant was identified in *STAB1* highlights the role

that problems during proper regulation of T-cell development to avoid autoimmunity reactions could play in the development of FPIES.

5.3.5 Extracellular matrix organisation

Two of the variants identified were in genes previously associated with extracellular matrix organization: *CAPN14* and *PPL*. The first one, *CAPN14*, has been associated with EoE and impairs epithelial barrier function by diminishing the expression of DSG1, a cadherin-like transmembrane glycoprotein that is major component of the desmosome [276]. The second one, *PPL*, is a component of desmosomes and of the epidermal cornified envelope in keratinocytes.

Desmosomes are cell-cell junctions that help resist shearing forces and are found in high concentrations in cells subject to mechanical stress. Impairments in the desmosome function can lead to extracellular matrix disorganization, specific cell type infiltrations, and cause an increased expression of proinflammatory extracellular matrix molecules. These results emphasize the importance of an appropriate extracellular matrix homeostasis, and how its impairment could lead to proinflammatory responses, including GI disorders.

5.3.6 Neuroimmune regulation and homeostasis

The immune system and nervous system are anatomically connected, mechanistically communicate and reciprocally influence the other's function. It has been suggested that enteric neurons and intestinal immune cells share common regulatory mechanisms and can coordinate their responses to specific challenges [294].

Melatonin has been associated with oesophagitis and chronic inflammation [266]. Furthermore, it is a Trp metabolite like serotonin, which was previously associated with FPIES [268, 129]. The variant identified in *GPR50* gene, which encodes for a protein that inhibits the melatonin receptor, is an interesting finding that reinforces the possible role of neuroimmune regulation in the pathogenesis of GI disorders, such as FPIES.

5.3.7 Gene expression and chromatin remodelling

Expression and/or repression of specific genes are important factors to consider when studying the pathogenesis of multiple diseases. For that reason, mutations identified in genes that encode for transcription factors (such as *GFI1*) or the chromatin remodelling complex (such as *INO80*) were of particular interest.

GFI1 encodes for a transcriptional repressor which is important for Th2 cell differentiation [295]. More specifically, Gfi1 plays an important role in the regulation of IL-5 and IFN-γ production in Th2 cells, as well as the regulation of GATA3. Therefore, the cooperation of transcriptional factors such as Gfi1 and GATA3 is required for the proper Th2 cell differentiation.

Similarly, *INO80* is proposed to bind DNA and be recruited by specific transcription factors to activate certain genes and repress inappropriate transcription at promoters in the opposite direction to the coding sequence. Although the molecular mechanism of *INO80* is uncertain, it appears to be associated with immunodeficiency. These data emphasise the role that transcription factors and/or chromatin remodelling proteins

have over Th2 cell differentiation and specific gene expression, possibly playing an important role in the pathogenesis of GI food allergies.

5.3.8 HLA variation and disease

Previously HLA alleles have been associated with diseases such as multiple sclerosis [296], T1D [297] and Coeliac disease [298]. Other works have found association of certain HLA alleles to peanut allergy [277]. However, the role of HLA in food allergy, especially GI food allergy, is not yet fully understood. It is suspected that HLA class I and II molecules play an important role in the pathogenesis of food allergy due to their crucial role in presenting a vast array of antigenic peptides to T cells [109].

The majority of autoimmune disease-HLA associations for which molecular mechanisms of actions have been identified are in *HLA-DR* and *HLA-DQ* alleles. There is not much known about a possible role of *HLA-B* and *HLA-C* in FPIES (although a specific *HLA-C* allele has been observed in individuals with Crohn's disease [109]). Therefore, the suggestive association of *HLA-C**07:02 and *HLA-B**07:02 alleles to FPIES identified in this work expands the concept about HLA variation and disease.

Nevertheless, there is a limitation in this analysis that needs to be taken into account: the low power of the association test due to the sample size. This could also be one of the reasons why the *p-values* from this work, although significant, are at the order of 1e-2, while previous large-scale cohort analysis have reported HLA associations with a *p-value* of the order of at least 1e-8 [296, 298, 297, 299]. Therefore, larger case-control studies would be required to confirm.

Solving this problem is not straight forward, since affected individuals with severe FPIES are very rare in the population. Therefore, it would be a challenge to recruit a large number of patients with this phenotype, avoiding those with similar symptoms but different aetiologies. Different projects have performed large-scale genome sequencing on patients with rare diseases [25, 53], but ideally this could be achieved at a national level, through the national health service.

Another possibility to consider is that specific HLA allele/s could be contributing to the manifestation of disease, in combination with other causal mechanisms such as the presence of SNVs or CNVs. This is one of the reasons why WES is a good technology for the study of patients with GI food allergies, because with only one experiment it is possible to perform a comprehensive analysis that allows consideration of multiple types of genome variation.

5.4 Gender bias

Gender differences in the development and prevalence of human diseases have long been recognised, and there is an increased interest in the understanding the different factors that may be responsible for this disparity in the homeostasis of immunity [189]. A slight male predominance of 60:40 has been reported in FPIES [300]. This is consistent with the sex disparity observed among children with food allergies (65:35). Interestingly, this ratio inverts in adulthood, were 65% are females, compared to 35% males [301]. Although different factors may be responsible for this disparity (including gender-specific behaviour or specific intake of medications), recent studies have focused on the study of the hormonal effects.

The direct effect of sex hormones has rarely been investigated in food allergies. However, it is well known that women show higher antibody responses against infections and vaccines [302]. This is because oestrogens can promote autoimmunity since they enhance humoral immune responses; on the contrary, androgens and progesterone have an immunosuppressive effect.

In this study, a candidate compound heterozygous variant in *ANKZF1* gene was identified in the proband of Family 1 and in the asymptomatic sister. A situation like this could be explained by many reasons, including 1) incomplete penetrance, 2) the variant is partially contributing to the phenotype, 3) the female has an additional protective variant, or if, as here suggested, 4) the response in the female is currently less severe due to the interplay of specific hormones. Hence, if pathogenicity of this mutation is demonstrated in the proband by functional assays, and the latter is true, its possible effect on the sister in a long term should be considered.

5.5 Effect of genetic variants in multiple genes

Oligogenic disorders are either caused or modulated by the action of a small number of loci. Some examples of oligogenic disorders are Usher syndrome type I and Nephrotic syndrome, among many others [303], where mutations in multiple genes from the same pathway or with similar functions contribute to disease.

The affected proband of Family 4 presented two different variants, one in *GFI1* gene inherited from the mother, and one in *INO80* gene inherited from the father. The role these two variants play in disease pathogenesis remains unknown, though a possible hypothesis could

be that impairment of both genes, which play important roles in transcriptional regulation, could be affecting expression of immune system genes.

Likewise, polygenic inheritance, involving many common genetic variants of small effect, can play a greater role than rare monogenic mutations for many common diseases [304]. This is based on the combination of multiple risk alleles, on the basis that there may be an accumulation of weak effects on the key genes and regulatory pathways that drive disease risk. Recent studies utilising large datasets have established polygenic risk predictors in different common diseases, as for example, in inflammatory bowel disease [304]. Therefore, it would not be surprising that polygenic risk can play an important role in the pathogenicity of GI food allergies.

For several diseases, when a specific gene is associated with disease, the study of genes with similar function or in the same pathway helps to highlight specific molecular processes, like the role of autophagy in Crohn's disease [305], and roles for adipocyte thermogenesis and central nervous system genes in obesity [306, 285, 307]. However, a recent hypothesis, proposed by Boyle *et al* [308], postulates that some complex disorders could be omnigenic. The omnigenic model posits the existence of a small number of core genes having biologically interpretable roles in disease, along with a much greater quantity of peripheral genes regulating the core genes. Because the number of peripheral genes is much greater than core genes, they account for a greater proportion of the variability than the core genes (Figure 5.1). This is a recent hypothesis, and separating these two classes will require more research.

The polygenic and omnigenic models are exciting fields of study. The main limitation is that a large number of data sets are required to

perform the analyses. Nevertheless, with the advent of HTS technologies, the investigation of the effect of multiples genes in the pathogenesis of GI food allergies might be facilitated.

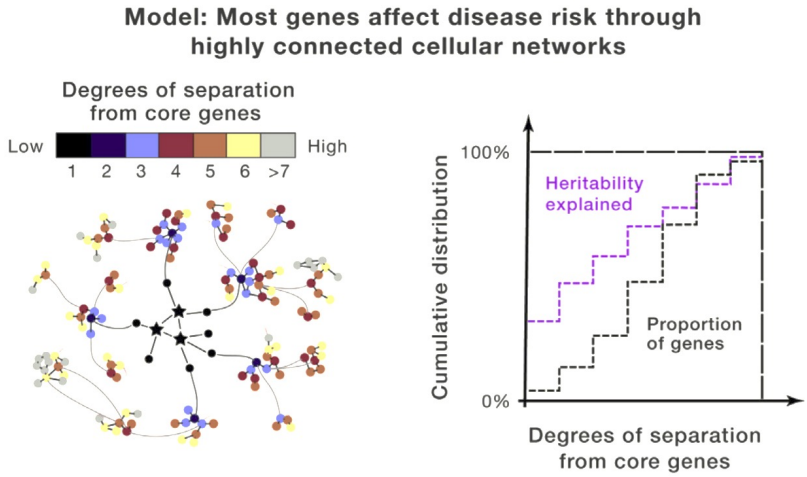


Fig. 5.1 Omnigenic model of complex traits. For any given disease phenotype, a limited number of genes have direct effects on disease risk. However, by the small world property of networks, most expressed genes are only a few steps from the nearest core gene and thus may have non-zero effects on disease. Since core genes only constitute a tiny fraction of all genes, most heritability comes from genes with indirect effects.

5.6 Translation into the clinic

Thanks to large-scale sequencing studies we are likely to see the rapid accumulation of known loci associated with complex traits like FPIES in the near future. It is hoped that geneticists will be able to complete the picture of missing heritability and explain the role of genetics in these kinds of disorders. However, it is important to interrogate the benefits these discoveries will provide to those individuals that are affected today.

First of all, providing a diagnosis would end with the diagnostic odyssey the patients and relatives are exposed to. Participants of this work have had a large number of investigations undertaken in an attempt to define and diagnose the cause of the symptoms, and some of them never even received a proper diagnosis. Therefore, knowing what is the cause of the disease, especially after many years of research, could be a relief for the family and ends the turmoil they are exposed to.

Moreover, a genetic diagnosis may lead to a specific treatment, if available. Unfortunately, there are no European Medicines Agency (EMA) approved treatment options for FPIES, and the current mainstay of treatment of food allergies is allergen avoidance. Nevertheless, the findings of this work expand the molecular biology of FPIES which could possibly lead to the development of new drugs, or even bring to light possible treatments that are already known to regulate certain pathways that now are associated with the pathogenesis of FPIES. For example, different variants in genes involved in the NF-$\kappa\beta$ pathway were identified in multiple families, and numerous drugs and substances have been seen to regulate the NF-$\kappa\beta$ pathway (listed in Table 5.2). A candidate variant was also identified in *IL13RA2*, a receptor involved in signalling mediated by IL-4/IL-13 cytokines, and dual blockade of IL-4 and IL-13 with *dupilumab* demonstrated significant efficacy in allergic diseases.

Additionally, anti-TNF treatments have been seen to be effective in different atopic diseases, and genetic variants were found in *TNFRSF1A*.

Table 5.2 Therapeutic strategies for NF-$\kappa\beta$ regulation. Adapted from [309].

Therapeutic strategy	Mechanism	References
IKK-β-dominant-negative gene therapy	Prevents TNF-α–mediated NF-$\kappa\beta$ nuclear translocation and proinflammatory gene expression in synoviocytes	[310]
NF-$\kappa\beta$ decoy oligonucleotides	By increasing apoptosis and suppressing cytokine gene expression - suggested	[311, 312]
T-cell specific NF-$\kappa\beta$ inhibitor	Significantly decreased arthritis severity in CIA in mice. NF-$\kappa\beta$-directed therapy is also effective in a model of inflammatory bowel disease induced by 2, 4, 6,-trinitrobenzene sulfonic acid	[313, 314]
Corticosteroids	Inhibition of NF-$\kappa\beta$ activation	[315]
Sulfasalazine	Block nuclear translocation of NF-$\kappa\beta$ through inhibition of IκBα degradation	[315]
5-aminosalicylic acid	Inhibit the production of cytokines and inflammatory mediators	[316, 317]
Aspirin	Function as a competitive inhibitor of IKK-β	[315]
Tepoxalin	Inhibit the production of cytokines and inflammatory mediators	[318]
Leflunomide	Block nuclear translocation of NF-$\kappa\beta$ through inhibition of IκBα degradation	[315]

Continued on next page

Table 5.2 – continued from previous page

Therapeutic strategy	Mechanism	References
Others	Curcumin suppresses IKK/IκB/NF-$\kappa\beta$ and c-Raf/MEK/ERK inflammatory cascades as well as prevents their translocation into the nucleus. Vanillin suppresses the expression of proteasome and other antioxidants, such as resveratrol, can inhibit the activities of NF-$\kappa\beta$ and I$\kappa\beta$ kinase.	[319–322]

Elucidating the central disease pathways in FPIES holds the potential to identify not only new therapies to provide temporary symptomatic relief, but also to investigate if benefits of already existing drugs or natural products is achievable.

For example, oral administration of TGF-β1 has been reported to protect the immature gut from injury by suppression of NF-$\kappa\beta$ signalling and proinflammatory cytokine production, and suggested to protect against gastrointestinal diseases [323].

A natural product is curcumin, which has numerous pharmacological benefits including anti-inflammatory activities. Previous studies observed that curcumin induces suppression of I$\kappa\kappa$/I$\kappa\beta$/NF-$\kappa\beta$ and c-Raf/MEK/ERK inflammatory cascades as well as prevents their translocation into the nucleus [320]. This suppression showed promising anti-inflammatory activity by significantly inhibited IL-6 production (which modulates allergic inflammation in skin) in HaCaT cells. Additionally, vanillin (4-hydroxy-3-methoxybenzaldehyde) has also been seen to play

a role in NF-$\kappa\beta$ pathway, a potent NF-$\kappa\beta$ inhibitor. Vanillin is a natural component which has been reported to have anti-inflammatory activities, improves and prevents colitis in mice and ameliorates the development of cancers in mice with induced colitis-associated colon cancer [322, 319]. It has been suggested that vanillin suppresses the expression of proteasome and subsequently alters NF-$\kappa\beta$ and MAPK pathways, which in turn suppress the proliferation of cells and the infiltration of immune cells.

This work emphasises the possible role that different signalling pathways may play in FPIES, and reveals possible therapeutic strategies that could be beneficial for the affected individuals.

5.7 The microbiome

The investigation of the interaction between an individual's genome and their environment is another area that offers particular promise for the translation of genetic findings. The GI microbiota plays an important role in disease pathogenesis, where the epithelial barrier and autophagy pathways are implicated [324]. Microbiome studies in individuals with allergic disease have reduced beneficial bifidobacterial species and increased numbers of clostridia and staphylococci compared to non-allergic infants [325]. Similarly, studies on infants with EoE demonstrated that distal oesophageal biopsies from healthy subjects are dominated by *Streptococcus* species, while affected individuals with oesophageal inflammation have predominantly gram-negative anaerobes or microaerophilic bacteria [326]. However, it is uncertain if the disturbed microbiome arises as a result of an extensive inflammatory response

caused by a different reason (such as genetic variation), or if it triggered the response.

It has been suggested that FPIES pathogenesis involves an interplay of environmental and genetic factors, so it could be possible that mutations in genes involved in maintenance of the epithelial barrier and autophagy could cause dysbiosis that might contribute to an aberrant or exaggerated inflammatory response. However, genetics is not the only factor that alters the microbiome; it can also be perturbed by maternal-foetal interaction, place and mode of delivery, early feedings strategies and the use of antibiotics, making it difficult to unravel cause and effect.

Understanding the role of the microbiome in FPIES is important due to the recent success of faecal microbiota transplants (FMTs) as treatment for allergic colitis [327]. FMTs aim to change the gut microbial composition of an affected individual and confer a health benefit by the administration of stool from a healthy donor [328]. FMTs have been used in gut microbiota dysbiosis, such as *Clostridium difficile* infection [329, 330], inflammatory bowel disease [331, 332], and irritable bowel syndrome [333, 334, 331]. Moreover, FMT is currently being investigated as a therapy for paediatric allergic disorders [335].

To determine whether the association between environmental factors such as microbiota dysbiosis and the presence of FPIES is consistent with a causal effect, a bidirectional Mendelian randomisation technique has to be considered [336]. Essentially, this is based on the fact that an individual genotype can affect the phenotype, and both phenotype and environment can interact with each other, but not with the genotype (except in somatic mutations) (Figure 5.2). Hence, genetic variation always acts as a causal 'anchor'. For that reason, its study is a useful

start for understanding the relationship between environmental factors and the development of disease.

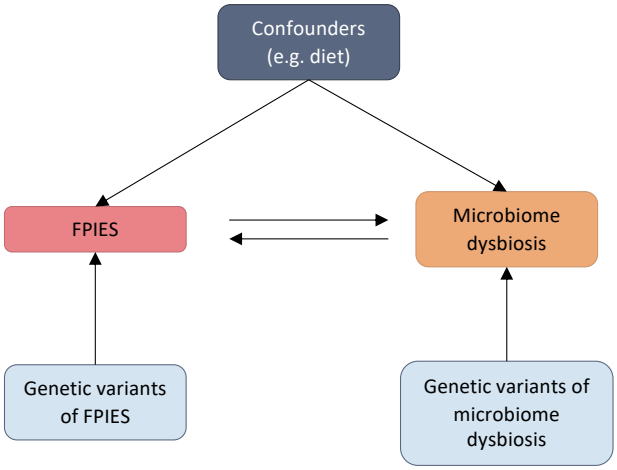


Fig. 5.2 Mendelian randomisation in FPIES. In this case, Mendelian randomisation can be used to infer a causal relationship between FPIES and the microbiome. If the correlation has arisen because FPIES causes microbiome dysbiosis, then any variable that affects FPIES (such as genetic variants) should also affect microbiome, but not vice versa.

Therefore, genetics could provide a valuable opportunity to unravel the role of the microbiome in FPIES, and exome sequencing is a powerful technology by which to achieve this. Microbiome study may even allow us to understand why individuals with susceptible genetic variants develop disease, while others do not.

5.8 Future perspectives

Exome sequencing has been successfully used in this dissertation to study rare variation and HLA haplotypes in individuals with severe FPIES and their relatives.

The full elucidation of the genetic basis of FPIES through this study was not possible because of the very small sample size, the variable quality of the sequencing data and the limitations of WES, that have been discussed. Research on this and similar phenotypes suggest that the disease is likely to be genetically heterogeneous, and it also is probably not a straightforward Mendelian phenotype so common variants and non-genetic factors may contribute, suggesting that large sample sizes may be required. This will be challenging given the low prevalence of the disease, but in the future, developments such as patient registries may make it possible.

Nevertheless, results from this work give insight into the pathogenesis of this disorder by the observation of inherited and sporadic mutations in genes which play an important role in regulation of the immune system. However, despite substantial progress in understanding the underlying mechanisms of FPIES, many questions in the field of food allergy remain to be answered.

First, functional assays of candidate variants would need to be performed to assess the pathogenicity of these mutations, especially the missense mutations. Design of the different types of assays, or the selection of an appropriate animal model, would depend on the consequence of the variants at cellular, tissular, physiological and immunological level.

Second, a different technology may be appropriate to study genetic variation in individuals with FPIES. Although WES has successfully identified SNVs and indels, CNVs and HLA haplotypes, WGS by short or long reads provides not only higher performance on the detection of these types of variants, but also the identification of others that could also be involved in the pathogenesis of the disease and are missed by WES. These include non-coding variants, mitochondrial variants, copy-neutral SVs, complex structural rearrangements and STR expansions. It is hoped that additional technological improvements and software development will lower costs and make these technologies accessible for the routine use in the scientific research.

