University of Genoa
Health Science Department
Biostatistics and Methods for Big Data Analysis

Study of susceptibility factors, modification and prognosis in the development of Hirschsprung's disease-associated enterocolitis via omics analysis of data from whole-exome sequencing studies, genome-wide association studies and proteomics.

Doctoral thesis

PhD candidate: Francesca Rosamilia

Dissertation Committee:

Professor Francesca Lantieri, Supervisor

Professor Suzanne M. Leal, co-Supervisor

Contents

Chapter 1

Abstract

Hirschsprung's disease (HSCR) is a congenital gut malformation caused by a lack of innervation. One of the most serious is HSCR associated enterocolitis (HAEC), a potentially lethal condition with 30% of incidence. However, the causes of HAEC are still unknown and the onset is difficult to predict. This thesis work focuses on the study of susceptibility factors, modifications, and prognosis in the development of Hirschsprung's (HSCR) disease-associated enterocolitis (HAEC) using different omics technologies. Genetic investigation of HAEC by Whole-Exome Sequencing (WES) on 24 HSCR patients affected (HAEC) or not affected (HSCR-only) by enterocolitis and replication of results on a larger panel of patients allowed the identification of the HAEC susceptibility variant p.H187Q in the Oncostatin-M receptor (OSMR) gene (14.6% in HAEC and 5.1% in HSCR-only,p= 0.0024). Proteomic analysis on the lymphoblastoid cell lines from one HAEC patient homozygote for this variant and one HAEC wild-type patient revealed two well distinct clusters of proteins significantly up or downregulated upon OSM stimulation. I have then carried out a transcriptome analysis on Intraepithelial lymphocytes (IEL) derived from gut biopsies and on Peripheral Blood Monocyte Cells (PBMCs) from HAEC, HSCR-only and pediatric patients affected by neither Hirschsprung, nor inflammatory related diseases. The analysis on the IELs showed a clear clustering between the groups of patients and an enrichment in immune and inflammatory pathways in the HAEC group. The results showed an interesting role for the gene Oncostatin M (OSMR) as a potential modifier from HSCR to HAEC. Additionally, a potential genetic connection between Inflammatory Bowel Disease and HSCR using the UKBioBank was explored. The finding of a shared genetic background with other inflammatory disorders affecting the gut, and the evaluation of gene networks and pathways involved in inflammation provides further knowledge on the mechanisms of gut inflammation.

Chapter 2

Introduction

Modern genetics has significantly influenced medicine by elucidating causes or identifying differential diagnosis for many diseases. This field has allowed the identification of large chromosomal defects associated with a disease and mutant genes responsible for "Mendelian" or "single gene" disorders, with more than 1,500 traits, as reported in the OMIM database (Online Mendelian Inheritance in Man, OMIM) [1]. These diseases, often rare, exhibit characteristic transmission patterns such as dominant, recessive, or X-linked (OMIM) [2].

In contrast, numerous common diseases, including chronic and adult onset disorders, such as coronary heart disease, hypertension, diabetes, obesity, cancers, Alzheimer's, and Parkinson's disease [3] display familial aggregation not following Mendelian patterns. These conditions result from multiple genes interacting with environmental factors [3]. The origin of complex diseases can be attributed to a single mutant gene transmitted through Mendelian inheritance for only a minority (less than 1%–7%) of affected individuals [2].

It is therefore possible to classify genetic diseases into Mendelian and complex disorders. Mendelian disorders arise from single genes rare variants, while complex diseases involve contributions from numerous independent or interacting common DNA variants, called polymorphisms, in several genes and other environmental factors. These polymorphisms are positions in the DNA where two or more different alleles can be present with a high frequency. The individual impact of each variant on polygenic and complex diseases can be modest, influencing the clinical diversity and therapeutic response. Polygenic disorders, more prevalent than monogenic ones in humans, pose significant social and economic burdens, yet are much more difficult to be investigated and their molecular genetics remain often largely not understood [4].

The search for the genetic causes of polygenic and complex diseases traditionally involves two strategies: candidate gene search, under the hypothesis that a few known genes are biologically likely to be involved in the disorder onset, or a hypothesis free approach, when the whole genome is scanned to search for the unknown genes and variants involved. Both approaches can rely on the genotyping of specific and already known, usually uniformly distributed, single nucleotide polymorphisms (SNP) or on the sequencing of entire DNA regions. The first method of genotyping, when applied at the whole genome scale, usually requires as the most adopted technique, the use of microscopic DNA spots containing hundreds of thousands of markers attached to a solid surface (called SNParray); the second method aim at the sequencing of specific DNA regions or even the entire exome or genome (Whole Exome or Whole Genome sequencing) to search for every variant possibly present. The earlier developed technology of marker genotyping has allowed the finding of several genes and variants involved in many diseases through the statistical approaches of linkage and association analysis. The linkage analysis, developed in the 50s, aimed to find the gene that caused a specific disease by mapping loci segregating with a disease within families through the characterization of genetic highly polymorphic markers. Association analyses were instead developed successively, when it became possible to easily screen SNPs in the DNA and allowed to detect specific alleles associated to a disease within a population, comparing their frequencies between individual affected by the disease and healthy controls (case-control studies), or within families, by investigating which allele was transmitted from the parents to the offspring, through statistical methods.

While genotyping methods were based on specific pre-selected and usually common markers, the newer developed technology of Next Generation Sequencing (NGS) has allowed the rapid and affordable sequencing of the whole exome or genome of an individual, opening up to the era of genomics, the study of the whole sequencing information [5]. NGS has also allowed more complex DNA variations than SNPs. The genome harbors numerous variants, ranging from benign and protective to potentially harmful. These variants include simple nucleotide variations (SNVs), such as single nucleotide changes and small insertions/deletions, and structural variations (SVs), encompassing larger indels, copy number variants (CNVs), and inversions. To navigate the intricate genetic architecture, alongside with genomics, other omics sciences have gained increasing prominence. The term ”omics” refers to disciplines focused on characterizing and quantifying pools of biological molecules. Researchers are increasingly shifting their focus from individual molecules to the totality of molecules [6].

Genomics, transcriptomics, proteomics: these terms are becoming more prevalent in recent literature, reflecting the central dogma of molecular biology, which passes from the DNA

to proteins through RNA. The DNA contains the information for an organism to develop and function, but this information must be "read" to take effect. DNA is first transcribed to RNA, in particular to messenger RNA (mRNA), which refers to the protein-coding part of an organism's genome, and to noncoding RNA molecules such as transfer RNA (tRNA), ribosomal RNA (rRNA) and other. This set of RNA molecules constitutes the transcriptome. While non-coding transcripts may serve to influence cell structure and to regulate genes, the mRNA transcripts are delivered to ribosomes, and translated to amino acids sequences that form the proteins. This set of proteins expressed by an organism is the proteome. While the genome is the same in every cell of an organism, the transcriptome and the proteome differ depending on the tissue and other factors that influence the level of the gene's expression. Moreover, proteins undergo diverse modifications post-formation, leading to their activity modulation by numerous factors beyond the expression level of the corresponding gene. The genome and the proteome are deemed to be more intricately linked to the ultimate phenotype, particularly concerning diseases. Genomics, transcriptomics, and proteomics respectively study the entirety of genetic information (DNA), messenger RNAs (mRNA), and cell proteins.

Genomic data have been generated with increasing speed and efficiency, allowing the transition from studies focused on individual genes to comparing genomes of whole populations [7]. Variants in coding regions may impact protein sequence, while those in non-coding regions likely influence gene expression and splicing processes. Transcriptome profiling, a pivotal technology for assessing gene expression within a transcriptome, is based on the count of the number of transcripts, performed on the cDNA converted from the RNA extracted from the cell for reasons of stability. While it was initially based on the hybridization of the cDNA on solid surfaces containing microscopic spots of specific DNA used as probes (called microarray), it has undergone a transformative shift with the advent of next-generation sequencing (NGS) technologies [8]. Ampliseq is the Ion Torrent (Thermo Fisher Scientific) approach for the target sequencing in RNA seq. The major difference between AmpliSeq and the whole transcriptome RNA sequencing methods is that AmpliSeq is designed to profile over 20,000 distinct human already known RNA targets using a highly multiplexed amplification method, therefore AmpliSeq will be denoted as RNA-seq targeted. Each amplicon represents a unique targeted gene. The average size of each amplicon is 150 bp. Because of the targeted nature and small amplicon size, the total number of raw reads needed for Differential Expression Gene (DEG) analysis for each library prepared with RNA-seq targeted is much smaller than typical whole-transcriptome RNA sequencing. Unlike earlier methods, which could provide information uniquely on already known genes, whole RNA-Seq enables to screen every transcript present in the cells. Proteomics data usually rely on mass-spectrometry technologies that measure the mass of the peptides obtained from the cell's lysate, giving information

on the proteins identified, their chemical modifications, and their structure. Molecular biology, driven by omics sciences, is increasingly shaping clinical practice towards the patient's disease-specific underlying biology. Omics sciences enable a transition from a "generalized" to an "individualized" approach, aligning medicine with the patient's unique needs through tailored care and more precise clinical management [6].

Pharmacogenomics, toxicogenomics, nutritional genomics: these are examples of omics "spin-offs" or translational medicine with growing significance for clinicians. To advance personalized or target therapy, a wider perspective is necessary, considering the patient's biological entirety—from the smallest molecular detail to an epidemiological outlook that situates them in reality. The opportunities provided by investigating health and disease at the omics scale come with the need for implementing a novel modus operandi to address data generation, analysis and sharing and new statistical challenges. It is critical to recognize that omics data need to be analyzed and interpreted as a whole through effective and integrative pipelines. This clearly requires the cooperation of multidisciplinary teams as well as the fundamental support of bioinformatics and biostatistics [9].

Genomics, Transcriptomics and Proteomics analysis necessitates judicious data reduction strategies due to the significantly higher number of variables/features than samples.

Many tools are available for handling genome-wide variant data such as other omics data (e.g. Plink [10], and a variety of R packages [11], including the Bioconductor project [12]) supporting the whole workflow from quality control (QC) of raw data to analysis, such as association, genetic risk scoring, case-control comparisons and burden analyses. Almost all omics applications produce a very large amount of data involving industrial-scale processes for which systems of quality control and quality assurance (QC/QA) are essential.

False negative results may be increased by failure to control various experimental factors, leading to 'noise' in the system and thereby reducing power. Such factors include low quality samples, poorly-performing assays, and errors in sample identification. It is becoming more and more essential to use rigorous statistical methods to dig among the huge amount of data. It is imperative to establish specific quality measures to eliminate potential noise.

All omics data undergo essential steps of specific data preprocessing, statistical analysis, and functional interpretation. The dimensionality of data requires careful consideration in feature selection approaches, as the abundance of variables demands substantial multiple tests correction for a reliable analysis.

The analysis of omics through case-control studies plays a crucial role in understanding

the molecular foundations of various pathological conditions. These studies help identify variants with different frequencies between cases and controls, and differentially expressed genes or proteins that may be associated with the specific conditions under investigation. However, to ensure robust conclusions from such studies, large cohort sizes and results validation and replication become imperative. The complexity and vastness of omics data necessitate numerous comparisons to achieve reliable and meaningful results. Obtaining samples of substantial sizes becomes, therefore, a significant yet essential challenge to ensure the statistical power required to identify significant differences and draw reliable conclusions regarding molecular associations in the various pathological conditions under study [13].

Data produced by omics strategies is enormous indeed, so that is often defined as "big data". Through world-wide collaboration and sharing, data generated by omics analyses are being quickly accumulating to very large sizes, necessitating appropriate storage and analysis infrastructures to handle this volume of information. Statistical analysis serves as a guiding force in identifying biologically relevant hits and generating hypotheses. Integration of various external databases, annotation sources, and multiple omics types enhances the power of statistical analysis, facilitating seamless data integration.

Moreover, it is necessary to have the instruments to understand what big data says. For instance, When considering genomic data, a whole human genome, which contains around millions base pairs, can be approximately 1 GB in size per person. The transcriptome, obtained through RNA-seq techniques, can generate several gigabytes (GB) of data per sample, depending on the coverage depth and read length used. An RNA-seq experiment on a single sample can occupy from tens of GB to several hundreds of GB. Concerning the proteome, the amount of data can vary depending on the technique used. Proteome analysis can generate data ranging from several megabytes (MB) to several gigabytes per individual sample, depending on the sample's complexity and the resolution of the analysis.

In summary, the evolving landscape of genomics, transcriptomics and proteomics data analysis underscores the importance of robust methodologies, sophisticated statistical analyses, and a deep understanding of data-driven approaches, leading advancements in biological and clinical research.

The potential of genomic technologies to identify individuals at risk of genetic disease is enormous, but incomplete penetrance and variable expressivity represent a challenge for doctors, especially when there is a different phenotype in terms of age of onset, making it difficult to know when a clinical phenotype will develop and, if so, how. Indeed, the cause of the variability in genotype-phenotype correlations may be difficult to elucidate. In

particular, there may be factors that influence penetrance and expressivity, such as specific causal variants with more distinctive effects up to so-called global modifiers that can have overall effects on gene expression. Furthermore, some mechanisms could underlie disease resistance processes that we could identify through sequencing of general population cohorts. However, ascertainment errors can occur with some study designs, with volunteer population cohorts tending to be healthier than the average individual [14] and clinical cohorts tending to have more severe phenotypes. Despite the growing number of sequenced individuals, the identification of genetic modifiers for diseases known as monogenic remains challenging. However, for the study of modifier variants as a global definition, two main models have been studied. For the disease threshold model, we know from the literature that some deleterious monogenic variants are sufficient to cause the disease to manifest on their own and do not require any genetic modifier to cause the phenotype while, there may be a threshold that must be met for disease manifestation of a clinical disease phenotype and other genetic factors may vary in their relative contribution to reaching this threshold for different diseases and in different individuals [15]. Furthermore, the implementation and use of polygenic risk scores in clinical utility have been widely debated recently. The penetrance and expressivity of genotypes can be altered through the accumulated impact of many common genetic variants throughout the genome. The "omnigenic" model proposes that, due to their interconnected nature, variants in gene regulatory networks that are expressed in disease-relevant cells or tissues can influence core disease functioning [16], suggesting that many unrelated disease variants contribute to the presentation of a phenotype.

Within this framework, I aspire to unravel novel disease mechanisms in Hirschsprung disease (HSCR) and Hirschsprung associated enterocolitis (HAEC), employing diverse Omics approaches to explore novel biological and pathological insights. Following the hypothesis of the polygenic and omnigenic models, I aimed to explore the affected core within regulatory network selecting patients with the same clinical and phenotipic signs, in order to study the phenotipic modifiers that would be optimistically genetic.

Our investigative journey began with the exploration of new susceptibility genes by the genomics approach of whole exome sequencing, followed by an examination of protein pathways through the proteomic strategy of mass spectrometry. Subsequently, I dig into enhancing our understanding of specific transcripts by RNA-seq targeted analysis of the whole transcriptome, contributing to the refinement of disease sub-classification.

Hirschsprung's disease (HSCR) is a chronic constipation resulting from the congenital absence of enteric ganglia [17]. HSCR is a classical example of a complex genetic dis-

ease, where the congenital absence of enteric ganglion cells causes failure to pass meconium, intestinal obstruction and colonic distension [18]. HSCR is typically diagnosed through a suction rectal biopsy, examining the rectal mucosa and submucosa for characteristics such as aganglionosis, thickened extrinsic nerve fibers, and an elevated expression of acetylcholinesterase. The primary treatment approach involves surgical intervention, specifically the resection of the aganglionic segment. Following this procedure, overall survival and functional prognosis are generally favorable. Severity of the disease depends on the length of the aganglionic tract, leading to classification of cases into short (S) and long-segment (L) forms, according to whether absence of innervation is confined below the recto-sigmoid junction or extended beyond it. The disease, with an incidence of 15 cases per 100,000 live births, is characterized by high heritability ($>$80%) and marked sex differences (male:female ratio, 4:1). HSCR is mostly isolated and sporadic, but 18% of cases are recurrent in the same family and occur in syndromic forms or together with other allied disorders in around one third of cases [17],[19]. The high (3 to 17%) sibling recurrence risk (i.e, the risk of being born with the disease, given that one full sibling is affected) is variable according to sex, segment length, and familiarity [19]. Predicting the risk and offering genetic counseling thus relies on factors such as family history, risk factors such as sex and segment length, assessment for syndromic features, and the genetics background. Hirschsprung's disease has multifactorial causes, although no environmental causes are known. Segregation analyses have refined this view by showing genetic heterogeneity according to the extent of aganglionosis. The long form is characterized by autosomal dominant inheritance and the short form by recessive or multifactorial inheritance, and the variants associated with both forms have incomplete penetrance [20].

Linkage studies allowed the identification of the RET proto-oncogene, located on 10q11.2, with 90% of the familial forms of HSCR linked to this locus. RET encodes a receptor tyrosine kinase expressed in human tissues of neural crest origin during embryogenesis, and is necessary for the migration and differentiation of neural crest derived cells. Loss-of-function RET mutations, scattered all over the gene, have been reported in 50% of familial and 7-35% of sporadic HSCR cases [21].

Advancements in the comprehension of Hirschsprung's disease's functional basis have been significant. While the RET proto-oncogene is the major gene implicated in HSCR, other genes and loci contribute to a minority of cases, including EDNRB, EDN3, GDNF, NRTN, SOX10, ECE1, L1CAM, PHOX2B and IHH [22],[23],[24]. Mutations in these genes are comparatively rare and have been mainly found in syndromic forms, often in combination with a RET germline mutation. A study on genotypes and exome sequences from 190 HSCR

patients aimed at investigating the role of different classes of variants in the risk of HSCR identified pathogenic alleles in 32 genes and loci, the majority of these playing a role in the development of the enteric nervous system [19]. Interestingly, 63.2% of patients exhibit pathogenic alleles primarily within the known RET regulatory network, resulting in decreased RET signaling. These mutations have been identified in 50% of familial (mostly L-HSCR, TCA) and up to 20% of sporadic (mostly S-HSCR) cases. In particular, Tilghman et al. [19] identified genetic causal factors in approximately 72% of HSCR cases, with a disease risk quantifiable based on the type of variants and their frequencies, and thus solely on sequence data. This explained a significant portion of the population attributable risk, ranging from 53.7% to 61.9%. About 21% of patients exhibited multiple risk factors, with the genotype-specific incidence dramatically increasing with the number of genotypic risk factors. This information could be crucial for addressing questions related to underlying causes, genetic architecture, and providing genetic counseling.

Consequently, although the identified genes have led to a deep understanding of the genetic basis of HSCR and are important to specific families, they are not a significant explanation of its incidence [25].

Despite surgical resolution for the majority of HSCR cases, the most severe complication, Hirschsprung Associated Enterocolitis (HAEC), remains life-threatening and affects around one-third of patients [26]. HAEC is the leading cause of morbidity and is responsible for half of deaths associated with Hirschsprung disease. Literature showed that the incidence of HAEC was 24% for infants diagnosed with HD after the first week of life compared to 11% if diagnosed within the first week [27].

HAEC presents with abdominal distention, fever, and diarrhea, with histological features including crypt dilatation, mucin retention, bacterial adherence to enterocytes, leukocyte infiltration, and epithelial damage. Initially attributed to intestinal mechanical obstruction, the occurrence of HAEC both before and after surgery challenges this hypothesis [28].

Multiple factors, involving the immune system, defense barrier, and microbiome, are implicated in HAEC pathogenesis, with altered goblet cell and mucus properties suggested [29]. Experimental colitis susceptibility and epithelial barrier alterations have been reported in mouse models with defective inflammasomes. Furthermore, distinct microbiome patterns have been observed in both HSCR and HAEC patients [30].

The marked HAEC susceptibility in HSCR, its occurrence before and after surgery, and the genetic basis of most HSCR related disorders underscore a potential genetic predisposi-

tion. Mouse models defective for genes implicated in HSCR, such as EdnrBNCC −/−[30] and Gfra1hypo/hypo mice [31], and showing symptoms resembling HAEC further support this notion. However, despite gut and fecal microbiome studies, comprehensive genetic screening and omics-based gene expression studies for HAEC remain unexplored to date [32], [33]. Furthermore, the HAEC low-grade bowel inflammation phenotype is similar in clinical features to inflammatory bowel diseases (IBDs), such as Crohn's disease and ulcerative colitis, and a possible association between HSCR and IBDs has been described [34], [35], [36]. However, a genetic overlap between HSCR and IBDs or between HAEC and IBDs, as far as we know, has never been explored.

For this purpose, this dissertation focused on the identification of new genetic factors in HSCR and HAEC. Our path articulated in those steps:

1. Whole Exome Sequencing on 12 HAEC patients and 12 HSCR patients without enterocolitis occurrence (HSCR) to identify rare causative variants and common susceptibility variants that might pinpoint to genes with a key role in HAEC occurrence;

2. Proteome sequencing on lymphoblastoid cell lines from HAEC patients carrying a variant possibly associated to HAEC to confirm the role of this genetic variant and to identify particular subset of protein involved in the pathological mechanisms of the disease;

3. Transcriptome analysis on IntraEpithelial Lymphocytes(IELs) derived from 6 HSCR, 6 HAEC and 2 control patients to detect differentially expressed genes (DEG) between HSCR and HAEC patients and the on Peripheral Blood Monocyte Cells (PBMCs) available from a subset of these patients to compare and/or confirm the identified DEGs;

4. Genetic association analyses to investigate the genetic overlap between HAEC and IBDs by a candidate genes association between HAEC and IBDs known genes, and between the HAEC candidate genes identified through the present project and IBDs data from the UK Biobank.

Chapter 3

Materials and Methods

3.1 Patients

We retrospectively checked for enterocolitis occurrence in all the consecutive patients admitted at the Gaslini Institute, Genova, Italy, since 1998, who were affected by Hirschsprung (HSCR) confirmed by biopsy. The Gaslini Institute has been a main and widely recognised hospital center for HSCR for decades, both under a clinical point of view, with surgical procedure performed here since 1960, and under a genetic and molecular point of view, with the HSCR major gene first discovered in the molecular genetics laboratory of Gaslini [37].

Between January 2010 and December 2012, HSCR patients have undergone an advanced diagnostic algorithm, through a complete phenotype screening, in the ambit of a previous prospective study carried out through a collaboration between the Gaslini Institute, the S. Raffaele Hospital in Milan, and the University of Genoa. This screening included renal ultrasound scan (US), cardiologic assessment with cardiac US, cerebral US, audiometry, ear, nose, throat and ophthalmologic assessments, plus further possible specialist evaluations to assess the prevalence of allied disorders and syndromes associated to HSCR [17]. Through this comprehensive approach, and the fact that patients have been generally followed up for several years, we could retrieve well established information about the presence of enterocolitis or other disorders associated to HSCR for a large amount of HSCR patients, and we selected HAEC patients to get sequenced by Whole Exome Sequencing (WES) based on:

- presence of enterocolitis;

- complete phenotypic screening resulted in the absence of additional anomalies;
- Italian ancestry (both parents Italian);
- sufficient and not degraded DNA.

Excluding patients for whom surgical complications and multiple surgeries might have increased HAEC risk, 12 HAEC patients were recruited.

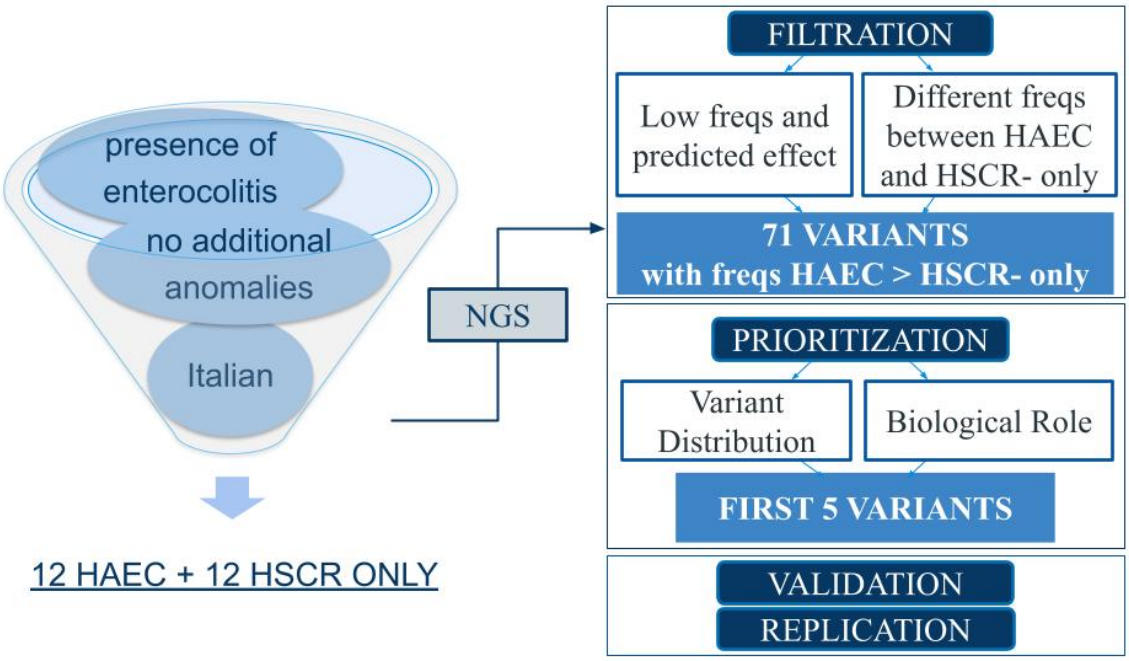
We chose the same number of HSCR patients without enterocolitis occurrence and with similar characteristics to the HAEC group in terms of gender, length of aganglionosis and familial occurrence of HSCR, following the same criteria. Since the challenges in recruiting patients, stratifying based on RET mutations or based on the predisposing allele was infeasible, we have thus opted not to include it in the inclusion criteria neither in the scope of the thesis. However, I have investigated the RET profile in our patients, and I haven't observed any enrichment of mutations or predisposing haplotypes in the samples with enterocolitis with respect to the HSCR-only patients. For the replicate analysis we also included patients with additional anomalies or incomplete phenotype screening, given that we were following up the HAEC complication testing only specific variants identified by WES, despite a putative more complex genetic background. Up to 72 HAEC and 108 HSCR patients, depending on the variant, were screened. For the proteome analysis, I followed up on the OSMR variant only, the most promising variant detected by WES, by analyzing the effect of the homozygous variant with respect to the wild type genotype (absence of the variant). For the analysis, I elected to use the lymphoblasts cells derived from the blood samples collected from patients at Gaslini Hospital for diagnostic purposes and obtained through proper transformation of lymphocytes. Since the low frequency of this variant and the limited blood availability, I could collect samples from one homozygous patient only, and thus I analyzed two patients, one homozygous for the OSMR G variant and one wild type TT carrier, and focused on four conditions corresponding to the two specific genotypes, both before and after treatment with OSM. For transcriptome analysis, samples were available thanks to the already mentioned prospective study in collaboration with the S. Raffaele Hospital. The surgeon collected biopsy fragments during the restorative surgery and dispatched the sample to S. Raffaele Hospital, where the collaborators performed cell extraction for subsequent analysis. The sole admission criterion for this process was the diagnosis of Hirschsprung. Consequently, a total of 92 patients with Hirschsprung disease (HSCR) and Hirschsprung Associated Enterocolitis (HAEC), along with 22 pediatric controls had been recruited. Subsequently, for the analysis described in this thesis, I carefully selected 6 HSCR patients, 6 HAEC patients, and 4 controls with no intestinal anomalies that could be potentially associated with Hirschsprung or

inflammation. The admission criteria for this selection mirrored the one applied for the WES analysis, considering the absence or presence of enterocolitis, confirmed through phenotypic screening, and ensuring Italian ancestry.

