UNIVERSITÀ DEGLI STUDI DI SALERNO

Màster Universitari
Anàlisi de Dades Òmiques /
Omics Data Analysis
FACULTAT DE CIÈNCIES I TECNOLOGIA
UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

Human gut derived-organoids provide model to study gluten response and effects of microbiota derived molecules in celiac disease

Dr. Davide Biondi

Supervisor: Prof. Roberto Tagliaferri,
DISA-MIS, University of Salerno (IT)
Tutor: Prof. Jordi Villà Freixa,
Departament de Biociències,
University of Vic (SP)

Biosciences Department

University of Vic – Central University of Catalonia

September 10, 2024

Index of Contents

1 Summary

Abstract

Celiac disease (CD) is an immune-mediated condition triggered by exposure to gluten. Although extensive research has been conducted on the role of the adaptive immune system in CD development, a lack of effective experimental models has limited our insight into the initial stages that lead to gluten intolerance. By utilizing intestinal organoids derived from duodenal biopsies of both non-celiac (NC) and celiac (CD) patients, I investigated the role of the gut epithelium in CD pathogenesis, focusing on how the epithelium responds to gluten. Compared to NC organoids, RNA sequencing of CD organoids demonstrated significant changes in the expression of genes related to the gut barrier and innate immune response. A total of 83 differentially expressed genes has been discovered, of which 45 were overexpressed in CD patients, and 38 were underexpressed.

Abbreviations

CD = Celiac disease NC = Non-celiac ; SE = Single-end ; PE = Paired-end ; GSEA = Gene Set Enrichment Analysis ; GOA = GO annotation analysis ; GFD = Gluten-free diet ; GCD = Gluten-containing diet ; DAVID = Database for Annotation, Visualization and Integrated Discovery ; STAR = Spliced Transcripts Alignment to a Reference ; TEER = Transepithelial electrical resistance ; PDTOs = Patient-derived tumor organoids ; iPSCs = Induced pluripotent stem cells ; TG2 = Transglutaminase 2 ; tTG = Tissue transglutaminase ; HLA = Human leukocyte antigen ; HP = Haptoglobin ; ReA = Reactive arthritis ; RA = Rheumatoid arthritis ; ACPA = Anti-citrullinated protein antibodies ; IFN = Interferon ; IL = Interleukin ; EGFR = Epidermal growth factor receptor ; TGF = Transforming growth factor ; Th = T helper cells ; SARA = SMAD anchor for receptor activation ; HMMs = Hidden Markov Models ; SAs = Suffix arrays ; PTPRK = Protein Tyrosine Phosphatase Receptor Type K ; DEFA6 = Defensin, alpha 6 ; HD6 = Human alpha defensin 6 ; TSP-1 = Thrombospondin-1 ; GPIV = Glycoprotein IV ; HVR = Hypervariable region ; SEP = shared epitope

2 Introduction

3D cell culture models, such as spheroids and organoids, were developed to better mimic the spatial and microenvironmental information of the in vivo situation. Organoids can be established from embryonic stem cells, adult stem cells, or induced pluripotent stem cells (iPSCs)[1]. Patient- derived tumor organoids (PDTOs) can be grown directly from patient tissue biopsies or surgically removed tumor tissues, the same applies for duodenal biopsies of NC and CD patients. Organoids can form organ-like structures reminiscent of the tissues they originated from and contain both stem and differentiated cell populations resembling patient tissues. Histological characteristics, such as crypt hyperplasia and villous blunting, are defining features of active celiac disease (CD). This condition is marked by an increase in the population of immature cells, accompanied by a simultaneous decrease in the mature