Third, the study of common variants and polygenic/omnigenic risk scores would also be required to assess pathogenicity, since previous studies observed that common variants may contribute to other types of GI disorders such as IBD [337]. However, in order to perform these analyses, a much larger cohort of affected individuals would be required, and this is a challenge due to the rareness of the disease and the overlapping phenotype spectrum of FPIES with other types of GI disorders.

Finally, due to the important weight of the environment in this disorder, the study of not only the genome, but also the epigenome, gene expression and/or microbiota in affected individuals may also be of interest to fully understand the disease pathogenesis of FPIES.

Chapter 6

Conclusions and final remarks

6.1 Conclusions

In this dissertation, the following has been accomplished:

- The development of a workflow to process the exome sequencing data from seven families affected with gastrointestinal food allergy induced by multiple food proteins. This has been released into the public domain (http://github.com/alsanju/wes-pipeline).
- The assembly of a list of candidate genes associated with immunological disorders, that future larger studies may be able to use to prioritise their own variants.
- The performance of thorough quality control, that showed i) good sequencing and variant quality, ii) good coverage of the exome, iii) that no relatedness between families or sex discrepancies were identified.

- The identification of candidate SNVs/indels and HLA haplotypes across multiple genes in all the families, supporting (with different levels of evidence) that rare genetic variants can be involved in the pathogenesis of the disease, and confirming that this disease is unlikely to be caused by rare mutations in a single gene.

- The identification of possibly associated pathways with the disease, which included i) interleukins signalling pathway, ii) NF-$\kappa\beta$ pathway, iii) T-cell development, iv) extracellular matrix organisation, v) mitochondrial dysfunction, vi) neuroimmune regulation and homeostasis and vii) gene expression and chromatin remodelling.

6.2 Final remarks

The worldwide prevalence of allergy, including FPIES, has increased dramatically over the last decades. Although not much is known about the pathogenesis of this disorder, genetic predispositions, environmental factors, and social behaviour interplay to orchestrate the scenario of allergy manifestation. Over the past few years, there have been dramatic advances in the genetic study of multiple disorders, especially thanks to WES technology, which gives us the ability to sequence large cohorts of individuals and perform analysis of rare variation at an affordable cost.

It is possible that the complete picture of heritability in FPIES will be resolved in the next decades. Studies like this one will be crucial in uncovering the biological mechanisms that underlie disease pathogenesis, and in offering insights that can be used for the development of new therapeutics. Ultimately, understanding the causes of GI food allergies

will lead to improvements in the lives of people suffering from these disorders.

References

[1] Portelli, M. A., Hodge, E. & Sayers, I. Genetic risk factors for the development of allergic disease identified by genome-wide association. *Clinical & Experimental Allergy* **45**, 21–31 (2014). URL https://doi.org/10.1111/cea.12327.

[2] Vicente, C. T., Revez, J. A. & Ferreira, M. A. R. Lessons from ten years of genome-wide association studies of asthma. *Clinical & Translational Immunology* **6**, e165 (2017). URL https://doi.org/10.1038/cti.2017.54.

[3] Marenholz, I. *et al.* Genome-wide association study identifies the SERPINB gene cluster as a susceptibility locus for food allergy. *Nature Communications* **8** (2017). URL https://doi.org/10.1038/s41467-017-01220-0.

[4] Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nature Communications* **8** (2017). URL https://doi.org/10.1038/s41467-017-00257-5.

[5] Khor, S.-S. *et al.* Genome-wide association study of self-reported food reactions in japanese identifies shrimp and peach specific loci in the HLA-DR/DQ gene region. *Scientific Reports* **8** (2018). URL https://doi.org/10.1038/s41598-017-18241-w.

[6] Waage, J. *et al.* Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nature Genetics* **50**, 1072–1080 (2018). URL https://doi.org/10.1038/s41588-018-0157-1.

[7] Matthews, A. G., Finkelstein, D. M. & Betensky, R. A. Analysis of familial aggregation studies with complex ascertainment schemes. *Stat Med* **27**, 5076–92 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18618413.

[8] Uhlig, H. H. Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. *Gut* **62**, 1795–805 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/24203055.

[9] Mitchell, K. J. What is complex about complex disorders? *Genome Biol* **13**, 237 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22269335.

[10] Peltonen, L. & McKusick, V. A. Genomics and medicine. dissecting human disease in the postgenomic era. *Science* **291**, 1224–9 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11233446.

[11] Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* **19**, 212–9 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19481926.

[12] 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). URL http://www.ncbi.nlm.nih.gov/pubmed/19474294.

[13] Stranger, B. E., Stahl, E. A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**, 367–83 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21115973.

[14] International HapMap, C. A haplotype map of the human genome. *Nature* **437**, 1299–320 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/16255080.

[15] Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–53 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19812666.

[16] Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5–23 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24995866.

[17] McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356–69 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18398418.

[18] Dipple, K. M. & McCabe, E. R. Phenotypes of patients with "simple" mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. *Am J Hum Genet* **66**, 1729–35 (2000). URL http://www.ncbi.nlm.nih.gov/pubmed/10793008.

[19] Riazuddin, S. A. *et al.* Missense mutations in tcf8 cause late-onset fuchs corneal dystrophy and interact with fcd4 on chromosome 9p. *Am J Hum Genet* **86**, 45–53 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20036349.

[20] Katsanis, N. The continuum of causality in human genetic disorders. *Genome Biol* **17**, 233 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27855690.

[21] Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333–351 (2016). URL https://doi.org/10.1038/nrg.2016.49.

[22] Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat Methods* **5**, 16–8 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18165802.

[23] Thompson, R., Drew, C. J. & Thomas, R. H. Next generation sequencing in the clinical domain: clinical advantages, practical, and ethical challenges. *Adv Protein Chem Struct Biol* **89**, 27–63 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/23046881.

[24] Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26432246.

[25] Consortium, U. K. *et al.* The uk10k project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26367797.

[26] Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* (2019). URL https://www.biorxiv.org/content/early/2019/01/30/531210. https://www.biorxiv.org/content/early/2019/01/30/531210.full.pdf.

[27] Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat Methods* **4**, 903–5 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17934467.

[28] Sanchis-Juan, A. *et al.* Complex structural variants in mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Medicine* **10** (2018). URL https://doi.org/10.1186/s13073-018-0606-6.

[29] Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-throughput sequencing technologies. *Mol Cell* **58**, 586–97 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26000844.

[30] Xue, Y., Ankala, A., Wilcox, W. R. & Hegde, M. R. Solving the molecular diagnostic testing conundrum for mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet Med* **17**, 444–51 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25232854.

[31] Neveling, K. *et al.* A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. *Hum Mutat* **34**, 1721–6 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/24123792.

[32] Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–6 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19684571.

[33] Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. & Jabado, N. What can exome sequencing do for you? *J Med Genet*

48, 580–9 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21730106.

[34] Ng, P. C. *et al.* Genetic variation in an individual human exome. *PLoS Genet* **4**, e1000160 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18704161.

[35] Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics* **33**, 228–237 (2003).

[36] Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences* **112**, 5473–5478 (2015). URL https://doi.org/10.1073/pnas.1418631112.

[37] Carss, K. J. *et al.* Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease. *The American Journal of Human Genetics* **100**, 75–90 (2017). URL https://doi.org/10.1016/j.ajhg.2016.12.003.

[38] Clark, M. M. *et al.* Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *npj Genomic Medicine* **3** (2018). URL https://doi.org/10.1038/s41525-018-0053-8.

[39] Monies, D. *et al.* The landscape of genetic diseases in saudi arabia based on the first 1000 diagnostic panels and exomes. *Human Genetics* **136**, 921–939 (2017). URL https://doi.org/10.1007/s00439-017-1821-8.

[40] Farnaes, L. *et al.* Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *npj Genomic Medicine* **3** (2018). URL https://doi.org/10.1038/s41525-018-0049-4.

[41] French, C. E. *et al.* Whole genome sequencing reveals that genetic conditions are frequent in intensively ill children. *Intensive*

Care Medicine **45**, 627–636 (2019). URL https://doi.org/10.1007/s00134-019-05552-x.

[42] Delio, M. *et al.* Development of a targeted multi-disorder high-throughput sequencing assay for the effective identification of disease-causing variants. *PLOS ONE* **10**, e0133742 (2015). URL https://doi.org/10.1371/journal.pone.0133742.

[43] Seleman, M., Hoyos-Bachiloglu, R., Geha, R. S. & Chou, J. Uses of next-generation sequencing technologies for the diagnosis of primary immunodeficiencies. *Frontiers in Immunology* **8** (2017). URL https://doi.org/10.3389/fimmu.2017.00847.

[44] Mak, T. S. H. *et al.* Coverage and diagnostic yield of whole exome sequencing for the evaluation of cases with dilated and hypertrophic cardiomyopathy. *Scientific Reports* **8** (2018). URL https://doi.org/10.1038/s41598-018-29263-3.

[45] Thiffault, I. *et al.* Clinical genome sequencing in an unbiased pediatric cohort. *Genetics in Medicine* **21**, 303–310 (2018). URL https://doi.org/10.1038/s41436-018-0075-8.

[46] Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human genome sequencing in health and disease. *Annu Rev Med* **63**, 35–61 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22248320.

[47] Brunham, L. R. & Hayden, M. R. Hunting human disease genes: lessons from the past, challenges for the future. *Hum Genet* **132**, 603–17 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23504071.

[48] Balaresque, P. L., Ballereau, S. J. & Jobling, M. A. Challenges in human genetic diversity: demographic history and adaptation. *Hum Mol Genet* **16 Spec No. 2**, R134–9 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17911157.

[49] Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat Genet* **51**, 88–95 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/30531870.

[50] Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30–5 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/19915526.

[51] Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H. & Inoue, I. Next-generation sequencing: impact of exome sequencing in characterizing mendelian disorders. *J Hum Genet* **57**, 621–32 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22832387.

[52] Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–9 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22604720.

[53] NIHR-BioResource. Whole-genome sequencing of rare disease patients in a national healthcare system (2019). URL https://doi.org/10.1101/507244.

[54] Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–8 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18988837.

[55] Monroe, G. R. *et al.* Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. *Genet Med* (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26845106.

[56] Zhu, X. *et al.* Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet Med* **17**, 774–81 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25590979.

[57] Sawyer, S. L. *et al.* Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clin Genet* **89**, 275–84 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26283276.

[58] MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–76 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24759409.

[59] Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **615**, 28–56 (2007). URL https://doi.org/10.1016/j.mrfmmm.2006.09.003.

[60] Han, F. & Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity* **70**, 42–54 (2010). URL https://doi.org/10.1159/000288704.

[61] Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93 (2011). URL https://doi.org/10.1016/j.ajhg.2011.05.029.

[62] Greene, D., NIHR BioResource, Richardson, S. & Turro, E. A fast association test for identifying pathogenic variants involved in rare diseases. *American Journal of Human Genetics* **101**, 104–114 (2017).

[63] Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311–321 (2008). URL https://doi.org/10.1016/j.ajhg.2008.06.024.

[64] Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics* **86**, 832–838 (2010). URL https://doi.org/10.1016/j.ajhg.2010.04.005.

[65] Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* **5**, e1000384 (2009). URL https://doi.org/10.1371/journal.pgen.1000384.

[66] Sul, J. H., Han, B., He, D. & Eskin, E. An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics* **188**, 181–188 (2011). URL https://doi.org/10.1534/genetics.110.125070.

[67] Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genetics* **7**, e1001322 (2011). URL https://doi.org/10.1371/journal.pgen.1001322.

[68] Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics* **91**, 224–237 (2012). URL https://doi.org/10.1016/j.ajhg.2012.06.007.

[69] Sun, J., Zheng, Y. & Hsu, L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology* **37**, 334–344 (2013). URL https://doi.org/10.1002/gepi.21717.

[70] King, C. R., Rathouz, P. J. & Nicolae, D. L. An evolutionary framework for association testing in resequencing studies. *PLoS Genetics* **6**, e1001202 (2010). URL https://doi.org/10.1371/journal.pgen.1001202.

[71] Chen, L. S., Hsu, L., Gamazon, E. R., Cox, N. J. & Nicolae, D. L. An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics* **91**, 977–986 (2012). URL https://doi.org/10.1016/j.ajhg.2012.09.017.

[72] Zhou, H., Sehl, M. E., Sinsheimer, J. S. & Lange, K. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* **26**, 2375–2382 (2010). URL https://doi.org/10.1093/bioinformatics/btq448.

[73] Ionita-Laza, I., Buxbaum, J. D., Laird, N. M. & Lange, C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics* **7**, e1001289 (2011). URL https://doi.org/10.1371/journal.pgen.1001289.

[74] Tuijnenburg, P. *et al.* Loss-of-function nuclear factor kappab subunit 1 (nfkb1) variants are the most common monogenic cause of common variable immunodeficiency in europeans. *J Allergy Clin Immunol* (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29477724.

[75] Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–90 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24463508.

[76] Grozeva, D. *et al.* Targeted next-generation sequencing analysis of 1,000 individuals with intellectual disability. *Hum Mutat* **36**, 1197–204 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26350204.

[77] O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–22 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/23160955.

[78] D'Alessandro, L. C. *et al.* Exome sequencing identifies rare variants in multiple genes in atrioventricular septal defect. *Genet Med* **18**, 189–98 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/25996639.

[79] Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**, 1095–106 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/23138309.

[80] Visscher, P. M. *et al.* 10 years of gwas discovery: Biology, function, and translation. *Am J Hum Genet* **101**, 5–22 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28686856.

[81] Grarup, N., Sandholt, C. H., Hansen, T. & Pedersen, O. Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia* **57**, 1528–41 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24859358.

[82] Massey, J. & Eyre, S. Rare variants and autoimmune disease. *Brief Funct Genomics* **13**, 392–7 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24817515.

[83] Jiang, T., Tan, M. S., Tan, L. & Yu, J. T. Application of next-generation sequencing technologies in neurology. *Ann Transl*

Med **2**, 125 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25568878.

[84] Cruchaga, C. *et al.* Rare coding variants in the phospholipase d3 gene confer risk for alzheimer's disease. *Nature* **505**, 550–4 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24336208.

[85] Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur J Hum Genet* **21**, 1158–62 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23386037.

[86] He, Z. *et al.* Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am J Hum Genet* **94**, 33–46 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24360806.

[87] Chen, H., Meigs, J. B. & Dupuis, J. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* **37**, 196–204 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23280576.

[88] Schifano, E. D. *et al.* Snp set association analysis for familial data. *Genet Epidemiol* **36**, 797–810 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22968922.

[89] Wang, X., Lee, S., Zhu, X., Redline, S. & Lin, X. Gee-based snp set association test for continuous and discrete traits in family-based association studies. *Genet Epidemiol* **37**, 778–86 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/24166731.

[90] He, X. *et al.* Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* **9**, e1003671 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23966865.

[91] Zhou, B. *et al.* Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *Journal of Medical*

Genetics **55**, 735–743 (2018). URL https://doi.org/10.1136/jmedgenet-2018-105272.

[92] Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20164920.

[93] Bochukova, E. G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666–70 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/19966786.

[94] Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–6 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18668039.

[95] Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–72 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20531469.

[96] Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**, 677–681 (2009). URL https://doi.org/10.1038/nmeth.1363.

[97] Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2015). URL https://doi.org/10.1093/bioinformatics/btv710.

[98] Roller, E., Ivakhno, S., Lee, S., Royce, T. & Tanner, S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics* **32**, 2375–2377 (2016). URL https://doi.org/10.1093/bioinformatics/btw163.

[99] Plagnol, V. *et al.* A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28**, 2747–54 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22942019.

[100] Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J*

Hum Genet **91**, 597–607 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/23040492.

[101] Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res* **22**, 1525–32 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22585873.

[102] Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome Res* **21**, 974–84 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21324876.

[103] Tan, R. *et al.* An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat* **35**, 899–907 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24599517.

[104] Miyatake, S. *et al.* Detecting copy-number variations in whole-exome sequencing data using the exome hidden markov model: an 'exome-first' approach. *J Hum Genet* **60**, 175–82 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25608832.

[105] Yamamoto, T. *et al.* Challenges in detecting genomic copy number aberrations using next-generation sequencing data and the exome hidden markov model: a clinical exome-first diagnostic approach. *Hum Genome Var* **3**, 16025 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27579173.

[106] Spataro, N. *et al.* Detection of genomic rearrangements from targeted resequencing data in parkinson's disease patients. *Mov Disord* **32**, 165–169 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28124432.

[107] Fromer, M. & Purcell, S. M. Using xhmm software to detect copy number variation in whole-exome sequencing data. *Curr Protoc Hum Genet* **81**, 7 23 1–21 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24763994.

[108] Amiel, J. L. & Schneider, M. [immunologic studies in hodgkin's disease]. *Bull Cancer* **58**, 9–20 (1971). URL http://www.ncbi.nlm.nih.gov/pubmed/5564288.

[109] Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. Hla variation and disease. *Nat Rev Immunol* **18**, 325–339 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29292391.

[110] Busch, R., Kollnberger, S. & Mellins, E. D. HLA associations in inflammatory arthritis: emerging mechanisms and clinical implications. *Nature Reviews Rheumatology* **15**, 364–381 (2019). URL https://doi.org/10.1038/s41584-019-0219-5.

[111] Major, E., Rigo, K., Hague, T., Berces, A. & Juhos, S. Hla typing from 1000 genomes whole genome and whole exome illumina data. *PLoS One* **8**, e78410 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/24223151.

[112] Dilthey, A. T. *et al.* High-accuracy hla type inference from whole-genome sequencing data using population reference graphs. *PLoS Comput Biol* **12**, e1005151 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27792722.

[113] Sicherer, S. H. & Sampson, H. A. Food allergy: Epidemiology, pathogenesis, diagnosis, and treatment. *J Allergy Clin Immunol* **133**, 291–307; quiz 308 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24388012.

[114] Eigenmann, P. A. *et al.* New visions for food allergy: an ipac summary and future trends. *Pediatr Allergy Immunol* **19 Suppl 19**, 26–39 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18665961.

[115] Turnbull, J. L., Adams, H. N. & Gorard, D. A. Review article: the diagnosis and management of food allergy and food intolerances. *Aliment Pharmacol Ther* **41**, 3–25 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25316115.

[116] Carrard, A., Rizzuti, D. & Sokollik, C. Update on food allergy. *Allergy* **70**, 1511–20 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26443043.

[117] Boyce, J. A. *et al.* Guidelines for the diagnosis and management of food allergy in the united states: Summary of the niaid-sponsored expert panel report. *J Allergy Clin Immunol* **126**, 1105–18 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/21134568.

[118] Bischoff, S. & Crowe, S. E. Gastrointestinal food allergy: New insights into pathophysiology and clinical perspectives. *Gastroenterology* **128**, 1089–1113 (2005).

[119] Sampson, H. A. Update on food allergy. *J Allergy Clin Immunol* **113**, 805–19; quiz 820 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15131561.

[120] Crowe, S. E. & Perdue, M. H. Gastrointestinal food hypersensitivity: basic mechanisms of pathophysiology. *Gastroenterology* **103**, 1075–95 (1992). URL http://www.ncbi.nlm.nih.gov/pubmed/1499910.

[121] Untersmayr, E. *et al.* Antacid medication inhibits digestion of dietary proteins and causes food allergy: a fish allergy model in balb/c mice. *J Allergy Clin Immunol* **112**, 616–23 (2003). URL http://www.ncbi.nlm.nih.gov/pubmed/13679824.

[122] Troncone, R., Caputo, N., Florio, G. & Finelli, E. Increased intestinal sugar permeability after challenge in children with cow's milk allergy or intolerance. *Allergy* **49**, 142–6 (1994). URL http://www.ncbi.nlm.nih.gov/pubmed/8198245.