3.2 Whole-Exome Sequencing and Filtering and Prioritization of Variants

Genomic DNA was extracted from lymphocytes or from immortalized lymphoblastoid cell lines by standard protocol. Whole-Exome Sequencing (WES) was performed on the HiSeq System (Illumina, San Diego, CA, USA) after enrichment library preparation with Nextera Rapid Capture Expanded Exome (Illumina). WES sequencing, the basic bioinformatics analysis and the Variant Calling Format (VCF) file generation were endowed by the Center for Translational Genomics and BioInformatics, San Raffaele Scientific Institute. The VCF file is a textual document that encloses genetic variants, commonly generated as an output file by variant calling programs associated with the employed sequencing method. Typically, the VCF file incorporates information obtained during the annotation step, presenting various details for each specific variant. This includes details such as population frequency, pathogenicity, conservation scores, and other relevant information. Filtering and case-control comparisons were performed with SnpSift. Filtering and prioritization steps followed a certain schema, here detailed in Figure 1 and all the steps were detailed below.

Figure 1: *WES Selection Criteria*



Variants with read depth ≥ 10 in all samples and a call quality ≥ 10 were filtered based on (i) low frequency (≤ 0.01) and the impact on the protein; and (ii) a very different distribution between cases (HAEC) and controls (HSCR):

1. "novel" variants (allele frequency below 0.01 or above 0.99 in the general population from public databases) with high impact on the protein (novel_high group), and "novel" variants with high or moderate impact, predicted to be deleterious and with evidence of selection constrains and conservation among vertebrates (novel_pred group). By this strategy we filtered respectively 8 and 23 variants, one of which in common to both groups, that were present in at least two cases' alleles (HAEC patients) and absent in controls (HSCR).

2. variants present in at least 6 alleles among cases and in no controls (or on 24 alleles in controls and less than 19 alleles in cases) (6_0 group) or showing p-value < 0.001 in an exploratory association analysis between cases and controls (p_001 group), thus selecting respectively 30 and 12 variants, one of which in common.

The minor allele frequency (MAF) was checked against the Non-Finnish European Gnomad database, or, when not available in the annotated vcf file, against 1000genomes Europeans or dbSNP database. The impact on the protein is the effect predicted by SnpEff as

described by the authors [38], "high" for deletions, duplications or inversions, frameshift variants, stop or start codon loss or gain, "moderate" for in frame inserts or deletions and missense variants; "low" for synonymous and "modifier" for non-coding variants. The deleterious prediction was based on Mutation Taster, Polyphen–2, and Sift, while the selective constraints and the conservation were annotated based on GERP++ RS score and phastCons100way vertebrate. To be more stringent, I filtered only those variants that were predicted to be damaging or probably damaging on at least one isoform of the protein by all three software tools and with both evidence of selective constraints and conservation among vertebrates, that I choose as GERP++ RS 2 and phastCons100way_vertebrate 0.13.

To prioritize the identified variants, I assigned a scale of relevance (WES score) based on the variants annotations (impact, deleteriousness prediction, frequency in the public databases, different distribution between HAEC and HSCR) and a scale of biological relevance (biological score) based on the genes' function and involvement in immunity and intestine reported in public databases and PubMed. In detail, to create the scale of relevance (WES score), I assigned an extra score to the variants evaluated as more relevant because:

- with high impact on the protein (except for the novel_high group);
- predicted as deleterious by one or more in silico methods or showing conservation or selection constrains evidences (except for the novel_pred group);
- present in more than 6 alleles among cases (except for the p_001 group);
- rare or never reported at all in the public databases (MAF) <0.05, except for the novel_high and novel_pred groups) ;
- with frequencies reported in public databases (here called MAFdb) similar to our HSCR sample (MAFcontrols) but different from HAEC (MAFcases) and in accordance with a role of the variant in predisposing to HAEC (i.e. MAFcases higher than both MAFcontrols and MAFdb, and not in between MAFcontrols and MAFdb);
- with p-value lower than the alpha 0.001 chosen to filter variants.

Variants detected by WES were compared between HAEC (cases) and HSCR (controls) patients by the Fisher exact test for allelic association and by the Cochran-Armitage trend test implemented in SnipSift (one-tail test, no multiple test correction). Frequencies of variants in cases and controls were compared to those reported on gnomAD v2.1.1 for the non-Finnish

European controls by the proportion test, also applying Bonferroni correction for the 77,396 variants detected by WES (i.e., $p \leq 6.5 \times 10^{-7}$).

In our analysis, we have taken into account synonymous variants as well, acknowledging their historical significance in diseases studied in our laboratory. This is evident firstly in the Hirschsprung disease (HSCR) haplotype [39], and subsequently in other disease mechanisms such as PHOX2B variants associated with life-threatening events (ALTE), Sudden Infant Death Syndrome (SIDS), or neuroblastoma [40].

To assign the biological score, I also searched Ensembl, genecard, OMIM, and PubMed for the biological role of the genes where the variants were localized. I considered as relevant the genes involved in inflammation, immune response, gut, and experimental colitis and IBDs. I have thus arbitrarily classified the genes as "very likely" (5 points), "likely" (4 points), "possible" (3 points), "unlikely" (2 points) and "no/unknown" (1 point). The WES and the biological scores were then multiplied to get a final ranking. Selected variants validation and follow-up genotyping have been performed by Sanger sequencing.

3.3 Validation and Replication of WES results: Sequencing and Genotyping

Selected variants were validated by Sanger sequencing first in the 12 HAEC cases, and, if confirmed, in the 12 HSCR patients. I have replicated the association analysis for the validated variants in a larger cohort of 65 HAEC and 105 HSCR patients for the most promising variant in OSMR and 23 HAEC and 42 HSCR patients for other promising variants. Regions spanning the selected variants were amplified by using the Accuprime GC RICH kit (Invitrogen, Life Technologies). In particular, all PCR reactions specific for GC rich templates were set up in 25 μl total volume PCR reactions containing 200 ng of genomic DNA, 400 nM primers, Accuprime buffer B and run for 35 cycles with 45s 95°C denaturation, 45s 57°C annealing, and 1min 72°C extension. PCR products were checked on 2% agarose gel, purified by ExoStar and subjected to Sanger sequencing analysis using an automated ABI-3730 Sequencer (Applied Biosystems; Thermo Fisher Scientific, Inc., Waltham, MA, USA). Sequences were visualized by FinchTV.

3.4 Validation and Replication of WES results: Statistical Analysis

The replicate genetic association was carried out using the Fisher exact test, two tails. Family-based association on available trios was performed by the Transmission Disequilibrium Test (TDT) implemented in PLINK 1.9 (https://zzz.bwh.harvard.edu/plink/, accessed on 11 March 2021) [10]. TDT test uses data from families with at least one affected child, and evaluate the transmission of the associated marker allele from a heterozygous parent to the affected offspring thanks to the method developed by Spielman et al [41],[42] apply the McNemar's test [43] for paired nominal data.

3.5 In silico analyses of *OSMR*

Protter (http://wlab.ethz.ch/protter/start/) [44] and Swiss Model (https://swissmodel.expasy.org/) [45] software were used to predict the OSMR protein secondary and tertiary structure and to localize the SNP in the model, also with respect to the hotspot binding sites of OSM-OSMR described by Du et al. [46]. The Protein Data Bank (PBD) 3l5h.1A model template was used as the best homolog for OSMR (corresponding to Interleukin-6 receptor subunit beta, with 21,84% identity). Three separate modeling servers further confirmed the 3l5h.1A model: Modeller (https://salilab.org/modeller/), PhyreRisk (http://phyrerisk.bc.ic.ac.uk/home) and Phyre2 (Protein Homology/analog Y Recognition Engine, http://www.sbg.bio.ic.ac.uk/ phyre2). In particular, Phyre2 reports that 561 residues of the query sequence (57%) were modeled with 100% confidence. The 3l5h.1A model was also used to design the Ramachandran plot and the results were investigated by the RAMPAGE software for the evaluation of the plot quality (http://mordred.bioc.cam.ac.uk/ rapper/rampage.php).

3.6 Cell Cultures Preparation for the proteome analysis and Western Blotting Assay

Lymphoblastoid cell lines, produced by Epstein–Barr virus (EBV) immortalization, and were available at the IRCCS Gaslini (Genova, Italy) from both HSCR and HAEC patients, as well as as fibroblast cell lines for a few samples. In detail, I used lymphoblasts from a patient affected by a disorder unrelated to HSCR as control (LY1765), an HSCR patient (LY3956) TT homozygous for the OSMR wild type allele of the p.H187Q SNP, an HAEC patient (LY3828) TT homozygous for the same SNP, an HAEC patient (LY4579) GG homozygous for the variant allele of the OSMR p.H187Q SNP, and fibroblasts from a patient affected by a hematological disease not related to HSCR (SC591). Cell cultures were grown by standard protocols. Cells were then plated and stimulated for 30 min with 50 ng/mL Oncostatin M (OSM) (Cat.PHC5015, no. L0216061917, Invitrogen) [47] to undergo immunofluorescence assay [48] and mass spectrometry (MS) run.

Cell lines were plated in flasks and expanded to 3×10^6 cells for lymphoblasts and to 1×10^6 for fibroblasts. After 48 h, cells were washed with Phosphate Buffered Saline (PBS) 1×, centrifuged, and lysed with RIPA buffer (Tris–HCl 50 mM pH 7.5, NaCl 150 mM, Triton-X 1%, SDS-20 0.1%, Na deoxycholate 1%) and Protease Inhibitor. Total cell lysates were quantified with BSA assay and equal amounts were electrophoresed using Mini-PROTEAN Precast Gels and then Trans-Blot Turbo Transfer System for transferring. Proteins were identified by probing the membrane with the rabbit anti-OSMR antibody (PA5100298, Life Technologies, dilution 1:1000), specifically addressed against the OSMR protein and then with a goat anti-rabbit IgG-488 (Sigma-Aldrich, Merck, Darmstadt, Germany) dilution 1:20,000. Signals were detected using the chemiluminescence reagent ECL advance (BIORAD- Segrate (MI)-Italy) and protein levels in each sample were evaluated by comparison with corresponding amounts of the housekeeping -tubulin with the mouse anti Tubulin antibody (Sigma-Aldrich T5168, dilution 1:2500).

3.7 Mass Spectrometer and proteome analysis

Mass Spectrometer (MS) run and analysis were performed by the Core Facility unit of the Giannina Gaslini Institute on the HAEC GG (LY4579) (var) and the HAEC TT (LY3828)

(wt) patients, four replicates for both the “untreated” (wt−and var−) and the “OSM treated” (wt+ and var+) conditions (16 runs in total). The intensity values were extracted and statistically evaluated using the ProteinGroup Table and Perseus software, Max Planck Institute of Biochemistry, Martinsried, Germany, version 1.6.10.50 [49]: the flowchart analysis is represented in Figure 1.

Figure 2: *Flowchart of the Perseus analysis.*

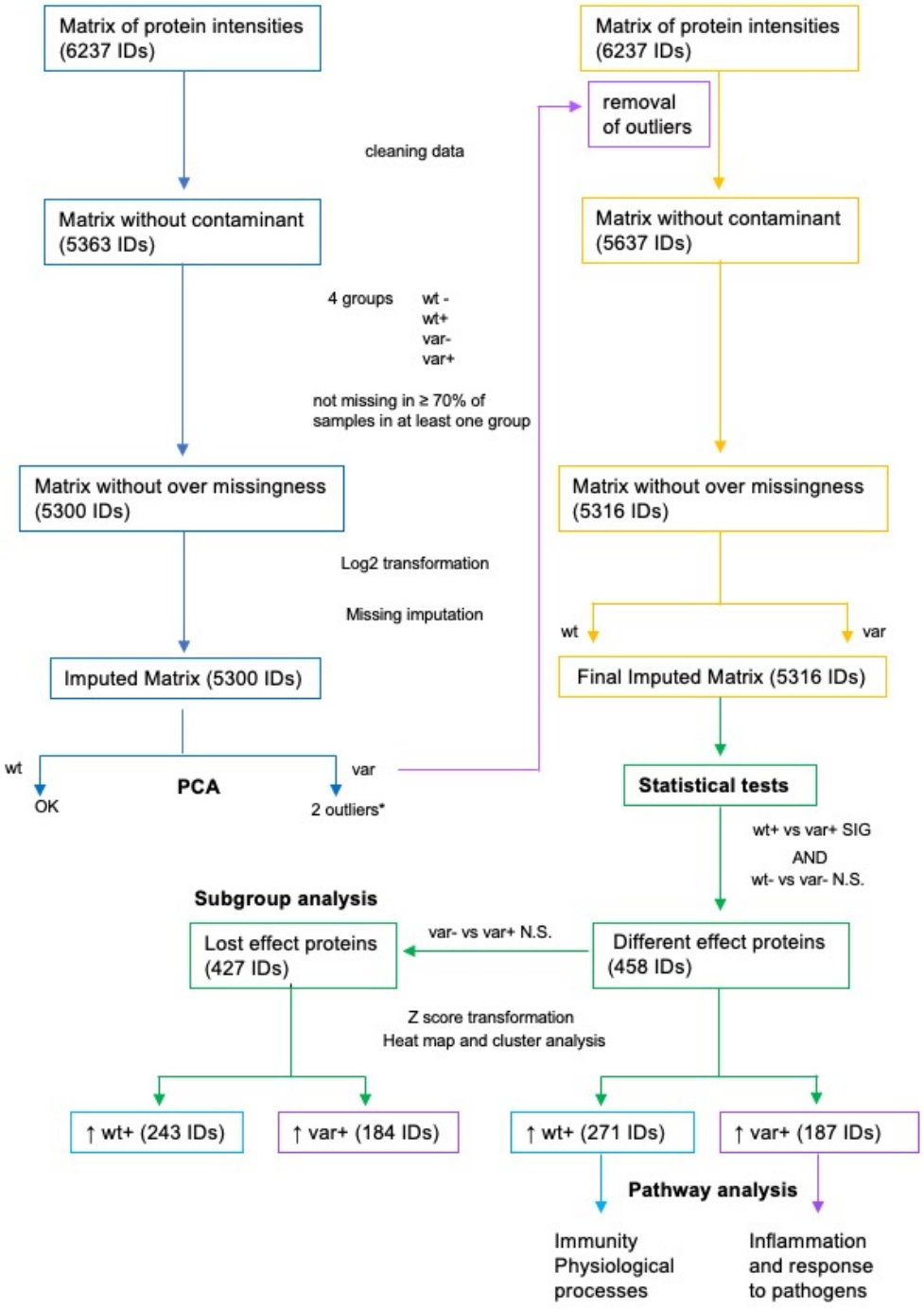


Figure 2: Flowchart of the analysis performed with Perseus to identify the protein differentially expressed between the wt TT cell line and the var GG HAEC cell lines following mass spectrometry.

3.8 Proteome Data Analysis

I performed the Proteome Data analysis using the Perseus Software. I excluded decoys and removed potential contaminants, employing a list that includes around 250 common contaminants, such as bovine proteins. After defining the four groups, wt-, wt+, var-, and var+, where "wt" is the OSMR TT genotype, "var" is the OSMR GG genotype, "-" stands for before treatment with OSM and "+" is after treatment, I excluded proteins with too many undetected values: by keeping only proteins with not missing values in at least 70% of replicates in each group, I have actually kept proteins detected in at least three replicates in one or more groups. I then log2 transformed the data and imputed missing data from normal distribution (Figure 1).

In detail, since missing values are associated with proteins with low levels of expression, I can substitute the missing values with numbers that are considered "small" in each sample. I can define this statistically by drawing from a normal distribution with a mean that is downshifted from the sample mean and a standard deviation that is a fraction of the standard deviation of the sample distribution. All the previous steps are part of an established protocol from the Core Facility Unit at Gaslini Hospital [50].

Outliers were checked by Principal Component Analysis (PCA) and excluded by the following analysis. The two outliers detected through the PCA could be reconducted to specific factual errors in the sample handling and preparation, as detailed below. The analysis resulted in 5 replicates for var- and 3 replicates for var+, instead of the planned 4 replicates each, because a label + was erroneously read as - during the sample run. Although the mistake was immediately recognized after the analysis, the sample, recognized as one of the two outliers has been excluded, together with the other outlier, for which I have afterward realized that an exchange between a pellet and a lysate had been made.

The protein intensities between the two samples, wt and variant for the OSMR SNP, were compared both before (wt- vs var-) and after (wt+ vs var+) OSM treatment by the two tails t test, applying the S0=0.1 parameter suggested by Perseus for more stringent analysis and

applying the Benjamini-Hochberg FDR correction for multiple testing (q-value). S0 is the cutoff for the level of difference in the actual measurement used to perform the standard t-test statistic. As described by Tusher et al., the S0 parameter is a positive constant added to the denominator of the "relative difference" in gene expression, calculated as the ratio of change to standard deviation for that gene. S0 is chosen to minimize the coefficient of variation of this ratio, to ensure that the "relative difference" in gene expression is independent of the level of gene expressions and comparable across all genes [51]. Proteins significantly differentially expressed in wt+ vs var+ ($q<0.05$) and not different before treatment with OSM (wt- vs var- n.s.) were considered differentially expressed after OSM activation. In an attempt to specifically select those proteins for which OSM activation leads to a loss of effect in the presence of the OSMR SNP variant, I further filtered proteins not significantly different between var- and var+, although only a slightly smaller group was obtained (427 out of 458 proteins) (Figure 1).

Hierarchical cluster analysis was performed to identify clusters of proteins overrepresented or underrepresented in wt+ with respect to var+. I run the hybrid hierarchical k-means clustering algorithm provided by Perseus on the z-scores normalized data applying the software default settings, which are the Euclidean distance method and average linkage agglomeration, pre-processing with k-means,number of clusters: 2; maximal number of iterations: 10. The Z-score is a form of standardization used for transforming normal variants to standard score form. Given a set of raw data Y, the Z-score standardization formula is defined as:

$$Z = \frac{X - \mu}{\sigma}$$

Where X is the raw score, μ is the population mean, and σ is the population standard deviation.

K-means clustering represents instead, given a set of observations (x1, x2, ..., xn), where each observation is a d-dimensional real vector, the partitioning of the n observations into k sets (kn) S = S1, S2, ..., Sk so as to minimize the Within-Cluster Sum of Squares (WCSS):

$$J(c,\mu) = \sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

where $J(c,\mu)$ represents the sum of squared distances between each point and its closest centroid [52]. The clusters were usually visualized by a color scaling heatmap.

Pathway analysis on the distinct clusters was then performed using HumanBase, Flatiron Institute, Simons Foundation, New York, NY, USA (https://hb.flatironinstitute.org/ ac-

cessed on 14 July 2020) [44], which reports the one-sided Fisher's exact tests with Benjamini–Hochberg False Discovery Rate corrections for multiple tests' q-value, searching the lymphocyte network. The Venn diagram was drawn by an online software (https://www.meta-chart.com/venn/display accessed on 3 Dicember 2020).

3.9 Preparation of the sample for Transcriptome sequencing

Transcriptome analysis was performed on Intraepithelial lymphocytes (IELs), which are the immune cells inside of the intestinal epithelium only, derived from gut biopsies extracted by the surgeons during the surgery prescribed by the normal clinical practice. IELs were obtained from gut biopsies of 6 HAEC, 6 HSCR patients and 4 pediatric controls and then rapidly processed in order to obtain mononuclear intraepithelial cells. Once obtained, the cells were lysed and cryopreserved at −80°C. RNA was then extracted and quantified in order to obtain 15 L of each diluted sample (10 ng) for an Ion AmpliSeq, i.e. 0.67 ng/μL diluted with Nuclease-free Water.

3.10 Transcriptome Sequencing

The RNA was first retro-transcribed using NGS Reverse Transcription Kit (A45003), and the resulting cDNA was amplified. The library was prepared with Ion AmpliSeq Transcriptome Human Gene Expression Kit, it was diluted to 100pM and then the chip was prepared for the sequencing using the Chef Instrument in house at Gaslini Hospital. First of all, I prepared 8 samples per run to 0.67 ng/μL with Nuclease-free Water, obtaining 15 µL of each diluted sample (10 ng). Afterward, I performed the two runs for the library preparation on the Ion Chef system, after setting up the barcodes names, the appropriate reference library and target region files. I quantified the obtained libraries with Qubit Fluorometer and Bioanalyzer Instrument. After quantification, I determined the dilution factor that results in a concentration of 100 pm, which was appropriate for template preparation using an Ion Template kit. Then, I used 25 µL of each diluted library to the bottom of the Ion Chef library sample tubes for the chip loading run on Chef system (the run takes about 15 hours). Once the chip was ready, I performed the sequencing step on the Ion S5 system, with the corresponding kit: the run

takes about 3 hours for each chip. After the chip loading, the sequencing was performed on Ion S5 with 540 chip.

3.11 Transcriptome Data Analysis

The raw results were analyzed first for quality control and then for data processing. Differential Expression (DE) Analysis to compare the gene expression levels between HSCR, HAEC and controls was performed using the DESeq2 package included in R [53]. I have excluded Batch effects assessing that the distribution of the mean of the transcripts missing call rate per batch was uniform without noticeable outliers [54]. To allow comparisons, taking into account transcripts length, total number of reads, and sequencing biases, the expression levels have been normalized. The normalization procedure implemented in DESeq2 is based on the ratio of the counts for a gene in each sample divided by the geometric mean of that gene across all samples. After normalization of the data to ensure comparability, Principal Component Analysis (PCA) is employed—a statistical technique used to reduce the dimensionality of complex datasets while retaining the essential variability. This method is crucial in uncovering patterns, commonly applied in transcriptomics to discern structures and relationships within datasets The data were then log 2 transformed to stabilize the variance across the range of expression values for downstream analyses. Following this, I selected the most variable genes retaining those with a variance exceeding 90% of the interquartile range (IQR) to generate the heatmap. Hierarchical clustering is a popular method for grouping objects based on expression similarity between samples.

An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models. The p-value in DESeq2 is calculated using the Wald test. The null hypothesis of the Wald test is that: for each gene, there is no differential expression across two sample groups. The formula for the test statistic used in the Wald test:

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\text{Var}(\hat{\theta})}$$

where W represents the test statistic, $\hat{\theta}$ is the maximum likelihood estimator of the parameter, θ_0 is the hypothesized value of the parameter under the null hypothesis, and $\text{Var}(\hat{\theta})$ is the

variance of the maximum likelihood estimator.

The Wald test is usually talked about in terms of Chi-Squared, because the sampling distribution (as n approaches infinity) is usually known. This variant of the test is sometimes called the Wald Chi-Squared Test.

Significance level was adjusted for the number of transcripts by Benjamini-Hochberg False Discovery Rate, setting significance at alpha=0.05. Results are reported in terms of log2 fold changes of pairwise comparisons between the HSCR, HAEC or controls, p values and adjusted p values, to identify transcripts significantly over or under expressed in the HAEC with respect to the HSCR samples and controls. Enrichment pathway analysis for genes over and under expressed was done using GSEA gene sets, including immunologic signature and regulatory target gene sets. We applied GSEA in order to identify gene sets and pathways that were significantly perturbed across conditions. Collections of genesets were downloaded from MiSigDB [55]. We used the Benjamini-Hochberg method to adjust gene set p-values, and set 0.05 as the significant threshold.

3.12 Validation for Gene differentially expressed between HSCR, HAEC and controls in IELs

The validation of gene expression results was carried out through quantitative PCR (qPCR) on some arbitrarily chosen differentially expressed genes. Total RNA from cells was isolated by a commercial RNA purification kit (RNeasy Mini kit, Qiagen, GmbH, Germany) according to the manufacturer's protocol. Total RNA was reverse transcribed by iScript cDNA synthesis kit (Bio-Rad Laboratories) according to the manufacturer's protocol. Amplification reactions were carried out in 20 μl total volume with 40 ng cDNA, 10 μl of 2 x TaqMan IQ SuperMix (Bio-Rad) and 1 μl of 20 x specific assay and the thermocycler protocol was as follows: initial denaturation at 95 °C for 2 min, then 40 cycles at 95 °C for 15 sec and 60 °C for 30 s. The primers were designated through Primer 3 plus software. Real time quantitative PCR was performed using inventoried Assays-on-Demand provided by Applied Biosystems. PCR reactions were performed using the iQTM5 Real Time PCR (Bio-Rad Laboratories). Once the primer amplification efficiency was tested by a standard curve, triplicates of each cDNA sample were amplified with Sybr Green SuperMix (Bio-Rad) and 0,5 μM of forward and reverse primers. Melting curves were calculated between 55 °C and 95 °C with an in-

crement of 0.5 every 15 s. The mRNA expression was normalized over the mean values of the reference genes. It is important to choose a suitable reference gene for relative quantification of gene expression for each experiment; I chose 2-Microglobulin and GAPDH which are strongly expressed in every cell type and thus defined as housekeeping genes. Three technical replicates were planned for each sample to control for manual errors. Analysis of gene expression was performed by Ct methods and data were elaborated by using the Bio-Rad IQ5 software [56].

3.13 Replication of transcriptome analysis on PBMCs

In a subsequent step, I sought to replicate the RNA sequencing (RNASeq) targeted approach using peripheral blood mononuclear cells (PBMCs). To this aim I strategically selected patients analyzed for IELs for whom also PBMCs were available to ensure a comprehensive dataset for IELs vs PBMCs comparison. Unfortunately, this criterion resulted in a relatively limited number of eligible samples. I performed the transcriptome analysis on PBMCs applying the same methodologies and techniques employed in the earlier transcriptomic analysis of IELs, ensuring consistency and comparability across the two datasets. I employed the same techniques for RNA extraction, sequencing runs, and statistical analysis using R and the Bioconductor package DESeq2. The analysis included principal component analysis for outlier detection, normalization, hierarchical clustering, heatmap generation, identification of differentially expressed genes, and pathway enrichment analysis using gene set enrichment analysis (GSEA).

To explore the feasibility of PBMCs as an alternative tissue for biomarker research, I performed a correlation analysis between the abundance of transcripts obtained from RNASeq on IELs and those obtained from PBMCs for each sample. Since the Shapiro-Wilk test revealed that the transcript abundances obtained for IELs and for PBMCs did not follow a normal distribution, I performed the non parametric Spearman correlation test.

The two most prevalent tests for assessing normality, namely Shapiro-Wilk's test and the Kolmogorov-Smirnov test, share the following hypotheses:

H_0: The data adhere to a normal distribution; H_1: The data deviate from a normal distribution

Given its superior statistical power, especially in smaller sample sizes, I opted for the

Shapiro-Wilk test over the Kolmogorov-Smirnov test. In detail, Shapiro–Wilk test is a more appropriate method for small sample sizes (<50 samples) although it can also be handled on larger sample sizes while Kolmogorov–Smirnov test is used for n $\geq$ 50[57].

The Shapiro Wilk test compares observed data to expected values from a normal distribution. A p value below 0.05 suggests rejecting the null hypothesis, indicating non normality in the data.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where:

- W is the test statistic for the Shapiro-Wilk test.
- n is the sample size.
- x_i are the ordered sample values.
- $\bar{x}$ is the sample mean.
- a_i are constants derived from the sample size and the expected values of order statistics under the null hypothesis.

The test has limitations, most importantly that it is biased by sample size: the larger the sample, the more likely you will get a statistically significant result.

Spearman correlation [58] is a non-parametric measure of rank correlation. It assesses the strength and direction of monotonic relationships between two variables. The Spearman correlation coefficient, denoted as ρ (rho), ranges from -1 to 1, with a ρ of 0 suggests no monotonic correlation. A positive ρ indicates a direct monotonic relationship, while a negative ρ signifies an inverse monotonic relationship. The Spearman correlation is robust to outliers and does not assume linearity in the relationship between variables.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Where: ρ is the Spearman rank correlation coefficient. n is the number of pairs of observations. d_i is the difference between the ranks of the paired observations.

The coefficient ranges from -1 to 1, where: $\rho = 1$ indicates a perfect increasing monotonic relationship. $\rho = -1$ indicates a perfect decreasing monotonic relationship. $\rho = 0$ indicates no monotonic relationship between the variables.

3.14 UK Biobank: an unprecedentedly large biobank

During my PhD, I visited the Statistical Genetics Laboratory at Columbia University in New York (US) to learn how to properly access the UK Biobank (UKBB) data and information. The UK Biobank stands as a colossal repository, providing genetic, imaging, and health data and information from over half a million participants with diverse ancestries. Each facet of this extensive dataset is meticulously categorized into specific fields, which needs to be carefully interrogated to get the desired information, such as demographics, medical examinations, genetic information, and more. Each of those fields has a particular code that can be used to interrogate the dataset, retrieve the trait of interest, whether it be a particular medical condition, a phenotypic trait, or a behavioral characteristic, and filter the samples of interest (i.e. ancestry). Next, the available data fields encompassing demographics, medical examinations, treatment information, or other relevant outcomes, in addition to genetic information, can be explored to retrieve information about the selected sample and trait of interest. For instance, navigating the International Disease Code (ICD) using data fields within the UK Biobank (UkB). The ICD-10, or the 10th revision of the International Statistical Classification of Diseases and Related Health Problems, is a medical classification list maintained by the World Health Organization. Each disease is assigned a specific code, such as K50 for Crohn's disease. In the ICD-10 reference website(https://icd.who.int/browse10/2019/en), it is possible to find the inclusion and exclusion criteria for each code (e.g., K50 includes granulomatous enteritis but excludes ulcerative colitis). In the UK Biobank, ICD-10 codes are recorded in data fields, the fundamental blocks of data held within the UK Biobank repository and identifying the results of a single question, measurement or result (i.e. 41270 for reported ICD10). It therefore provides the ability to filter specific traits in or out of your cohort. The scale of the data is monumental, rendering traditional downloading infeasible due to its size and the more and more larger number of users who want to approach that kind of data, in addition to privacy issues related to the genomic data. Instead, all operations, from cohort selection to precise cohort definition and diverse statistical analyses, are efficiently conducted through the command-line interventions or through the Research Analysis Platform (RAP).

The RAP, a powerful computational engine, serves as the first step for manipulating and exploring the data provided by the UKBB and allows to include only instances that closely match the conditions of interest in the search, as well as extra criteria to add more layers to the selection process to get a well refined sample. Its architecture, comprising various nodes and memory configurations, let the user perform different kinds of operations with a large library of different softwares (PLINK, bcftools, etc.). During my experience at Columbia University's Department of Statistical Genetics, I got a full understanding of what is UKBB, how I can interrogate it to explore the data and I got a good knowledge about programming languages such as R and Python, along with their libraries, to manage such big data. These tools are indispensable for navigating and extracting meaningful insights from the vast seas of data within the platform.

Among the other techniques I learnt, a pivotal aspect involved genotype imputation for untyped variants, optimizing genetic coverage. Through the utilization of IMPUTE 5 software and the Haplotype Reference Consortium (HRC), I mastered the reconstruction of haplotypes. This proficiency ensured a comprehensive analysis of genetic variations, unlocking hidden dimensions within our study. The genotyping and sequencing data collected in the UKBB are suitable for various genetic analyzes such as, among the others, imputation of untyped variants, reconstruction of haplotypes and genetic association analysis. One approach to control that the results of association analysis are plausible, not affected by errors, is to check the inflation in the observed test statistic with respect to those expected. This check involves calculating the ratio between the median of the observed chi-squared test statistic divided by the expected median of the corresponding chi-squared distribution, which is called the Genomic Inflation Factor , and should thus be close to 1. Furthermore, it would be highly beneficial to estimate relatedness, ancestry, and population structure within the sample through techniques such as Kinship calculation and Principal Component Analysis. This is particularly crucial, as varying degrees of sample relatedness or distinct population origins may contribute to inflated results and high Genomic Inflation Factor . The association analysis can be then performed by regression analysis, including linear mixed models and generalized linear mixed models.

3.15 Phenotypic selection on UKB

In details, I have selected patients with Inflammatory Bowel Diseases from the UK Biobank according to baseline records, hospital diagnosis or clinical reports following these filters:

For Chron's Disease:

- White Europeans, British, Irish and Any other white origin as Ethnic Background.
- Diagnosis ICD10 K50 corresponding to Chron's disease between January 1st, 1950 and December 31st, 2050 (as default settings).
- Source of diagnosis including: "Primary Care" or "Primary Care and other sources", "Hospital admission data" or "Hospital admission data and other sources", "Data death registry" or "Death registry and other sources".

For Ulcerative Colitis:

- White Europeans, British, Irish and Any other white origin as Ethnic Background.
- Diagnosis ICD10 K51 corresponding to Ulcerative Colitis between 1950-1-1 and 2050-12-31 (as default settings).
- Source of diagnosis including: "Primary Care" or "Primary Care and other sources", "Hospital admission data" or "Hospital admission data and other sources", "Data death registry" or "Death registry and other sources".

All those patients for which the type of IBD diagnosis was left unspecified, or genetic information was unavailable, were excluded. Additional information collected in the baseline questionnaires including age, sex, medications, dietary factors and comorbidities were collected to be used as covariates.

3.16 Gene candidate association analysis

I aimed to explore potential gene candidates for Hirschsprung's disease (HSCR) and Hirschsprung-associated enterocolitis (HAEC) within the UK Biobank dataset of patients with Inflammatory Bowel Diseases (IBDs). Initially, our focus was on variants with established associations with HSCR and HAEC. Specifically, I prioritized polymorphisms in key genes implicated in HSCR. Starting from Tilghman's paper summarizing the genetic profile of Hirschsprung's disease, reporting 17 associated genes with a particular emphasis on RET and EDNRB [19], I proceeded to select a subset of variants from the studies cited by Tilghman and colleagues

[59], [60], [39], [61]). This selection, outlined in Table 1, aimed to focus on variants primarily associated with Hirschsprung's disease, considering the pivotal genes highlighted in the aforementioned studies, including RET and EDNRB.

Table 1: *List of the selected SNP implicated in Hirschsprung Disease*

Chr	Pos	Gene	AminoAcidChange	rs	Reference
7	84515886	SEMA3D	-	rs11766001	Jiang et al., 2015 [61]
7	84720526	SEMA3D	-	rs80227144	Tang et al., 2016 [62]
8	32552791	NRG1	-	rs4541858	Jiang et al., 2015 [61]
10	43086608	RET	-	rs2435357	Emison et al., 2005 [63]
10	42952399	RET	-	rs2506030	Kapoor et al., 2015 [59]
10	43057447	RET	-	rs7069590	Chatterjee et al., 2016 [64]
10	43114671	RET	G691S	rs1799939	Lantieri et al., 2016 [39]
10	43118395	RET	L797L	rs1800861	Lantieri et al., 2016 [39]
10	43119646	RET	S836S	rs1800862	Lantieri et al., 2016 [39]
10	43120185	RET	S904S	rs1800863	Lantieri et al., 2016 [39]
10	43126769	RET	-	rs2075912	Lantieri et al., 2016 [39]
10	43130238	RET	-	rs3026785	Lantieri et al., 2016 [39]
10	43100520	RET	A45A	rs1800858	Lantieri et al., 2016 [39]
10	43080177	RET	-	rs1864410	Lantieri et al., 2016 [39]
10	43111408	RET	D489N	rs9282834	Tang et al., 2016 [62]
13	77918405	EDNRB	G57S	rs1801710	Amiel et al., 1996 [60]
13	77900651	EDNRB	R319W	rs200363611	Amiel et al., 1996 [60]

Similarly, for HAEC, I selected the susceptibility SNP rs34675408 in OSMR, which showed associations in our cohort through exome sequencing, along with other SNPs associated within our cohort but not replicated in subsequent analyses (Table 2)

Table 2: *List of the selected SNP for HAEC*

Chr	Pos	Gene	AminoAcidChange	rs	Reference
5	38884071	OSMR	H187Q	rs34675408	Bachetti et al., 2021 [65]
19	17940842	JAK3	A1094A	rs3212780	Bachetti et al., 2021 [65]
21	48069682	PRMT2	R229W	rs76937225	Bachetti et al., 2021 [65]
2	209190632	PIKFYVE	S1033A	rs999890	Bachetti et al., 2021 [65]
11	7059981	NLRP14	R55G	rs61063081	Bachetti et al., 2021 [65]
11	7091569	NLRP14	L1010F	rs17280682	Bachetti et al., 2021 [65]

3.17 Quality Control and Association Study

Once the cohort was defined, we proceeded with a quality check analysis to ensure the reliability of our data. Since the fundamental goal of a case-control association study is to test for an allelic frequency difference between cases and controls to find SNPs that affect disease susceptibility, even small artifactual differences in allelic frequency between cases and controls can generate false-positive results. Therefore, it is particularly important to avoid experimental confounding factors that have potential effects on allele frequency differences. At this aim we have selected only "White Europeans", "British", "Irish" and "Any other white origin as Ethnic Background" for the UKBB sample to retrieve a sample with a genetic background possibly more homogeneous for what concern the ancestry. We extracted the genotype information for a curated list of SNPs previously implicated in HSCR and HAEC disease through literature or prior studies. I performed association tests by chi square test, due to the nature of the phenotype (binary, case and control) and because I did not adjust for covariates in this primary analysis. The results include association statistics such as p-values and odds ratios for each SNP. I conducted multiple testing corrections by False Discovery Rate (FDR) on the list of variants to address the issue of inflated type I errors due to the testing of multiple SNPs.

Chapter 4

Results

4.1 Variants Detection by Whole-Exome Sequencing

Whole-Exome Sequencing (WES) was conducted on a total of 24 individuals, including 12 with HAEC (referred to as cases) and 12 HSCR (referred to as controls). This analysis yielded a total of 77,396 variants meeting the criteria of a read depth of ≥ 10 and a call quality of ≥ 10.

Among these variants, 14,781 were unique to HAEC patients, 14,316 were exclusive to HSCR patients, and 48,299 were found in both groups. Despite the extensive variant analysis, it was not possible to attribute all HAEC cases to a single variant or a single gene combination of rare variants that were absent in HSCR patients. To further refine our analysis, I employed two distinct filtering strategies. The first strategy focused on variants with predicted protein-level effects and low prevalence in the general population. The second strategy considered the differential distribution of variants between HAEC and HSCR patients. Thirty and 41 variants were identified by these two strategies, respectively, for a total of 71 variants located in 64 genes. A ranking of relevance to HAEC was created based on:

- putative effect on the protein and comparison between the allele frequency in cases and controls detected by Next Generation Sequencing (NGS) and those reported in the genome aggregation database (gnomAD, https://gnomad.broadinstitute.org/ accessed on 14 December 2020) (WES score),
- biological role of the gene, as supposed from literature data (biological score).

Of note, although the biological relevance to HAEC has been assigned in a blinded way with respect to the WES relevance, 36.8% of variants in genes likely relevant to HAEC had a high WES score, in comparison to 15.4% of variants in genes only possibly or unlikely to be related to HAEC [65]. Subsequently, I selected the top five most promising variants, which were located within the JAK3, OSMR, PIKFYVE, and NLRP14 genes, for validation and replication via Sanger sequencing. We decided to also consider JAK3 even if it is a synonymous variant, since the latter appear to be important in the onset of various disorders, precisely in a modifier type mechanism [66]. After confirming all genotypes, PRMT2 was excluded from further investigations as it was not confirmed in one HAEC heterozygote patient (as outlined in Table 3). Performing genotyping on an additional set of 23 HAEC and 42 HSCR patients yielded similar allele frequencies between HAEC and HSCR patients for the JAK3 and NLRP14, while I observed a higher frequency in HAEC cases compared to the control group for the two variants in OSMR and PIKFYVE variants, a difference that was statistically significant for the rs34675408 variant in OSMR (data not shown).

Table 3: *Variants that ranked at the top after Next Generation Sequencing (NGS) filtering and prioritization.*

Rank	Chr:Pos	Gene	AA	rs	Ref/Alt	gnomAD °	NGS MAF	Replicate MAF		Total Geno (omo/HT/WT)		Total MAF		pvalue*
							Cases	Cases	Controls	Cases	Controls	Cases	Controls	
1	19:17940842	JAK3	A1094A	rs3212780	G/A	0.281	0.375	0.250	0.321	3/14/17	3/21/30	0.294	0.250	0.6002
2	21:48069682	PRMT2	R229W	rs76937225 §	C/T	0.039	0.292	-	-	-	-	-	-	-
3-4	5:38884071	OSMR	H187Q	rs34675408	T/G	0.070	0.292	0.117	0.057	2/17/53	0/11/97	0.146	0.051	0.0024
3–4	2:209190632	PIKFYVE	p.S1033A	rs999890	T/G	0.140	0.292	0.150	0.143	2/9/21	0/6/27	0.203	0.091	0.0850
5–6	11:7059981	NLRP14	R55G	rs61063081	G/A	0.207	0.333	0.206	0.268	1/13/15	5/12/36	0.259	0.208	0.5581
5–6	11:7091569	NLRP14	p.L1010F	rs17280682 §	C/T	0.207	0.333	-	-	-	-	-	-	-

°Non Finnish European controls; Minor allele frequency (MAF) in controls was 0 for all the selected variants; Fisher's exact test p-values on NGS + replicate data; rs76937225 (PRMT2) was not confirmed in one patients, rs17280682 (NLRP14) was not validated because in complete linkage disequilibrium (LD) with the selected variant rs61063081.

Further enlarging the screening of the OSMR variant to a total of 72 HAEC and 108 HSCR patients, I determined an overall minor allele frequency (MAF) for the G allele of 14.6% for HAEC and 5.1% for HSCR controls (p=0.0024), as detailed in Table 3. Additionally, I conducted genotyping for the rs34675408 variant in OSMR for both parents of the 64 samples (38 HAEC and 26 HSCR) for which DNA was available. The Transmission Disequilibrium Test (TDT) provided support for an over-transmission of the G variant allele from the parents to the HAEC-affected patients, with the allele being transmitted 11 times out of 13 (p = 0.0126). In contrast, this over-transmission was not observed in HSCR patients, with only 2 alleles transmitted out of 7, as indicated in Table 4.

Table 4: *Transmission Disequilibrium Test (TDT) on Hirschsprung (HSCR) Associated Enterocolitis (HAEC) and HSCR available trios*

Cases (HAEC)							
Gene	SNP	REF	ALT	T:U*	OR(95%CI)	Chi-Square	pvalue
OSMR	rs34675408	T	G	11:2	5.5 (1.2-24.8)	6.231	0.0126
				Controls (HSCR)			
OSMR	rs34675408	T	G	2:5	0.4 (0.1-2.1)	1.286	0.257

**Transmitted (T) and Untransmitted (U) ratio Family-based association performed on the Oncostatin-M receptor (OSMR) Single Nucleotide Polymorphism (SNP) rs34675408 using the Transmission Disequilibrium Test (TDT).*

4.2 In silico analysis of OSMR variant

The OSMR protein is a component of type I and type II cytokine receptor families expressed in several tissues, including colon and rectum. The Single Nucleotide Polymorphism (SNP) rs34675408 c.561T G is a missense variant in the exon 5 of OSMR leading to p.H187Q. The localization of p.H187Q in the OSM binding region of the large extracellular domain suggests that this variant could exert its susceptibility effect by modulating the OSMR activation (Figure 3).

Figure 3: *The figure illustrates the secondary and tertiary structure of the OSMR protein.*

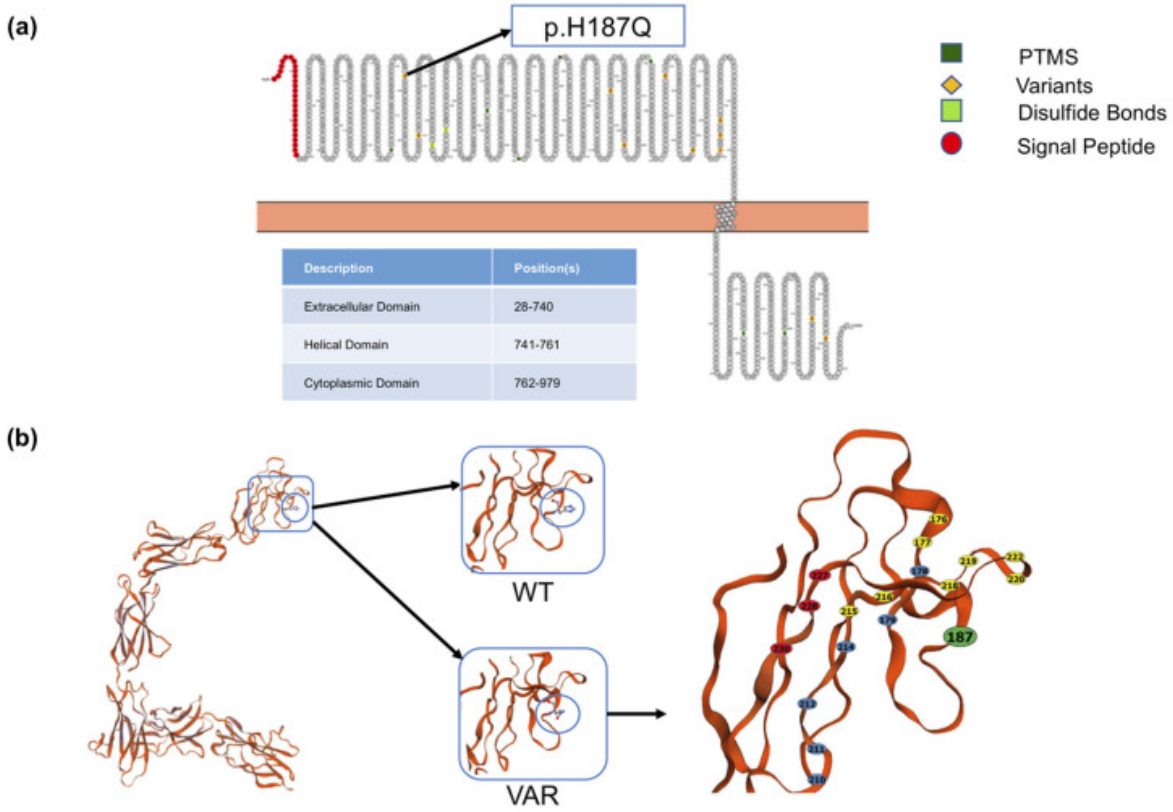


Figure 3: (a) Displays the extracellular, transmembrane, and intracellular domains, along with the amino acid (AA) sequence, post-translational modifications (PTMs), and known variants, including the p.H187Q SNP indicated by an arrow. (b) Shows the 3D model of the OSMR protein structure from Swiss Model server [45], with an enlargement of the region around H187. The wild-type (wt) and p.H187Q variant sequences are distinguished. A zoomed-in view highlights three hotspot sites around H187 (yellow, blue, and red), with Tyr214 identified as a common hotspot binding residue reported by Du et al [46].

To investigate the variant effect on the protein conformation, I compared the OSMR proteins with and without the amino acid change using in silico tools. The tertiary structure designed by Swiss Model [45] exhibits some divergences in the model parameters between the H187 and Q187 structures. In particular, a slightly different conformation was visible (Figure 3), with the H187 allele showing, differentially from the Q187 allele, a distorted angle at the Ramachandran Plot (Figure 4) [65].

Figure 4: *The Ramachandran Plot was investigated for a geometric validation of the OSMR tertiary structure model.*

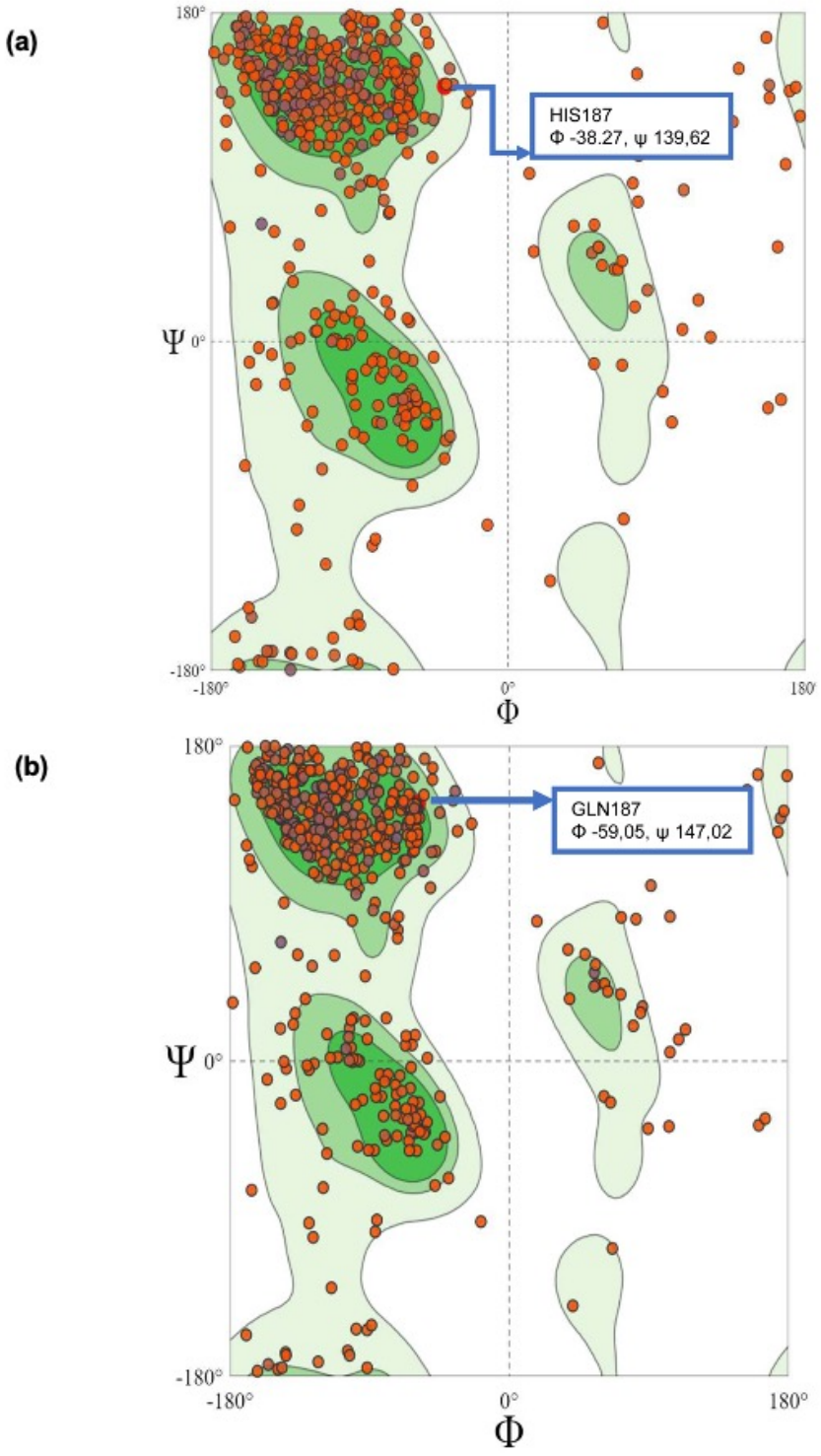


Figure 4: The plot resolution correlates with the residues found inside the most favored , regions. From our Ramachandran plot analysis, the residue H187 was estimated to create a bad angle (a), while the variant p.H187Q did not (b).

4.3 Proteomics and Pathway Analyses Reveal Pathway Perturbations Driven by the OSMR Variant

I evaluated the effect of the rs34675408 variant (c.561T>G, p.His187Gln) detected in the OSM receptor (OSMR) on the OSM-OSMR cascade through a proteome analysis. To this end, I compared two cell lines derived from two HAEC patients, TT wt and GG homozygous for the rs34675408 variant, before and after stimulation with the ligand OSM. Principal Component Analysis (PCA) evidenced two outliers that were then excluded by the following analysis (Figure 5).