[123] Brandtzaeg, P. E. Current understanding of gastrointestinal immunoregulation and its relation to food allergy. *Ann N Y Acad Sci* **964**, 13–45 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/12023193.

[124] Morita, H., Nomura, I., Matsuda, A., Saito, H. & Matsumoto, K. Gastrointestinal food allergy in infants. *Allergol Int* **62**, 297–307 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23974876.

[125] Sathe, S. K., Liu, C. & Zaffran, V. D. Food allergy. *Annu Rev Food Sci Technol* **7**, 191–220 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26934173.

[126] Eigenmann, P. A. & Frossard, C. P. The t lymphocyte in food-allergy disorders. *Curr Opin Allergy Clin Immunol* **3**, 199–203 (2003). URL http://www.ncbi.nlm.nih.gov/pubmed/12840703.

[127] Katz, Y., Goldberg, M. R., Rajuan, N., Cohen, A. & Leshno, M. The prevalence and natural course of food protein-induced enterocolitis syndrome to cow's milk: a large-scale, prospective population-based study. *J Allergy Clin Immunol* **127**, 647–53 e1–3 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21377033.

[128] Sopo, S. M. *et al.* A multicentre retrospective study of 66 italian children with food protein-induced enterocolitis syndrome: different management for different phenotypes. *Clin Exp Allergy* **42**, 1257–65 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22805473.

[129] Nowak-Wegrzyn, A., Katz, Y., Mehr, S. S. & Koletzko, S. Non-ige-mediated gastrointestinal food allergy. *J Allergy Clin Immunol* **135**, 1114–24 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25956013.

[130] Bischoff, S. C., Mayer, J. H. & Manns, M. P. Allergy and the gut. *Int Arch Allergy Immunol* **121**, 270–83 (2000). URL http://www.ncbi.nlm.nih.gov/pubmed/10828717.

[131] Valenta, R., Vrtala, S., Ebner, C., Kraft, D. & Scheiner, O. Diagnosis of grass pollen allergy with recombinant timothy grass (phleum pratense) pollen allergens. *Int Arch Allergy Immunol* **97**, 287–94 (1992). URL http://www.ncbi.nlm.nih.gov/pubmed/1597349.

[132] Rabjohn, P. *et al.* Modification of peanut allergen ara h 3: effects on ige binding and t cell stimulation. *Int Arch Allergy Immunol* **128**, 15–23 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/12037397.

[133] American gastroenterological association medical position statement: Guidelines for the evaluation of food allergies. *Gastroenterology* **120**, 1023–1025 (2001). URL https://doi.org/10.1053/gast.2001.23417.

[134] Bischoff, S. C. *et al.* Colonoscopic allergen provocation (colap): a new diagnostic approach for gastrointestinal food allergy. *Gut* **40**, 745–53 (1997). URL http://www.ncbi.nlm.nih.gov/pubmed/9245928.

[135] Nomura, I. *et al.* Four distinct subtypes of non-ige-mediated gastrointestinal food allergies in neonates and infants, distinguished by their initial symptoms. *J Allergy Clin Immunol* **127**, 685–8 e1–8 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21377037.

[136] Chen, M. & Land, M. The current state of food allergy therapeutics. *Hum Vaccin Immunother* 0 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28846472.

[137] Wood, R. A. Food allergen immunotherapy: Current status and prospects for the future. *J Allergy Clin Immunol* **137**, 973–982 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27059725.

[138] Vickery, B. P. *et al.* Sustained unresponsiveness to peanut in subjects who have completed peanut oral immunotherapy. *J Allergy Clin Immunol* **133**, 468–75 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24361082.

[139] Anagnostou, K. *et al.* Assessing the efficacy of oral immunotherapy for the desensitisation of peanut allergy in children (stop ii): a phase 2 randomised controlled trial. *Lancet* **383**, 1297–304 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24485709.

[140] Blumchen, K. *et al.* Oral peanut immunotherapy in children with peanut anaphylaxis. *J Allergy Clin Immunol* **126**, 83–91 e1 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20542324.

[141] Burks, A. W. *et al.* Oral immunotherapy for treatment of egg allergy in children. *N Engl J Med* **367**, 233–43 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22808958.

[142] Jones, S. M. *et al.* Long-term treatment with egg oral immunotherapy enhances sustained unresponsiveness that persists after cessation of therapy. *J Allergy Clin Immunol* **137**, 1117–1127 e10 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26924470.

[143] Skripak, J. M. *et al.* A randomized, double-blind, placebo-controlled study of milk oral immunotherapy for cow's milk allergy. *J Allergy Clin Immunol* **122**, 1154–60 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18951617.

[144] Kim, E. H. *et al.* Sublingual immunotherapy for peanut allergy: clinical and immunologic evidence of desensitization. *J Allergy Clin Immunol* **127**, 640–6 e1 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21281959.

[145] Fleischer, D. M. *et al.* Sublingual immunotherapy for peanut allergy: a randomized, double-blind, placebo-controlled multicenter trial. *J Allergy Clin Immunol* **131**, 119–27 e1–7 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23265698.

[146] Enrique, E. *et al.* Sublingual immunotherapy for hazelnut food allergy: a randomized, double-blind, placebo-controlled study with a standardized hazelnut extract. *J Allergy Clin Immunol* **116**, 1073–9 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/16275379.

[147] Fernandez-Rivas, M. *et al.* Randomized double-blind, placebo-controlled trial of sublingual immunotherapy with a pru p 3 quantified peach extract. *Allergy* **64**, 876–83 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19183164.

[148] Jones, S. M. *et al.* Epicutaneous immunotherapy for the treatment of peanut allergy in children and young adults. *J Allergy Clin Immunol* **139**, 1242–1252 e9 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28091362.

[149] Jones, S. M. *et al.* Safety of epicutaneous immunotherapy for the treatment of peanut allergy: A phase 1 study using the viaskin patch. *J Allergy Clin Immunol* **137**, 1258–1261 e10 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26920463.

[150] Dupont, C. *et al.* Cow's milk epicutaneous immunotherapy in children: a pilot trial of safety, acceptability, and impact on allergic reactivity. *J Allergy Clin Immunol* **125**, 1165–7 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20451043.

[151] Burks, A. W., King, N. & Bannon, G. A. Modification of a major peanut allergen leads to loss of ige binding. *Int Arch Allergy Immunol* **118**, 313–4 (1999). URL http://www.ncbi.nlm.nih.gov/pubmed/10224425.

[152] MacGlashan, J., D. W. *et al.* Down-regulation of fc(epsilon)ri expression on human basophils during in vivo treatment of atopic patients with anti-ige antibody. *J Immunol* **158**, 1438–45 (1997). URL http://www.ncbi.nlm.nih.gov/pubmed/9013989.

[153] Leung, D. Y. *et al.* Effect of anti-ige therapy in patients with peanut allergy. *N Engl J Med* **348**, 986–93 (2003). URL http://www.ncbi.nlm.nih.gov/pubmed/12637608.

[154] Wenzel, S. *et al.* Dupilumab in persistent asthma with elevated eosinophil levels. *N Engl J Med* **368**, 2455–66 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23688323.

[155] Beck, L. A. *et al.* Dupilumab treatment in adults with moderate-to-severe atopic dermatitis. *N Engl J Med* **371**, 130–9 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25006719.

[156] Wechsler, J. B. & Hirano, I. Biological therapies for eosinophilic gastrointestinal diseases. *J Allergy Clin Immunol* **142**, 24–31 e2 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29859203.

[157] Hall, I. P. *et al.* Efficacy of bi 671800, an oral crth2 antagonist, in poorly controlled asthma as sole controller and in the presence of inhaled corticosteroid treatment. *Pulm Pharmacol Ther* **32**, 37–44 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25861737.

[158] Krug, N. *et al.* Efficacy of the oral chemoattractant receptor homologous molecule on th2 cells antagonist bi 671800 in patients with seasonal allergic rhinitis. *J Allergy Clin Immunol* **133**, 414–9 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24332218.

[159] Holbrook, T., Keet, C. A., Frischmeyer-Guerrerio, P. A. & Wood, R. A. Use of ondansetron for food protein-induced enterocolitis syndrome. *J Allergy Clin Immunol* **132**, 1219–20 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23890754.

[160] Miceli Sopo, S., Battista, A., Greco, M. & Monaco, S. Ondansetron for food protein-induced enterocolitis syndrome. *Int Arch Allergy Immunol* **164**, 137–9 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24993542.

[161] Miceli Sopo, S. *et al.* Ondansetron in acute food protein-induced enterocolitis syndrome, a retrospective case-control study. *Allergy* **72**, 545–551 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/27548842.

[162] Ding, W. *et al.* Interleukin-33: Its emerging role in allergic diseases. *Molecules* **23** (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29987222.

[163] Chu, D. K. *et al.* Il-33, but not thymic stromal lymphopoietin or il-25, is central to mite and peanut allergic sensitization. *J Allergy Clin Immunol* **131**, 187–200 e1–8 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23006545.

[164] Gao, C., Holscher, C., Liu, Y. & Li, L. Gsk3: a key target for the development of novel treatments for type 2 diabetes mellitus and alzheimer disease. *Rev Neurosci* **23**, 1–11 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/22718609.

[165] Jope, R. S., Yuskaitis, C. J. & Beurel, E. Glycogen synthase kinase-3 (gsk3): inflammation, diseases, and therapeutics. *Neurochem Res* **32**, 577–95 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/16944320.

[166] Pineton de Chambrun, G. *et al.* Diagnosis, natural history and treatment of eosinophilic enteritis: a review. *Curr Gastroenterol Rep* **20**, 37 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29968127.

[167] Grela, F. *et al.* The tlr7 agonist r848 alleviates allergic inflammation by targeting invariant nkt cells to produce ifn-gamma. *J Immunol* **186**, 284–90 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21131420.

[168] Creticos, P. S. *et al.* Immunotherapy with a ragweed-toll-like receptor 9 agonist vaccine for allergic rhinitis. *N Engl J Med* **355**, 1445–55 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/17021320.

[169] Patel, P., Holdich, T., Fischer von Weikersthal-Drachenberg, K. J. & Huber, B. Efficacy of a short course of specific immunotherapy in patients with allergic rhinoconjunctivitis to ragweed pollen. *J Allergy Clin Immunol* **133**, 121–9 e1–2 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/23870670.

[170] Helm, R. M. & Burks, A. W. Animal models of food allergy. *Curr Opin Allergy Clin Immunol* **2**, 541–6 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/14752339.

[171] Hogan, S. P. *et al.* A pathological function for eotaxin and eosinophils in eosinophilic gastrointestinal inflammation. *Nat Immunol* **2**, 353–60 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11276207.

[172] Rautava, S., Kalliomaki, M. & Isolauri, E. Probiotics during pregnancy and breast-feeding might confer immunomodulatory protection against atopic disease in the infant. *J Allergy Clin Immunol* **109**, 119–21 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/11799376.

[173] Helin, T., Haahtela, S. & Haahtela, T. No effect of oral treatment with an intestinal bacterial strain, lactobacillus rhamnosus (atcc 53103), on birch-pollen allergy: a placebo-controlled double-blind study. *Allergy* **57**, 243–6 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/11906339.

[174] Hong, X. *et al.* Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in us children. *Nat Commun* **6**, 6304 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25710614.

[175] Li, J. *et al.* Copy number variations in ctnna3 and rbfox1 associate with pediatric food allergy. *J Immunol* **195**, 1599–607 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26188062.

[176] DeWan, A. T. *et al.* Whole-exome sequencing of a pedigree segregating asthma. *BMC Med Genet* **13**, 95 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/23046476.

[177] Torgerson, D. G. *et al.* Resequencing candidate genes implicates rare variants in asthma susceptibility. *Am J Hum Genet* **90**, 273–81 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22325360.

[178] Peters, R. L., Neeland, M. R. & Allen, K. J. Primary prevention of food allergy. *Curr Allergy Asthma Rep* **17**, 52 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28634899.

[179] Garcia-Compean, D., Gonzalez-Gonzalez, J. A., Gonzalez-Moreno, E. I. & Maldonado-Garza, H. J. Eosinophilic esophagitis. the north against the south? a bio-economic-social mechanistic approach and clinical implications. *Rev Gastroenterol Mex* **82**, 328–336 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28676192.

[180] Jensen, E. T. *et al.* Early-life environmental exposures interact with genetic susceptibility variants in pediatric patients with eosinophilic esophagitis. *J Allergy Clin Immunol* **141**, 632–637 e5 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29029802.

[181] Jensen, E. T., Kuhl, J. T., Martin, L. J., Rothenberg, M. E. & Dellon, E. S. Prenatal, intrapartum, and postnatal factors are associated with pediatric eosinophilic esophagitis. *J Allergy Clin Immunol* **141**, 214–222 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/28601683.

[182] Khor, B., Gardet, A. & Xavier, R. J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474**, 307–17 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21677747.

[183] Rachid, R. & Chatila, T. A. The role of the gut microbiota in food allergy. *Curr Opin Pediatr* **28**, 748–753 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27749359.

[184] Kirjavainen, P. V. *et al.* Characterizing the composition of intestinal microflora as a prospective treatment target in infant allergic

disease. *FEMS Immunol Med Microbiol* **32**, 1–7 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11750215.

[185] Martin, P. E. *et al.* Which infants with eczema are at risk of food allergy? results from a population-based cohort. *Clin Exp Allergy* **45**, 255–64 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25210971.

[186] Gupta, M. & Sicherer, S. H. Timing of food introduction and atopy prevention. *Clin Dermatol* **35**, 398–405 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28709571.

[187] Lack, G. Update on risk factors for food allergy. *J Allergy Clin Immunol* **129**, 1187–97 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22464642.

[188] Liu, A. H. *et al.* National prevalence and risk factors for food allergy and relationship to asthma: results from the national health and nutrition examination survey 2005-2006. *J Allergy Clin Immunol* **126**, 798–806 e13 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20920770.

[189] Chen, W., Mempel, M., Schober, W., Behrendt, H. & Ring, J. Gender difference, sex hormones, and immediate type hypersensitivity reactions. *Allergy* **63**, 1418–27 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18925878.

[190] Benede, S., Blazquez, A. B., Chiang, D., Tordesillas, L. & Berin, M. C. The rise of food allergy: Environmental factors and emerging treatments. *EBioMedicine* **7**, 27–34 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27322456.

[191] Roduit, C. *et al.* Increased food diversity in the first year of life is inversely associated with allergic diseases. *J Allergy Clin Immunol* **133**, 1056–64 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24508301.

[192] Tan, T. H., Ellis, J. A., Saffery, R. & Allen, K. J. The role of genetics and environment in the rise of childhood food allergy. *Clin Exp Allergy* **42**, 20–9 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/21771119.

[193] Shahali, Y. & Dadar, M. Plant food allergy: Influence of chemicals on plant allergens. *Food and Chemical Toxicology* **115**, 365–374 (2018). URL https://doi.org/10.1016/j.fct.2018.03.032.

[194] Melen, E. *et al.* Interactions between glutathione s-transferase p1, tumor necrosis factor, and traffic-related air pollution for development of childhood allergic disease. *Environ Health Perspect* **116**, 1077–84 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18709160.

[195] Gern, J. E. *et al.* Effects of dog ownership and genotype on immune development and atopy in infancy. *J Allergy Clin Immunol* **113**, 307–14 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/14767447.

[196] Leynaert, B. *et al.* Association between farm exposure and atopy, according to the cd14 c-159t polymorphism. *J Allergy Clin Immunol* **118**, 658–65 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16950285.

[197] Simpson, A. *et al.* Endotoxin exposure, cd14, and allergic disease: an interaction between genes and the environment. *Am J Respir Crit Care Med* **174**, 386–92 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16614348.

[198] Kabesch, M. *et al.* Il-4/il-13 pathway genetics strongly influence serum ige levels and childhood asthma. *J Allergy Clin Immunol* **117**, 269–74 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16461126.

[199] Syed, A. *et al.* Peanut oral immunotherapy results in increased antigen-induced regulatory t-cell function and hypomethylation of forkhead box protein 3 (foxp3). *J Allergy Clin Immunol* **133**, 500–10 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24636474.

[200] Martino, D. *et al.* Epigenome-wide association study reveals longitudinally stable dna methylation differences in cd4+ t cells from children with ige-mediated food allergy. *Epigenetics* **9**, 998–1006 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24762976.

[201] Martino, D. *et al.* Blood dna methylation biomarkers predict clinical reactivity in food-sensitized infants. *J Allergy Clin Immunol* **135**, 1319–28 e1–12 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25678091.

[202] McWilliam, V., Heine, R., Tang, M. L. & Allen, K. J. Multiple food protein intolerance of infancy or severe spectrum of non-ige-mediated cow's milk allergy?–a case series. *J Allergy Clin Immunol Pract* **4**, 324–6 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26755099.

[203] Chen, R., Im, H. & Snyder, M. Whole-exome enrichment with the agilent sureselect human all exon platform. *Cold Spring Harb Protoc* **2015**, 626–33 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25762417.

[204] Caruccio, N. Preparation of next-generation sequencing libraries using nextera technology: simultaneous dna fragmentation and adaptor tagging by in vitro transposition. *Methods Mol Biol* **733**, 241–55 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21431775.

[205] Chen, F. *et al.* The history and advances of reversible terminators used in new generations of sequencing technology. *Genomics Proteomics Bioinformatics* **11**, 34–40 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23414612.

[206] Van der Auwera, G. A. *et al.* From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 1–33 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/25431634.

[207] Li, H. & Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**, 1754–60 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19451168.

[208] Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and snp calling from next-generation sequencing data. *Nat Rev Genet* **12**, 443–51 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21587300.

[209] Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports* **8** (2018).

[210] McLaren, W. *et al.* Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. *Bioinformatics* **26**, 2069–70 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20562413.

[211] Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47**, D886–D894 (2018). URL https://doi.org/10.1093/nar/gky1016.

[212] Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nature Genetics* **48**, 1107–1111 (2016). URL https://doi.org/10.1038%2Fng.3638.

[213] Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R. & McVean, G. Improved genome inference in the mhc using a population reference graph. *Nat Genet* **47**, 682–8 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25915597.

[214] Li, L. H. *et al.* Long contiguous stretches of homozygosity in the human genome. *Hum Mutat* **27**, 1115–21 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16955415.

[215] Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–73 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20926424.

[216] Hinds, D. A. *et al.* Whole-genome patterns of common dna variation in three human populations. *Science* **307**, 1072–9 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15718463.

[217] Romanel, A., Zhang, T., Elemento, O. & Demichelis, F. Ethseq: ethnicity annotation from whole exome sequencing data. *Bioinformatics* **33**, 2402–2404 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28369222.

[218] Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. Gemini: integrative exploration of genetic variation and genome annotations. *PLoS Comput Biol* **9**, e1003153 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23874191.

[219] Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011). URL https://doi.org/10.1038/nbt.1754.

[220] Fan, Y. & Song, Y. Q. Pyhla: tests for the association between hla alleles and diseases. *BMC Bioinformatics* **18**, 90 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28166716.

[221] Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Research* **8**, 175–185 (1998). URL https://www.ncbi.nlm.nih.gov/pubmed/9521921.

[222] Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Research* **8**, 186–194 (1998). URL https://www.ncbi.nlm.nih.gov/pubmed/9521922.

[223] Kong, S. W., Lee, I. H., Liu, X., Hirschhorn, J. N. & Mandl, K. D. Measuring coverage and accuracy of whole-exome sequencing in clinical context. *Genet Med* (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29789557.

[224] Ku, C. S. *et al.* Exome sequencing: dual role as a discovery and diagnostic tool. *Ann Neurol* **71**, 5–14 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22275248.

[225] Do, R., Kathiresan, S. & Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet* **21**, R1–9 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22983955.

[226] Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* **20**, 490–7 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22258526.

[227] Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* **112**, 5473–8 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25827230.