Figure 5: *Principal component analysis (PCA) was performed to investigate the samples run by Mass spectrometry.*

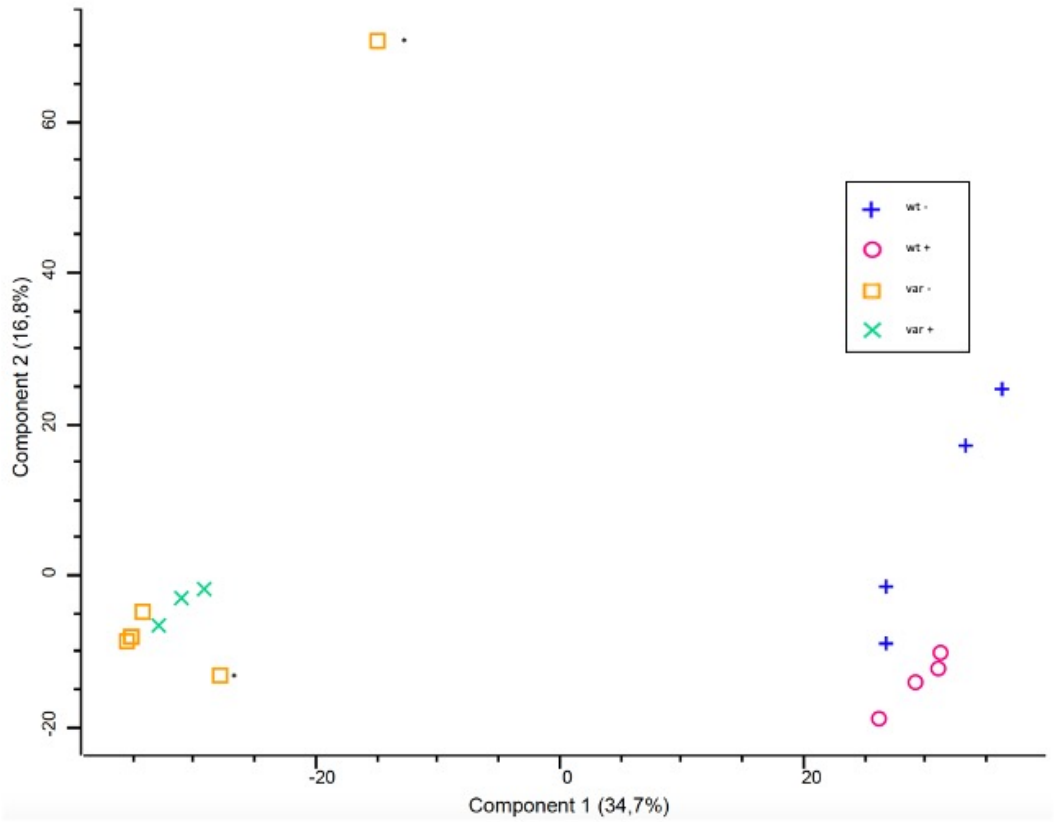


Figure 5: Principal component analysis (PCA) was performed to investigate the samples run by Mass spectrometry and check for outliers. The two samples mistaken during handling and therefore excluded from subsequent analyzes are indicated by an asterisk. The dot on the right of the symbol indicate that the sasmple was considered as an outlier, the value was reconducted to an experimental error and the samples excluded from the downstream analysis

Mass spectrometry analysis identified a total of 458 proteins that exhibited significantly different intensities after OSM treatment between lymphoblasts carrying the T allele (wt+) and those carrying the G allele (var+) and that had no different intensities before the treat-

ment. Hierarchical clustering analysis and a visual heatmap revealed two distinct protein clusters: 271 proteins were more highly expressed in the wt+ sample compared to the var+ sample, and 187 proteins exhibited higher expression in the var+ sample than in the wt+ sample (Figure 6).

Figure 6: *Protein cluster analysis.*

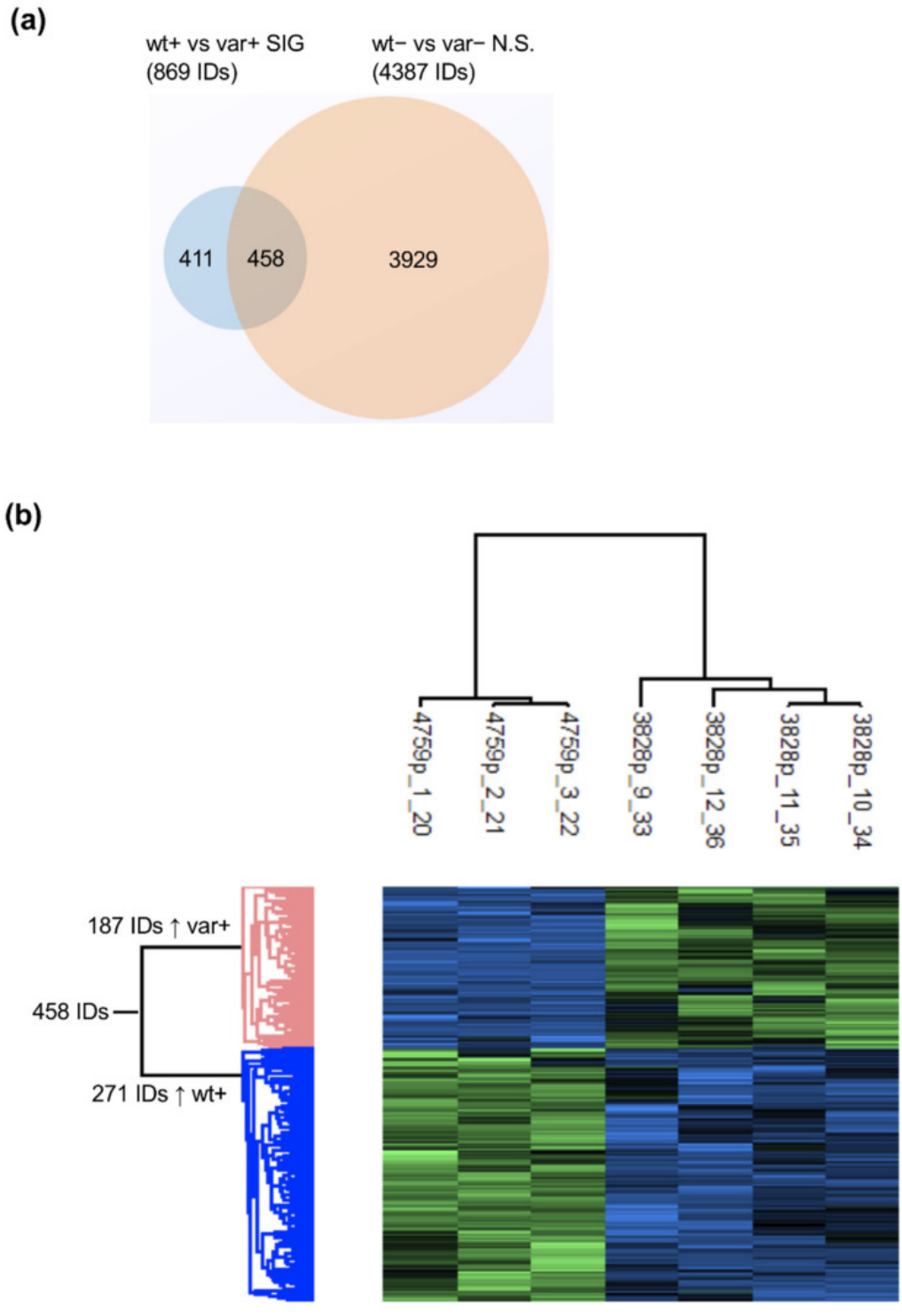


Figure 6: The analysis focuses on proteins exhibiting differential expression between the OSMR H187 patient's cell line (LY3828, four replicates) and the p.H187Q homozygous pa-

tient's cell line (LY4759, in triplicate) following OSM stimulation. The two samples are denoted as wt+ and var+, respectively. (a) The Venn diagram depicts the 458 proteins showing differential expression between wt and var cell lines after OSM treatment (indicated by the "plus sign") that were not significantly different between wt and var before treatment ("minus sign"). (b) The hierarchical analysis reveals two distinct clusters in the heatmap for these 458 proteins: 187 proteins were overexpressed in var+ compared to wt+ (shown in blue on the heatmap color scale), while 271 proteins were overexpressed in wt+ compared to var+ (in green).

Pathway analysis of the proteins overrepresented in the wt+ set indicated a significant enrichment of immune response pathways, a feature entirely absent in the cluster of proteins overrepresented in var+ (Figure 7). Specifically, module M1 $q < 10^{-4}$ comprised 245 terms, including several terms related to immune response, antigen receptor-mediated signaling, T cell activation and regulation, toll-like receptor 3 (TLR3) signaling, viral processes, defense response to organisms and peptides, response to interferon-alpha and -gamma, ERK1/ERK2 cascades, autophagy, and response to wounding. It's worth noting that the last three pathways were borderline in terms of significance.

Figure 7: *Pathway analysis in H187 (LY3828) and p.H187Q homozygous variant (LY4759) cells.*

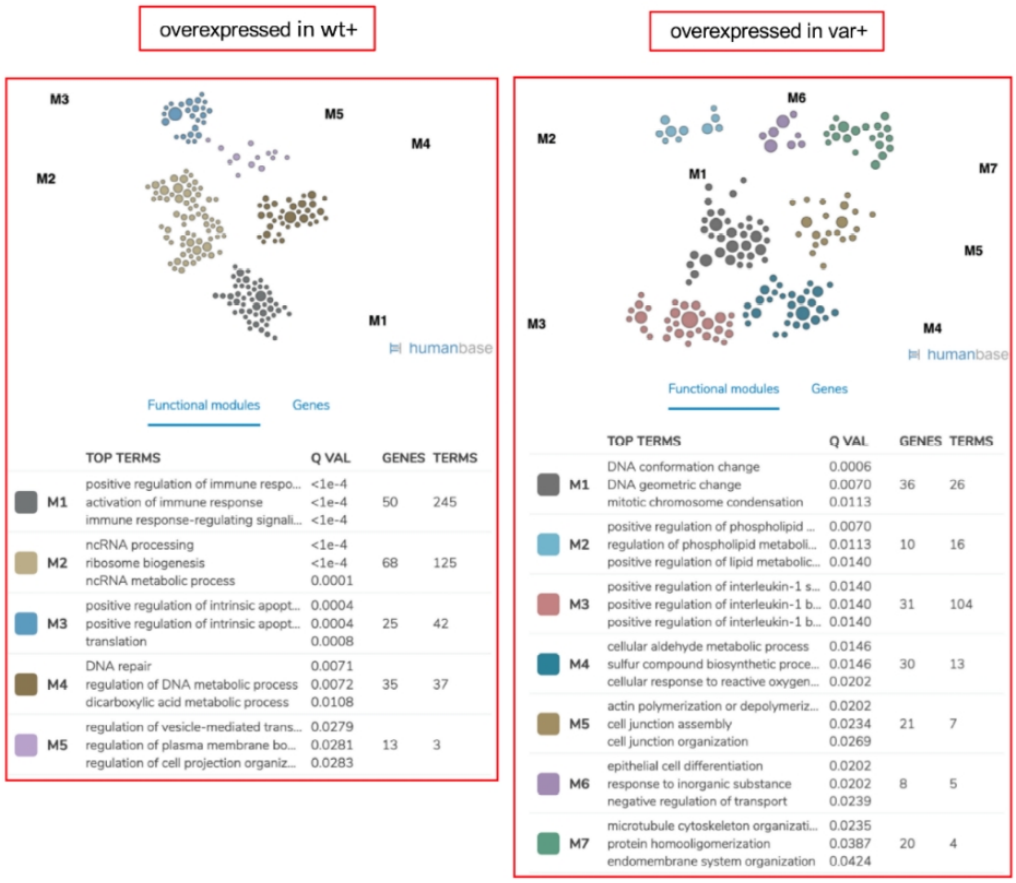


Figure 7: The screenshot illustrates the pathway analysis conducted using HumanBase on the protein clusters identified through Nano-LC mass spectrometry. On the left side are the 5 modules of pathways enriched based on the cluster of proteins overexpressed in wt+ compared to var+, while on the right are the 7 enriched pathway modules in the cluster of proteins overexpressed in var+ compared to wt+.

The second most notable module, M2, was associated with RNA and cell mitosis, while the M3 module predominantly encompassed proteins involved in regulating apoptosis, responding to DNA damage stimuli, and catabolic processes. Additionally, pathways related to DNA repair, replication, protein complex disassembly, and cell polarity (M4) were enriched, along with pathways associated with the regulation of vesicle-mediated transport and cell projection organization (M5) within the wt+ vs. var+ protein cluster.

On the other hand, among the proteins overexpressed in the var+ sample, there was significant enrichment in pathways related to DNA conformation (also present in the wt+ cluster

but to a lesser extent) (M1), lipid metabolism, and kinase activity (M2). Furthermore, enrichment was observed in pathways related to interleukin-1 production and regulation, the tumor necrosis factor-mediated signaling pathway (TNF), the -Jun N-terminal Kinase (JNK) cascade, and the positive regulation of the stress-activated Mitogen-Activated Protein Kinase (MAPK) cascade (M3). Other pathways enriched in var+ sample included those related to responses to oxidative stress, sulfur compound biosynthetic and metabolic processes (M4), actin and cell junction organization (M5), negative regulation of transport, epithelium differentiation, and development (M6), and microtubule cytoskeleton organization (M7).

In an effort to specifically identify proteins unaffected by OSM activation in the presence of the OSMR SNP variant (hereinafter referred to as "lost-effect proteins"), I further refined the analysis by filtering proteins from both the wt+ overexpressed and var+ overexpressed clusters that showed no differential expression between var- and var+. The wt+ overrepresented cluster yielded results similar to those obtained with the larger panel, with 243 out of 271 proteins showing this characteristic. Notably, a new enriched pathway related to the unsaturated fatty acid biosynthetic process emerged in this context (Figure 8). The var+ overrepresented cluster, although representing a slightly smaller group (184 out of 187 proteins), revealed additional significantly enriched pathways associated with the glutathione metabolic process, detoxification, endoplasmic reticulum (ER), particularly ER stress, transport, retrograde transport, and degradation.

Figure 8: *Pathway analysis of the OSM-lost effect proteins.*

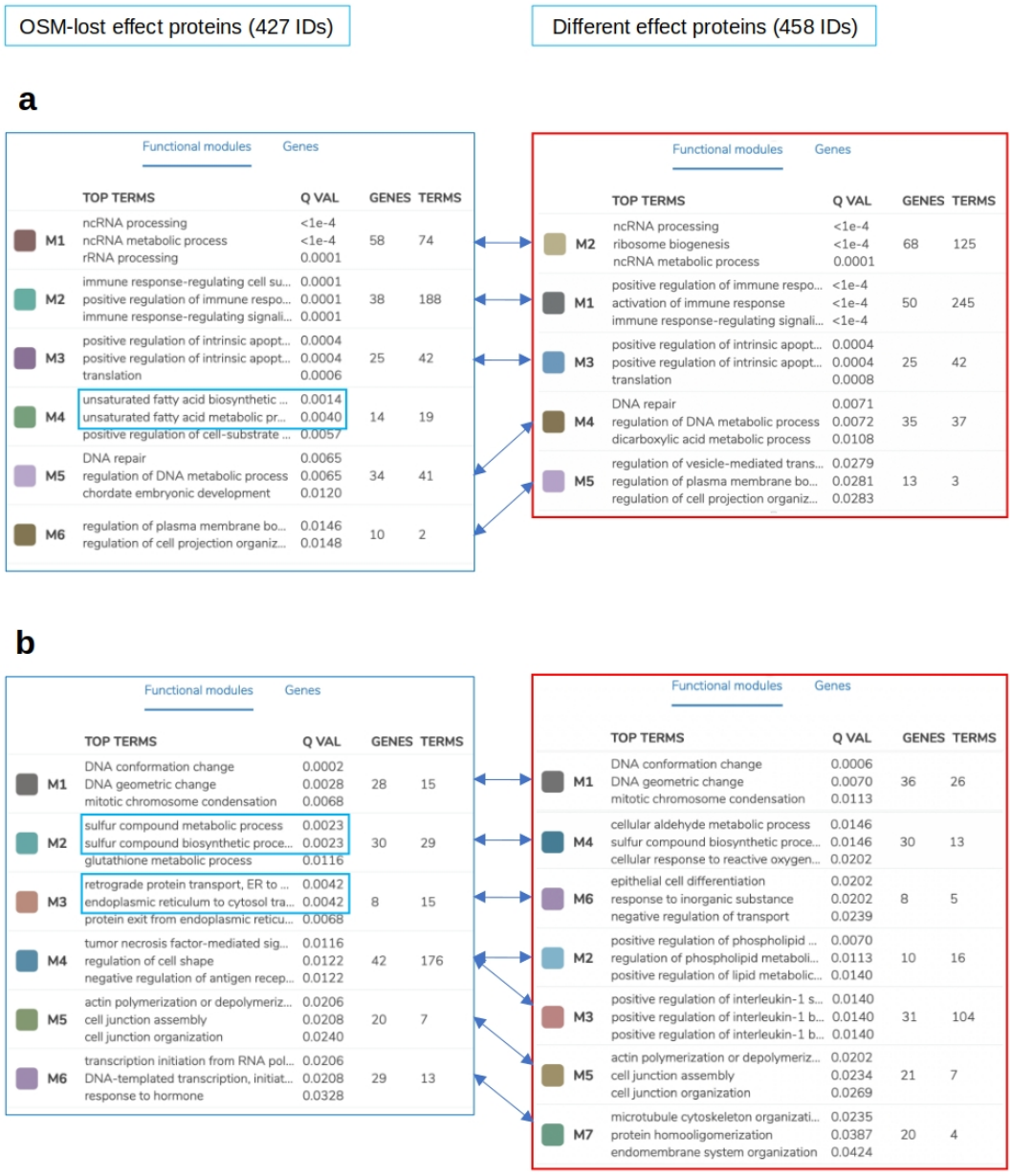


Figure 8: The OSM-lost effect proteins are those proteins part of the 458 differentially expressed between wt and var cell lines after OSM stimulation that are specifically not dif-

ferentially expressed between var- and var+. a) the proteins more expressed in the wt+ than in the var+ cell line and b) the proteins more expressed in the var+ than in the wt+ cells are shown. The original 458 different-effect proteins analysis is represented on the right for comparison. Correspondence between the modules is indicated by the arrows, while the additional pathways that have emerged are highlighted by the surrounding rectangles.

4.4 OSMR Expression in HAEC Lymphoblastoid Cell Lines

To assess if OSMR was expressed in lymphoblastoid cell and thus if it was reasonable to perform the OSM stimulation in the proteome analysis, I checked its expression by Western blot assay after OSM stimulation. The OSMR expression was confirmed in lymphoblasts from HSCR and HAEC patients, in addition to fibroblasts, used as positive control as known to express the receptor. Moreover, the OSMR expression seemed to be even higher in the two HAEC than in the HSCR-only cell lines (Figure 9).

Figure 9: *Western Blot performed to confirm the OSMR expression in Lymphoblast cell lines.*

Figure 9: OSMR was also expressed in lymphoblasts from HSCR (LY3956) and HAEC patients (LY3828 and LY4759). It was also confirmed to be expressed in cell line from a patient affected by Duchenne muscular dystrophy (LY1765), a disease not related to HSCR but in which OSMR is known to be expressed [67], and from Fibroblasts (SC591) used as positive controls. The housekeeping expression was confirmed with a mouse anti Tubulin antibody.

4.5 Transcriptome and Pathways Analysis:

I performed a transcriptome analysis to explore the molecular mechanisms underlying the distinct phenotypic spectrum between HAEC and HSCR patients at the transcriptome level. I sequenced Intestinal Epithelial Lymphocytes (IELs) from 16 samples (6 HAEC, 6 HSCR, and 4 CTRL patients), 8 per chip of sequencing. The data size within each chip was approximately 10.5 GB and 10.01 GB, respectively. The average number of reads per chip was 90,715,357.5, with 66% of these reads deemed usable, while the remaining 34% were discarded due to factors such as empty wells, polyclonal reads, primer dimers, or low quality. In the first chip, the minimum number of reads per sample was 4,911,173, and the maximum was 14,959,430. In the second chip, the minimum was 7,685,695, and the maximum was 12,717,130. Regarding alignment, it has been achieved an average of 7 GB of reads aligned to the hg19 AmpliSeq transcriptome v1.1, with 95% of reads correctly aligned for both chips. RNA-seq targeted was employed to explore the molecular mechanisms underlying the distinct phenotypic spectrum between HAEC and HSCR patients at the transcriptome level.

I used R for the quality control steps and the Differential Expression Analysis. I started with the creation of a heatmap using normalized data. I visualized a heat map on log transformed data using only the most variable set of genes, as described in the method section. This heatmap (Figure 10) visually represents the expression patterns of the selected genes, offering insights into their differential regulation.

Figure 10: *The hierarchical analysis and Heatmap.*

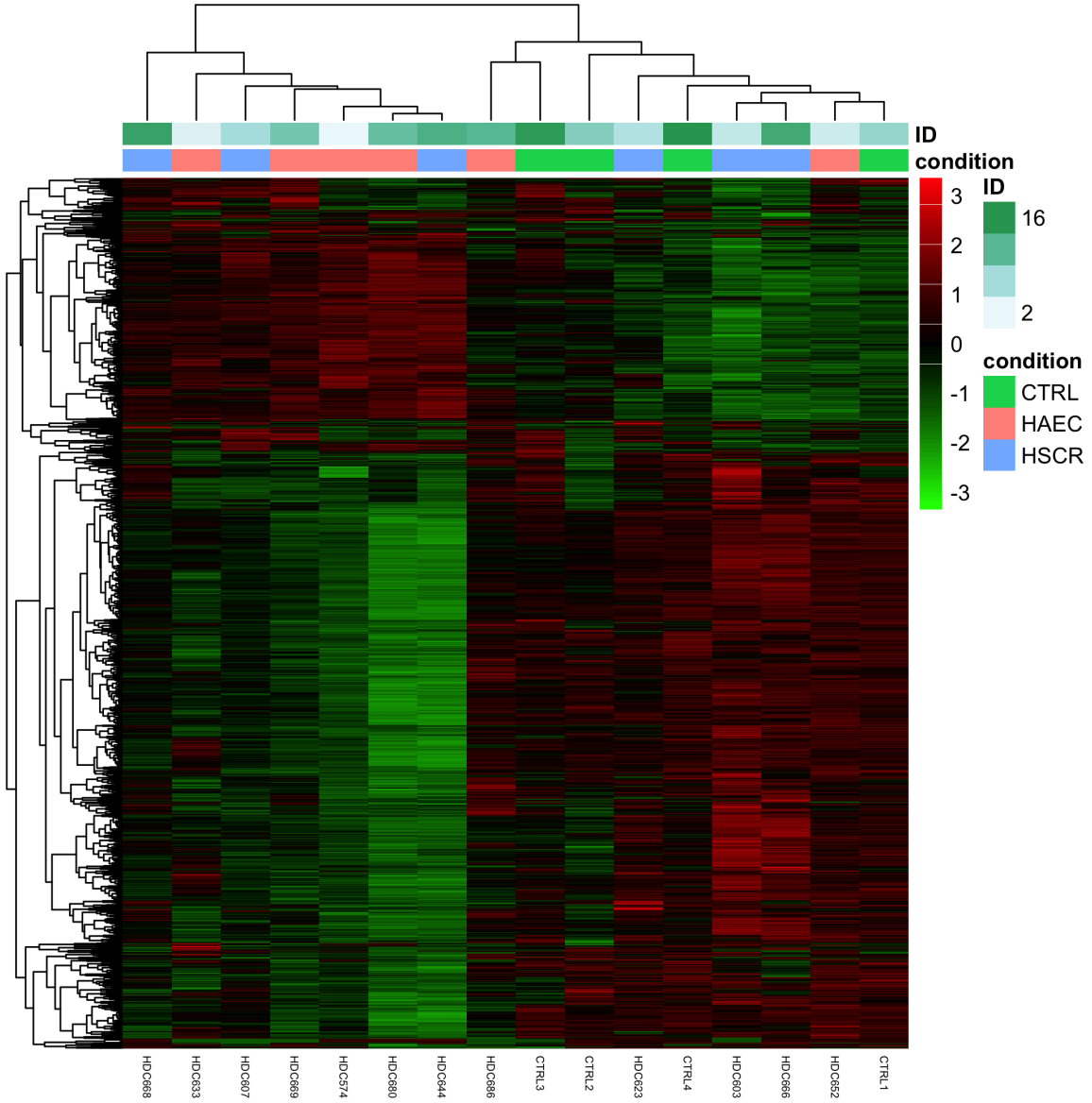


Figure 10: The hierarchical analysis reveals two distinct clusters of transcript that were overexpressed (shown in red on the heatmap color scale), while underexpressed proteins are shown in green. Among the hierarchical clustering I can note 4 out 6 HAEC patients are pretty stacked together and far away from controls.

The principal component analysis (PCA) plot (Figure 11) revealed pretty homogeneous groups with no outliers, exhibiting great similarity between HAEC, HSCR and CTRL groups. I obtained comparable results performing the PCA after batch correction for the two different runs of sequencing.

Figure 11: *The Principal Component plots from IELs Transcriptome Analysis.*

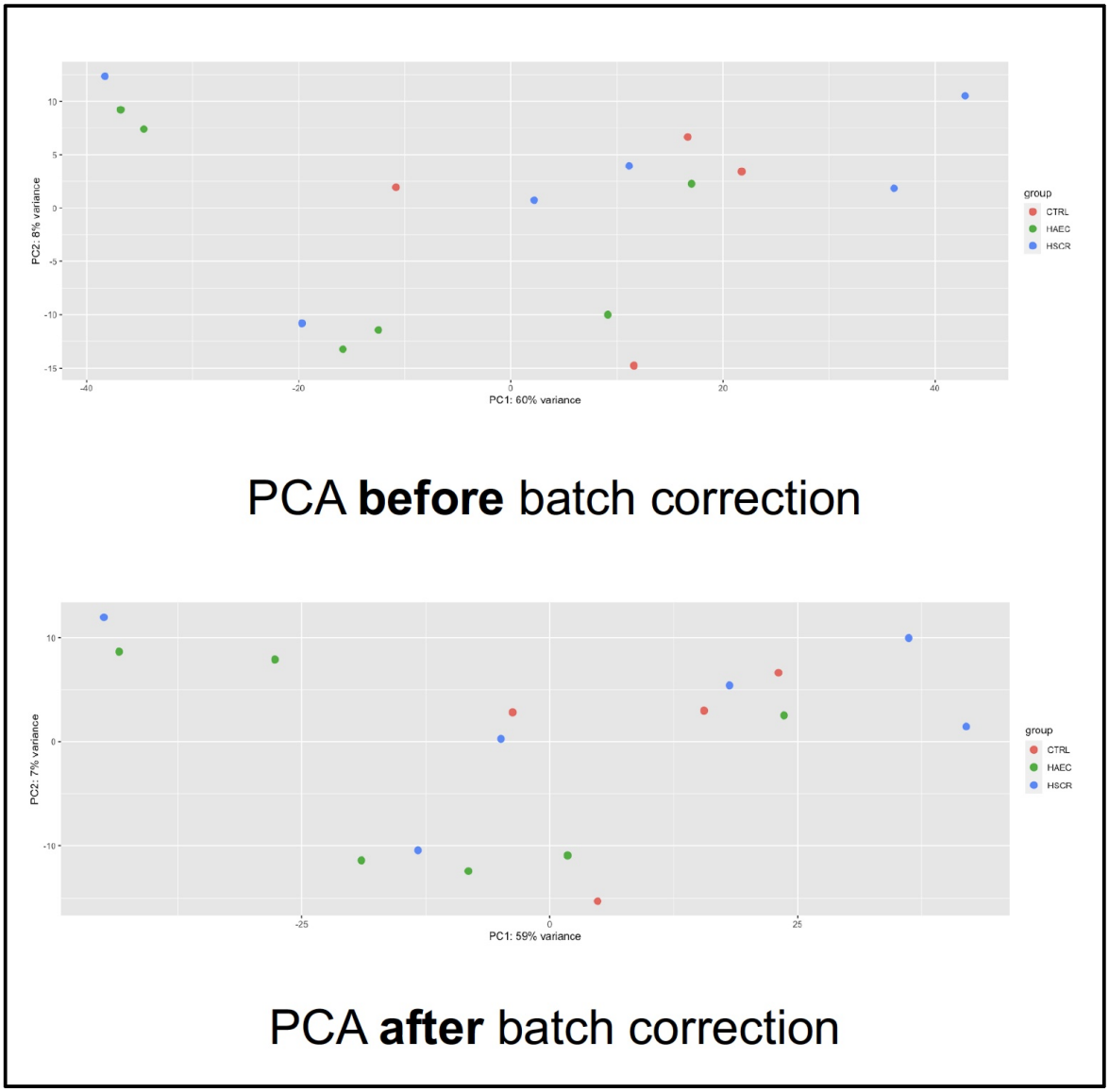


Figure 11: The Principal Component plots revealed the two most variable components, indicating a lack of clear clusters between cases and controls and suggesting significant inter-patient variability. Even after applying batch corrections for the two different sequencing runs, the plot does not show noticeable clustering between groups.

Through the analysis of differential gene expression, a crucial step in understanding how gene activity varies between different experimental conditions, I compared HAEC, HSCR and not inflammatory controls, performing DE analysis and to visualize the most differentially expressed genes I found 1342 transcripts that were nominally significantly over- or under-expressed in the HAEC vs HSCR patients, 1670 differentially expressed (DE) in the

HAEC vs CTRL comparison and 1389 transcripts in the HSCR vs CTRL comparison. Three, 10 and 1 transcripts were still differentially expressed after multiple test corrections (FDR) for the statistical significance in HAEC vs CTRL, HSCR vs CTRL and HAEC vs HSCR respectively. Of the 3 DE genes comparing samples from patients with Hirschsprung Associated Enterocolitis (HAEC) to those from control patients (CTRL), one gene was found to be upregulated, while two genes were downregulated. Similarly, when comparing patients with Hirschsprung's disease (HSCR) to control patients, four genes showed increased expression, while six genes exhibited a decreased expression. Lastly, comparing patients with HAEC to those with HSCR revealed one upregulated gene, corresponding to HLA-C, and no down regulated genes.

Table 5: *List of transcripts resulted to be still significant after False discovery rate correction.*

	Gene	BaseMean	log2FC	lfcSE	Stat	pvalue	padj
HAEC_vs_CTRL	CCNI2	65	-1.5	0.25	-6	1.50E-09	0.000015
HAEC_vs_CTRL	SYNM	21	2.3	0.46	5	5.80E-07	0.0038
HAEC_vs_CTRL	HLA-B	690	-6.8	1.1	-6.3	2.30E-10	4.40E-06
HSCR_vs_CTRL	HLA-B	690	-6.8	1.1	-6.4	1.90E-10	2.50E-06
HSCR_vs_CTRL	HLA-C	2100	-8	1.3	-6	2.50E-09	0.000016
HSCR_vs_CTRL	NARF	340	0.72	0.15	4.8	1.50E-06	0.0056
HSCR_vs_CTRL	TRIP6	200	-1.6	0.33	-4.8	1.70E-06	0.0056
HSCR_vs_CTRL	CD248	56	1.4	0.31	4.7	2.50E-06	0.0065
HSCR_vs_CTRL	KIAA0913	860	0.74	0.16	4.6	3.60E-06	0.0078
HSCR_vs_CTRL	EYA2	31	-3.5	0.79	-4.4	9.60E-06	0.018
HSCR_vs_CTRL	SYNM	21	2	0.47	4.3	0.000018	0.029
HSCR_vs_CTRL	ASPH	38	-1.5	0.36	-4.2	0.000029	0.039
HSCR_vs_CTRL	MLEC	850	-1.1	0.27	-4.2	3.00E-05	0.039
HAEC_vs_HSCR	HLA-C	2100	7.4	1.2	6.1	8.80E-10	0.000017

Table 5: In the table are reported the Gene Name (rownames,res), the mean number of probes for transcripts (Base Mean), the value log2 transformed (log2FoldChange), the standard error of the log2fold change values (lfcSE), the p value calculated with the Wald test and the p value adjusted with False Discovery Rate (padj).

Among the transcripts identified as statistically significant after false discovery rate (FDR) correction, several genes stand out with potential implications for HSCR/HAEC and related processes. Notably, NARF, recognized as a hypoxia-induced coactivator for OCT4, where the current understanding enlighten the role of the hypoxia-responsive (hypoxia-inducible factor) and inflammatory (nuclear factor-B) transcription factor families in rheumatoid arthritis, inflammatory bowel disease and colorectal cancer [68]. SYNM, an intermediate filament family member, forms a linkage between desmin and the extracellular matrix, providing essential structural support in muscle. Desmin and intermediate filament in general have been reported several times in the molecular pathogenesis mechanisms in Hirschsprung disease

[69], [70]. The gene CD248, predicted to have extracellular matrix binding activity and extracellular matrix protein binding activity, emerged as a potential player in cell migration and to act upstream of or within several processes, including anatomical structure regression, lymph node development; and positive regulation of endothelial cell apoptotic process [71]. This function implies its significance in processes related to tissue development, repair and inflammation, which may have relevance in the context of HSCR. Additionally, ASPH, implicated in calcium homeostasis, presents an intriguing connection to Hirschsprung's disease. Studies have suggested the importance of calcium homeostasis in the context of HSCR and IBDs, linking ASPH to potential roles in disease mechanisms [72], [73], [74].

In particular, among the nominally significantly differentially expressed genes between HAEC vs HSCR, 46 transcripts are also involved in IBDs pathogenesis, although not significant after correction for multiple tests.

From the GSEA analysis I pointed out an enrichment in immune and inflammatory pathways. In particular, I found 1369 enriched pathways in the immunologic signature gene sets, among which 1005 in the HAEC vs CTRL group.

Table 6: Transcripts nominally significant through Differential expression analysis between HAEC vs HSCR group and already known to be involved in Intestinal Bowel Disease genetics or pathogenesis. Log2fold change shows the change from HSCR to HAEC.

Table 6: *Transcripts nominally significant through DEGs analysis.*

gene	log2FoldChange	pvalue	reference
HLA-C	7,37	8,80E-10	Jung Eun Suk et al., 2016
IL22	3,54016	0,00041	Li LJ et al., 2014
CASP9	-0,5746	0,00126	Linares-Pineda TM et al., 2018
MYC	0,90221	0,0018	Macpherson AJ et al., 1992
JOSD2	-1,012	0,00297	Qiuyun Xu et al., 2021
SMAD7	-1,0375	0,00298	Monteleone Giovanni et al., 2023
CCL7	2,85008	0,00415	He J et al., 2019
PUS10	-1,0948	0,00538	Festen EAM et al., 2011
CYP2C19	-2,308	0,00712	Yamamoto R et al., 2018
CYP3A5	-1,9959	0,00895	Yamamoto Y et al., 2020
SULT1A2	-2,0507	0,0098	Zhou Ting et al., 2021
SKIV2L	0,58902	0,01026	Kammermeier J et al., 2014
SOX11	-3,7273	0,01148	Tsang SM et al., 2020
SBNO2	0,7754	0,01152	Ntunzwenimana JC et al., 2021
HTR3C	-3,6668	0,01154	Motavallian A et al., 2013
EIF2AK4	-0,6613	0,01272	Bretin A et al., 2016
TFAP4	0,77089	0,01352	Bhat AA et al., 2022
TNFAIP3	0,89494	0,01505	Zheng C et al., 2018
WDR78	-1,5878	0,01561	Hu W et al., 2022
NLRP3	1,48666	0,01606	Zhen Y et al., 2019
GSTM1	-2,8322	0,02047	Zhou YJ et al., 2019
TLR1	1,25221	0,02174	Lu Y et al., 2018
SLCO2A1	-2,9103	0,02313	Ito N et al., 2023
SULT1A1	-1,4295	0,02394	Zhou Ting et al., 2021
ALPI	-1,8165	0,02425	Parlato M et al., 2018
ATG2A	-0,3392	0,0247	Brinar M et al., 2012
CDKN2B	-1,5932	0,02622	Rankin CR et al., 2019
OTUD3	0,63494	0,0277	Ntunzwenimana JC et al., 2021
SLC16A3	-0,9546	0,0287	He L et al., 2018
S100A8	1,95515	0,02917	Okada K et al., 2019
TNF	1,13275	0,02977	Peyrin-Biroulet L et al., 2021
TNFSF10	-1,2044	0,03277	Ślebioda TJ et al., 2014
CYP3A4	-3,1871	0,03381	Wilson A et al., 2019
SLC9A3	-2,1475	0,0359	Fonseca-Camarillo G et al., 2012
SLC26A3	-2,4011	0,03668	Kumar A et al., 2021
IPMK	-0,8144	0,03868	Park SE et al., 2022
MYO9B	0,42717	0,03894	Wang MJ et al., 2016
CD44	0,86123	0,0391	Hankard GF et al., 1998
CCL2	2,61675	0,0424	Luo X et al., 2023
HCAR2	2,40523	0,04395	Nuzzo A et al., 2021
SLC5A12	-3,8232	0,04427	Michaels M et al., 2023
GSTA1	-2,0585	0,04594	Liu H et al., 2015
ACE2	-2,1773	0,04752	Potdar AA et al., 2021
TNFSF4	1,05124	0,04816	Cooke J et al., 2012
FKBP5	0,66139	0,04838	Skrzypczak-Zielinska M et al., 2021
ELF4	-0,4893	0,0497	Tyler PM et al., 2021

Figure 12: *Enriched pathways of Gene Set Enrichment Analysis (GSEA) related to immunologic signatures gene sets.*

The figure 12 depicts enriched pathways within the gene sets of Gene Set Enrichment Analysis (GSEA) related to immunologic signatures. The comparison focuses on the prevalence of enriched pathways in the HAEC group compared to the control (CTRL) group. The findings underscore a notable dysregulation of immune-related pathways in the HAEC group, highlighting potential immunologic abnormalities associated with this condition.

4.6 Validation by qPCR

I moved forward to validate the identified differentially expressed genes (DEGs) between the distinct biological conditions of HAEC, HSCR and not inflammatory controls. To confirm the true-positive DEGs, I opted for validating 6 genes already involved in IBDs or in Hirschsprung pathogenesis, by high-throughput quantitative reverse-transcription PCR (qPCR) on technical replicates.

Figure 13: *qPCR validation graph.*

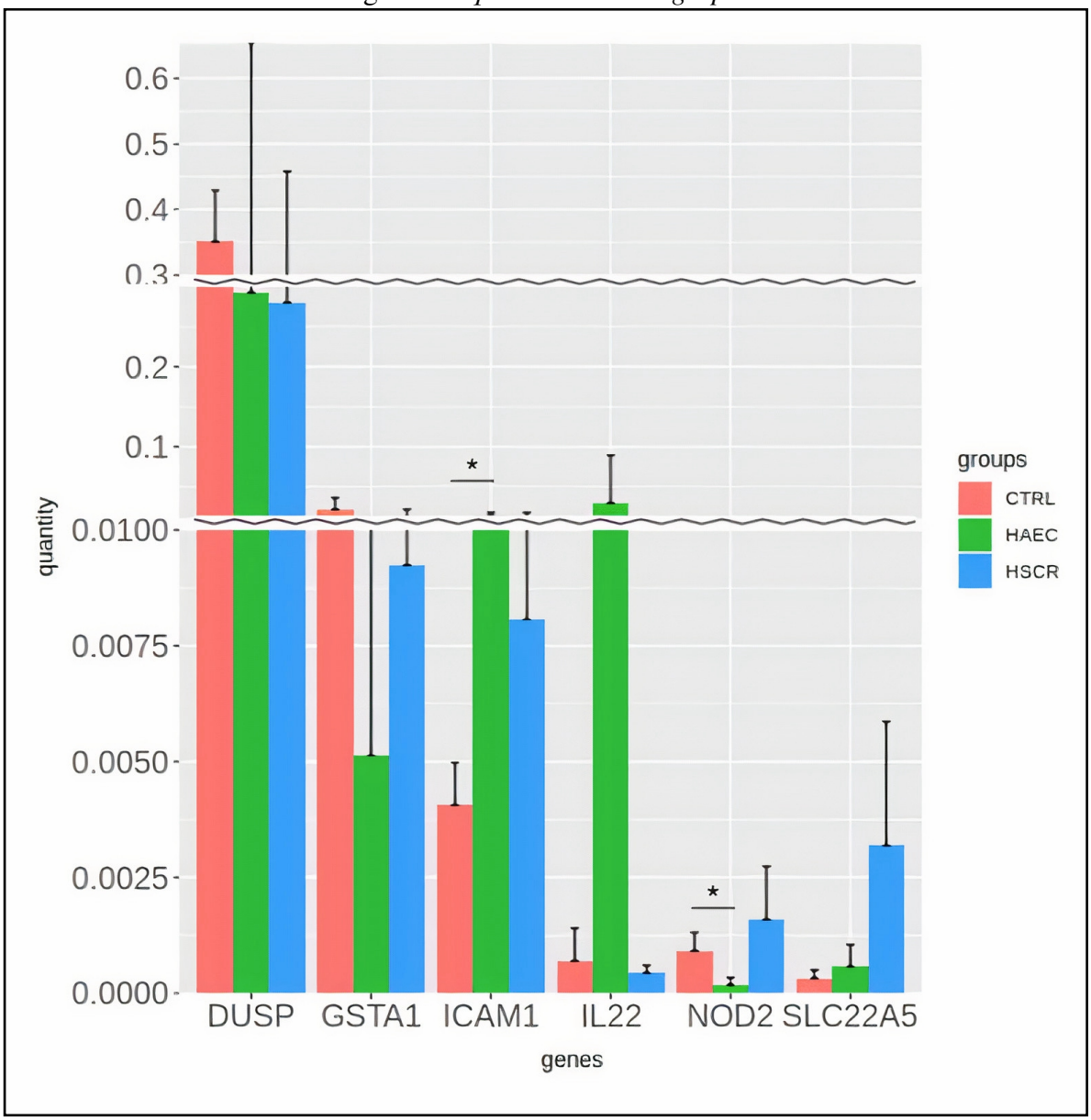


Figure 13: The expression trends of GSTA1, ICAM1, and IL22 were validated through qPCR assays. These six transcripts were selected based on their molecular and pathogenic relevance to either Inflammatory Bowel Diseases (IBDs) or Hirschsprung's disease (HSCR). The color-coded representation (red, green, and blue) distinguishes between various disease groups.

The confirmation of the gene expression results assessed by RNA seq targeted analysis, was achieved for only three out of the six selected transcripts. This limited success can be attributed to several challenges encountered during the experimental process. Firstly, the

available biological material was severely reduced, and only a minimal amount of material remained for analysis. Additionally, the RNA concentration derived from the Isolated Enteric Lymphocytes (IELs) is notably low, further complicating the analysis.

Debates persist in the scientific community regarding the transition from RNA sequencing to quantitative polymerase chain reaction (qPCR). Numerous studies [75], [76], [77], [78] have demonstrated a strong correlation between the results of these two methods. However, when disparities occur, they are attributed to biases inherent in qPCR experiments, influenced by probe biases and the amplified region of cDNA. Consequently, the use of qPCR following RNAseq may not always offer novel insights and, in some instances, may provide information of lower quality compared to RNAseq data. These considerations highlight the necessity of critically evaluating chosen methodologies to ensure the reliability and robustness of experimental outcomes.

To further emphasize the validation of the previous methods, I analyzed the correlation of the expression levels measured by RNA sequencing with those obtained through qPCR experiments. This analysis confirmed a substantial correlation between the two analytical approaches, showing a Spearman's rho coefficient of 0.84 ($p<0.01$). This strong correlation reinforced the reliability of the RNA sequencing results but also demonstrated the consistency and accuracy of the findings obtained through the qPCR methodology and thus confirmed the different expression level between HAEC and HSCR.

4.7 Transcriptome Analysis Replication In Peripheral Blood Monocytes Cells

I performed the transcriptome analysis on the Peripheral blood mononuclear cells (PBMCs) from a subset of the same patients tested on the Intestinal Epithelium Lymphocytes to evaluate this source as a promising substitute of intestinal biopsies. I conducted the transcriptome analysis on PBMCs exclusively from the HAEC and HSCR cohorts. Unfortunately, obtaining a control cohort for PBMCs was infeasible due to the original experimental design, which aimed to collect material from biopsies in HSCR cases. Regrettably, the limited availability of healthy patients with both biopsy and PBMC data restricted our options. Consequently, adhering to stringent filtering criteria, I was constrained to utilize a small sample size of six patients, three with HAEC and three with HSCR, for whom I already possessed expression

data on IELs. I performed the same analysis pipeline used for the IELs: normalization, hierarchical clustering, principal component analysis and differential expression analysis.

Figure 14: *The Principal Component plots from PBMCs Transcriptome Analysis.*

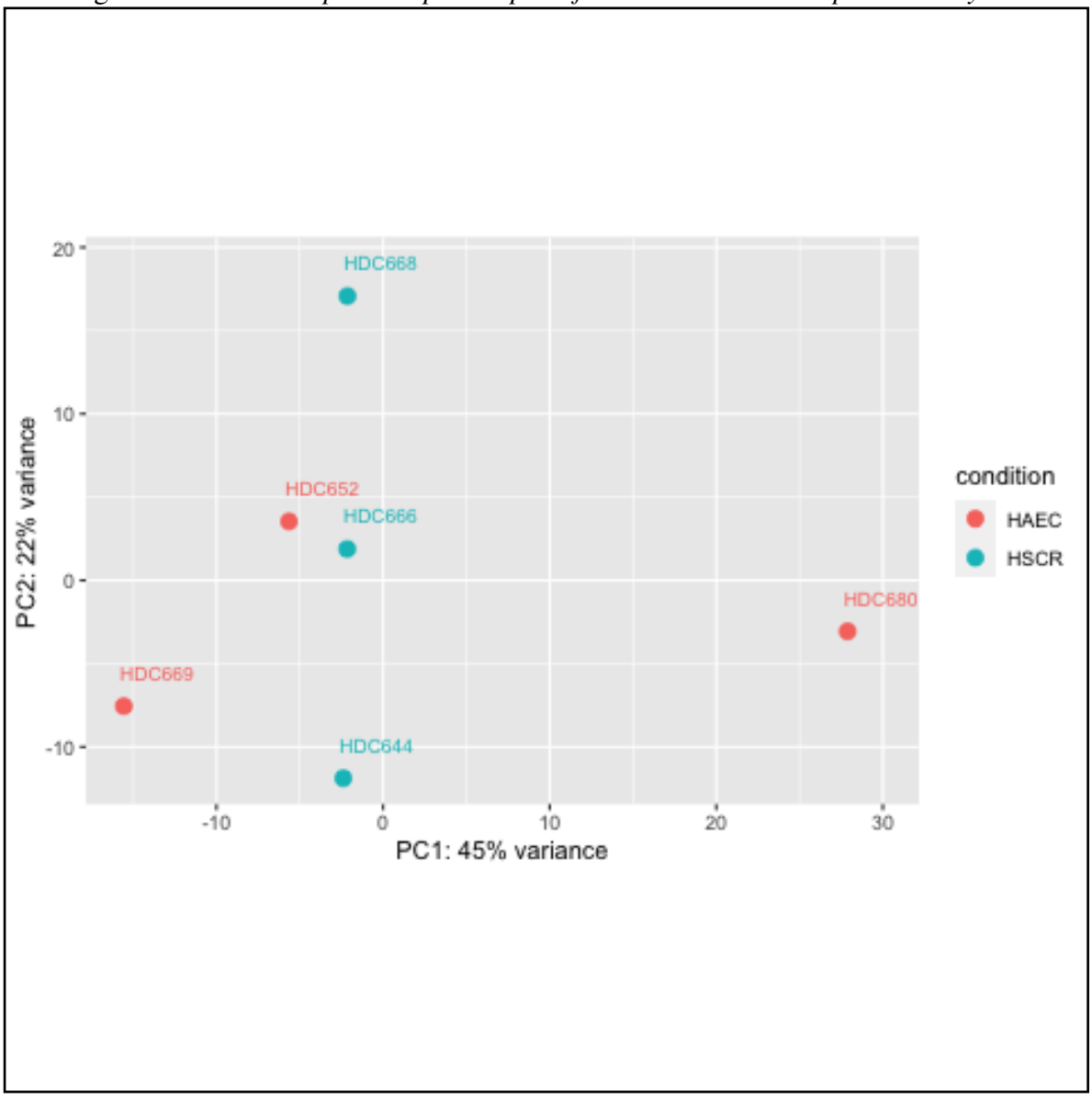


Figure 14: In the principal component analysis (PCA) plot, clear clusters are not evident, which was expected given the inherent complexity and interpatient variability of the disease. The absence of well-defined clusters in PCA underscores the difficulty in categorizing patients into homogeneous groups based on transcriptome analyses.

In comparing HSCR and HAEC, a limited number of significantly expressed transcripts emerged after correction for false discovery rate (FDR). However, no substantial differences

in any pathway enrichment were observed.

Figure 15: *Hierarchical clustering, heat map and sample distance plots from PBMCs.*

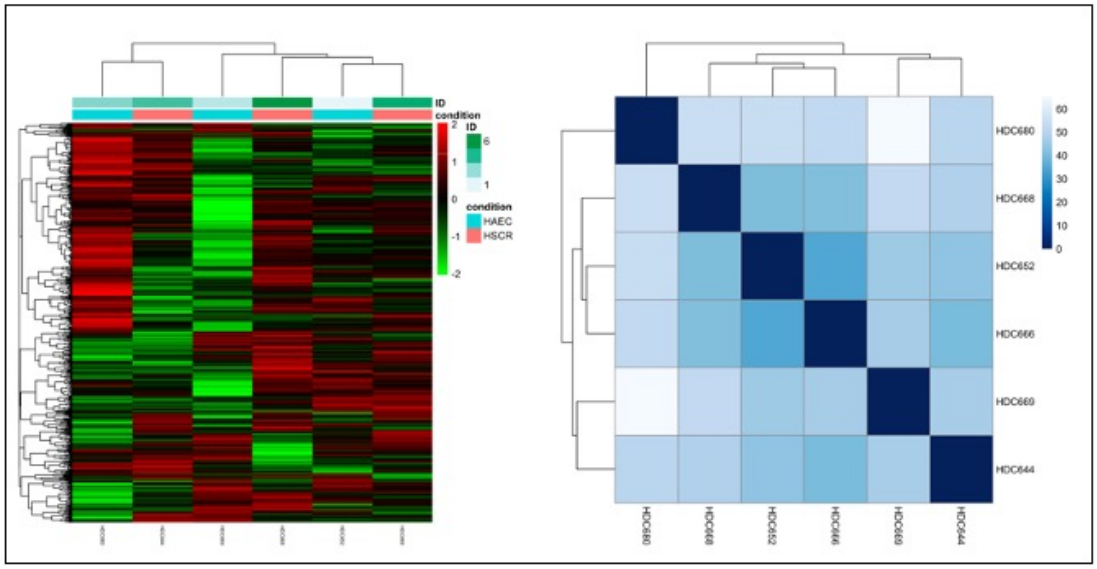


Figure 15: Hierarchical clustering, heat map and sample distance plot after analysis of HAEC versus HSCR, did not identify any clustering among groups, confirming what already observed by PCA

Nevertheless, our primary focus was to conduct an initial exploratory analysis on PBMCs to determine if and how these cells could serve as representative substitutes for other tissues, particularly biopsies, which are more invasive, especially in pediatric cases.

To assess the robustness of our analysis, I investigated the correlation between the expression levels of each transcript of paired samples across the two different runs on IELs and PBMCs (the same 3 HAEC and 3 HSCR sample in common between the IELs and PBMCs compared between the two tissues). Since the Shapiro-Wilk test indicated a non normal distribution of the expression level for each transcript in our small sample (p-value < 0.05 for all six samples), I used the Spearman correlation analysis to systematically correlate the expression levels between the Intraepithelial Lymphocytes (IELs) and the Peripheral Blood Mononuclear Cells (PBMCs) for the six samples (3 HAEC and 3 HSCR). The correlation results revealed strong and significant correlations, suggesting that PBMCs could serve as reliable indicators for the exploration of gene expression levels in HSCR and HAEC samples. The average correlation coefficient computed across these comparisons was 0.868, with

a median of 0.8633. The correlation coefficients ranged from a minimum of 0.830 to a maximum of 0.912. Some transcripts, exhibiting a significant alteration in expression, emerged as outliers in the data in all the samples (Figure 16) We thus repeated the analysis excluding these outliers. The non-normal distribution of these transcripts was confirmed through the use of QQ plots (data not shown), reinforcing the decision to remove them to ensure the reliability of the conducted analyses.