[228] Huss, W. J. *et al.* Comparison of SureSelect and nextera exome capture performance in single-cell sequencing. *Human Heredity* **83**, 153–162 (2018). URL https://doi.org/10.1159/000490506.

[229] Shigemizu, D. *et al.* Performance comparison of four commercial human whole-exome capture platforms. *Scientific Reports* **5** (2015). URL https://doi.org/10.1038/srep12742.

[230] Guo, Y., Ye, F., Sheng, Q., Clark, T. & Samuels, D. C. Three-stage quality control strategies for dna re-sequencing data. *Brief Bioinform* **15**, 879–89 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24067931.

[231] Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–91 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27535533.

[232] Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **44**, D481–7 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26656494.

[233] Sharif-Askari, E. *et al.* Zinc finger protein gfi1 controls the endotoxin-mediated toll-like receptor inflammatory response by antagonizing nf-kappab p65. *Mol Cell Biol* **30**, 3929–42 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20547752.

[234] Allen, I. C. *et al.* Nlrp12 suppresses colon inflammation and tumorigenesis through the negative regulation of noncanonical nf-kappab signaling. *Immunity* **36**, 742–54 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22503542.

[235] Franz, S. *et al.* Mumps virus sh protein inhibits nf-kappab activation by interacting with tumor necrosis factor receptor 1, interleukin-1 receptor 1, and toll-like receptor 3 complexes. *J Virol* **91** (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28659487.

[236] Greco, E. *et al.* The novel s59p mutation in the tnfrsf1a gene identified in an adult onset tnf receptor associated periodic syndrome (traps) constitutively activates nf-kappab pathway. *Arthritis Res Ther* **17**, 93 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25888769.

[237] van Haaften-Visser, D. Y. *et al.* Ankyrin repeat and zinc-finger domain-containing 1 mutations are associated with infantile-onset inflammatory bowel disease. *J Biol Chem* **292**, 7904–7920 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28302725.

[238] Ashton, J. J., Ennis, S. & Beattie, R. M. Early-onset paediatric inflammatory bowel disease. *Lancet Child Adolesc Health* **1**, 147–158 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/30169204.

[239] Borte, S. *et al.* Novel nlrp12 mutations associated with intestinal amyloidosis in a patient diagnosed with common variable immunodeficiency. *Clin Immunol* **154**, 105–11 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25064839.

[240] Firth, H. V. *et al.* DECIPHER: Database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics* **84**, 524–533 (2009). URL https://doi.org/10.1016/j.ajhg.2009.03.010.

[241] Lukens, J. R. *et al.* The nlrp12 sensor negatively regulates autoinflammatory disease by modulating interleukin-4 production in t cells. *Immunity* **42**, 654–64 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25888258.

[242] Xia, X. *et al.* Identification of a novel nlrp12 nonsense mutation (trp408x) in the extremely rare disease fcas by exome sequencing. *PLoS One* **11**, e0156981 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27314497.

[243] Vitale, A. *et al.* Rare nlrp12 variants associated with the nlrp12-autoinflammatory disorder phenotype: an italian case series. *Clin Exp Rheumatol* **31**, 155–6 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/24064030.

[244] Gandhi, N. A., Pirozzi, G. & Graham, N. M. H. Commonality of the il-4/il-13 pathway in atopic diseases. *Expert Rev Clin Immunol* **13**, 425–437 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28277826.

[245] Ashley, S. E. *et al.* Genetic variation at the th2 immune gene il13 is associated with ige-mediated paediatric food allergy. *Clin Exp Allergy* **47**, 1032–1037 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28544327.

[246] Sherrill, J. D. *et al.* Whole-exome sequencing uncovers oxidoreductases dhtkd1 and ogdhl as linkers between mitochondrial dysfunction and eosinophilic esophagitis. *JCI Insight* **3** (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29669943.

[247] Tu, M. *et al.* Il-13 receptor alpha2 stimulates human glioma cell growth and metastasis through the src/pi3k/akt/mtor signaling pathway. *Tumour Biol* **37**, 14701–14709 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27623944.

[248] Liu, Y. Q. *et al.* Human ring finger protein znf645 is a novel testis-specific e3 ubiquitin ligase. *Asian J Androl* **12**, 658–66 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20657603.

[249] Li, J. *et al.* Laminin-10 is crucial for hair morphogenesis. *EMBO J* **22**, 2400–10 (2003). URL http://www.ncbi.nlm.nih.gov/pubmed/12743034.

[250] Wegner, J. *et al.* Laminin alpha5 in the keratinocyte basement membrane is required for epidermal-dermal intercommunication. *Matrix Biol* **56**, 24–41 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27234307.

[251] Sampaolo, S. *et al.* Identification of the first dominant mutation of lama5 gene causing a complex multisystem syndrome due to dysfunction of the extracellular matrix. *J Med Genet* **54**, 710–720 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28735299.

[252] Kaji, T. *et al.* Ask3, a novel member of the apoptosis signal-regulating kinase family, is essential for stress-induced cell death

in hela cells. *Biochem Biophys Res Commun* **395**, 213–8 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20362554.

[253] Tartey, S., Gurung, P., Dasari, T. K., Burton, A. & Kanneganti, T. D. Ask1/2 signaling promotes inflammation in a mouse model of neutrophilic dermatosis. *J Clin Invest* **128**, 2042–2047 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29629899.

[254] Ben-Horin, S., Kopylov, U. & Chowers, Y. Optimizing anti-tnf treatments in inflammatory bowel disease. *Autoimmun Rev* **13**, 24–30 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/23792214.

[255] St-Pierre, J. & Chadee, K. How the discovery of tnf-alpha has advanced gastrointestinal diseases and treatment regimes. *Dig Dis Sci* **59**, 712–5 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24504593.

[256] Vassalli, P. The pathophysiology of tumor necrosis factors. *Annu Rev Immunol* **10**, 411–52 (1992). URL http://www.ncbi.nlm.nih.gov/pubmed/1590993.

[257] Stojanov, S. & McDermott, M. F. The tumour necrosis factor receptor-associated periodic syndrome: current concepts. *Expert Rev Mol Med* **7**, 1–18 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/16216134.

[258] Fabre, A. *et al.* Skiv2l mutations cause syndromic diarrhea, or trichohepatoenteric syndrome. *Am J Hum Genet* **90**, 689–92 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22444670.

[259] Morgan, N. V. *et al.* A combination of mutations in akr1d1 and skiv2l in a family with severe infantile liver disease. *Orphanet J Rare Dis* **8**, 74 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23679950.

[260] van den Heuvel, A. P. J. Binding of protein kinase b to the plakin family member periplakin. *Journal of Cell Science* **115**, 3957–3966 (2002). URL https://doi.org/10.1242/jcs.00069.

[261] Thomson, P. A. *et al.* Sex-specific association between bipolar affective disorder in women and gpr50, an x-linked orphan g protein-coupled receptor. *Mol Psychiatry* **10**, 470–8 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15452587.

[262] Ryan, J., Carriere, I., Ritchie, K. & Ancelin, M. L. Involvement of gpr50 polymorphisms in depression: independent replication in a prospective elderly cohort. *Brain Behav* **5**, e00313 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25798330.

[263] Chaste, P. *et al.* Genetic variations of the melatonin pathway in patients with attention-deficit and hyperactivity disorders. *J Pineal Res* **51**, 394–9 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21615493.

[264] *Melatonin* (Bethesda (MD), 2006). URL http://www.ncbi.nlm.nih.gov/pubmed/30000923.

[265] Acuna-Castroviejo, D. *et al.* Extrapineal melatonin: sources, regulation, and potential functions. *Cell Mol Life Sci* **71**, 2997–3025 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24554058.

[266] Majka, J. *et al.* Melatonin in prevention of the sequence from reflux esophagitis to barrett's esophagus and esophageal adenocarcinoma: Experimental and clinical perspectives. *Int J Mol Sci* **19** (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/30011784.

[267] Gao, J. *et al.* Impact of the gut microbiota on intestinal immunity mediated by tryptophan metabolism. *Front Cell Infect Microbiol* **8**, 13 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29468141.

[268] Heine, R. G. Gastrointestinal food allergies. *Chem Immunol Allergy* **101**, 171–80 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26022877.

[269] Kondo, M. *et al.* Satb1 plays a critical role in establishment of immune tolerance. *J Immunol* **196**, 563–72 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26667169.

[270] Suzuki, J. *et al.* Gfi1, a transcriptional repressor, inhibits the induction of the t helper type 1 programme in activated cd4 t cells. *Immunology* **147**, 476–87 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26749286.

[271] Noval Rivas, M. *et al.* Regulatory t cell reprogramming toward a th2-cell-like lineage impairs oral tolerance and promotes food allergy. *Immunity* **42**, 512–23 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25769611.

[272] Lafon, A. *et al.* Ino80 chromatin remodeler facilitates release of rna polymerase ii from chromatin for ubiquitin-mediated proteasomal degradation. *Mol Cell* **60**, 784–796 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26656161.

[273] Kracker, S. *et al.* An inherited immunoglobulin class-switch recombination deficiency associated with a defect in the ino80 chromatin remodeling complex. *J Allergy Clin Immunol* **135**, 998–1007 e6 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25312759.

[274] Dear, T. N., Meier, N. T., Hunn, M. & Boehm, T. Gene structure, chromosomal localization, and expression pattern of capn12, a new member of the calpain large subunit gene family. *Genomics* **68**, 152–60 (2000). URL http://www.ncbi.nlm.nih.gov/pubmed/10964513.

[275] Litosh, V. A. *et al.* Calpain-14 and its association with eosinophilic esophagitis. *J Allergy Clin Immunol* **139**, 1762–1771 e7 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28131390.

[276] Davis, B. P. *et al.* Eosinophilic esophagitis-linked calpain 14 is an il-13-induced protease that mediates esophageal epithelial barrier impairment. *JCI Insight* **1**, e86355 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27158675.

[277] Madore, A. M. *et al.* Hla-dqb1*02 and dqb1*06:03p are associated with peanut allergy. *Eur J Hum Genet* **21**, 1181–4 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23443026.

[278] Boehncke, W. H. *et al.* Identification of hla-dr and -dq alleles conferring susceptibility to pollen allergy and pollen associated food allergy. *Clin Exp Allergy* **28**, 434–41 (1998). URL http://www.ncbi.nlm.nih.gov/pubmed/9641569.

[279] Torgerson, T. R. *et al.* Severe food allergy as a variant of IPEX syndrome caused by a deletion in a noncoding region of the FOXP3 gene. *Gastroenterology* **132**, 1705–1717 (2007). URL https://doi.org/10.1053/j.gastro.2007.02.044.

[280] Asai, Y. *et al.* Genome-wide association study and meta-analysis in multiple populations identifies new loci for peanut allergy and establishes c11orf30/EMSY as a genetic risk factor for food allergy. *Journal of Allergy and Clinical Immunology* **141**, 991–1001 (2018). URL https://doi.org/10.1016/j.jaci.2017.09.015.

[281] Griffin, H. R. *et al.* Accurate mitochondrial DNA sequencing using off-target reads provides a single test to identify pathogenic point mutations. *Genetics in Medicine* **16**, 962–971 (2014). URL https://doi.org/10.1038/gim.2014.66.

[282] Ellingford, J. M. *et al.* Whole genome sequencing increases molecular diagnostic yield compared with current diagnostic testing for inherited retinal disease. *Ophthalmology* **123**, 1143–1150 (2016). URL https://doi.org/10.1016/j.ophtha.2016.01.009.

[283] Chaitidis, P. *et al.* Gene expression alterations of human peripheral blood monocytes induced by medium-term treatment with the th2-cytokines interleukin-4 and -13. *Cytokine* **30**, 366–77 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15869882.

[284] Yakubenko, V. P., Bhattacharjee, A., Pluskota, E. & Cathcart, M. K. alphambeta(2) integrin activation prevents alternative activation of human and murine macrophages and impedes foam cell formation. *Circ Res* **108**, 544–54 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21252155.

[285] Chung, K. F. Targeting the interleukin pathway in the treatment of asthma. *Lancet* **386**, 1086–96 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26383000.

[286] Hirano, I. *et al.* RPC4046, a monoclonal antibody against IL13, reduces histologic and endoscopic activity in patients with eosinophilic esophagitis. *Gastroenterology* **156**, 592–603.e10 (2019). URL https://doi.org/10.1053/j.gastro.2018.10.051.

[287] Wechsler, J. B. & Hirano, I. Biological therapies for eosinophilic gastrointestinal diseases. *Journal of Allergy and Clinical Immunology* **142**, 24–31.e2 (2018). URL https://doi.org/10.1016/j.jaci.2018.05.018.

[288] Akdis, M. *et al.* Interleukins (from il-1 to il-38), interferons, transforming growth factor beta, and tnf-alpha: Receptors, functions, and roles in diseases. *J Allergy Clin Immunol* **138**, 984–1010 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27577879.

[289] Brown, S. D. *et al.* Characterization of a high tnf-alpha phenotype in children with moderate-to-severe asthma. *J Allergy Clin Immunol* **135**, 1651–4 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25725987.

[290] Zimmermann, M. *et al.* Tnf-like weak inducer of apoptosis (tweak) and tnf-alpha cooperate in the induction of keratinocyte apoptosis. *J Allergy Clin Immunol* **127**, 200–7, 207 e1–10 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21211655.

[291] Ather, J. L., Hodgkins, S. R., Janssen-Heininger, Y. M. & Poynter, M. E. Airway epithelial nf-kappab activation promotes allergic sensitization to an innocuous inhaled antigen. *Am J Respir Cell Mol Biol* **44**, 631–8 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/20581095.

[292] Poynter, M. E. *et al.* Nf-kappa b activation in airways modulates allergic inflammation but not hyperresponsiveness. *J Immunol* **173**, 7003–9 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15557197.

[293] Malinin, N. L., Boldin, M. P., Kovalenko, A. V. & Wallach, D. Map3k-related kinase involved in nf-kappab induction by tnf, cd95 and il-1. *Nature* **385**, 540–4 (1997). URL http://www.ncbi.nlm.nih.gov/pubmed/9020361.

[294] Veiga-Fernandes, H. & Pachnis, V. Neuroimmune regulation during intestinal development and homeostasis. *Nat Immunol* **18**, 116–122 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28092371.

[295] Shinnakasu, R. *et al.* Gfi1-mediated stabilization of gata3 protein is required for th2 cell differentiation. *J Biol Chem* **283**, 28216–25 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/18701459.

[296] Madsen, L. S. *et al.* A humanized model for multiple sclerosis using hla-dr2 and a human t-cell receptor. *Nat Genet* **23**, 343–7 (1999). URL http://www.ncbi.nlm.nih.gov/pubmed/10610182.

[297] Pugliese, A. *et al.* The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the ins vntr-iddm2 susceptibility locus for type 1 diabetes. *Nat Genet* **15**, 293–7 (1997). URL http://www.ncbi.nlm.nih.gov/pubmed/9054945.

[298] Molberg, O. *et al.* Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived t cells in celiac disease. *Nat Med* **4**, 713–7 (1998). URL http://www.ncbi.nlm.nih.gov/pubmed/9623982.

[299] Mosley, J. D. *et al.* Identifying genetically driven clinical phenotypes using linear mixed models. *Nat Commun* **7**, 11433 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27109359.

[300] Nowak-Wegrzyn, A. Food protein-induced enterocolitis syndrome and allergic proctocolitis. *Allergy Asthma Proc* **36**, 172–84 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25976434.

[301] Kelly, C. & Gangur, V. Sex disparity in food allergy: Evidence from the pubmed database. *J Allergy (Cairo)* **2009**, 159845 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/20975795.

[302] Afify, S. M. & Pali-Scholl, I. Adverse reactions to food: the female dominance - a secondary publication and update. *World Allergy Organ J* **10**, 43 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/29308110.

[303] Badano, J. L. & Katsanis, N. Beyond mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet* **3**, 779–89 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/12360236.

[304] Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219–1224 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/30104762.

[305] Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–24 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/23128233.

[306] Claussnitzer, M. *et al.* Fto obesity variant circuitry and adipocyte browning in humans. *N Engl J Med* **373**, 895–907 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26287746.

[307] Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25673413.

[308] Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28622505.

[309] Tak, P. P. & Firestein, G. S. Nf-kappab: a key role in inflammatory diseases. *J Clin Invest* **107**, 7–11 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11134171.

[310] Aupperle, K. R. *et al.* Nf-kappa b regulation by i kappa b kinase in primary fibroblast-like synoviocytes. *J Immunol* **163**, 427–33 (1999). URL http://www.ncbi.nlm.nih.gov/pubmed/10384145.

[311] Miagkov, A. V. *et al.* Nf-kappab activation provides the potential link between inflammation and hyperplasia in the arthritic joint. *Proc Natl Acad Sci U S A* **95**, 13859–64 (1998). URL http://www.ncbi.nlm.nih.gov/pubmed/9811891.

[312] Neurath, M. F., Pettersson, S., Meyer zum Buschenfelde, K. H. & Strober, W. Local administration of antisense phosphorothioate

oligonucleotides to the p65 subunit of nf-kappa b abrogates established experimental colitis in mice. *Nat Med* **2**, 998–1004 (1996). URL http://www.ncbi.nlm.nih.gov/pubmed/8782457.

[313] Gerlag, D. M. *et al.* The effect of a t cell-specific nf-kappa b inhibitor on in vitro cytokine production and collagen-induced arthritis. *J Immunol* **165**, 1652–8 (2000). URL http://www.ncbi.nlm.nih.gov/pubmed/10903776.

[314] Tomita, T. *et al.* Suppressed severity of collagen-induced arthritis by in vivo transfection of nuclear factor kappab decoy oligodeoxynucleotides as a gene therapy. *Arthritis Rheum* **42**, 2532–42 (1999). URL http://www.ncbi.nlm.nih.gov/pubmed/10615998.

[315] Yamamoto, Y. & Gaynor, R. B. Therapeutic potential of inhibition of the nf-kappab pathway in the treatment of inflammation and cancer. *J Clin Invest* **107**, 135–42 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11160126.

[316] Rousseaux, C. *et al.* Intestinal antiinflammatory effect of 5-aminosalicylic acid is dependent on peroxisome proliferator-activated receptor-gamma. *J Exp Med* **201**, 1205–15 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15824083.

[317] Siddique, I. & Khan, I. Mechanism of regulation of na-h exchanger in inflammatory bowel disease: role of tlr-4 signaling mechanism. *Dig Dis Sci* **56**, 1656–62 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21221801.

[318] Fiebich, B. L. *et al.* The non-steroidal anti-inflammatory drug tepoxalin inhibits interleukin-6 and alpha1-anti-chymotrypsin synthesis in astrocytes by preventing degradation of ikappab-alpha. *Neuropharmacology* **38**, 1325–33 (1999). URL http://www.ncbi.nlm.nih.gov/pubmed/10471086.

[319] Li, J. M. *et al.* Vanillin-ameliorated development of azoxymethane/dextran sodium sulfate-induced murine colorectal cancer: The involvement of proteasome/nuclear factor-kappab/mitogen-activated protein kinase pathways. *J Agric Food*

Chem **66**, 5563–5573 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29790745.

[320] Razali, N. A., Nazarudin, N. A., Lai, K. S., Abas, F. & Ahmad, S. Curcumin derivative, 2,6-bis(2-fluorobenzylidene)cyclohexanone (ms65) inhibits interleukin-6 production through suppression of nf-kappab and mapk pathways in histamine-induced human keratinocytes cell (hacat). *BMC Complement Altern Med* **18**, 217 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/30012134.

[321] Ren, Z. *et al.* Resveratrol inhibits nf-kb signaling through suppression of p65 and ikappab kinase activities. *Pharmazie* **68**, 689–94 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/24020126.

[322] Wu, S. L. *et al.* Vanillin improves and prevents trinitrobenzene sulfonic acid-induced colitis in mice. *J Pharmacol Exp Ther* **330**, 370–6 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19423842.