Figure 16: *Scatter plots illustrating the correlation analyses conducted between IELs and PBMCs.*

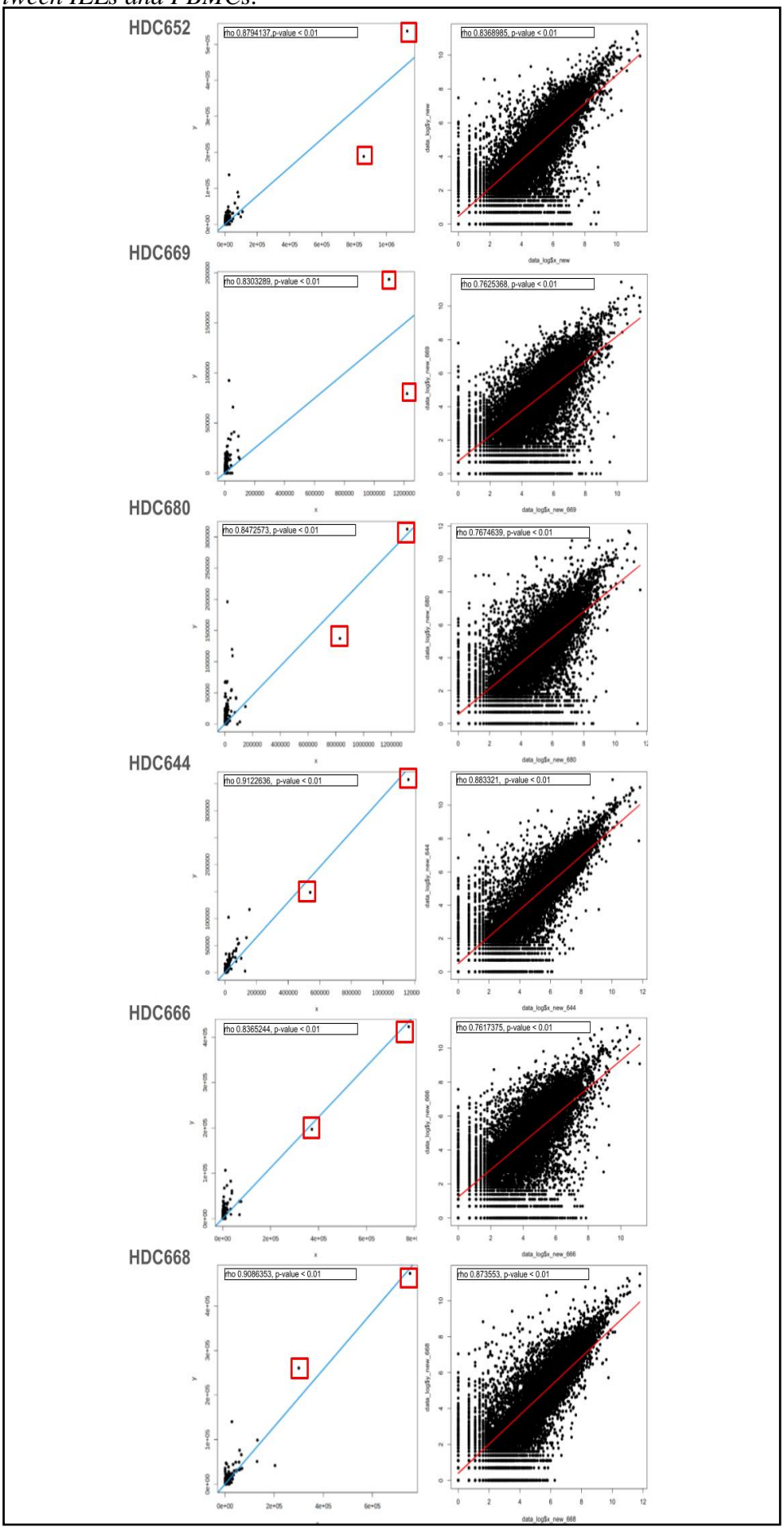


Figure 16 presents a comprehensive panel of scatter plots illustrating the correlation analyses conducted between Isolated Enteric Lymphocytes (IELs) and Peripheral Blood Mononuclear Cells (PBMCs) for each patient. The left section of the figure displays scatter plots with accompanying linear regression lines, highlighting the presence of overexpressed transcripts identified as outliers. On the right side, the same comparisons are depicted with log2-transformed values, aiming to achieve a more dispersed distribution for a comprehensive understanding of the correlation patterns. This detailed exploration provides insights into the interplay between IELs and PBMCs at the transcriptomic level, emphasizing specific transcripts that exhibit notable expression patterns

After the removal of the two outliers the Spearman correlations were anyway high and significant, with an average correlation of rho=0.814, with rho ranging from 0.762 to 0.883 and p-values were all still significant at $p<0.01$. Thus, despite variations observed in pathway enrichment and the number of differentially expressed genes, the individual gene expression levels and abundance metrics remain comparable across the two experimental conditions of targeted RNA-seq and qPCR. This insight underscored the reliability of our data, providing confidence in the robustness of our findings.

Given the previous results, I confirmed the possibility for the use of PBMCs as an alternative tissue to IELs for biomarker research. Nevertheless, I will need a larger cohort of HAEC and HSCR cases and a large and better selected cohort of controls to better evaluate PBMCs as a good proxy for genetic expression in HSCR.

4.8 UK Biobank and Gene candidate association analysis

During my experience at Columbia University in New York, I gained proficiency in phenotype selection by systematically exploring the vast dataset comprising over half a million participants. This involved defining specific phenotypes of interest and managing data fields to identify variables associated with the target phenotype, including medical diagnoses, treatment information, or pertinent outcomes.

Additionally, I learned to employ filtering mechanisms to refine the search, applying criteria such as demographics or specific diagnostic test results.

In this way, I could apply what I learnt in my project at Columbia University to the cohort definition of the IBDs phenotypes on the RAP system. This process resulted in two

complementary cohorts with white origin background, 3,648 samples for Ulcerative Colitis and 2,023 for Chron, with 444 samples exhibiting occurrences of both conditions.

Figure 17: *Venn diagram illustrates distinct phenotype definitions.*

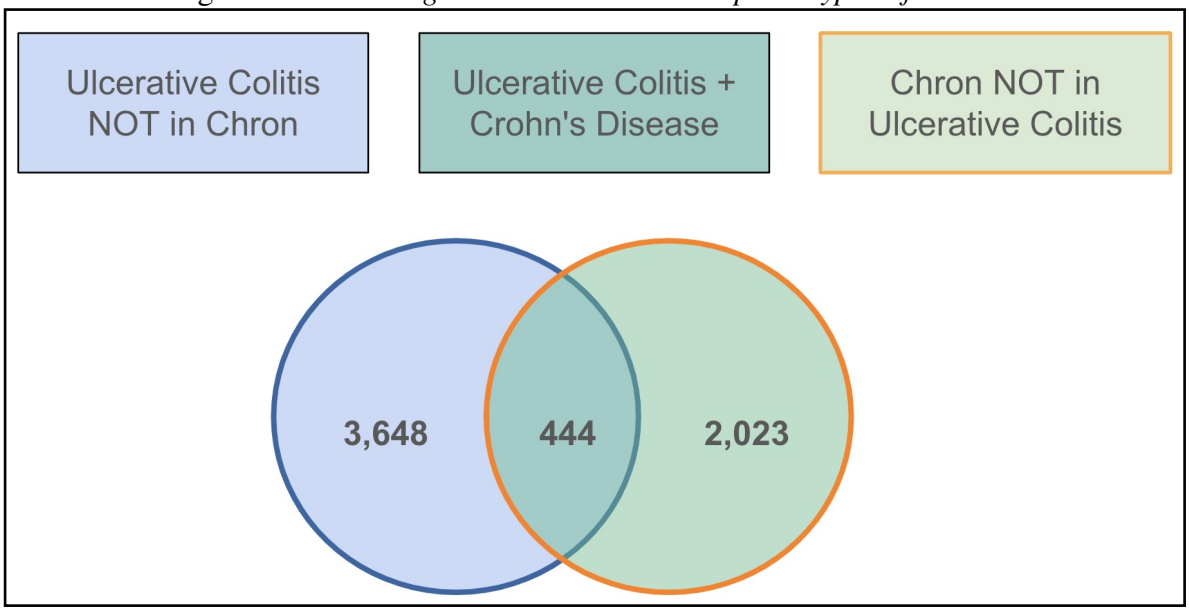


Figure 17: The Venn diagram illustrates three distinct phenotype definitions: Crohn's disease (CD) phenotype with 3,648 patients shown in blue, ulcerative colitis (UC) phenotype with 2,023 patients represented in green, and the shared phenotype comprising 444 patients common to both CD and UC groups, depicted in greenish-blue.

In the Chron Cohort, there are 1,102 female samples (54.47%) and 921 males (45.53%), spanning birth years from 1937 to 1969. A total of 275 samples have recorded dates of death.

Figure 18: *Descriptive characteristics of the Crohn's disease (CD) patient cohort.*

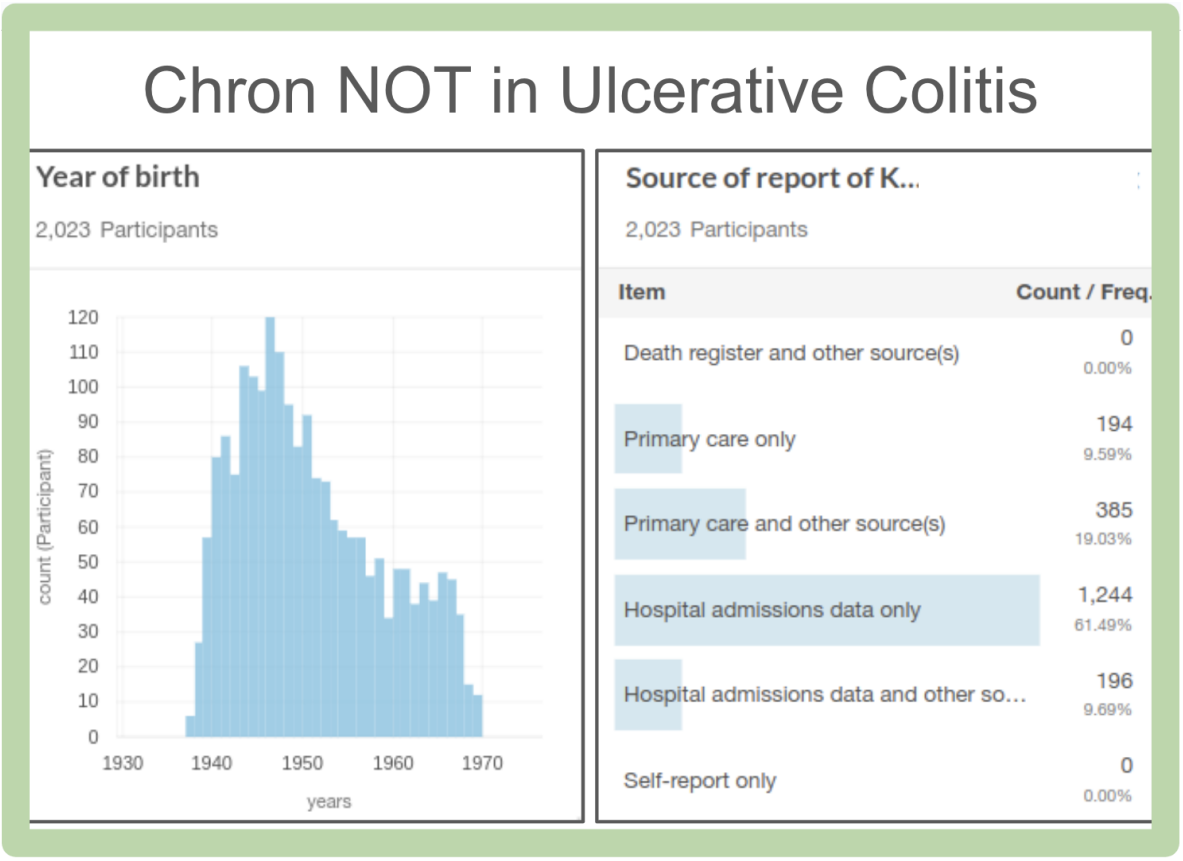


Figure 18: The panel displays descriptive characteristics of the Crohn's disease (CD) patient cohort. The left bar chart represents the distribution of birth years, while the right horizontal bar chart illustrates the source of disease reports.

The predominant source of Chron reports is "Hospital Admission" files (61.85%), followed by "Primary Care and other sources" records (19.03%), "Hospital admission and other sources" (9.69%), "Primary Care only" (9.59%), and "Death register" (0.20%).

The Ulcerative Colitis Cohort comprises 2,088 female samples (57.24%) and 1,560 males (42.76%), born between 1937 and 1969.

For 402 samples, there are recorded dates of death, and all samples belong to the white ethnicity. Predominant sources of Ulcerative Colitis ICD10 code reports include "Hospital Admission" files (59.95%), "Primary Care and other sources" records (20.94%), "Primary Care only" (10.83%), and "Hospital admission and other sources" (8.28%).

Figure 19: *Overview of the Ulcerative Colitis (UC) patient cohort.*

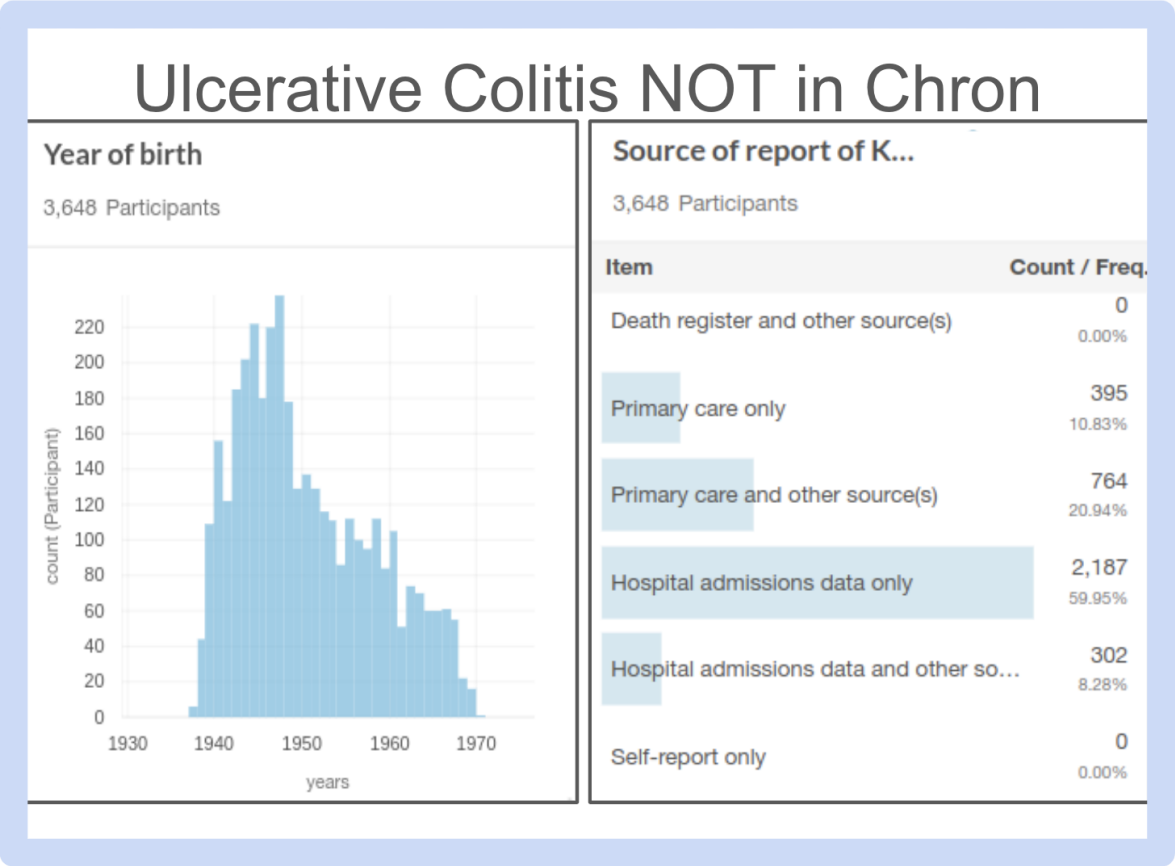


Figure 19: The panel provides an overview of the Ulcerative Colitis (UC) patient cohort, showcasing the distribution of birth years in the left bar chart and detailing the source of disease reports in the right horizontal bar chart.

In the mixed group, 220 samples are females (49.55%) and 224 are males, born between 1937 and 1969. For 65 samples, there are recorded dates of death, all from a white background. Primary sources for K50 and K51 codes are “Hospital Admission” data, followed by “Primary care” sources, and almost equally from other hospital admission and primary care sources.

Figure 20: *Combined cohort of patients with Ulcerative Colitis and Crohn's Disease*

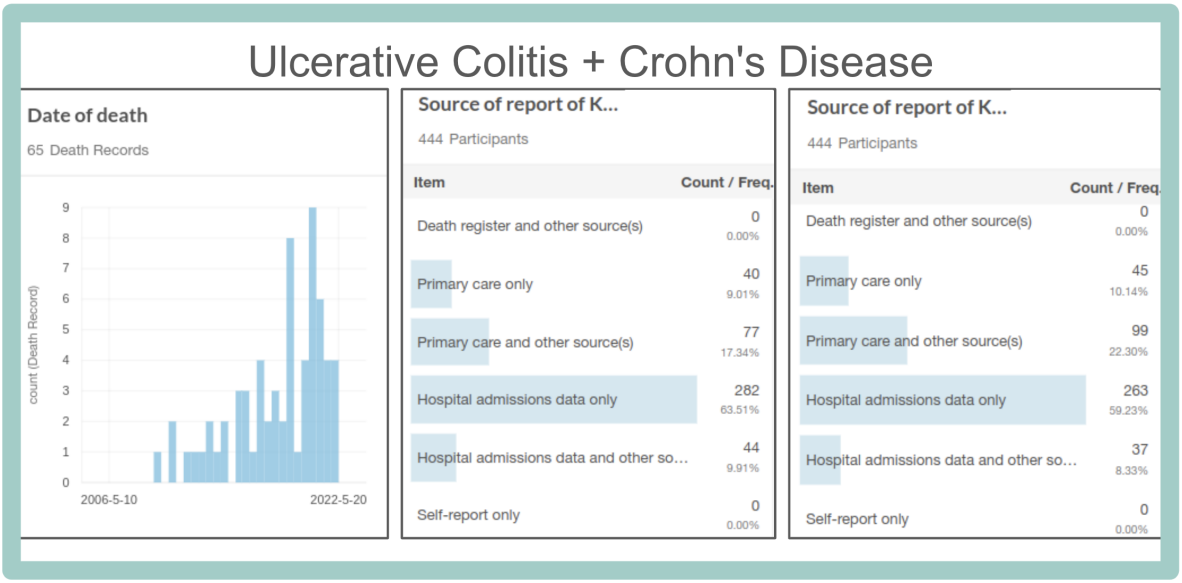


Figure 20: The panel illustrates a combined cohort of patients with Ulcerative Colitis (UC) and Crohn's Disease (CD). On the left, the bar chart depicts the age distribution. In the center, the sources of reports specifying the Crohn's trait are highlighted, while on the right, the sources of reports indicating the Ulcerative Colitis trait are presented.

I have performed a candidate gene association of the variants already implicated in HSCR and the variants identified in the research reported in this thesis as implicated in HAEC with the Inflammatory Bowel Diseases (IBDs) as shown in Table 5. I selected the cohort of IBDs patients through the Research Analysis Platform provided by the UK Biobank and focusing on Crohn Disease and Ulcerative Colitis.

Table 7: *Variants that ranked at the top after Next Generation Sequencing (NGS) filtering and prioritization.*

Trait	Chr	Pos	Gene	AminoAcidChange	rs	Effect Allele	MAF	pvalue	pvalue_adj	OR	OR(95% CI)
CH	2	209190632	PIKFYVE	S1033A	rs999890	G	0,1478	0.3945	0.7326	0.9619491	[0.8817939 - 1.04939]
CH	5	38884071	OSMR	H187Q	rs34675408	G	0,0707	0.8972	0.9457	0.9897863	[0.873655 - 1.121354]
CH	10	43100520	RET	A45A	rs1800858	G	0,7557	0.4899	0.7490	0.9743483	[0.9067807 - 1.046951
CH	10	43114671	RET	G691S	rs1799939	A	0,166	0.1743	0.4855	1,060108	[0.9757134 - 1.151803]
CH	10	43118395	RET	L797L	rs1800861	T	0,7859	0.01398	0.1363	0.9093111	[0.8433788 - 0.9803976]
CH	10	43119646	RET	S836S	rs1800862	T	0,0429	0.06512	0.2783	1.157.055	[0.99374 - 1.34721]
CH	10	43120185	RET	S904S	rs1800863	G	0,1673	0.1478	0.4491	1.063.907	[0.9794505 - 1.155646]
CH	10	43126769	RET	-	rs2075912	C	0,8355	0.2134	0.5548	0.9476793	[0.8719739 - 1.029958]
CH	11	7059981	NLRP14	R55G	rs61063081	A	0,2271	0.06458	0.2783	0.9322086	[0.8659769 - 1.003506]
CH	11	7091569	NLRP14	L1010F	rs17280682	T	0,2266	0.07137	0.2783	0.9337384	[0.8673488 - 1.00521]
CH	13	77918405	EDNRB	G57S	rs1801710	T	0,0074	0.8846	0.9457	0.9570681	[0.6677659 - 1.371707]
CH	19	17940842	JAK3	A1094A	rs3212780	A	0.2733	2.2e-16	4.2e-15	0.106301	[0.09900436 - 0.1141354]
CH	21	48069682	PRMT2	R229W	rs76937225	T	0,0411	0.8446	0.9457	0.9814227	[0.8399467 - 1.146728]
UC	2	209190632	PIKFYVE	S1033A	rs999890	A	0,1544	0.005244	0.0681	0.9127426	[0.8563793 - 0.9728156]
UC	5	38884071	OSMR	H187Q	rs34675408	G	0,0674	0.5439	0.7856	1,029954	[0.939588 - 1.129012]
UC	10	43114671	RET	G691S	rs1799939	A	0,1773	0.4994	0.7491	0.9789679	[0.9216845 - 1.039812]
UC	10	43100520	RET	A45A	rs1800858	G	0,7562	0.312	0.6404	0.9723555	[0.9215804 - 1.025928]
UC	10	43118395	RET	L797L	rs1800861	T	0,7718	0.6655	0.8651	0.9875819	[0.934849 - 1.043289]
UC	10	43119646	RET	S836S	rs1800862	T	0,0493	1	1,0	1,001243	[0.9002415 - 1.113576]
UC	10	43120185	RET	S904S	rs1800863	G	0,1795	0.4466	0.7491	0.9765061	[0.9196324 - 1.036897]
UC	10	43126769	RET	-	rs2075912	C	0,8304	0.5722	0.7969	0.9819921	[0.9235365 - 1.044148]
UC	11	7091569	NLRP14	L1010F	rs17280682	T	0,225	0.03528	0.1966	0.9421122	[0.8915587 - 0.9955323]
UC	11	7059981	NLRP14	R55G	rs61063081	T	0,2253	0.03402	0.1965	0.9417461	[0.8912334 - 0.9951217]
UC	13	77918405	EDNRB	G57S	rs1801710	T	0,0071	1	1,0	0.9961199	[0.7575512 - 1.309819]
UC	19	17940842	JAK3	A1094A	rs3212780	A	0.2654	2.2e-16	4.2e-15	9,7866260	[9.279192 - 10.32181]
UC	21	48069682	PRMT2	R229W	rs76937225	G	0,0396	0.7868	0.9298	1,018255	[0.9048554 - 1.145865]
CH+UC	2	209190632	PIKFYVE	S1033A	rs999890	G	0,1525	0.471	0.7491	0.9306529	[0.774806 - 1.117847]
CH+UC	5	38884071	OSMR	H187Q	rs34675408	G	0,0677	0.8908	0.9456	1,027581	[0.7906122 - 1.335577]
CH+UC	10	43114671	RET	G691S	rs1799939	A	0,1549	0.1497	0.4491	1.147.379	[0.9568185 - 1.375891]
CH+UC	10	43100520	RET	A45A	rs1800858	G	0,7566	0.7203	0.8778	0.9694213	[0.8315816 - 1.130109]
CH+UC	10	43118395	RET	L797L	rs1800861	T	0,7832	0.3764	0.7326	0.9275299	[0.7907507 - 1.087968]
CH+UC	10	43119646	RET	S836S	rs1800862	T	0,046	0.7138	0.8778	1,073631	[0.7846594 - 1.469023]
CH+UC	10	43120185	RET	S904S	rs1800863	G	0,1561	0.1369	0.4491	1,151871	[0.9610777 - 1.380541]
CH+UC	10	43126769	RET	-	rs2075912	C	0,8341	0.6358	0.8550	0.954284	[0.7994448 - 1.139113]
CH+UC	11	7091569	NLRP14	L1010F	rs17280682	T	0,23	0.3004	0.6404	0.9177056	[0.7848137 - 1.0731]
CH+UC	11	7059981	NLRP14	R55G	rs61063081	A	0,23	0.3074	0.6404	0.918795	[0.7857453 - 1.074374]
CH+UC	13	77918405	EDNRB	G57S	rs1801710	T	0,0036	0.2621	0.6388	2,110443	[0.6791832 - 6.557831]
CH+UC	19	17940842	JAK3	A1094A	rs3212780	A	0.2917	0.02616	0.1965	0.8431023	[0.7271303 - 0.977571]
CH+UC	21	48069682	PRMT2	R229W	rs76937225	T	0,046	0.421	0.7463	0.8673747	[0.6339003 - 1.186841]

Table presenting the results of the chi-square test for selected SNPs(as described above) associated with HSCR (Hirschsprung's disease) and HAEC (Hirschsprung-associated enterocolitis) tested in the IBD cohorts. The cohorts include CH (Chron), UC (Ulcerative Colitis), and the combined UC+UC cohort. *Abbreviations: Pos: Position; MAF : Frequency in the selected cohort; pvalue_adj= False Discovery Rate adjusted pvalue; OR : Odd Ratio; OR(95% CI): 95% confidence interval (CI) for Odd Ratio Estimation.*

In our association analysis, we initially directed our focus on the RET gene, a major player in Hirschsprung disease (HSCR), responsible for approximately 70% of the phenotypic variability associated with the condition. Additionally, I explored the OSMR gene, currently identified as the sole susceptibility variant identified as associated with HAEC and then I compiled a list of SNPs for testing, as outlined earlier. Unfortunately, I could only assess a portion of them due to their absence in the UK Biobank, presumably ungenotyped. Notably, the investigation did not reveal any significant associations, between the selected HSCR and HAEC associated variants and ulcerative colitis (UC), Crohn's disease (Chron), or the combination of both (UC+Chron), except for JAK3, which was one of the first ranked genes in WES analysis on HSCR and HAEC patients. To note, also the variant rs999890 is borderline with significance. JAK3 and PIKFYVE have already been recognized for their significant role in Inflammatory Bowel Diseases (IBDs), the identification of these variants, found to be associated with enterocolitis in Hirschsprung's disease (HAEC), offers potential insights into shared genetics between HAEC and IBDs. Despite the lack of confirmation in the WES replication study (rs3212780 p=0.6, rs999890 p=0.0850), the presence of these variants suggest avenues for further exploration into the genetic connections between HAEC and IBD. It is plausible that other variants in these genes and/or other genes not yet explored here, could prove a genetic overlap between HSCR/HAEC and IBDs. Further exploration into these potential genetic factors is essential for a comprehensive understanding of the complex genetic landscape associated with Hirschsprung and its complications.

Chapter 5

Discussion

My thesis has been an exciting journey into the world of omics and big data, where the future of scientific research is unfolding with extraordinary potential. The era of omics and big data represents a revolution in genetics and biology, requiring significant statistical, bioinformatic, and data management efforts. These impressive tasks are not just a challenge but a true scientific adventure.

Throughout my research, I undertook the task of managing and approaching three different omics: genome, transcriptome, and proteome. This triad of omics provides a comprehensive overview of an individual's genetic information, offering a glimpse into the molecular world. The genome, with its vast assortment of genetic variants, served as the start point. Subsequently, I delved into the Proteome, adding a layer of complexity, unveiling how the composition of proteins is crucial to biological processes. Finally, the study of the transcriptome provided me an open window into genetic activities through gene transcript analysis [79].

Handling big data was as challenging as it was thrilling. The vastness of databases, such as the UK Biobank, required not only statistical skills to extract meaningful insights but also a solid data preparation and quality control foundation. Navigating through the seas of data demanded a deep commitment to understanding the technical details and the bioinformatic challenges associated.

This journey highlighted the importance of a solid preparation in using programming languages, essential for processing and interpreting data effectively. Bioinformatic languages have become indispensable tools for probing the hidden secrets within omics data [80].

In summary, my thesis has represented a fascinating dive into the ocean of omics and big data, combining in-depth knowledge of genetics and molecular biology with advanced skills in statistics and bioinformatics. This multidisciplinary approach has proven fundamental to successfully apply the acquired knowledge to the study of Hirschsprung-associated enterocolitis (HAEC), laying the groundwork for future scientific explorations in the omics and big data era.

In this study, I employed diverse omics approaches on patients with Hirschsprung Associated Enterocolitis (HAEC) and Hirschsprung's disease (HSCR) without enterocolitis. This comprehensive analysis initially led to the identification of Oncostatin M Receptor (OSMR) as a candidate gene for HAEC, through Whole Exome Sequencing (WES), a discovery subsequently explored through proteomics and reported in our paper by Bachetti, Rosamilia et al. [65]. The OSMR gene, expressed across various tissues, plays a pivotal role in diverse biological functions, including cell growth, neuronal development, and inflammatory responses.

Oncostatin M (OSM), belonging to the interleukin-6 (IL-6) family of cytokines, interacts with two receptors—Leukemia inhibitory factor receptor (LIFR) and Oncostatin M receptor (OSMR). This interaction, facilitated by the common glycoprotein 130 (gp130) subunit, triggers downstream signaling events. OSMR binding initiates Janus kinases JAK1 and JAK2, leading to the activation of signal transducers and transcription factors, known as STATs. The unique Src homology 2 (SH2) domain of SHC Adaptor Protein 1 (Shc1), specific to OSMR, undergoes phosphorylation, activating mitogen-activated protein kinase (MAPK) cascades, including extracellular signal-related kinase ERK, p38, and c-Jun N-terminal kinase JNK. Additionally, the activation of phosphatidylinositol-3-kinase/AKT (PI3K/AKT) and protein kinase C delta (PKC) pathways adds further complexity to OSM signaling [81],[82]. OSM plays a role in several physiological and pathological conditions, among which cancer, mesenchymal stem cell differentiation, inflammation, and metabolism. [82], [83], [84], [85]. The diverse roles of Oncostatin M (OSM) have been explored in various physiological contexts. Bailey et al. found that OSM induces pro-inflammatory responses in adipocytes, contributing to insulin resistance and adipose tissue inflammation through the OSM Receptor (OSMR) and the p66Shc-MEK/ERK pathway [81]. Lorchner et al. engineered a human-like OSM protein to replicate the effects of mouse OSM (mOSM) and mLIF in cardiomyocytes [86]. The human-like OSM demonstrated superior efficacy in STAT3 activation and cardiomyocyte survival under hypoxic conditions, highlighting its potential clinical applications. Jakob et al. investigated mOSM signaling in bone formation and resorption, revealing dual effects on fibroblast and bone marrow stromal cells [87]. OSMR predominantly mediated proliferative effects, whole LIFR had synergistic influences. Transcriptome analysis showed differential

regulation of STAT family members and enrichment in gene sets associated with inflammatory phenotypes. Nummenmaa et al. explored IL-6 upregulation in osteoarthritis (OA) and its connection to Transient Receptor Potential Ankyrin 1 (TRPA1) [88]. Their study, using RNAseq analysis and comparing chondrocytes from wild-type and TRPA1 knockout mice, identified TRPA1 as a potential drug target in osteoarthritis, highlighting its therapeutic implications. In an animal model, genetic deletion or pharmacological blockade of OSM significantly attenuated colitis, indicating its crucial role in anti-TNF-resistant intestinal inflammation [89].

In our inaugural publication within this field [65], we confirm the inflammatory role of Oncostatin M Receptor (OSMR), as previously documented in the literature. Notably, we identify OSMR, a gene already associated with Inflammatory Bowel Diseases (IBDs), as a gene predisposing to Hirschsprung Associated Enterocolitis (HAEC) susceptibility through variant SNP rs34675408 located in the OSM binding site. Both case-control and family-based association analyses support OSMR SNP rs34675408 as an HAEC susceptibility variant. The proteomic analysis conducted on lymphoblasts from a wild-type (wt) patient and a patient carrying this predisposing OSMR variant in homozygous state revealed distinct protein expression patterns in response to OSM treatment between the two samples.

Through pathway analysis carried out with the tissue-specific network-based functional analysis implemented in HumanBase, interesting insights were discerned. Among the proteins overrepresented in the OSMR TT wt sample there was a substantial enrichment of immune response pathways, including also ERK1/ERK2 cascades, autophagy, and response to wounding, apoptosis, catabolic processes, DNA repair and replication. Caveolin-1 (CAV1) and ERK1/2, implicated in inflammation and tissue repair, showed altered expression patterns in TT cells. The differential localization of pERK in response to OSM treatment implies consequences for integrin pathways and cell motility, impacting inflammation and infections [65]. The dysregulated endocytosis pathway in HAEC OSMR variant cells, suggested by CAV1 down-regulation, implies consequences for recycling and immune response. Conversely, for the OSMR variant carrier, no immune response pathway was enriched. Diverse pathways lead by protein overexpressed after OSM stimulation were enriched, encompassing lipid metabolism, interleukin-1 production and regulation, tumor necrosis factor-mediated signaling pathway (TNF), Jun N-terminal Kinase (JNK) cascade, positive regulation of stress-activated Mitogen-Activated Protein Kinase (MAPK) cascade, response to oxidative stress, epithelium differentiation and development, actin and cell junction organization, and microtubule cytoskeleton organization. The negative regulation of NF-kB and the positive regulation of MAPK, JNK, and IL-1 in the OSMR SNP variant homozygous patient suggests an

inflammation cascade, potentially due to the overexpression of ASC proteins, PYCARD and CARD8. Among the "lost-effect" proteins, autophagy and fatty acid process pathways enriched in TT cells were down-regulated in HAEC GG cells. Sulfur biosynthesis was more represented in GG lymphocytes, with implications for colon inflammation. The altered SCFA profile reported in HAEC [28] and the impaired functions of goblet cells in HSCR [29] might precede immune involvement in the HAEC mice model [31].

This comprehensive proteomic analysis and pathways analysis substantiate the pivotal role of the OSM–OSMR axis in HAEC susceptibility initially identified through Whole Exome Sequencing (WES). These findings suggest a potential mechanism involving an impaired immune system response and an imbalance in gut homeostasis, previously implicated in the acute inflammation observed in HAEC.

These outcomes collectively underscore the significance of Oncostatin M (OSM) in orchestrating signaling pathways crucial for maintaining tissue health, emphasizing its dual impact across various physiological processes. Significantly, the confirmation of OSM/OSMR involvement in preserving gut homeostasis underscores its relevance in preventing enterocolitis episodes in Hirschsprung's disease (HSCR). However, in recent investigations, the intricate relationship between Oncostatin M (OSM) and its receptor OSMR type II has been brought to light, challenging our conventional understanding of their impact on inflammation. A recent study [90] suggests that OSM signaling via the OSMR type II receptor exerts a dual modulatory effect on both pro- and anti-inflammatory responses, in accordance with the dual effect exerted by OSM/OSMR reported by others [81], [86], [87], [88] as described more above. This discovery unveils a nuanced and context-specific influence on immune function, wherein the cumulative effects of OSM/OSMR signaling may vary based on the specific conditions and cellular context.

Growing evidence shows a role for the OSM-OSMR pathway in gut inflammation [89], [91]. OSM and OSMR expression are reported to be higher in inflammatory bowel diseases (IBDs) patients than in controls and in the inflamed than non-inflamed lesions of IBDs patients [85].

IBDs share common clinical manifestations and abnormal intestinal mucosal barrier function with HAEC, suggesting shared pathogenic mechanisms [85]. Moreover, the H187 residue, positioned within the OSM-OSMR binding site, is crucial as it connects three hotspot sites, yet found as potential targets for IBDs [46]. The substitution of basic histidine with polar glutamine, though not directly impacting a binding hot spot, may alter the conformation of active sites in OSM-OSMR interaction. The HAEC-associated G allele appeared to rectify

the distorted angle in the protein conformation caused by the T allele, potentially amplifying the OSM effect, as seen in type 2 inflammation for gain-of-function SNPs in OSM and OSMR [84].

WES results and follow-up case-control and family-based association support OSMR SNP rs34675408 as an HAEC susceptibility variant.

The advent of omics technologies has transformed biomedical research, enabling the examination of molecular intricacies underlying complex diseases on an unprecedented scale. The analysis of Omics data presents vast opportunities for applications in biomarker discovery, patient stratification, disease classification, and drug development, propelling advancements in precision medicine strategies.

I extended our investigation by transcriptome analysis. Notably, these integrated approaches of transcriptome, together with WES and proteomics, were employed to explore the genetic susceptibility of HAEC for the first time in the studies part of the present dissertation. Our decision to sequence the transcriptome is underscored by the admission that genomics alone may not always yield a diagnosis or a comprehensive understanding of the genetic basis of a trait. Many studies utilizing whole-blood RNA-seq, combined with comprehensive variant analyses and gene filters relevant to the phenotype, have successfully identified causal genes and variants in a notable percentage of individuals with the disease [92]. The authors suggest employing stringent filters, particularly focusing on splicing and expression outliers. The parallels with exome sequencing underscore the importance of leveraging control RNA-seq to pinpoint aberrant expression events in candidate rare-disease genes. The choice to proceed with targeted RNA-seq of IELs and PBMCs in the cohort of HSCR, HAEC, and controls aligned with this paradigm, offering insights into the intricate molecular landscape of these conditions.

I used the IELs in the first place on 16 samples, performed the Differential Expression Analysis and then I moved forward to Gene Set Enrichment Analysis (GSEA), widely applied to determine whether a predefined gene set showed statistically significant difference between two biological states, enhancing the potential for deeper insights into biological pathways. The process of identifying and visualizing gene expression differences was crucial to interpret and draw meaningful conclusions from transcriptomic data.

The identification of differentially expressed transcripts, particularly in the HAECvsCTRL group, gained paramount significance as these transcripts have already been implicated in Inflammatory Bowel Diseases (IBDs), sharing certain pathological characteristics

with HAEC. The observed dysregulation in immune pathways, particularly prominent in the HAEC group, underscored the dysfunctional immune system in these patients, potentially justifying the inflammatory grading that can escalate to sepsis in certain individuals. The availability of Intestinal Epithelial Lymphocytes (IELs) derived from biopsies provided a crucial snapshot of the local immune system, yet the use of Peripheral Blood Mononuclear Cells (PBMCs) as a replication set was equally important. As emphasized in previous studies, RNA-seq on peripheral blood specimens, readily available in clinical practice, presents a practical advantage. PBMCs, besides being easier to collect than biopsies, demonstrated potential in detecting proteins expressed at lower levels, particularly in tissues where reaching the correct biopsy site is challenging [93]. The preference for PBMCs is further supported by the established higher tissue specificity of network edges (transcription factor to target gene connections) compared to genes, suggesting that tissue specificity is driven by context-dependent regulatory paths [94]. The high correlation between IELs and PBMCs in RNA-seq abundance enhanced confidence in utilizing PBMCs for patient stratification based on molecular insights. The use of PBMCs might contribute significantly to understanding patient heterogeneity, paving the way for more refined and personalized medical stratification approaches in the future.

The validation step involved the use of the same RNA samples, emphasizing its superiority over in silico analyses, online databases, or simulated datasets. The validation of dysregulated transcripts within the transcriptomic profile of Intestinal Epithelial Lymphocytes (IELs) has been of paramount importance in the present study, shedding light on key genes such as GSTA1. This gene, responsible for encoding enzymes facilitating the addition of glutathione to various electrophilic compounds, has garnered attention due to its role in metabolizing drugs. Notably, polymorphisms in GSTA1 have been associated with altered drug metabolism in Crohn's disease [95]. Specifically, the authors found that GSTA1 variants, identified in 12.8% of patients, displayed a trend for an association with decreased clinical efficacy, particularly in response to azathioprine treatment.

The follow up also included ICAM1, a gene encoding a cell surface glycoprotein prominently expressed on endothelial cells and immune system cells. Investigations have revealed that inflamed intestinal tissues from patients with Inflammatory Bowel Diseases (IBD) express elevated levels of ICAM1. Notably, the ICAM1 production, together with the generation of other proinflammatory molecules, including interleukin (IL)-6 and chemokines attracting immune cells, is induced by OSMR response to OSM [96]. Moreover, the authors reported intestinal stromal cells were found to be enriched in IL-6 expression relative to their OSMR. Lastly, our focus extended to IL22, a member of the IL10 family of cytokines, con-

tributing to cellular inflammatory responses. This cytokine's dual role in antimicrobial defense and pro-inflammatory processes aligns with its involvement in the pathogenesis of various intestinal diseases [97]. Collectively, our validation experiments confirmed dysregulation of these transcripts in Hirschsprung Associated Enterocolitis (HAEC) and Hirschsprung disease (HSCR), providing also evidence, given the involvement of these same genes in both Inflammatory Bowel Diseases (IBDs) and HAEC and HSCR, of a possible shared pathogenesis between these conditions.

IELs were an optimal tissue to investigate gene expression, however they needed to be yielded by intestinal biopsy. The potential interchangeability of Peripheral Blood Mononuclear Cells (PBMCs) as a source for biomarker discovery would boost the search for DEGs. While my attempt to replicate the results in IELs faced challenges due to the very small sample size and the absence of a control cohort available, the observed high correlation between IELs and PBMCs encouraged further exploration. My study has some limitations. First of all, the sample size was very small. The WES analysis relied on 24 samples only, thus the possibility exists that tho OSMR variant detected is a confounding factor rather than predisposing to HAEC. Nevertheless, proteomic analysis comparing the consequences of OSM treatment on the two alleles conducted on two HAEC patients, one homozygous for the OSMR G allele and the other a T homozygous HAEC patient, revealed distinct protein clusters after OSM stimulation. The functional enriched pathways detected, related to immune response for the proteins overexpressed in the wt TT sample and to intestinal inflammation in the GG variant samples, were in line with the HAEC clinics. Also, the relatively low frequency of the SNP, the low incidence of HSCR, and challenges in collecting proper HAEC samples complicate its genetic investigation. The strength of our WES analysis, nonetheless, lies in the careful selection of the sample. I made a concerted effort to choose individuals with Italian ancestry, free from additional complications, following a comprehensive phenotypic screening. While this approach ensured purity in our sample, it unfortunately led to the selection of a limited number of patients who met our specific criteria. The prevalence of the disease in the general population, particularly for the specific condition under study, HAEC, further constrained the availability of a substantial number of patients, occurring in approximately one-third of children with HSCR. In addition, I was conscious of the Winner's Curse, a phenomenon according to which newly discovered genetic effects tend to be consistently overestimated and thus fail to replicate in subsequent studies. This phenomenon casts doubts on the credibility of the Association Studies, particularly discussed in Genome Wide Association Studies (GWAS) [98]. The GWAS, which are constituted by a vast array of genetic variants that have to undergo rigorous statistical tests, are likened to athletes in a time trial. The statistical score of the winning variants tends to overestimate their actual effect on the examined trait due to

a degree of luck, akin to an athlete's exceptional performance on a particular day. To address this, replication studies, especially in GWAS but also in all the genetic association studies, have been proposed to 'recalibrate' the associated SNPs [99]. With the comprehensive validations conducted in this thesis work, I believe that the potential impact of the Winner's Curse on the performed tests has been effectively mitigated. Another pitfall is the retrospective nature of the study, which inherently lacks tailored controls on participant selection, necessitating cautious interpretation. Although the selection of Intestinal Epithelial Lymphocytes (IELs) was prospective, it was not initially intended for this study, and thus it lacked the same initial filtering criteria as performed for WES's patients selection, leading to a subsequent cleaning which, together with the low incidence of HAEC cases, significantly limited tissue availability.

A direct correlation between proteomic and transcriptomic data was not feasible due to the distinct designs of the two assays. The proteomic assay was cell-based and stimulus-specific, designed with two immortalized patient cell lines with and without the homozygous OSMR variant, stimulated with OSM ligand. On the other hand, the transcriptomic assay aimed to investigate the transcriptomic profile not only of a specific patient type, HAEC or HSCR, but also to compare them, involving 6 HAEC patients, 6 HSCR patients, and 4 pediatric controls for intraepithelial lymphocyte transcriptome analysis, and 3 HAEC patients, 3 HSCR patients, and 2 controls for PBMC transcriptome analysis.

Therefore, directly comparing the two assays could potentially introduce confounding factors due to the differing experimental setups. However, analysis revealed that 20% of the dysregulated pathways identified in the transcriptomic assay overlapped with those identified in the proteomic assay, particularly involving key modules related to T cell activation, immune response, and immune effector processes. This underscores a dysregulation of the immune system, suggesting it as a potential target for therapeutic modulation.

In conclusion, although the small sample size, the robustness of my dissertation is however underpinned also by the meticulous replication and validation process that accompanied every step of our research: post-WES analysis (with Sanger sequencing), post-proteomic analysis (using Western blot), and transcriptomic analysis (via rtPCR). The research conducted for my thesis has surely enlightened the role of genes and pathways involved in several biological processes, among which immunity response and inflammation are the most prominent, in HAEC, and the possible link with other inflammatory disorders of the intestine, such as the IBDs.

The need for increasingly larger samples in recent years has given impetus to the cre-

ation of large databases, including the UK Biobank (UKBB). Navigating the UKBB can be perceived as a colossal and intricate task, posing challenges in terms of its size and complexity. However, through my experience in New York, I acquired the skills to comprehend, manage, and interrogate this extensive resource effectively. Moreover, I am equipped with the knowledge of which tools to employ to overcome these challenges, ensuring a fruitful analytical process once the desired dataset is obtained. I focused on a few genes, including: RET, the major gene in Hirschsprung disease, OSMR and JAK3, the genes that we have primarily discovered in HAEC susceptibility through the WES. Even though my association analysis did not yield significant findings for RET and OSMR, JAK3 was anyway found to be associated to Chron and Ulcerative Colitis in the UkB. Moreover, the absence of signals associated with Inflammatory Bowel Diseases (IBDs) in the OSMR and RET genes does not preclude the possibility of such signals existing in other variants in these genes or in other loci that we have not investigated thoroughly yet. Based on these findings, the future project aims at Whole Genome Sequencing (WGS), total RNA sequencing, and analysis of gene co-expression networks on a larger panel of PBMCs from HSCR and HAEC patients to further investigate the genetic bases of HAEC. WGS will allow to investigate also non-coding regions, and total RNA-seq, differentially from RNA-seq targeted, will offer insights into both coding RNAs and non-coding RNAs (e.g., lncRNAs and miRNAs), providing valuable information about regulatory regions, overall transcript expression levels, splicing patterns, and the identification of exons, introns, and their boundaries. These future endeavors aim to enhance the understanding of the intricate genetic landscape of HAEC, paving the way for more refined and personalized medical strategies in the management of these conditions. In addition, I mean to explore the genetic overlap between HAEC and IBDs by a more comprehensive analysis of UKBB and by performing a more accurate control for the sample by checking for population structure that might have distorted our association results.

The need for increasingly larger samples in recent years has given impetus to the creation of large databases, including the UK Biobank (UKBB). Navigating the UKBB can be perceived as a colossal and intricate task, posing challenges in terms of its size and complexity. However, through my experience in New York, I acquired the skills to comprehend, manage, and interrogate this extensive resource effectively. Moreover, I am equipped with the knowledge of which tools to employ to overcome these challenges, ensuring a fruitful analytical process once the desired dataset is obtained. I focused on a few genes, including RET, the major gene in Hirschsprung disease, OSMR and JAK3 the genes that we have primarily discovered in HAEC susceptibility through the WES. Even though my association analysis did not yield significant findings, JAK3 and PIKFYVE were anyway found to be associated to IBDs cohorts in the UKB. Furtherly, the absence of signals associated with Inflammatory

Bowel Diseases (IBDs) in the OSMR and RET genes does not preclude the possibility of such signals existing in other variants in these genes or in other loci that we have not investigated thoroughly yet.

I hope that the study carried out for my dissertation and the planned next steps will guide future research in Hirschsprung disease and Hirschsprung associated Enterocolitis, offering a roadmap for potential therapeutic interventions.

Bibliography

[1] A. Hamosh et al. “Online Mendelian Inheritance in Man (OMIM)”. eng. In: *Human Mutation* 15.1 (2000), pp. 57–61. ISSN: 1059-7794. DOI: 10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G.

[2] Arno G. Motulsky. “Genetics of complex diseases”. In: *Journal of Zhejiang University. Science. B* 7.2 (Feb. 2006), pp. 167–168. ISSN: 1673-1581. DOI: 10.1631/jzus.2006.B0167. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363767/ (visited on 12/16/2023).

[3] George Davey Smith et al. “Genetic epidemiology and public health: hope, hype, and future prospects”. en. In: 366 (2005).

[4] D. Lvovs et al. “A Polygenic Approach to the Study of Polygenic Diseases”. en. In: *Acta Naturae* 4.3 (Sept. 2012). Number: 3, pp. 59–71. ISSN: 2075-8251. DOI: 10.32607/20758251-2012-4-3-59-71. URL: https://doi.org/10.32607/20758251-2012-4-3-59-71 (visited on 12/16/2023).

[5] H. P. J. Buermans and J. T. den Dunnen. “Next generation sequencing technology: Advances and applications”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. From genome to function 1842.10 (Oct. 2014), pp. 1932–1941. ISSN: 0925-4439. DOI: 10.1016/j.bbadis.2014.06.015. URL: https://www.sciencedirect.com/science/article/pii/S092544391400180X (visited on 12/16/2023).

[6] Bruce Alberts et al. *Molecular Biology of the Cell*. 4th. Garland Science, 2002. ISBN: 978-0-8153-3218-3 978-0-8153-4072-0.

[7] Teri A. Manolio, Lisa D. Brooks, and Francis S. Collins. “A HapMap harvest of insights into the genetics of common disease”. en. In: *The Journal of Clinical Investigation* 118.5 (May 2008). Publisher: American Society for Clinical Investigation, pp. 1590–1605. ISSN: 0021-9738. DOI: 10.1172/JCI34772. URL: https://www.jci.org/articles/26160 (visited on 12/16/2023).

[8] Kai-Oliver Mutz et al. “Transcriptome analysis using next-generation sequencing”. eng. In: *Current Opinion in Biotechnology* 24.1 (Feb. 2013), pp. 22–30. ISSN: 1879-0429. DOI: 10.1016/j.copbio.2012.09.004.

[9] Claudia Manzoni et al. “Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences”. In: *Briefings in Bioinformatics* 19.2 (Mar. 2018), pp. 286–302. ISSN: 1477-4054. DOI: 10.1093/bib/bbw114. URL: https://doi.org/10.1093/bib/bbw114 (visited on 12/16/2023).

[10] Shaun Purcell et al. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses”. In: *American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. ISSN: 0002-9297. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/ (visited on 12/16/2023).

[11] *R: The R Project for Statistical Computing*. URL: https://www.r-project.org/ (visited on 12/16/2023).

[12] Federico Marini, Jan Linke, and Harald Binder. “ideal: an R/Bioconductor package for Interactive Differential Expression Analysis”. en. In: ().

[13] Xiaojing Chu et al. “Multi-Omics Approaches in Immunological Research”. In: *Frontiers in Immunology* 12 (2021). ISSN: 1664-3224. URL: https://www.frontiersin.org/articles/10.3389/fimmu.2021.668045 (visited on 12/26/2023).

[14] Anna Fry et al. “Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population”. eng. In: *American Journal of Epidemiology* 186.9 (Nov. 2017), pp. 1026–1034. ISSN: 1476-6256. DOI: 10.1093/aje/kwx246.

[15] Roddy Walsh, Rafik Tadros, and Connie R. Bezzina. “When genetic burden reaches threshold”. eng. In: *European Heart Journal* 41.39 (Oct. 2020), pp. 3849–3855. ISSN: 1522-9645. DOI: 10.1093/eurheartj/ehaa269.

[16] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. “An expanded view of complex traits: from polygenic to omnigenic”. In: *Cell* 169.7 (June 2017), pp. 1177–1186. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.05.038. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5536862/ (visited on 04/23/2024).

[17] Alessio Pini Prato et al. “A prospective observational study of associated anomalies in Hirschsprung’s disease”. In: *Orphanet Journal of Rare Diseases* 8.1 (Nov. 2013), p. 184. ISSN: 1750-1172. DOI: 10.1186/1750-1172-8-184. URL: https://doi.org/10.1186/1750-1172-8-184 (visited on 12/16/2023).

[18] *The Online Metabolic and Molecular Bases of Inherited Disease — OMMBID — McGraw Hill Medical*. URL: https://ommbid.mhmedical.com/book.aspx?bookID=2709 (visited on 12/16/2023).

[19] Joseph M. Tilghman et al. "Molecular Genetic Anatomy and Risk Profile of Hirschsprung's Disease". In: *New England Journal of Medicine* 380.15 (Apr. 2019). Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1706594, pp. 1421–1432. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1706594. URL: https://doi.org/10.1056/NEJMoa1706594 (visited on 12/16/2023).

[20] J A Badner et al. "A genetic study of Hirschsprung disease". eng. In: *American journal of human genetics* 46.3 (Mar. 1990), pp. 568–580. ISSN: 1537-6605. URL: https://europepmc.org/articles/PMC1683643 (visited on 12/16/2023).