[323] Shiou, S. R. *et al.* Oral administration of transforming growth factor-beta1 (tgf-beta1) protects the immature gut from injury via smad protein-dependent suppression of epithelial nuclear factor kappab (nf-kappab) signaling and proinflammatory cytokine production. *J Biol Chem* **288**, 34757–66 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/24129565.

[324] Tsabouri, S., Priftis, K. N., Chaliasos, N. & Siamopoulou, A. Modulation of gut microbiota downregulates the development of food allergy in infancy. *Allergol Immunopathol (Madr)* **42**, 69–77 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/23827644.

[325] Kalliomaki, M. *et al.* Distinct patterns of neonatal gut microflora in infants in whom atopy was and was not developing. *J Allergy Clin Immunol* **107**, 129–34 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11150002.

[326] Yang, L. *et al.* Inflammation and intestinal metaplasia of the distal esophagus are associated with alterations in the microbiome. *Gastroenterology* **137**, 588–97 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19394334.

[327] Liu, S. X. *et al.* Fecal microbiota transplantation induces remission of infantile allergic colitis through gut microbiota reestablishment. *World J Gastroenterol* **23**, 8570–8581 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/29358865.

[328] Gupta, S., Allen-Vercoe, E. & Petrof, E. O. Fecal microbiota transplantation: in perspective. *Therap Adv Gastroenterol* **9**, 229–39 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26929784.

[329] Kelly, C. R. *et al.* Fecal microbiota transplant for treatment of clostridium difficile infection in immunocompromised patients. *Am J Gastroenterol* **109**, 1065–71 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24890442.

[330] Khoruts, A. & Sadowsky, M. J. Understanding the mechanisms of faecal microbiota transplantation. *Nat Rev Gastroenterol Hepatol* **13**, 508–16 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27329806.

[331] Colman, R. J. & Rubin, D. T. Fecal microbiota transplantation as therapy for inflammatory bowel disease: a systematic review and meta-analysis. *J Crohns Colitis* **8**, 1569–81 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25223604.

[332] Suskind, D. L., Singh, N., Nielson, H. & Wahbeh, G. Fecal microbial transplant via nasogastric tube for active pediatric ulcerative colitis. *J Pediatr Gastroenterol Nutr* **60**, 27–9 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25162366.

[333] Holvoet, T. *et al.* Assessment of faecal microbial transfer in irritable bowel syndrome with severe bloating. *Gut* **66**, 980–982 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/27511198.

[334] Millan, B., Laffin, M. & Madsen, K. Fecal microbiota transplantation: Beyond clostridium difficile. *Curr Infect Dis Rep* **19**, 31 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28770495.

[335] Kelly, C. R. *et al.* Update on fecal microbiota transplantation 2015: Indications, methodologies, mechanisms, and outlook. *Gastroenterology* **149**, 223–37 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25982290.

[336] Emdin, C. A., Khera, A. V. & Kathiresan, S. Mendelian randomization. *JAMA* **318**, 1925–1926 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/29164242.

[337] Andreoletti, G. *et al.* Exome analysis of rare and common variants within the nod signaling pathway. *Sci Rep* **7**, 46454 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28422189.

[338] Köhler, S. *et al.* Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Research* **47**, D1018–D1027 (2018). URL https://doi.org/10.1093/nar/gky1105.

[339] Torgerson, D. G. *et al.* Resequencing candidate genes implicates rare variants in asthma susceptibility. *The American Journal of Human Genetics* **90**, 273–281 (2012). URL https://doi.org/10.1016/j.ajhg.2012.01.008.

[340] Cepika, A.-M. *et al.* Tregopathies: Monogenic diseases resulting in regulatory t-cell deficiency. *Journal of Allergy and Clinical Immunology* **142**, 1679–1695 (2018). URL https://doi.org/10.1016/j.jaci.2018.10.026.

[341] Tan, T. H.-T., Ellis, J. A., Saffery, R. & Allen, K. J. The role of genetics and environment in the rise of childhood food allergy. *Clinical & Experimental Allergy* **42**, 20–29 (2011). URL https://doi.org/10.1111/j.1365-2222.2011.03823.x.

[342] Litosh, V. A. *et al.* Calpain-14 and its association with eosinophilic esophagitis. *Journal of Allergy and Clinical Immunology* **139**, 1762–1771.e7 (2017). URL https://doi.org/10.1016/j.jaci.2016.09.027.

[343] DeWan, A. T. *et al.* Whole-exome sequencing of a pedigree segregating asthma. *BMC Medical Genetics* **13** (2012). URL https://doi.org/10.1186/1471-2350-13-95.

[344] Hong, X., Tsai, H.-J. & Wang, X. Genetics of food allergy. *Current Opinion in Pediatrics* **21**, 770–776 (2009). URL https://doi.org/10.1097/mop.0b013e32833252dc.

[345] Li, J. *et al.* Copy number variations in CTNNA3 and RBFOX1 associate with pediatric food allergy. *The Journal of Immunology* **195**, 1599–1607 (2015). URL https://doi.org/10.4049/jimmunol.1402310.

[346] Ullemar, V. *et al.* Heritability and confirmation of genetic association studies for childhood asthma in twins. *Allergy* **71**, 230–238 (2015). URL https://doi.org/10.1111/all.12783.

[347] Krishnamurthy, P. & Kaplan, M. H. STAT6 and PARP family members in the development of t cell-dependent allergic inflammation. *Immune Network* **16**, 201 (2016). URL https://doi.org/10.4110/in.2016.16.4.201.

[348] Bernstein, D. I. *et al.* Genetic variants with gene regulatory effects are associated with diisocyanate-induced asthma. *Journal of Allergy and Clinical Immunology* **142**, 959–969 (2018). URL https://doi.org/10.1016/j.jaci.2018.06.022.

Chapter 7

Appendix

7.1 Software

The versions of the featured software used to perform the WES analysis can be found below. Additional details on the reproducible *conda* environment file can be found in the github repository: http://github.com/alsanju/wes-pipeline

- bcftools=1.2
- bedtools=2.25.0
- bwa=0.7.12
- cutadapt=1.9.1
- fastqc=0.10.1
- gatk=3.4-46
- gemini=0.17.2
- htslib=1.2
- parallel=20150922
- picard=1.140
- python=2.7.6
- r=3.2.3
- samtools=1.2
- vcftools=0.1.14
- vep=84

7.2 Gene information

The MIM and ensembl IDs for all the genes mentioned in this dissertation can be found in the next table.

Gene symbol	MIM	Ensembl Gene ID
AGT	106150	ENSG00000135744
AKR1D1	604741	ENSG00000122787
ANKZF1	617541	ENSG00000163516
CAPN14	610229	ENSG00000214711
CBLB	251110	ENSG00000114423
CD14	158120	ENSG00000170458
CD40L	300386	ENSG00000102245
CEP290	610142	ENSG00000198707
CTNNA3	607667	ENSG00000183230
DHODH	126064	ENSG00000102967
PARK7	602533	ENSG00000116288
DPP10	608209	ENSG00000175497
EOMES	604615	ENSG00000163508
FCGR3A	146740	ENSG00000203747
FLG	135940	ENSG00000143631
FOXP3	300292	ENSG00000049768
GBA	606463	ENSG00000177628
GFI1	600871	ENSG00000162676
GPR50	300207	ENSG00000102195
GSTP1	134660	ENSG00000084207
HLA-A	142800	ENSG00000206503
HLA-B	142830	ENSG00000234745
HLA-C	142840	ENSG00000204525
HLA-DPA1	142880	ENSG00000231389
HLA-DPB1	142858	ENSG00000223865
HLA-DQA1	146880	ENSG00000196735
HLA-DQA2	613503	ENSG00000237541
HLA-DQB1	604305	ENSG00000179344
HLA-DQB2	615161	ENSG00000232629
HLA-DRA	142860	ENSG00000204287
HLA-DRB1	142857	ENSG00000196126
HLA-DRB2	604776	ENSG00000227442
HLA-DRB3	612735	ENSG00000196101
HLA-DRB4	142857	ENSG00000227357
HLA-DRB5	604776	ENSG00000198502
HLA-E	143010	ENSG00000204592
HLA-F	143110	ENSG00000204642
HLA-G	142871	ENSG00000204632
IKBKAP	603722	ENSG00000070061
IL-10	124092	ENSG00000136634
IL12A	161560	ENSG00000168811
IL13	147683	ENSG00000169194

Gene symbol	MIM	Ensembl Gene ID
IL4	147780	ENSG00000113520
IL5	147850	ENSG00000113525
IL12RB1	601604	ENSG00000096996
IL13RA2	300130	ENSG00000123496
INO80	610169	ENSG00000128908
KALRN	604605	ENSG00000160145
KMT2B	606834	ENSG00000272333
LAMA5	601033	ENSG00000130702
MAP3K15	300820	ENSG00000180815
NFKB1	164011	ENSG00000109320
NRK	300791	ENSG00000123572
NLRP12	609648	ENSG00000142405
PARK2	600116	ENSG00000185345
PDE4DIP	608117	ENSG00000178104
PPL	602871	ENSG00000118898
RBFOX1	605104	ENSG00000078328
RUNX2	600211	ENSG00000124813
SKIV2L	600478	ENSG00000204351
STAB1	608560	ENSG00000010327
TBX21	604895	ENSG00000073861
TCF4	602272	ENSG00000196628
TNFRSF1A	191190	ENSG00000067182
ZNF645	-	ENSG00000175809

7.3 Gene list

This gene list has been assembled from literature searches for allergy and immunodeficiency, as well as associated Human Phenotype Ontology (HPO) terms [338] (accessed March 2018), comprising a total number of 1,346 genes. The HPO terms considered were those containing the words: *allerg*, *asth*, *immun*, *food*, *diarr*.

If the reason for inclusion was literature searches, the reference is specified in the following table. If it was by HPO term inference, the *minimal set* in the sense of the ontology's directed acyclic graph is included in the table. Abbreviations are as follows:

Abnormality of the immune system=AIS, Abnormality of immune system physiology=AISP, Diarrhea=DIA, Abnormality of cellular immune system=ACIS, Immunodeficiency=IDEF, Autoimmune thrombocytopenia=AT, Cellular immunodeficiency=CEI, Combined immunodeficiency=COI, Immune dysregulation=IDYS, Allergy=ALL, Asthma=AST, Autoimmune antibody positivity=AAP, Autoimmune hemolytic anemia=AHA, Chronic diarrhea=CD, Immunoglobulin IgG2 deficiency=IID, Severe combined immunodeficiency=SCI, Abnormal immunoglobulin level=AIL, Autoimmunity=AUTO, Food intolerance=FI, Intermittent diarrhea=INTD, Protracted diarrhea=PD, Abnormality of humoral immunity=AHI, Immunologic hypersensitivity=IH, Intractable diarrhea=INTD, Autoimmune neutropenia=AN, Allergic rhinitis=AR, Severe T-cell immunodeficiency=STI, Cow milk allergy=CMA, Secretory diarrhea= SECD, Abnormality of immune serum protein physiology=AISPP, Aspirininduced asthma=AIA.

Gene symbol	Ensembl gene ID	Source
A2ML1	ENSG00000166535	AIS
ABCA1	ENSG00000165029	AIS
ABCA12	ENSG00000144452	AISP
ABCB11	ENSG00000073734	AIS;DIA
ABCB4	ENSG00000005471	AISP;DIA
ABCC6	ENSG00000091262	AISP
ABCC8	ENSG00000006071	AIS;DIA
ABCC9	ENSG00000069431	ACIS
ABCD3	ENSG00000117528	AIS
ABCD4	ENSG00000119688	ACIS
ABCG5	ENSG00000138075	AIS
ABCG8	ENSG00000143921	AIS
ABL1	ENSG00000097007	ACIS
ACD	ENSG00000102977	ACIS;IDEF
ACP5	ENSG00000102575	ACIS;AT;CEI;COI
ACSF3	ENSG00000176715	DIA
ACTA1	ENSG00000143632	AISP
ACTB	ENSG00000075624	IDEF
ACTC1	ENSG00000159251	ACIS
ACTG2	ENSG00000163017	AISP
ACTN2	ENSG00000077522	ACIS
ACVR2B	ENSG00000114739	AIS
ACVRL1	ENSG00000139567	DIA
ADA	ENSG00000196839	ALL;AST;CD;IID;SCI
ADAM17	ENSG00000151694	ACIS;AISP;DIA
ADAMTS2	ENSG00000087116	AISP
ADAMTS3	ENSG00000156140	AIL
ADNP	ENSG00000101126	AISP
AFF4	ENSG00000072364	AISP
AGA	ENSG00000038002	ACIS;AISP;DIA
AGL	ENSG00000162688	IDEF
AGPAT2	ENSG00000169692	IDEF
AGT	ENSG00000135744	[339]
AGXT	ENSG00000172482	AISP
AICDA	ENSG00000111732	AIL;IDEF
AIP	ENSG00000110711	AISP
AIRE	ENSG00000160224	AUTO;DIA
AK2	ENSG00000004455	AIL;CEI;DIA;SCI
AKR1D1	ENSG00000122787	AIS;DIA
AKT1	ENSG00000142208	CEI
AKT2	ENSG00000105221	AIS
ALAD	ENSG00000148218	DIA

Gene symbol	Ensembl gene ID	Source
ALAS2	ENSG00000158578	AIS
ALDH3A2	ENSG00000072210	AISP
ALDOA	ENSG00000149925	AIS
ALG1	ENSG00000033011	AISP
ALG12	ENSG00000182858	AISP
ALG13	ENSG00000101901	AISP
ALG3	ENSG00000214160	DIA;FI
ALG9	ENSG00000086848	AIS
ALMS1	ENSG00000116127	AST
ALOX12B	ENSG00000179477	AISP
ALOXE3	ENSG00000179148	AISP
ALPL	ENSG00000162551	AISP
AMACR	ENSG00000242110	AISP
ANK1	ENSG00000029534	AIS
ANKRD1	ENSG00000148677	ACIS
ANKRD11	ENSG00000167522	AISP
ANKRD55	ENSG00000164512	AAP
ANO5	ENSG00000171714	AISP
ANTXR2	ENSG00000163297	CD;IDEF
AP1S1	ENSG00000106367	DIA
AP2S1	ENSG00000042753	AISP
AP3B1	ENSG00000132842	ACIS;AISP
AP3D1	ENSG00000065000	ACIS;IDEF
APC	ENSG00000134982	AISP
APC2	ENSG00000115266	AISP
APOA1	ENSG00000118137	AISP
APOC2	ENSG00000234906	AISP
APOE	ENSG00000130203	ACIS;AISP
APRT	ENSG00000198931	AISP
ARHGAP26	ENSG00000145819	ACIS
ARHGAP31	ENSG00000031081	ACIS
ARID1A	ENSG00000117713	AISP
ARID1B	ENSG00000049618	AISP
ARID2	ENSG00000189079	AISP
ARMC4	ENSG00000169126	AISP
ARSB	ENSG00000113273	AISP
ARVCF	ENSG00000099889	AST;AUTO;IDEF
ARX	ENSG00000004848	DIA
ASAH1	ENSG00000104763	ACIS;AISP
ATL3	ENSG00000184743	AISP
ATM	ENSG00000149311	CEI;IID
ATP6AP1	ENSG00000071553	AIL

Gene symbol	Ensembl gene ID	Source
ATP6V0A2	ENSG00000185344	AISP
ATP7A	ENSG00000165240	AISP;CD
ATP7B	ENSG00000123191	AISP
ATP8B1	ENSG00000081923	AISP;DIA
ATRX	ENSG00000085224	ACIS;IDEF;INTD;PD
AVP	ENSG00000101200	DIA
AXIN1	ENSG00000103126	AISP
B2M	ENSG00000166710	AIL;CD;INTD
B9D1	ENSG00000108641	AIS
B9D2	ENSG00000123810	AIS
BACH2	ENSG00000112182	[340]
BAG3	ENSG00000151929	ACIS
BAP1	ENSG00000163930	AISP
BAZ1B	ENSG00000009954	AISP
BBS1	ENSG00000174483	AST
BBS4	ENSG00000140463	AST
BCKDHA	ENSG00000248098	AISP
BCKDHB	ENSG00000083123	AISP
BCL10	ENSG00000142867	AIL;IDEF
BCL11B	ENSG00000127152	SCI
BCL2	ENSG00000171791	AISP
BCL6	ENSG00000113916	AISP
BCOR	ENSG00000183337	AISP
BCR	ENSG00000186716	ACIS;IDEF
BCS1L	ENSG00000074582	AISP
BIRC3	ENSG00000023445	AISP
BLM	ENSG00000197299	AIL;DIA
BLNK	ENSG00000095585	AIL;DIA;IDEF
BLOC1S6	ENSG00000104164	ACIS
BMPR1A	ENSG00000107779	DIA
BPGM	ENSG00000172331	AIS
BRAF	ENSG00000157764	AISP
BRCA1	ENSG00000012048	AISP;INTD
BRCA2	ENSG00000139618	ACIS;AISP;INTD
BRIP1	ENSG00000136492	ACIS;AISP
BSCL2	ENSG00000168000	IDEF
BTD	ENSG00000169814	AISP;DIA
BTK	ENSG00000010671	AIL;AUTO;CD;IDEF
BTNL2	ENSG00000204290	ACIS;AISP
BUB1	ENSG00000169679	ACIS;AISP
BUB1B	ENSG00000156970	ACIS;COI
BUB3	ENSG00000154473	ACIS;AISP

Gene symbol	Ensembl gene ID	Source
C11orf30	ENSG00000158636	[341]
C15orf41	ENSG00000186073	AIS
C1GALT1C1	ENSG00000171155	AUTO
C1orf172	ENSG00000175707	AISP
C1QA	ENSG00000173372	AHI;AUTO
C1QB	ENSG00000173369	AHI;AUTO
C1QC	ENSG00000159189	AHI;AUTO
C1R	ENSG00000159403	AHI;AUTO
C1S	ENSG00000182326	AHI;AUTO
C2	ENSG00000166278	AUTO
C21orf2	ENSG00000160226	AISP
C21orf59	ENSG00000159079	AISP
C3	ENSG00000125730	AHI
C4A	ENSG00000244731	AHI;AUTO;IH
C4B	ENSG00000224389	AHI
C5	ENSG00000106804	AHI;INTD
C5orf42	ENSG00000197603	AIS
C6	ENSG00000039537	AHI
C6orf25	ENSG00000204420	AIS
C7	ENSG00000112936	AHI
C8A	ENSG00000157131	AHI;AUTO
C8B	ENSG00000021852	AHI
C9	ENSG00000113600	AHI
CA2	ENSG00000104267	AIS
CACNA1C	ENSG00000151067	AISP
CALR	ENSG00000179218	ACIS
CAPN14	ENSG00000214711	[342]
CAPN3	ENSG00000092529	ACIS
CAPN5	ENSG00000149260	AISP
CARD11	ENSG00000198286	AIL;IDEF
CARD14	ENSG00000141527	AISP
CARD9	ENSG00000187796	IDEF
CASP10	ENSG00000003400	AIL;AAP;AHA;AN;AT
CASP8	ENSG00000064012	ACIS;AST;CD
CASR	ENSG00000036828	ACIS;AISP
CAV1	ENSG00000105974	AUTO;IDEF
CBL	ENSG00000110395	ACIS
CBLB	ENSG00000114423	[343]
CBS	ENSG00000160200	AISP
CC2D2A	ENSG00000048342	AIS
CCBE1	ENSG00000183287	AIL
CCDC103	ENSG00000167131	AISP

Gene symbol	Ensembl gene ID	Source
CCDC114	ENSG00000105479	AISP
CCDC115	ENSG00000136710	AIS
CCDC151	ENSG00000198003	AST
CCDC22	ENSG00000101997	AISP
CCDC39	ENSG00000145075	AISP
CCDC40	ENSG00000141519	AISP
CCDC65	ENSG00000139537	AISP
CCND1	ENSG00000110092	AIL
CCNO	ENSG00000152669	AISP
CCR1	ENSG00000163823	IH
CCR6	ENSG00000112486	AUTO
CCT5	ENSG00000150753	AISP
CD14	ENSG00000170458	[344]
CD151	ENSG00000177697	AISP
CD19	ENSG00000177455	AIL;AT;DIA;IDEF
CD247	ENSG00000198821	ACIS;AAP;IDEF;PD
CD27	ENSG00000139193	AIL
CD28	ENSG00000178562	AIL;IDEF
CD3D	ENSG00000167286	ACIS;DIA;IDEF
CD3E	ENSG00000198851	ACIS;IDEF
CD3G	ENSG00000160654	ACIS;AHA;IDEF
CD4	ENSG00000010610	ACIS
CD40	ENSG00000101017	AIL;IDEF
CD40LG	ENSG00000102245	AIL;DIA;IDEF
CD55	ENSG00000196352	DIA
CD79A	ENSG00000105369	AIL;DIA;IDEF
CD79B	ENSG00000007312	AIL;DIA;IDEF
CD81	ENSG00000110651	AIL;AT;IDEF
CD8A	ENSG00000153563	ACIS;AISP
CD96	ENSG00000153283	AISP
CDAN1	ENSG00000140326	AIS
CDC73	ENSG00000134371	AISP
CDCA7	ENSG00000144354	AIL;CEI
CDH23	ENSG00000107736	IDEF
CDH3	ENSG00000062038	IH
CDK4	ENSG00000135446	AIS
CDKN1A	ENSG00000124762	AISP;DIA
CDKN1B	ENSG00000111276	AISP;DIA
CDKN2A	ENSG00000147889	AIS;INTD
CDKN2B	ENSG00000147883	AISP;DIA
CDKN2C	ENSG00000123080	AISP;DIA
CDKN2D	ENSG00000129355	AIS

Gene symbol	Ensembl gene ID	Source
CDON	ENSG00000064309	AIS
CDSN	ENSG00000204539	AIL;AST
CEBPA	ENSG00000245848	ACIS
CEBPE	ENSG00000092067	ACIS;AISP
CEP290	ENSG00000198707	AIS
CEP57	ENSG00000166037	ACIS;AISP
CERS3	ENSG00000154227	AISP
CFB	ENSG00000243649	AHI
CFC1	ENSG00000136698	AIS
CFD	ENSG00000197766	AHI
CFH	ENSG00000000971	AHI
CFHR5	ENSG00000134389	AISP
CFI	ENSG00000205403	AHI
CFP	ENSG00000126759	AHI
CFTR	ENSG00000001626	AIL;AST;IDEF
CHAMP1	ENSG00000198824	AISP
CHAT	ENSG00000070748	AIS
CHD7	ENSG00000171316	ACIS;AUTO;CD;SCI
CHRM3	ENSG00000133019	AISP
CHRNE	ENSG00000108556	AIS
CHST14	ENSG00000169105	AISP
CIDEC	ENSG00000187288	AISP
CIITA	ENSG00000179583	AIL;PD
CISD2	ENSG00000145354	AISP
CLCA4	ENSG00000016602	AIL;IDEF
CLCN7	ENSG00000103249	ACIS;AISP
CLDN1	ENSG00000163347	AISP
CLDN16	ENSG00000113946	AISP
CLDN19	ENSG00000164007	AISP
CLEC7A	ENSG00000172243	AISP
CLIP2	ENSG00000106665	AISP
CLMP	ENSG00000166250	CD
CLN3	ENSG00000188603	ACIS
CLPB	ENSG00000162129	ACIS;AISP
CMA1	ENSG00000092009	[344]
CNBP	ENSG00000169714	AIL
COG2	ENSG00000135775	AIS
COG4	ENSG00000103051	AISP;CD;INTD
COG6	ENSG00000133103	AIL;CD
COG7	ENSG00000168434	AISP
COL11A2	ENSG00000204248	AISP
COL13A1	ENSG00000197467	AISP

Gene symbol	Ensembl gene ID	Source
COL18A1	ENSG00000182871	AIS
COL1A1	ENSG00000108821	AIL
COL2A1	ENSG00000139219	AISP
COL3A1	ENSG00000168542	AISP
COL4A3	ENSG00000169031	AISP
COL4A4	ENSG00000081052	AISP
COL4A5	ENSG00000188153	AISP
COL5A1	ENSG00000130635	AISP
COL5A2	ENSG00000204262	AISP
COL6A1	ENSG00000142156	AISP
COL6A2	ENSG00000142173	AISP
COL6A3	ENSG00000163359	AISP
COL7A1	ENSG00000114270	AISP
COLQ	ENSG00000206561	AIS
COMT	ENSG00000093010	AST;AUTO;IDEF
CORO1A	ENSG00000102879	ACIS;IDEF
COX4I2	ENSG00000131055	AR;AST
CPA1	ENSG00000091704	ACIS;AISP
CPLX1	ENSG00000168993	IDEF
CPOX	ENSG00000080819	AIS;DIA
CPT1A	ENSG00000110090	DIA
CR2	ENSG00000117322	AIL;AT;CD;IDEF
CREBBP	ENSG00000005339	AISP
CRIPT	ENSG00000119878	AISP
CRKL	ENSG00000099942	IDEF
CRYAB	ENSG00000109846	ACIS
CSF3R	ENSG00000119535	ACIS;AISP
CSNK2A1	ENSG00000101266	AIL
CSPP1	ENSG00000104218	AIS
CSRP3	ENSG00000129170	ACIS
CSTA	ENSG00000121552	ALL
CTBP1	ENSG00000159692	IDEF
CTC1	ENSG00000178971	ACIS;CEI
CTLA4	ENSG00000163599	AIL;AHA;AT;DIA
CTNNA3	ENSG00000183230	[345]
CTNNB1	ENSG00000168036	AISP
CTNS	ENSG00000040531	AIS
CTPS1	ENSG00000171793	IDEF;IID
CTRC	ENSG00000162438	ACIS;AISP
CTSA	ENSG00000064601	AIS
CTSB	ENSG00000164733	AISP
CTSC	ENSG00000109861	AISP

Gene symbol	Ensembl gene ID	Source
CTSK	ENSG00000143387	AISP
CUL4B	ENSG00000158290	IDEF
CXCR4	ENSG00000121966	AIL
CYBA	ENSG00000051523	ACIS;AISP
CYBB	ENSG00000165168	ACIS;AISP
CYP26C1	ENSG00000187553	ACIS
CYP27A1	ENSG00000135929	DIA
CYP4F22	ENSG00000171954	AISP
CYP7A1	ENSG00000167910	AISP
CYP7B1	ENSG00000172817	AISP;DIA
CYSLTR2	ENSG00000152207	AISP
DAXX	ENSG00000204209	AIS;INTD;PD
DBT	ENSG00000137992	AISP
DCDC2	ENSG00000146038	AISP
DCLRE1C	ENSG00000152457	AIL;AUTO;CD;SCI
DCTN4	ENSG00000132912	AIL;IDEF
DDB2	ENSG00000134574	AISP
DDC	ENSG00000132437	DIA
DDOST	ENSG00000244038	AISP
DDR2	ENSG00000162733	AISP
DDRGK1	ENSG00000198171	AIS
DEAF1	ENSG00000177030	AISP
DENND1B	ENSG00000213047	[341]
DES	ENSG00000175084	ACIS;DIA
DGAT1	ENSG00000185000	DIA
DGCR14	ENSG00000100056	AISP
DGCR2	ENSG00000070413	AISP
DGCR6	ENSG00000183628	AISP
DGCR8	ENSG00000128191	AISP
DGUOK	ENSG00000114956	AIS
DHCR24	ENSG00000116133	AIS
DHCR7	ENSG00000172893	AISP
DIS3L2	ENSG00000144535	AIS
DKC1	ENSG00000130826	ACIS;CEI
DLEC1	ENSG00000008226	AIS
DLL3	ENSG00000090932	AISP
DLL4	ENSG00000128917	ACIS
DMD	ENSG00000198947	ACIS
DNAAF1	ENSG00000154099	AISP
DNAAF2	ENSG00000165506	AISP
DNAAF3	ENSG00000167646	AISP
DNAH1	ENSG00000114841	AISP

Gene symbol	Ensembl gene ID	Source
DNAH11	ENSG00000105877	AISP
DNAH5	ENSG00000039139	AISP
DNAI1	ENSG00000122735	AISP
DNAI2	ENSG00000171595	AISP
DNAJB13	ENSG00000187726	AISP
DNAJC21	ENSG00000168724	ACIS;AISP
DNAL1	ENSG00000119661	AISP
DNASE1L3	ENSG00000163687	AHI;AUTO;DIA;IH
DNMT1	ENSG00000130816	AISP
DNMT3B	ENSG00000088305	AIL;CEI;DIA
DOCK2	ENSG00000134516	IDEF
DOCK6	ENSG00000130158	ACIS
DOCK8	ENSG00000107099	AIL;AST
DOK7	ENSG00000175920	AIS
DOLK	ENSG00000175283	ACIS
DPM1	ENSG00000000419	AIS
DPP10	ENSG00000175497	[339]
DRC1	ENSG00000157856	AISP
DSG1	ENSG00000134760	AISP
DSG2	ENSG00000046604	ACIS
DYNC2LI1	ENSG00000138036	ACIS
DYX1C1	ENSG00000256061	AISP
EBP	ENSG00000147155	AISP
ECE1	ENSG00000117298	AISP;DIA
ECM1	ENSG00000143369	AISP
EDA	ENSG00000158813	AISP
EDAR	ENSG00000135960	AISP
EDARADD	ENSG00000186197	AISP
EDN3	ENSG00000124205	AISP;DIA
EDNRB	ENSG00000136160	AISP;DIA
EFEMP2	ENSG00000172638	AISP
EFTUD1	ENSG00000140598	ACIS;AISP
EGFR	ENSG00000146648	AISP
EHMT1	ENSG00000181090	AISP
EIF2AK3	ENSG00000172071	ACIS
ELANE	ENSG00000197561	AIL
ELN	ENSG00000049540	IH
EMP2	ENSG00000213853	AISP
ENG	ENSG00000106991	DIA
ENPP1	ENSG00000197594	AISP
EOGT	ENSG00000163378	ACIS
EP300	ENSG00000100393	AISP

Gene symbol	Ensembl gene ID	Source
EPCAM	ENSG00000119888	INTD
EPG5	ENSG00000152223	CEI;IID
ERAP1	ENSG00000164307	IH
ERCC1	ENSG00000012061	AISP
ERCC2	ENSG00000104884	AIL;AST;CD
ERCC3	ENSG00000163161	AISP
ERCC4	ENSG00000175595	ACIS;AISP
ERCC5	ENSG00000134899	AISP
ERCC6	ENSG00000225830	AISP
ERCC6L2	ENSG00000182150	ACIS
ERCC8	ENSG00000049167	AIS
ERF	ENSG00000105722	AISP
ESCO2	ENSG00000171320	AIS
ESR1	ENSG00000091831	AISP
ETHE1	ENSG00000105755	CD
ETV6	ENSG00000139083	ACIS
EVC	ENSG00000072840	ACIS
EVC2	ENSG00000173040	ACIS
EWSR1	ENSG00000182944	AIS
EXD3	ENSG00000187609	ACIS;IDEF
EXT1	ENSG00000182197	AISP
EXTL3	ENSG00000012232	AIL
EYA4	ENSG00000112319	AISP
F5	ENSG00000198734	AISP
FADD	ENSG00000168040	AAP
FAH	ENSG00000103876	AIS
FAM105B	ENSG00000154124	ACIS;DIA
FAM111A	ENSG00000166801	AIS
FAM111B	ENSG00000189057	AISP
FAM134B	ENSG00000154153	AISP
FANCA	ENSG00000187741	ACIS;AISP
FANCB	ENSG00000181544	ACIS;AISP
FANCC	ENSG00000158169	ACIS;AISP
FANCD2	ENSG00000144554	ACIS;AISP
FANCE	ENSG00000112039	ACIS;AISP
FANCF	ENSG00000183161	ACIS;AISP
FANCG	ENSG00000221829	ACIS;AISP
FANCI	ENSG00000140525	ACIS;AISP
FANCL	ENSG00000115392	ACIS;AISP
FANCM	ENSG00000187790	ACIS;AISP
FAS	ENSG00000026103	AIL;AAP;AHA;AN
FASLG	ENSG00000117560	AIL;AAP;AHA;AN

Gene symbol	Ensembl gene ID	Source	Gene symbol	Ensembl gene ID	Source
FAT4	ENSG00000196159	AIL	***GABRD***	ENSG00000187730	AIS
FBLN5	ENSG00000140092	AISP	***GALC***	ENSG00000054983	AT
FBXL4	ENSG00000112234	ACIS;AISP	***GALE***	ENSG00000117308	AIS
FCER1A	ENSG00000179639	[341]	***GALNS***	ENSG00000141012	AISP
FCGR2C	ENSG00000244682	AAP	***GALT***	ENSG00000213930	AISP;DIA
FCGR3A	ENSG00000203747	IDEF	***GAS8***	ENSG00000141013	AISP
FCN3	ENSG00000142748	IDEF	***GATA1***	ENSG00000102145	ACIS;IDEF
FECH	ENSG00000066926	AISP	***GATA2***	ENSG00000179348	ACIS;IDEF
FERMT1	ENSG00000101311	AISP	***GATA3***	ENSG00000107485	AISP
FERMT3	ENSG00000149781	ACIS;AISP	***GATA6***	ENSG00000141448	INTD
FGA	ENSG00000171560	AISP	***GATAD1***	ENSG00000157259	ACIS
FGB	ENSG00000171564	AIS	***GBA***	ENSG00000177628	AIL
FGF3	ENSG00000186895	AISP	***GBE1***	ENSG00000114480	AIS
FGFR2	ENSG00000066468	AISP	***GCK***	ENSG00000106633	AIS
FGFR3	ENSG00000068078	AISP	***GDF1***	ENSG00000130283	AIS
FGFRL1	ENSG00000127418	IDEF	***GDNF***	ENSG00000168621	AISP;DIA
FGG	ENSG00000171557	AIS	***GFI1***	ENSG00000162676	ACIS
FHL2	ENSG00000115641	ACIS	***GFI1B***	ENSG00000165702	AIS
FKTN	ENSG00000106692	ACIS	***GFPT1***	ENSG00000198380	AIS
FLG	ENSG00000143631	AST	***GH1***	ENSG00000259384	AIS
FLI1	ENSG00000151702	AISP	***GIF***	ENSG00000134812	AIS
FLII	ENSG00000177731	AISP	***GJA1***	ENSG00000152661	AISP
FLNA	ENSG00000196924	AISP	***GJB2***	ENSG00000165474	AISP
FLT3	ENSG00000122025	ACIS	***GJB3***	ENSG00000188910	AISP
FLT4	ENSG00000037280	AIS	***GJB4***	ENSG00000189433	AISP
FLVCR1	ENSG00000162769	AISP	***GJB6***	ENSG00000121742	AISP
FMO3	ENSG00000007933	ACIS;AISP	***GJC2***	ENSG00000198835	AISP
FMR1	ENSG00000102081	AISP	***GLA***	ENSG00000102393	DIA
FOS	ENSG00000170345	IDEF	***GLB1***	ENSG00000170266	ACIS;AISP
FOXC2	ENSG00000176692	AISP	***GLI1***	ENSG00000111087	ACIS
FOXE1	ENSG00000178919	AIS	***GLI3***	ENSG00000106571	AISP
FOXF1	ENSG00000103241	AIS	***GLIS3***	ENSG00000107249	AISP
FOXN1	ENSG00000109101	ACIS;STI	***GLRA1***	ENSG00000145888	AISP
FOXP1	ENSG00000114861	AISP	***GLRB***	ENSG00000109738	AISP
FOXP3	ENSG00000049768	ACIS;AHA;DIA;IDYS	***GLRX5***	ENSG00000182512	AIS
FRAS1	ENSG00000138759	STI	***GLUL***	ENSG00000135821	AISP
FREM2	ENSG00000150893	STI	***GMNN***	ENSG00000112312	AISP
FTCD	ENSG00000160282	ACIS	***GNA11***	ENSG00000088256	AISP
FUCA1	ENSG00000179163	ACIS;AISP	***GNAQ***	ENSG00000156052	AISP
G6PC	ENSG00000131482	AISP;INTD	***GNAS***	ENSG00000087460	AISP
G6PC3	ENSG00000141349	ACIS;AISP	***GNB1***	ENSG00000078369	ACIS
GAA	ENSG00000171298	AISP	***GNE***	ENSG00000159921	AIS

Gene symbol	Ensembl gene ID	Source
GNPTAB	ENSG00000111670	AISP
GNS	ENSG00000135677	AISP;DIA
GORAB	ENSG00000120370	AISP
GP1BA	ENSG00000185245	AIS
GP1BB	ENSG00000203618	AST;AUTO;IDEF
GPC3	ENSG00000147257	AIL
GPC4	ENSG00000076716	AIL
GPD1	ENSG00000167588	AIS
GPHN	ENSG00000171723	AISP
GPI	ENSG00000105220	ACIS
GPIHBP1	ENSG00000182851	AISP
GPR101	ENSG00000165370	AISP
GPR35	ENSG00000178623	AIL;AUTO;IH
GRHL2	ENSG00000083307	AST
GRHPR	ENSG00000137106	AISP
GRIP1	ENSG00000155974	STI
GSS	ENSG00000100983	ACIS
GSTP1	ENSG00000084207	[341]
GTF2H5	ENSG00000272047	AST
GTF2I	ENSG00000077809	AISP
GTF2IRD1	ENSG00000006704	AISP
GUCY2C	ENSG00000070019	DIA
GUSB	ENSG00000169919	AISP
H19	ENSG00000130600	AIS
H6PD	ENSG00000049239	AISP
HABP2	ENSG00000148702	AIS
HADH	ENSG00000138796	DIA
HAMP	ENSG00000105697	AIS
HAX1	ENSG00000143575	ACIS;AISP
HBA1	ENSG00000206172	AISP
HBA2	ENSG00000188536	AISP
HBB	ENSG00000244734	ACIS;IDEF
HBG1	ENSG00000213934	AIS
HBG2	ENSG00000196565	AIS
HDAC4	ENSG00000068024	AISP
HDAC8	ENSG00000147099	AISP
HEATR2	ENSG00000164818	AISP
HELLS	ENSG00000119969	AIL;CEI
HERC2	ENSG00000128731	AISP
HES7	ENSG00000179111	AISP
HEXB	ENSG00000049860	AIS;CD
HFE	ENSG00000010704	AIS