[21] Laura E. Kuil et al. "Size matters: Large copy number losses in Hirschsprung disease patients reveal genes involved in enteric nervous system development". In: *PLoS Genetics* 17.8 (Aug. 2021), e1009698. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1009698. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8372947/ (visited on 12/16/2023).

[22] *Impaired Cellular Immunity in the Murine Neural Crest Conditional Deletion of Endothelin Receptor-B Model of Hirschsprung's Disease — PLOS ONE*. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128822 (visited on 12/16/2023).

[23] Joannie M. Allaire et al. "The Intestinal Epithelium: Central Coordinator of Mucosal Immunity". eng. In: *Trends in Immunology* 39.9 (Sept. 2018), pp. 677–696. ISSN: 1471-4981. DOI: 10.1016/j.it.2018.04.002.

[24] Clara Sze-man Tang et al. "Genetics of Hirschsprung's disease". en. In: *Pediatric Surgery International* 39.1 (Feb. 2023), p. 104. ISSN: 1437-9813. DOI: 10.1007/s00383-022-05358-x. URL: https://doi.org/10.1007/s00383-022-05358-x (visited on 12/28/2023).

[25] Eileen Sproat Emison et al. "Differential contributions of rare and common, coding and noncoding Ret mutations to multifactorial Hirschsprung disease liability". eng. In: *American Journal of Human Genetics* 87.1 (July 2010), pp. 60–74. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2010.06.007.

[26] Alessio Pini Prato et al. "Hirschsprung's disease: what about mortality?" eng. In: *Pediatric Surgery International* 27.5 (May 2011), pp. 473–478. ISSN: 1437-9813. DOI: 10.1007/s00383-010-2848-2.

[27] R. Surana, F. M. J. Quinn, and P. Puri. “Evaluation of risk factors in the development of enterocolitis complicating Hirschsprung’s disease”. en. In: *Pediatric Surgery International* 9.4 (1994), pp. 234–236. ISSN: 0179-0358, 1437-9813. DOI: 10.1007/BF00832245. URL: http://link.springer.com/10.1007/BF00832245 (visited on 12/26/2023).

[28] Farokh R. Demehri et al. “Altered fecal short chain fatty acid composition in children with a history of Hirschsprung-associated enterocolitis”. eng. In: *Journal of Pediatric Surgery* 51.1 (Jan. 2016), pp. 81–86. ISSN: 1531-5037. DOI: 10.1016/j.jpedsurg.2015.10.012.

[29] Jay R. Thiagarajah et al. “Altered Goblet Cell Differentiation and Surface Mucus Properties in Hirschsprung Disease”. en. In: *PLOS ONE* 9.6 (June 2014). Publisher: Public Library of Science, e99944. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0099944. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0099944 (visited on 12/16/2023).

[30] Ankush Gosain et al. “Guidelines for the Diagnosis and Management of Hirschsprung-Associated Enterocolitis”. In: *Pediatric surgery international* 33.5 (May 2017), pp. 517–521. ISSN: 0179-0358. DOI: 10.1007/s00383-017-4065-8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5395325/ (visited on 12/16/2023).

[31] L. Lauriina Porokuokka et al. “Gfra1 Underexpression Causes Hirschsprung’s Disease and Associated Enterocolitis in Mice”. eng. In: *Cellular and Molecular Gastroenterology and Hepatology* 7.3 (2019), pp. 655–678. ISSN: 2352-345X. DOI: 10.1016/j.jcmgh.2018.12.007.

[32] Alessio Pini Prato et al. “A Metagenomics Study on Hirschsprung’s Disease Associated Enterocolitis: Biodiversity and Gut Microbial Homeostasis Depend on Resection Length and Patient’s Clinical History”. In: *Frontiers in Pediatrics* 7 (2019). ISSN: 2296-2360. URL: https://www.frontiersin.org/articles/10.3389/fped.2019.00326 (visited on 12/16/2023).

[33] Yuqing Li et al. “Characterization of Intestinal Microbiomes of Hirschsprung’s Disease Patients with or without Enterocolitis Using Illumina-MiSeq High-Throughput Sequencing”. en. In: *PLOS ONE* 11.9 (Sept. 2016). Publisher: Public Library of Science, e0162079. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0162079. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0162079 (visited on 12/16/2023).

[34] H. Nakamura, T. Lim, and P. Puri. “Inflammatory bowel disease in patients with Hirschsprung’s disease: a systematic review and meta-analysis”. eng. In: *Pediatric Surgery International* 34.2 (Feb. 2018), pp. 149–154. ISSN: 1437-9813. DOI: 10.1007/s00383-017-4182-4.

[35] A. Pini Prato et al. “Uncommon causes of postoperative chronic diarrhoea mimicking enterocolitis in Hirschsprung’s disease: is there a role for digestive endoscopy?” en. In: *Pediatric Surgery International* 24.3 (Mar. 2008), pp. 389–389. ISSN: 1437-9813. DOI: 10.1007/s00383-007-2093-5. URL: https://doi.org/10.1007/s00383-007-2093-5 (visited on 12/16/2023).

[36] Charles N. Bernstein et al. “Increased Incidence of Inflammatory Bowel Disease After Hirschsprung Disease: A Population-based Cohort Study”. en. In: *The Journal of Pediatrics* 233 (June 2021), 98–104.e2. ISSN: 00223476. DOI: 10.1016/j.jpeds.2021.01.060. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022347621000949 (visited on 12/16/2023).

[37] G. Romeo et al. “Point mutations affecting the tyrosine kinase domain of the RET proto-oncogene in Hirschsprung’s disease”. eng. In: *Nature* 367.6461 (Jan. 1994), pp. 377–378. ISSN: 0028-0836. DOI: 10.1038/367377a0.

[38] Pablo Cingolani et al. “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff”. In: *Fly* 6.2 (Apr. 2012), pp. 80–92. ISSN: 1933-6934. DOI: 10.4161/fly.19695. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3679285/ (visited on 12/27/2023).

[39] F. Lantieri et al. “Haplotypes of the human RET proto-oncogene associated with Hirschsprung disease in the Italian population derive from a single ancestral combination of alleles”. eng. In: *Annals of Human Genetics* 70.Pt 1 (Jan. 2006), pp. 12–26. ISSN: 0003-4800. DOI: 10.1111/j.1529-8817.2005.00196.x.

[40] Tiziana Bachetti and Isabella Ceccherini. “Causative and common PHOX2B variants define a broad phenotypic spectrum”. eng. In: *Clinical Genetics* 97.1 (Jan. 2020), pp. 103–113. ISSN: 1399-0004. DOI: 10.1111/cge.13633.

[41] R. S. Spielman, R. E. McGinnis, and W. J. Ewens. “Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)”. eng. In: *American Journal of Human Genetics* 52.3 (Mar. 1993), pp. 506–516. ISSN: 0002-9297.

[42] Frank Dudbridge et al. “Unbiased Application of the Transmission/Disequilibrium Test to Multilocus Haplotypes”. English. In: *The American Journal of Human Genetics* 66.6 (June 2000). Publisher: Elsevier, pp. 2009–2012. ISSN: 0002-9297, 1537-6605. DOI: 10.1086/302915. URL: https://www.cell.com/ajhg/abstract/S0002-9297(07)63557-5 (visited on 12/16/2023).

[43] Joshua Henrina Sundjaja, Rijen Shrestha, and Kewal Krishan. “McNemar And Mann-Whitney U Tests”. eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK560699/ (visited on 12/16/2023).

[44] Ulrich Omasits et al. “Protter: interactive protein feature visualization and integration with experimental proteomic data”. In: *Bioinformatics* 30.6 (Mar. 2014), pp. 884–886. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt607. URL: https://doi.org/10.1093/bioinformatics/btt607 (visited on 12/16/2023).

[45] Marco Biasini et al. “SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information”. eng. In: *Nucleic Acids Research* 42.Web Server issue (July 2014), W252–258. ISSN: 1362-4962. DOI: 10.1093/nar/gku340.

[46] Qingqing Du, Yan Qian, and Weiwei Xue. “Molecular Simulation of Oncostatin M and Receptor (OSM–OSMR) Interaction as a Potential Therapeutic Target for Inflammatory Bowel Disease”. In: *Frontiers in Molecular Biosciences* 7 (2020). ISSN: 2296-889X. URL: https://www.frontiersin.org/articles/10.3389/fmolb.2020.00029 (visited on 12/16/2023).

[47] Kiyoshi Migita et al. “CP690,550 inhibits oncostatin M-induced JAK/STAT signaling pathway in rheumatoid synoviocytes”. In: *Arthritis Research & Therapy* 13.3 (May 2011), R72. ISSN: 1478-6354. DOI: 10.1186/ar3333. URL: https://doi.org/10.1186/ar3333 (visited on 12/16/2023).

[48] *Protocol for adhesion and immunostaining of lymphocytes and other non-adherent cells in culture — BioTechniques*. URL: https://www.future-science.com/doi/full/10.2144/000114610?rfr_dat=cr_pub++0pubmed&url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org (visited on 12/16/2023).

[49] Stefka Tyanova et al. “The Perseus computational platform for comprehensive analysis of (prote)omics data”. en. In: *Nature Methods* 13.9 (Sept. 2016). Number: 9 Publisher: Nature Publishing Group, pp. 731–740. ISSN: 1548-7105. DOI: 10.1038/nmeth.3901. URL: https://www.nature.com/articles/nmeth.3901 (visited on 12/16/2023).

[50] Federica Raggi et al. “Proteomic profiling of extracellular vesicles in synovial fluid and plasma from Oligoarticular Juvenile Idiopathic Arthritis patients reveals novel immunopathogenic biomarkers”. In: *Frontiers in Immunology* 14 (2023). ISSN: 1664-3224. URL: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1134747 (visited on 12/16/2023).

[51] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. “Significance analysis of microarrays applied to the ionizing radiation response”. In: *Proceedings of the National Academy of Sciences* 98.9 (Apr. 2001). Publisher: Proceedings of the National Academy of Sciences, pp. 5116–5121. DOI: 10.1073/pnas.091062498. URL: https://www.pnas.org/doi/full/10.1073/pnas.091062498 (visited on 12/16/2023).

[52] Ismail Bin Mohamad and Dauda Usman. “Standardization and Its Effects on K-Means Clustering Algorithm”. en. In: *Research Journal of Applied Sciences, Engineering and Technology* 6.17 (Sept. 2013), pp. 3299–3303. ISSN: 20407459, 20407467. DOI: 10.19026/rjaset.6.3638. URL: http://maxwellsci.com/jp/mspabstract.php?jid=RJASET&doi=rjaset.6.3638 (visited on 12/16/2023).

[53] Michael I. Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12 (Dec. 2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8 (visited on 12/16/2023).

[54] Cathy C. Laurie et al. “Quality control and quality assurance in genotypic data for genome-wide association studies”. eng. In: *Genetic Epidemiology* 34.6 (Sept. 2010), pp. 591–602. ISSN: 1098-2272. DOI: 10.1002/gepi.20516.

[55] Anthony S. Castanza et al. “Extending support for mouse data in the Molecular Signatures Database (MSigDB)”. eng. In: *Nature Methods* 20.11 (Nov. 2023), pp. 1619–1620. ISSN: 1548-7105. DOI: 10.1038/s41592-023-02014-7.

[56] Eleonora Di Zanni et al. “Identification of novel pathways and molecules able to down-regulate PHOX2B gene expression by in vitro drug screening approaches in neuroblastoma cells”. eng. In: *Experimental Cell Research* 336.1 (Aug. 2015), pp. 43–57. ISSN: 1090-2422. DOI: 10.1016/j.yexcr.2015.03.025.

[57] Prabhaker Mishra et al. “Descriptive Statistics and Normality Tests for Statistical Data”. In: *Annals of Cardiac Anaesthesia* 22.1 (2019), pp. 67–72. ISSN: 0971-9784. DOI: 10.4103/aca.ACA_157_18. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350423/ (visited on 12/16/2023).

[58] "Spearman Rank Correlation Coefficient". en. In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer, 2008, pp. 502–505. ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_379. URL: https://doi.org/10.1007/978-0-387-32833-1_379 (visited on 12/16/2023).

[59] Ashish Kapoor et al. "Population variation in total genetic risk of Hirschsprung disease from common RET, SEMA3 and NRG1 susceptibility polymorphisms". In: *Human Molecular Genetics* 24.10 (May 2015), pp. 2997–3003. ISSN: 0964-6906. DOI: 10.1093/hmg/ddv051. URL: https://doi.org/10.1093/hmg/ddv051 (visited on 12/30/2023).

[60] J. Amiel et al. "Heterozygous endothelin receptor B (EDNRB) mutations in isolated Hirschsprung disease". eng. In: *Human Molecular Genetics* 5.3 (Mar. 1996), pp. 355–357. ISSN: 0964-6906. DOI: 10.1093/hmg/5.3.355.

[61] Qian Jiang et al. "Functional Loss of Semaphorin 3C and/or Semaphorin 3D and Their Epistatic Interaction with Ret Are Critical to Hirschsprung Disease Liability". In: *The American Journal of Human Genetics* 96.4 (Apr. 2015), pp. 581–596. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2015.02.014. URL: https://www.sciencedirect.com/science/article/pii/S0002929715000671 (visited on 12/30/2023).

[62] Clara Sze-man Tang et al. "Trans-ethnic meta-analysis of genome-wide association studies for Hirschsprung disease". In: *Human Molecular Genetics* 25.23 (Dec. 2016), pp. 5265–5275. ISSN: 0964-6906. DOI: 10.1093/hmg/ddw333. URL: https://doi.org/10.1093/hmg/ddw333 (visited on 12/30/2023).

[63] Eileen Sproat Emison et al. "A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk". en. In: *Nature* 434.7035 (Apr. 2005). Number: 7035 Publisher: Nature Publishing Group, pp. 857–863. ISSN: 1476-4687. DOI: 10.1038/nature03467. URL: https://www.nature.com/articles/nature03467 (visited on 12/30/2023).

[64] Sumantra Chatterjee et al. "Enhancer Variants Synergistically Drive Dysfunction of a Gene Regulatory Network In Hirschsprung Disease". In: *Cell* 167.2 (Oct. 2016), 355–368.e10. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.09.005. URL: https://www.sciencedirect.com/science/article/pii/S0092867416312351 (visited on 12/30/2023).

[65] Tiziana Bachetti et al. "The OSMR Gene Is Involved in Hirschsprung Associated Enterocolitis Susceptibility through an Altered Downstream Signaling". eng. In: *International Journal of Molecular Sciences* 22.8 (Apr. 2021), p. 3831. ISSN: 1422-0067. DOI: 10.3390/ijms22083831.

[66] Avik Sarkar, Kalpana Panati, and Venkata Ramireddy Narala. “Code inside the codon: The role of synonymous mutations in regulating splicing machinery and its impact on disease”. In: *Mutation Research/Reviews in Mutation Research* 790 (July 2022), p. 108444. ISSN: 1383-5742. DOI: 10.1016/j.mrrev.2022.108444. URL: https://www.sciencedirect.com/science/article/pii/S1383574222000345 (visited on 04/24/2024).

[67] John D. Porter et al. “A chronic inflammatory response dominates the skeletal muscle molecular signature in dystrophin-deficient mdx mice”. eng. In: *Human Molecular Genetics* 11.3 (Feb. 2002), pp. 263–272. ISSN: 0964-6906. DOI: 10.1093/hmg/11.3.263.

[68] John Biddlestone, Daniel Bandarra, and Sonia Rocha. “The role of hypoxia in inflammatory disease (Review)”. In: *International Journal of Molecular Medicine* 35.4 (Apr. 2015). Publisher: Spandidos Publications, pp. 859–869. ISSN: 1107-3756. DOI: 10.3892/ijmm.2015.2079. URL: https://www.spandidos-publications.com/10.3892/ijmm.2015.2079 (visited on 12/30/2023).

[69] Laszlo Nemeth, Udo Rolle, and Prem Puri. “Altered cytoskeleton in smooth muscle of aganglionic bowel”. eng. In: *Archives of Pathology & Laboratory Medicine* 126.6 (June 2002), pp. 692–696. ISSN: 0003-9985. DOI: 10.5858/2002-126-0692-ACISMO.

[70] Jin Zhu et al. “Dysmorphic Neurofilament-Positive Ganglion Cells in the Myenteric Plexus at the Proximal Resection Margin Indicate Worse Postoperative Prognosis in Hirschsprung’s Disease”. en. In: *Pediatric and Developmental Pathology* 23.3 (June 2020). Publisher: SAGE Publications Inc, pp. 222–229. ISSN: 1093-5266. DOI: 10.1177/1093526619878083. URL: https://doi.org/10.1177/1093526619878083 (visited on 12/30/2023).

[71] Melissa Payet et al. “Inflammatory Mesenchymal Stem Cells Express Abundant Membrane-Bound and Soluble Forms of C-Type Lectin-like CD248”. eng. In: *International Journal of Molecular Sciences* 24.11 (May 2023), p. 9546. ISSN: 1422-0067. DOI: 10.3390/ijms24119546.

[72] Melanie I. Morris et al. “A study of calretinin in Hirschsprung pathology, particularly in total colonic aganglionosis”. eng. In: *Journal of Pediatric Surgery* 48.5 (May 2013), pp. 1037–1043. ISSN: 1531-5037. DOI: 10.1016/j.jpedsurg.2013.02.026.

[73] Zhi-Hua Chen et al. “Characterization of Interstitial Cajal Progenitors Cells and Their Changes in Hirschsprung’s Disease”. en. In: *PLOS ONE* 9.1 (Jan. 2014). Publisher: Public Library of Science, e86100. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0086100. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0086100 (visited on 12/30/2023).

[74] Jing Wang et al. “Identification and validation of the common pathogenesis and hub biomarkers in Hirschsprung disease complicated with Crohn’s disease”. In: *Frontiers in Immunology* 13 (Sept. 2022), p. 961217. ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.961217. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9555215/ (visited on 12/30/2023).

[75] Malachi Griffith et al. “Alternative expression analysis by RNA sequencing”. en. In: *Nature Methods* 7.10 (Oct. 2010). Number: 10 Publisher: Nature Publishing Group, pp. 843–847. ISSN: 1548-7105. DOI: 10.1038/nmeth.1503. URL: https://www.nature.com/articles/nmeth.1503 (visited on 12/27/2023).

[76] Yan W. Asmann et al. “3’ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer”. In: *BMC Genomics* 10.1 (Nov. 2009), p. 531. ISSN: 1471-2164. DOI: 10.1186/1471-2164-10-531. URL: https://doi.org/10.1186/1471-2164-10-531 (visited on 12/27/2023).

[77] Angela R. Wu et al. “Quantitative assessment of single-cell RNA-sequencing methods”. en. In: *Nature Methods* 11.1 (Jan. 2014). Number: 1 Publisher: Nature Publishing Group, pp. 41–46. ISSN: 1548-7105. DOI: 10.1038/nmeth.2694. URL: https://www.nature.com/articles/nmeth.2694 (visited on 12/27/2023).

[78] Yu Shi and Maoxian He. “Differential gene expression identified by RNA-Seq and qPCR in two sizes of pearl oyster (Pinctada fucata)”. In: *Gene* 538.2 (Apr. 2014), pp. 313–322. ISSN: 0378-1119. DOI: 10.1016/j.gene.2014.01.031. URL: https://www.sciencedirect.com/science/article/pii/S0378111914000523 (visited on 12/27/2023).

[79] Xiaofeng Dai and Li Shen. “Advances and Trends in Omics Technology Development”. In: *Frontiers in Medicine* 9 (2022). ISSN: 2296-858X. URL: https://www.frontiersin.org/articles/10.3389/fmed.2022.911861 (visited on 12/16/2023).

[80] Bin Hu et al. “Challenges in Bioinformatics Workflows for Processing Microbiome Omics Data at Scale”. In: *Frontiers in Bioinformatics* 1 (2022). ISSN: 2673-7647. URL: https://www.frontiersin.org/articles/10.3389/fbinf.2021.826370 (visited on 12/16/2023).

[81] Jennifer L. Bailey et al. “Oncostatin M Induces Lipolysis and Suppresses Insulin Response in 3T3-L1 Adipocytes”. In: *International Journal of Molecular Sciences* 23.9 (Apr. 2022), p. 4689. ISSN: 1422-0067. DOI: 10.3390/ijms23094689. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9104719/ (visited on 12/31/2023).

[82] Michitaka Matsuda et al. “Oncostatin M causes liver fibrosis by regulating cooperation between hepatic stellate cells and macrophages in mice”. eng. In: *Hepatology (Baltimore, Md.)* 67.1 (Jan. 2018), pp. 296–312. ISSN: 1527-3350. DOI: 10.1002/hep.29421.

[83] Zhenjia Yu et al. “Oncostatin M receptor, positively regulated by SP1, promotes gastric cancer growth and metastasis upon treatment with Oncostatin M”. eng. In: *Gastric Cancer: Official Journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association* 22.5 (Sept. 2019), pp. 955–966. ISSN: 1436-3305. DOI: 10.1007/s10120-019-00934-y.

[84] Kathryn L. Pothoven and Robert P. Schleimer. “The barrier hypothesis and Oncostatin M: Restoration of epithelial barrier function as a novel therapeutic strategy for the treatment of type 2 inflammatory disease”. In: *Tissue Barriers* 5.3 (June 2017), e1341367. ISSN: 2168-8362. DOI: 10.1080/21688370.2017.1341367. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5571776/ (visited on 12/16/2023).

[85] Nathaniel R. West et al. “Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor–neutralizing therapy in patients with inflammatory bowel disease”. en. In: *Nature Medicine* 23.5 (May 2017). Number: 5 Publisher: Nature Publishing Group, pp. 579–589. ISSN: 1546-170X. DOI: 10.1038/nm.4307. URL: https://www.nature.com/articles/nm.4307 (visited on 12/16/2023).

[86] Holger Lörchner et al. “Concomitant Activation of OSM and LIF Receptor by a Dual-Specific hlOSM Variant Confers Cardioprotection after Myocardial Infarction in Mice”. eng. In: *International Journal of Molecular Sciences* 23.1 (Dec. 2021), p. 353. ISSN: 1422-0067. DOI: 10.3390/ijms23010353.

[87] Lena Jakob et al. “Murine Oncostatin M Has Opposing Effects on the Proliferation of OP9 Bone Marrow Stromal Cells and NIH/3T3 Fibroblasts Signaling through the OSMR”. In: *International Journal of Molecular Sciences* 22.21 (Oct. 2021), p. 11649. ISSN: 1422-0067. DOI: 10.3390/ijms222111649. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8584221/ (visited on 12/16/2023).

[88] Elina Nummenmaa et al. “Transient Receptor Potential Ankyrin 1 (TRPA1) Is Involved in Upregulating Interleukin-6 Expression in Osteoarthritic Chondrocyte Models”. eng. In: *International Journal of Molecular Sciences* 22.1 (Dec. 2020), p. 87. ISSN: 1422-0067. DOI: 10.3390/ijms22010087.

[89] Yao Li et al. “Feiyangchangweiyan capsule protects against ulcerative colitis in mice by modulating the OSM/OSMR pathway and improving gut microbiota”. eng. In: *Phytomedicine: International Journal of Phytotherapy and Phytopharmacology* 80 (Jan. 2021), p. 153372. ISSN: 1618-095X. DOI: 10.1016/j.phymed.2020.153372.

[90] Saad Y. Salim et al. “Oncostatin M Receptor Type II Knockout Mitigates Inflammation and Improves Survival from Sepsis in Mice”. In: *Biomedicines* 11.2 (). ISSN: 2227-9059. DOI: 10.3390/biomedicines11020483. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9953488/ (visited on 12/31/2023).

[91] Heng Li et al. “Intervention of oncostatin M-driven mucosal inflammation by berberine exerts therapeutic property in chronic ulcerative colitis”. en. In: *Cell Death & Disease* 11.4 (Apr. 2020). Number: 4 Publisher: Nature Publishing Group, pp. 1–17. ISSN: 2041-4889. DOI: 10.1038/s41419-020-2470-8. URL: https://www.nature.com/articles/s41419-020-2470-8 (visited on 12/16/2023).

[92] Laure Frésard et al. “Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts”. en. In: *Nature Medicine* 25.6 (June 2019). Number: 6 Publisher: Nature Publishing Group, pp. 911–919. ISSN: 1546-170X. DOI: 10.1038/s41591-019-0457-8. URL: https://www.nature.com/articles/s41591-019-0457-8 (visited on 12/16/2023).

[93] Orietta Pansarasa et al. “Biomarkers in Human Peripheral Blood Mononuclear Cells: The State of the Art in Amyotrophic Lateral Sclerosis”. In: *International Journal of Molecular Sciences* 23.5 (Feb. 2022), p. 2580. ISSN: 1422-0067. DOI: 10.3390/ijms23052580. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8910056/ (visited on 12/16/2023).

[94] Abhijeet Rajendra Sonawane et al. “Understanding Tissue-Specific Gene Regulation”. In: *Cell reports* 21.4 (Oct. 2017), pp. 1077–1088. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.10.001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5828531/ (visited on 12/16/2023).

[95] Marianna Lucafò et al. “Azathioprine Biotransformation in Young Patients with Inflammatory Bowel Disease: Contribution of Glutathione-S Transferase M1 and A1 Variants”. In: *Genes* 10.4 (Apr. 2019), p. 277. ISSN: 2073-4425. DOI: 10.3390/

genes10040277. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6523194/ (visited on 12/31/2023).

[96] Nathaniel R. West. "Coordination of Immune-Stroma Crosstalk by IL-6 Family Cytokines". In: *Frontiers in Immunology* 10 (2019). ISSN: 1664-3224. URL: https://www.frontiersin.org/articles/10.3389/fimmu.2019.01093 (visited on 12/31/2023).

[97] Nan Zhao et al. "Role of Interleukin-22 in ulcerative colitis". In: *Biomedicine & Pharmacotherapy* 159 (Mar. 2023), p. 114273. ISSN: 0753-3322. DOI: 10.1016/j.biopha.2023.114273. URL: https://www.sciencedirect.com/science/article/pii/S0753332223000616 (visited on 12/31/2023).

[98] Erik W. van Zwet and Eric A. Cator. "The significance filter, the winner's curse and the need to shrink". en. In: *Statistica Neerlandica* 75.4 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pd pp. 437–452. ISSN: 1467-9574. DOI: 10.1111/stan.12241. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/stan.12241 (visited on 12/18/2023).

[99] Konstantinos C. M. Siontis, Nikolaos A. Patsopoulos, and John P. A. Ioannidis. "Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies". en. In: *European Journal of Human Genetics* 18.7 (July 2010). Number: 7 Publisher: Nature Publishing Group, pp. 832–837. ISSN: 1476-5438. DOI: 10.1038/ejhg.2010.26. URL: https://www.nature.com/articles/ejhg201026 (visited on 12/20/2023).