Gene symbol	Ensembl gene ID	Source
HFE2	ENSG00000168509	AIS
HGD	ENSG00000113924	AISP
HGSNAT	ENSG00000165102	AISP;DIA
HIRA	ENSG00000100084	AST;AUTO;IDEF
HK1	ENSG00000156515	AIS
HLA-A	ENSG00000206503	AISP
HLA-B	ENSG00000234745	ACIS;DIA;IH
HLA-DPB1	ENSG00000223865	ACIS;AUTO
HLA-DQB1	ENSG00000179344	AUTO
HLA-DRB1	ENSG00000196126	ACIS;AUTO
HLCS	ENSG00000159267	AISP
HMBS	ENSG00000256269	DIA
HMGA2	ENSG00000149948	AIS
HMGCS2	ENSG00000134240	DIA
HNF1A	ENSG00000135100	DIA
HNF4A	ENSG00000101076	DIA
HNRNPA2B1	ENSG00000122566	AISP
HOXA13	ENSG00000106031	AISP
HPGD	ENSG00000164120	AISP
HPS1	ENSG00000107521	AISP;DIA
HPSE2	ENSG00000172987	AISP
HSD3B2	ENSG00000203859	AISP
HSD3B7	ENSG00000099377	AISP;DIA
HSPA9	ENSG00000113013	AISP
HSPG2	ENSG00000142798	AISP
HTRA2	ENSG00000115317	ACIS
HYAL1	ENSG00000114378	AISP
HYDIN	ENSG00000157423	AISP
HYLS1	ENSG00000198331	AIS
ICOS	ENSG00000163600	AIL;AN;AT;DIA
IDH1	ENSG00000138413	AIS
IDH2	ENSG00000182054	AIS
IDS	ENSG00000010404	AST;DIA
IDUA	ENSG00000127415	AISP;CD
IER3IP1	ENSG00000134049	AISP
IFIH1	ENSG00000115267	AIL
IFNGR1	ENSG00000027697	IDEF
IFNGR2	ENSG00000159128	IDEF
IFT172	ENSG00000138002	AISP
IGF2R	ENSG00000197081	AISP
IGHM	ENSG00000211899	AIL;DIA;IDEF
IGKC	ENSG00000211592	AIL;DIA

Gene symbol	Ensembl gene ID	Source
IGLL1	ENSG00000128322	AIL;DIA;IDEF
IGSF3	ENSG00000143061	AISP
IKBKAP	ENSG00000070061	AISP;DIA
IKBKB	ENSG00000104365	AIL;CD;IDEF
IKBKG	ENSG00000073009	AIL;IDEF
IKZF1	ENSG00000185811	AIL;DIA
IL10	ENSG00000136634	IH
IL10RA	ENSG00000110324	AISP;DIA
IL10RB	ENSG00000243646	AISP
IL12A	ENSG00000168811	AIL;AAP;IH
IL12A-AS1	ENSG00000244040	IH
IL12B	ENSG00000113302	IDEF
IL12RB1	ENSG00000096996	AIL;AAP;IDEF;IH
IL13	ENSG00000169194	[341]
IL17F	ENSG00000112116	AISP
IL17RA	ENSG00000177663	AISP
IL17RC	ENSG00000163702	AISP
IL18	ENSG00000150782	[344]
IL1RL1	ENSG00000115602	[346]
IL1RN	ENSG00000136689	AISP
IL21	ENSG00000138684	AIL;CD;IDEF
IL21R	ENSG00000103522	CD;IDEF
IL23R	ENSG00000162594	IH
IL2RA	ENSG00000134460	AIL;AAP;AHA;CD
IL2RB	ENSG00000100385	AAP
IL2RG	ENSG00000147168	AIL;AUTO;CD;SCI
IL36RN	ENSG00000136695	AISP
IL4	ENSG00000113520	[341]
IL4R	ENSG00000077238	[344]
IL6	ENSG00000136244	AUTO
IL7R	ENSG00000168685	ACIS;AUTO;CD;SCI
INPP5E	ENSG00000148384	AIS
INPPL1	ENSG00000165458	AISP
INS	ENSG00000254647	AIS
INSR	ENSG00000171105	AISP
IQSEC2	ENSG00000124313	AISP
IRAK4	ENSG00000198001	ACIS;IDEF
IRF5	ENSG00000128604	AIL;AAP;IH
IRF7	ENSG00000185507	IDEF
IRF8	ENSG00000140968	IDEF
ISG15	ENSG00000187608	IDEF
ISL1	ENSG00000016082	AISP

Gene symbol	Ensembl gene ID	Source
ITCH	ENSG00000078747	AUTO;CD
ITGA3	ENSG00000005884	AISP
ITGA6	ENSG00000091409	AISP;INTD
ITGA7	ENSG00000135424	AISP
ITGB2	ENSG00000160255	ACIS;AISP
ITGB4	ENSG00000132470	AISP;INTD
ITK	ENSG00000113263	AIL;AUTO
IVD	ENSG00000128928	ACIS
JAGN1	ENSG00000171135	ACIS;AISP
JAK2	ENSG00000096968	ACIS;AISP
JAK3	ENSG00000105639	AIL;DIA;SCI
JMJD1C	ENSG00000171988	AST;AUTO;IDEF
KALRN	ENSG00000160145	[343]
KANSL1	ENSG00000120071	AISP
KAT6B	ENSG00000156650	AISP
KCNAB2	ENSG00000069424	AIS
KCNH1	ENSG00000143473	AIS
KCNJ1	ENSG00000151704	DIA
KCNJ11	ENSG00000187486	AIS;DIA
KCNJ6	ENSG00000157542	AISP
KCNN4	ENSG00000104783	AIS
KCTD1	ENSG00000134504	AISP
KDM6A	ENSG00000147050	AT
KDSR	ENSG00000119537	AISP
KIAA0196	ENSG00000164961	AISP
KIAA0319L	ENSG00000142687	AUTO
KIAA0556	ENSG00000047578	AISP
KIAA1377	ENSG00000110318	AIS
KIF11	ENSG00000138160	ACIS;AISP
KIF1A	ENSG00000130294	AISP
KIF23	ENSG00000137807	DIA
KIT	ENSG00000157404	ACIS;AISP
KLF1	ENSG00000105610	AIS
KLLN	ENSG00000227268	CEI
KLRC4	ENSG00000183542	IH
KMT2A	ENSG00000118058	AISP
KMT2D	ENSG00000167548	AT
KRAS	ENSG00000133703	ACIS;INTD
KRT1	ENSG00000167768	AIL
KRT10	ENSG00000186395	AISP
KRT14	ENSG00000186847	AISP
KRT16	ENSG00000186832	AISP

Gene symbol	Ensembl gene ID	Source
KRT17	ENSG00000128422	AISP
KRT5	ENSG00000186081	AISP
KRT9	ENSG00000171403	AIL
LACC1	ENSG00000179630	AUTO
LAGE3	ENSG00000196976	AISP
LAMA2	ENSG00000196569	AISP
LAMA3	ENSG00000053747	AISP
LAMA4	ENSG00000112769	ACIS
LAMB3	ENSG00000196878	AISP
LAMC2	ENSG00000058085	AISP
LAMTOR2	ENSG00000116586	AIL;IDEF
LBR	ENSG00000143815	ACIS;AISP
LCAT	ENSG00000213398	AIS
LCK	ENSG00000182866	AUTO;DIA;IDEF
LCT	ENSG00000115850	DIA
LDB3	ENSG00000122367	ACIS
LEMD3	ENSG00000174106	AIS
LEP	ENSG00000174697	ACIS;AISP
LEPR	ENSG00000116678	ACIS;IDYS
LETM1	ENSG00000168924	IDEF
LFNG	ENSG00000106003	AISP
LHCGR	ENSG00000138039	AISP
LIFR	ENSG00000113594	AST
LIG4	ENSG00000174405	ACIS;AST;AUTO;CD
LIMK1	ENSG00000106683	AISP
LIPA	ENSG00000107798	ACIS;DIA
LIPN	ENSG00000204020	AISP
LMBRD1	ENSG00000168216	ACIS;AISP
LMF1	ENSG00000103227	AISP
LMNA	ENSG00000160789	AHI
LMNB2	ENSG00000176619	ACIS;AHI;AUTO
LMOD1	ENSG00000163431	AISP
LMX1B	ENSG00000136944	AISP
LPIN2	ENSG00000101577	ACIS;AISP
LPL	ENSG00000175445	AISP
LRBA	ENSG00000198589	AIL;AST;AHA;CD
LRIG2	ENSG00000198799	AISP
LRRC32	ENSG00000137507	[341]
LRRC6	ENSG00000129295	AISP
LRRC8A	ENSG00000136802	AIL;DIA;IDEF
LYST	ENSG00000143669	ACIS;IDEF
LYZ	ENSG00000090382	AISP

Gene symbol	Ensembl gene ID	Source
LZTR1	ENSG00000099949	AIS
MAD2L2	ENSG00000116670	ACIS;AISP
MAF	ENSG00000178573	AISP
MAGEL2	ENSG00000254585	AISP
MAGT1	ENSG00000102158	ACIS;IDEF
MALT1	ENSG00000172175	IDEF
MAN2B1	ENSG00000104774	AIL
MANBA	ENSG00000109323	AISP
MAP2K2	ENSG00000126934	AIS
MAP3K7	ENSG00000135341	AISP
MAPK1	ENSG00000100030	IDEF
MASP2	ENSG00000009724	AHI;AUTO
MBTPS2	ENSG00000012174	IDEF
MC1R	ENSG00000258839	AIS
MC2R	ENSG00000185231	AISP
MCCC2	ENSG00000131844	AISP
MCIDAS	ENSG00000234602	AISP
MCM4	ENSG00000104738	AISP
MCM6	ENSG00000076003	DIA
MECOM	ENSG00000085276	ACIS
MECP2	ENSG00000169057	AISP
MED13L	ENSG00000123066	AISP
MEFV	ENSG00000103313	ACIS;DIA;IH
MEGF8	ENSG00000105429	AIS
MEIS2	ENSG00000134138	IDEF
MEN1	ENSG00000133895	AISP;DIA
MESP2	ENSG00000188095	AISP
MET	ENSG00000105976	AISP
MGME1	ENSG00000125871	AISP;DIA
MGMT	ENSG00000170430	AIS
MGP	ENSG00000111341	AISP
MIF	ENSG00000240972	AUTO
MINPP1	ENSG00000107789	AIS
MITF	ENSG00000187098	AIS
MKKS	ENSG00000125863	AST
MKRN3	ENSG00000179455	AISP
MKS1	ENSG00000011143	AIS
MLH1	ENSG00000076242	ACIS
MLLT11	ENSG00000213190	ACIS
MLX	ENSG00000108788	AISP
MLXIPL	ENSG00000009950	IH
MLYCD	ENSG00000103150	DIA

Gene symbol	Ensembl gene ID	Source
MMAA	ENSG00000151611	ACIS
MMAB	ENSG00000139428	ACIS
MMACHC	ENSG00000132763	ACIS
MMEL1	ENSG00000142606	AIL;AAP;IH
MMP1	ENSG00000196611	AISP
MMP2	ENSG00000087245	AAP
MMP21	ENSG00000154485	AIS
MNX1	ENSG00000130675	AISP
MOGS	ENSG00000115275	AIL
MPDU1	ENSG00000129255	AISP
MPI	ENSG00000178802	DIA
MPL	ENSG00000117400	ACIS
MPLKIP	ENSG00000168303	AISP
MPO	ENSG00000005381	AIS
MPV17	ENSG00000115204	AISP;DIA
MPZ	ENSG00000158887	AIS
MS4A1	ENSG00000156738	AIL;AT;IDEF
MS4A2	ENSG00000149534	[344]
MSH2	ENSG00000095002	ACIS
MSH6	ENSG00000116062	ACIS
MSMO1	ENSG00000052802	AISP
MSN	ENSG00000147065	AIL
MST1	ENSG00000173531	AIL;AUTO;IH
MT-CO1	ENSG00000198804	AISP
MT-CO2	ENSG00000198712	AISP
MT-CO3	ENSG00000198938	AISP
MT-ND1	ENSG00000198888	AISP
MT-ND4	ENSG00000198886	AISP
MT-ND5	ENSG00000198786	AISP
MT-ND6	ENSG00000198695	AISP
MT-TF	ENSG00000210049	AISP
MT-TH	ENSG00000210176	AISP
MT-TL1	ENSG00000209082	AISP
MT-TQ	ENSG00000210107	AISP
MT-TS1	ENSG00000210151	AISP
MT-TS2	ENSG00000210184	AISP
MT-TW	ENSG00000210117	AISP
MTOR	ENSG00000198793	AIL
MVK	ENSG00000110921	AIL;DIA
MYBPC3	ENSG00000134571	ACIS
MYC	ENSG00000136997	ACIS
MYCN	ENSG00000134323	AIS

Gene symbol	Ensembl gene ID	Source
MYD88	ENSG00000172936	DIA;IDEF
MYH11	ENSG00000133392	AISP
MYH6	ENSG00000197616	ACIS
MYH7	ENSG00000092054	ACIS
MYH9	ENSG00000100345	ACIS;AISP
MYL2	ENSG00000111245	AISP
MYLK	ENSG00000065534	AISP
MYO5A	ENSG00000197535	AISP
MYO5B	ENSG00000167306	PD
MYPN	ENSG00000138347	ACIS
NAA10	ENSG00000102030	AISP
NAGLU	ENSG00000108784	AISP;DIA
NAT2	ENSG00000156006	[344]
NBEAL2	ENSG00000160796	AIS
NBN	ENSG00000104320	AIL;AHA;CD
NCF1	ENSG00000158517	ACIS;AISP
NCF2	ENSG00000116701	ACIS;AISP
NCF4	ENSG00000100365	ACIS;AISP;DIA
NCSTN	ENSG00000162736	AISP
NDN	ENSG00000182636	AISP
NDNL2	ENSG00000185115	AISP
NDP	ENSG00000124479	AISP
NEBL	ENSG00000078114	ACIS
NEK8	ENSG00000160602	AIS
NEK9	ENSG00000119638	AST
NELFA	ENSG00000185049	AISP
NEU1	ENSG00000204386	ACIS
NEUROG3	ENSG00000122859	DIA
NEXN	ENSG00000162614	ACIS
NF1	ENSG00000196712	ACIS
NFIX	ENSG00000008441	AISP
NFKB1	ENSG00000109320	AIL;AT;IDEF
NFKB2	ENSG00000077150	AIL;AST;AT;IDEF
NFKBIA	ENSG00000100906	AISP
NGLY1	ENSG00000151092	AISP
NHEJ1	ENSG00000187736	AIL;AUTO;IDEF
NHP2	ENSG00000145912	ACIS;CEI
NIPAL4	ENSG00000172548	AISP
NIPBL	ENSG00000164190	AISP
NKX2-1	ENSG00000136352	AST
NLRC4	ENSG00000091106	ACIS;AISP
NLRP1	ENSG00000091592	AIL;AAP;AHA

Gene symbol	Ensembl gene ID	Source
NLRP12	ENSG00000142405	AISP
NLRP3	ENSG00000162711	ACIS;AISP
NME1	ENSG00000239672	DIA
NME8	ENSG00000086288	AISP
NOD2	ENSG00000167207	AISP
NOP10	ENSG00000182117	ACIS;CEI
NOTCH1	ENSG00000148400	ACIS
NOTCH2	ENSG00000134250	AISP
NOTCH3	ENSG00000074181	AISP
NPAP1	ENSG00000185823	AISP
NPC1	ENSG00000141458	ACIS
NPC2	ENSG00000119655	ACIS
NPHP3	ENSG00000113971	AIS
NPHS1	ENSG00000161270	AISP
NR3C1	ENSG00000113580	AISP
NR3C2	ENSG00000151623	DIA
NRAS	ENSG00000213281	AIL;AT
NRTN	ENSG00000171119	AISP;DIA
NSD1	ENSG00000165671	AISP
NSMCE2	ENSG00000156831	AISP
NSUN2	ENSG00000037474	ACIS;AST;CD
NTRK1	ENSG00000198400	AISP
NUMA1	ENSG00000137497	ACIS
NUP107	ENSG00000111581	AISP
NUP214	ENSG00000126883	ACIS
NXN	ENSG00000167693	AISP
OCLN	ENSG00000197822	AIS
OCRL	ENSG00000122126	AISP
OFD1	ENSG00000046651	AISP
OPLAH	ENSG00000178814	AISP;DIA
ORAI1	ENSG00000182500	IDEF
ORC6	ENSG00000091651	AISP
ORMDL3	ENSG00000172057	[341]
OSGEP	ENSG00000092094	AISP
OSTM1	ENSG00000081087	ACIS
OTC	ENSG00000036473	AIS
PAH	ENSG00000171759	AISP
PALB2	ENSG00000083093	ACIS;AISP;INTD
PALLD	ENSG00000129116	AIS;INTD
PAPSS2	ENSG00000198682	AISP
PARN	ENSG00000140694	ACIS;CEI
PARP14	ENSG00000173193	[347]

Gene symbol	Ensembl gene ID	Source
PAX6	ENSG00000007372	AISP
PCCA	ENSG00000175198	ACIS;AISP
PCCB	ENSG00000114054	ACIS;AISP
PCNT	ENSG00000160299	AISP
PCSK1	ENSG00000175426	DIA
PCYT1A	ENSG00000161217	AISP
PDE4D	ENSG00000113448	AISP
PDE4DIP	ENSG00000178104	[343]
PDGFRA	ENSG00000134853	ACIS;AISP
PDGFRB	ENSG00000113721	ACIS
PDGFRL	ENSG00000104213	AISP
PDX1	ENSG00000139515	AIS
PEPD	ENSG00000124299	AST;AUTO
PEX2	ENSG00000164751	AIS
PEX5	ENSG00000139197	AST
PEX7	ENSG00000112357	AIS
PGM3	ENSG00000013375	ACIS;AR;AST;IDEF
PHKB	ENSG00000102893	DIA
PHKG2	ENSG00000156873	AIS
PHYH	ENSG00000107537	AIS
PIEZO1	ENSG00000103335	AISP
PIGA	ENSG00000165195	ACIS;AISP
PIGL	ENSG00000108474	ACIS
PIGM	ENSG00000143315	AIS
PIGT	ENSG00000124155	DIA
PIH1D3	ENSG00000080572	AISP
PIK3CA	ENSG00000121879	ACIS;CEI
PIK3CD	ENSG00000171608	AIL;IDEF
PIK3R1	ENSG00000145675	AIL;DIA;IDEF
PKD2	ENSG00000118762	AISP
PKHD1	ENSG00000170927	AIS
PKLR	ENSG00000143627	AIS
PKP1	ENSG00000081277	CD;IDEF
PLCD1	ENSG00000187091	AISP
PLCG2	ENSG00000197943	AIL;AR;AST;AUTO
PLEC	ENSG00000178209	AISP;INTD
PLG	ENSG00000122194	AISP
PLN	ENSG00000198523	ACIS
PLOD1	ENSG00000083444	AISP
PLP1	ENSG00000123560	AISP
PLXND1	ENSG00000004399	AISP
PMM2	ENSG00000140650	AIL;DIA

Gene symbol	Ensembl gene ID	Source
PMP22	ENSG00000109099	AIS
PMS2	ENSG00000122512	ACIS
PNLIP	ENSG00000175535	CD
PNP	ENSG00000198805	AHA;AN;AT
PNPLA1	ENSG00000180316	AISP
POLA1	ENSG00000101868	ACIS;AISP;DIA
POLE	ENSG00000177084	IDEF
POLG	ENSG00000140521	INTD
POLH	ENSG00000170734	AISP
POLR1C	ENSG00000171453	AIS
POLR1D	ENSG00000186184	AIS
POR	ENSG00000127948	AISP
POT1	ENSG00000128513	AIS
POU2AF1	ENSG00000110777	AIL;AAP;IH
POU6F2	ENSG00000106536	AIS
PPARG	ENSG00000132170	IDEF
PPP2R5D	ENSG00000112640	CD
PRDM16	ENSG00000142611	ACIS
PRF1	ENSG00000180644	ACIS;AISP
PRG4	ENSG00000116690	AISP
PRKACA	ENSG00000072062	AISP
PRKAR1A	ENSG00000108946	AISP
PRKCD	ENSG00000163932	AIL;AT;IDEF
PRKDC	ENSG00000253729	SCI
PRPS1	ENSG00000147224	IDEF
PRSS1	ENSG00000204983	ACIS;AISP
PRSS2	ENSG00000262739	ACIS;AISP
PRTN3	ENSG00000196415	ACIS;AUTO
PSAP	ENSG00000197746	AISP
PSEN1	ENSG00000080815	ACIS;AISP
PSEN2	ENSG00000143801	ACIS
PSENEN	ENSG00000205155	AISP
PSMB8	ENSG00000204264	AIL
PSTPIP1	ENSG00000140368	AIL
PTEN	ENSG00000171862	AUTO;CEI;DIA
PTH1R	ENSG00000160801	AIS
PTPLA	ENSG00000165996	AISP
PTPN11	ENSG00000179295	ACIS
PTPN2	ENSG00000175354	AAP
PTPN22	ENSG00000134242	ACIS;AAP
PTPRC	ENSG00000081237	ACIS;DIA;SCI
PTRF	ENSG00000177469	AIL

Gene symbol	Ensembl gene ID	Source
PVRL1	ENSG00000110400	AISP
PWRN1	ENSG00000259905	AISP
RAB23	ENSG00000112210	AIS
RAB27A	ENSG00000069974	ACIS;IDEF
RAB3GAP2	ENSG00000118873	AISP
RAB7A	ENSG00000075785	AISP
RAC2	ENSG00000128340	ACIS;IDEF
RAD21	ENSG00000164754	AISP
RAD50	ENSG00000113522	[341]
RAD51	ENSG00000051180	ACIS;AISP
RAD51C	ENSG00000108384	ACIS;AISP
RAF1	ENSG00000132155	ACIS
RAG1	ENSG00000166349	AIL;AHA;AN;CD;SCI
RAG2	ENSG00000175097	AIL;AUTO;CD;SCI
RAI1	ENSG00000108557	AISP
RAP1A	ENSG00000116473	AISP
RAP1B	ENSG00000127314	AISP
RARA	ENSG00000131759	ACIS
RARB	ENSG00000077092	AIS
RASA2	ENSG00000155903	AIS
RB1	ENSG00000139687	ACIS
RBCK1	ENSG00000125826	IDEF
RBFOX1	ENSG00000078328	[345]
RBM20	ENSG00000203867	ACIS
RBM8A	ENSG00000131795	AIL;CMA
RBP4	ENSG00000138207	AISP
RBPJ	ENSG00000168214	ACIS
RECQL4	ENSG00000160957	ACIS;DIA
RERE	ENSG00000142599	AIS
REST	ENSG00000084093	AIS
RET	ENSG00000165731	AISP;DIA
REV3L	ENSG00000009413	AISP
RFC2	ENSG00000049541	AISP
RFWD3	ENSG00000168411	ACIS;AISP
RFX5	ENSG00000143390	AIL;PD
RFX6	ENSG00000185002	DIA
RFXANK	ENSG00000064490	AIL;PD
RFXAP	ENSG00000133111	AIL;PD
RHAG	ENSG00000112077	AIS
RIPPLY2	ENSG00000203877	AISP
RIT1	ENSG00000143622	AIS
RMRP	ENSG00000269900	AIL;AUTO;CEI;CD

Gene symbol	Ensembl gene ID	Source
RNASEH2A	ENSG00000104889	AIS
RNASEH2C	ENSG00000172922	AIS
RNF113A	ENSG00000125352	AIL;CD
RNF125	ENSG00000101695	AISP
RNF168	ENSG00000163961	AIL;IDEF
RNF6	ENSG00000127870	AIS
RNU4ATAC	ENSG00000264229	ACIS;AISP
ROR2	ENSG00000169071	AISP
RORC	ENSG00000143365	AIS
RPGR	ENSG00000156313	AISP
RPGRIP1	ENSG00000092200	AIS
RPGRIP1L	ENSG00000103494	AIS
RPL10	ENSG00000147403	AISP
RPL11	ENSG00000142676	ACIS;AISP
RPL15	ENSG00000174748	ACIS
RPL18	ENSG00000063177	ACIS
RPL26	ENSG00000161970	ACIS
RPL27	ENSG00000131469	ACIS
RPL35	ENSG00000136942	ACIS
RPL35A	ENSG00000182899	ACIS
RPL5	ENSG00000122406	ACIS
RPS10	ENSG00000124614	ACIS
RPS17	ENSG00000184779	ACIS
RPS19	ENSG00000105372	ACIS
RPS24	ENSG00000138326	ACIS
RPS26	ENSG00000197728	ACIS
RPS27	ENSG00000177954	ACIS
RPS28	ENSG00000233927	ACIS
RPS29	ENSG00000213741	ACIS
RPS7	ENSG00000171863	ACIS
RPSA	ENSG00000168028	AIS
RRAS	ENSG00000126458	AIS
RREB1	ENSG00000124782	AST;AUTO;IDEF
RRM2B	ENSG00000048392	DIA
RSPH1	ENSG00000160188	AISP
RSPH3	ENSG00000130363	AISP
RSPH4A	ENSG00000111834	AISP
RSPH9	ENSG00000172426	AISP
RTEL1	ENSG00000258366	AIL;CEI
RUNX1	ENSG00000159216	ACIS
RUNX2	ENSG00000124813	AISP
RYR1	ENSG00000196218	AISP

Gene symbol	Ensembl gene ID	Source
SAA1	ENSG00000173432	CD
SALL4	ENSG00000101115	ACIS
SAMD9	ENSG00000205413	ACIS;AISP;CD
SAMD9L	ENSG00000177409	AIL
SAR1B	ENSG00000152700	DIA
SARS2	ENSG00000104835	ACIS
SAT1	ENSG00000130066	AISP
SBDS	ENSG00000126524	ACIS;AISP
SC5D	ENSG00000109929	AIS
SCARB2	ENSG00000138760	AIL
SCN11A	ENSG00000168356	DIA
SCN4A	ENSG00000007314	AST
SCN5A	ENSG00000183873	ACIS
SCN9A	ENSG00000169432	ACIS;AISP;DIA
SCNN1A	ENSG00000111319	AISP;DIA
SCNN1B	ENSG00000168447	AISP;DIA
SCNN1G	ENSG00000166828	AISP;DIA
SCYL1	ENSG00000142186	AIS
SDCCAG8	ENSG00000054282	AST
SDHA	ENSG00000073578	ACIS;AISP
SDHB	ENSG00000117118	CEI
SDHC	ENSG00000143252	CEI
SDHD	ENSG00000204370	AST;CEI;PD
SEC23B	ENSG00000101310	AUTO;CEI
SEC24C	ENSG00000176986	AST;AUTO;IDEF
SEC61A1	ENSG00000058262	ACIS
SEMA3C	ENSG00000075223	AISP;DIA
SEMA3D	ENSG00000153993	AISP;DIA
SEMA3E	ENSG00000170381	ACIS;AISP
SEPN1	ENSG00000162430	AISP
SERAC1	ENSG00000122335	AISP
SERPINA1	ENSG00000197249	AISP
SERPING1	ENSG00000149131	AUTO;DIA
SETD2	ENSG00000181555	AISP
SETD5	ENSG00000168137	AISP
SETX	ENSG00000107290	AIL
SF3B1	ENSG00000115524	AISP
SFTPA2	ENSG00000185303	AIL
SFTPC	ENSG00000168484	AIL
SGCD	ENSG00000170624	ACIS
SGCG	ENSG00000102683	AISP
SGSH	ENSG00000181523	AISP;DIA

Gene symbol	Ensembl gene ID	Source
SH2B3	ENSG00000111252	ACIS
SH2D1A	ENSG00000183918	AIL;CEI
SH3PXD2B	ENSG00000174705	AISP
SHANK3	ENSG00000251322	IDEF
SHH	ENSG00000164690	AST
SHOC2	ENSG00000108061	AISP
SHPK	ENSG00000197417	AISP
SI	ENSG00000090402	DIA
SIK1	ENSG00000142178	AISP
SIN3A	ENSG00000169375	IDEF
SKI	ENSG00000157933	AIS
SKIV2L	ENSG00000204351	DIA;IDEF
SLC10A2	ENSG00000125255	CD
SLC12A1	ENSG00000074803	DIA
SLC17A5	ENSG00000119899	ACIS
SLC19A2	ENSG00000117479	DIA
SLC25A13	ENSG00000004864	AISP
SLC25A15	ENSG00000102743	AISP
SLC25A22	ENSG00000177542	AISP
SLC26A2	ENSG00000155850	AISP
SLC26A3	ENSG00000091138	DIA
SLC27A4	ENSG00000167114	ACIS
SLC29A3	ENSG00000198246	ACIS;AISP
SLC2A1	ENSG00000117394	AIS
SLC2A10	ENSG00000197496	AISP
SLC30A2	ENSG00000158014	AISP
SLC35A1	ENSG00000164414	ACIS;AISP
SLC35A2	ENSG00000102100	AISP
SLC35C1	ENSG00000181830	ACIS;AISP
SLC37A4	ENSG00000137700	ACIS;AISP
SLC39A4	ENSG00000147804	AISP;CD
SLC39A8	ENSG00000138821	AISP
SLC3A1	ENSG00000138079	AISP
SLC46A1	ENSG00000076351	AIL;DIA;IDEF
SLC4A1	ENSG00000004939	AIS
SLC4A11	ENSG00000088836	AIS
SLC52A3	ENSG00000101276	AISP
SLC5A1	ENSG00000100170	CD
SLC6A19	ENSG00000174358	AISP
SLC6A5	ENSG00000165970	AISP
SLC7A7	ENSG00000155465	ACIS;AISP;DIA
SLC7A9	ENSG00000021488	AISP

Gene symbol	Ensembl gene ID	Source
SLC9A3	ENSG00000066230	AISP;SECD
SLCO2A1	ENSG00000174640	AISP
SLX4	ENSG00000188827	ACIS;AISP
SMAD4	ENSG00000141646	AIS;INTD
SMARCA2	ENSG00000080503	AISP
SMARCA4	ENSG00000127616	AISP
SMARCAD1	ENSG00000163104	AISP
SMARCAL1	ENSG00000138375	AIL;CEI
SMARCB1	ENSG00000099956	AISP
SMARCE1	ENSG00000073584	AISP
SMC1A	ENSG00000072501	AISP
SMC3	ENSG00000108055	AISP
SMN1	ENSG00000172062	AISP
SMPD1	ENSG00000166311	ACIS;AISP
SNORD115-1	ENSG00000201831	AISP
SNORD116-1	ENSG00000207063	AISP
SNRPN	ENSG00000128739	AISP
SNX10	ENSG00000086300	AISP
SOS1	ENSG00000115904	AIS
SOS2	ENSG00000100485	AIS
SOX10	ENSG00000100146	AIS
SOX11	ENSG00000176887	AISP
SOX18	ENSG00000203883	AISP
SP110	ENSG00000135899	AIL;IDEF
SPAG1	ENSG00000104450	AISP
SPATA5	ENSG00000145375	IDEF
SPIB	ENSG00000269404	AIL;AAP;IH
SPINK1	ENSG00000164266	ACIS;AISP
SPINK5	ENSG00000133710	AIL;AR;AST
SPINT2	ENSG00000167642	SECD
SPTB	ENSG00000070182	AIS
SPTLC1	ENSG00000090054	AISP
SPTLC2	ENSG00000100596	AISP
SRCAP	ENSG00000080603	IH
SRD5A3	ENSG00000128039	AISP
SRP54	ENSG00000100883	ACIS;AISP
SRY	ENSG00000184895	AISP
STAT1	ENSG00000115415	AISPP;AAP;AHA;AN
STAT3	ENSG00000168610	AIL;AHA;AT;IH
STAT4	ENSG00000138378	AAP;IH
STAT6	ENSG00000166888	[341]
STEAP3	ENSG00000115107	AIS

Gene symbol	Ensembl gene ID	Source
STIM1	ENSG00000167323	AHA;IDEF
STK36	ENSG00000163482	AISP
STK4	ENSG00000101109	ACIS;IDEF
STOM	ENSG00000148175	AIS
STRA6	ENSG00000137868	AIS
STS	ENSG00000101846	ACIS
STX11	ENSG00000135604	ACIS
STX16	ENSG00000124222	AISP
STX1A	ENSG00000106089	AIL;IDEF
STX3	ENSG00000166900	DIA
STXBP2	ENSG00000076944	ACIS;AISP
SUGCT	ENSG00000175600	DIA
SULT2B1	ENSG00000088002	AISP
SUMF1	ENSG00000144455	AIS
SUOX	ENSG00000139531	AISP
TACR1	ENSG00000115353	[348]
TADA2A	ENSG00000108264	AIL;AAP;IDEF
TAF1	ENSG00000147133	AISP
TAL1	ENSG00000162367	ACIS
TAL2	ENSG00000186051	ACIS
TALDO1	ENSG00000177156	AST
TAP1	ENSG00000168394	AISP
TAP2	ENSG00000204267	AISP
TAPBP	ENSG00000231925	AISP
TAZ	ENSG00000102125	ACIS;AISP
TBCE	ENSG00000116957	CEI
TBL2	ENSG00000106638	AISP
TBX1	ENSG00000184058	AST;AUTO;IDEF
TBX19	ENSG00000143178	AISP
TBX21	ENSG00000073861	AIA
TBX4	ENSG00000121075	AISP
TBX6	ENSG00000149922	AISP
TBXAS1	ENSG00000059377	ACIS;AISP
TCAP	ENSG00000173991	ACIS
TCF3	ENSG00000071564	AIL;DIA;IDEF
TCF4	ENSG00000196628	AIL;AUTO;IH
TCIRG1	ENSG00000110719	AISP
TCN2	ENSG00000185339	AIL;DIA
TCOF1	ENSG00000070814	AIS
TCTN2	ENSG00000168778	AIS
TCTN3	ENSG00000119977	AISP
TEK	ENSG00000120156	AISP

Gene symbol	Ensembl gene ID	Source
TERC	ENSG00000270141	ACIS;CEI
TERF2IP	ENSG00000166848	AIS
TERT	ENSG00000164362	AIL;CEI
TET2	ENSG00000168769	ACIS
TF	ENSG00000091513	AISP
TFAP2A	ENSG00000137203	AIS
TFR2	ENSG00000106327	ACIS
TFRC	ENSG00000072274	AIL;CD
TGDS	ENSG00000088451	AISP
TGFB1	ENSG00000105329	AIL;IDEF
TGFBR2	ENSG00000163513	AIS
TGM1	ENSG00000092295	AISP
TGM5	ENSG00000104055	ALL
THOC6	ENSG00000131652	AISP
THPO	ENSG00000090534	ACIS
TINF2	ENSG00000092330	ACIS;CEI
TKT	ENSG00000163931	AISP
TLR4	ENSG00000136869	IH
TMC6	ENSG00000141524	AISP
TMC8	ENSG00000167895	AISP
TMEM107	ENSG00000179029	AIS
TMEM173	ENSG00000184584	AIL
TMEM216	ENSG00000187049	AIS
TMEM231	ENSG00000205084	AIS
TMEM67	ENSG00000164953	AIS
TMPO	ENSG00000120802	ACIS
TMPRSS15	ENSG00000154646	DIA
TNF	ENSG00000232810	[341]
TNFAIP3	ENSG00000118503	ACIS;AAP
TNFRSF11A	ENSG00000141655	AIL
TNFRSF13B	ENSG00000240505	AIL;AT;DIA;IDEF
TNFRSF13C	ENSG00000159958	AIL;AT;DIA;IDEF
TNFRSF1A	ENSG00000067182	ACIS;AISP;DIA
TNFRSF1B	ENSG00000028137	AIL;IDEF
TNFRSF4	ENSG00000186827	IDEF
TNFSF11	ENSG00000120659	AISP
TNFSF12	ENSG00000239697	AIL;AT;IDEF
TNFSF15	ENSG00000181634	AIL;AAP;IH
TNNC1	ENSG00000114854	ACIS
TNNI3	ENSG00000129991	ACIS
TNNT2	ENSG00000118194	ACIS
TNPO3	ENSG00000064419	AIL;AAP;IH

Gene symbol	Ensembl gene ID	Source
TNXB	ENSG00000168477	AISP
TP53	ENSG00000141510	ACIS;AISP;INTD
TP53RK	ENSG00000172315	AISP
TP63	ENSG00000073282	AISP
TPI1	ENSG00000111669	AISP
TPM1	ENSG00000140416	ACIS
TPM2	ENSG00000198467	AISP
TPM3	ENSG00000143549	AISP
TPP2	ENSG00000134900	ACIS;AHA;AT
TPRKB	ENSG00000144034	AISP
TRAC	ENSG00000229164	AUTO
TRAF3IP2	ENSG00000056972	AISP
TRAF6	ENSG00000175104	AISP
TRAIP	ENSG00000183763	AST
TREH	ENSG00000118094	DIA
TREM2	ENSG00000095970	ACIS
TREX1	ENSG00000213689	AAP
TRIO	ENSG00000038382	AISP
TRIP13	ENSG00000071539	ACIS;AISP
TRNT1	ENSG00000072756	AIL
TRPM1	ENSG00000134160	AISP
TRPS1	ENSG00000104447	AISP
TSC1	ENSG00000165699	AISP
TSC2	ENSG00000103197	AISP
TSHR	ENSG00000165409	DIA
TSR2	ENSG00000158526	ACIS
TTC25	ENSG00000204815	AISP
TTC37	ENSG00000198677	AIS;INTD
TTC7A	ENSG00000068724	AHA;DIA;SCI
TTN	ENSG00000155657	ACIS
TTR	ENSG00000118271	DIA
TXNRD2	ENSG00000184470	ACIS
TYK2	ENSG00000105397	AIL;AAP;IDEF
TYMP	ENSG00000025708	INTD
TYROBP	ENSG00000011600	ACIS
UBAC2	ENSG00000134882	IH
UBE2T	ENSG00000077152	ACIS;AISP
UCP2	ENSG00000175567	DIA
UFD1L	ENSG00000070010	AST;AUTO;IDEF
UMPS	ENSG00000114491	AISP
UNC119	ENSG00000109103	ACIS;IDEF
UNC13D	ENSG00000092929	ACIS;AISP

Gene symbol	Ensembl gene ID	Source
UNG	ENSG00000076248	AIL;IDEF
UROC1	ENSG00000159650	AISP
UROS	ENSG00000188690	IDEF
USB1	ENSG00000103005	ACIS;CEI
USP8	ENSG00000138592	IDEF
USP9X	ENSG00000124486	AISP
VANGL1	ENSG00000173218	AISP
VCL	ENSG00000035403	ACIS
VHL	ENSG00000134086	AIS
VIPAS39	ENSG00000151445	AISP
VPS13A	ENSG00000197969	AISP
VPS13B	ENSG00000132549	ACIS
VPS33A	ENSG00000139719	AISP
VPS33B	ENSG00000184056	AISP
VPS45	ENSG00000136631	AIL
WAS	ENSG00000015285	AIL;AUTO;CD
WDPCP	ENSG00000143951	AIS
WDR19	ENSG00000157796	AISP
WDR34	ENSG00000119333	AISP
WDR73	ENSG00000177082	AISP
WFS1	ENSG00000109501	AISP
WHSC1	ENSG00000109685	IDEF
WIPF1	ENSG00000115935	ACIS;AUTO;CD
WISP2	ENSG00000064205	AUTO
WNT3	ENSG00000108379	AIS
WNT4	ENSG00000162552	AISP
WRAP53	ENSG00000141499	ACIS;CEI
WT1	ENSG00000184937	AIS
WWOX	ENSG00000186153	AIS
XDH	ENSG00000158125	AISP
XIAP	ENSG00000101966	AIL;CEI
XK	ENSG00000047597	AIS
XPA	ENSG00000136936	AISP
XPC	ENSG00000154767	AISP
XPNPEP3	ENSG00000196236	AISP
XRCC2	ENSG00000196584	ACIS;AISP
XRCC4	ENSG00000152422	ACIS;SCI
ZAP70	ENSG00000115085	AIL;DIA
ZBTB24	ENSG00000112365	AIL;CEI
ZIC3	ENSG00000156925	AIS
ZMPSTE24	ENSG00000084073	AISP
ZMYND10	ENSG00000004838	AISP

Gene symbol	Ensembl gene ID	Source
ZNF750	ENSG00000141579	AISP
ZNHIT3	ENSG00000108278	AISP