cells of the villi.[2]. These pathological changes in the intestinal mucosa associated with CD are believed to result from a defective stem cell compartment that fails to adequately support tissue regeneration and/or from heightened epithelial apoptosis[3]. A study highlighted the role of inflammation in CD, noting that even in the absence of gluten, there is a low-grade inflammation present in CD cells and organoids. This suggests that organoids can be a valuable tool for studying the inflammatory processes in CD, even before gluten exposure[4]. Another study suggested that the inflammatory markers pNF-κB, pERK, IL-1β, and IL-6 were increased and persistent in CD organoids[5]. Celiac disease is also considered a "Western" pathology, because in general, the Western diet and lifestyle have been long linked to low-grade inflammation that represents the "common background" of several different diseases[6]. Celiac disease is an immune-mediated enteropathy triggered in genetically susceptible individuals by a group of wheat proteins and related prolamins from cereals[7]. It is induced by gluten consumption in individuals genetically predisposed with HLA-DQ2 or HLA-DQ8 haplotypes[8]. As regarding the molecular pathology aspect of the disease, we know that the enzyme transglutaminase 2 (TG2) plays a key role in celiac disease (CD) pathogenesis. Active TG2 is located mainly extracellularly in the lamina propria but also in the villous enterocytes of the duodenum. The TG2 inhibitor ZED1227 is a promising drug candidate for treating CD and is designed to block the TG2-catalyzed deamidation and crosslinking of gliadin peptides[9]. This role of TG2 has as a consequence the proliferation of gluten-specific and tissue transglutaminase (tTG) antibodies expressed during the acute phase[10]. Another pivotal role in the pathogenesis of CD is played by the HLA (human leukocyte antigen)-restricted gliadin-specific intestinal T cell response[7,11]. Although activation of T cells has been studied in depth, the central question still remains unanswered, namely, why a pro-inflammatory T cell response is generated toward gliadin instead of a regulatory response, which normally promotes oral tolerance to dietary protein antigens. When, how and where this inflammation is generated in CD is not clear. The evaluation of the haptoglobin (HP) genotype of the organoids employed in the study is also important. Haptoglobin encodes for two alleles, of which HP2 carries a duplication of the 3$^{\text{rd}}$ and 4$^{\text{th}}$ exons[12]. The allele HP2, encodes for the HP2 pre-haptoglobin also known as zonulin, that has been shown to be a physiological regulator of tight junctions and it is released upon gluten stimulation[13]. All the organoids in this study had at least one copy of the HP2 allele (HP2-1 or HP2-2 genotype). Based on available evidence, it is suggested that the onset of CD follows this sequence of events: after ingestion, undigested gliadin peptides stimulate the release of zonulin, resulting in increased intestinal permeability[14,15]. The compromised barrier permits the translocation of gliadin peptides into the lamina propria, where they interact with macrophages and other immune cells[16]. Gliadin peptides trigger IL8 and IL15 secretion from enterocytes, which in turn recruits neutrophils[17] and intraepithelial lymphocytes[18]. Ultimately, the engagement of T cells with cells presenting gliadin peptides in a pro-inflammatory environment results in the breakdown of oral tolerance and the activation of the Th1/Th17 adaptive immune response[19]. In this study we focused on spotting the aforementioned genes in our differentially expression analysis, based on the hypothesis that organoids can reproduce the epression patterns of CD patients.

3 Materials and Methods

Hardware specs influence on results

Most of the pre-processing (till the rRNA sequence removal) has been done on a Linux Mint machine, with the following system specifications: System: Kernel: 5.4.0-135-generic x86_64 bits: 64 compiler: gcc v: 9.4.0 Desktop: Cinnamon 4.6.7 wm: muffin dm: LightDM Distro: Linux Mint 20 Ulyana base: Ubuntu 20.04 focal Machine: Type: Laptop System: Dynabook product: SATELLITE PRO C50-E-10R v: PYS20E-00L0D2IT

Mobo: Dynabook model: DBIIP300 UEFI: American Megatrends v: KN16RV100 date: 05/26/2020 CPU: Topology: Dual Core model: Intel Core i3-8130U bits: 64 type: MT MCP arch: Kaby Lake rev: A L2 cache: 4096 KiB flags: avx avx2 lm nx pae sse sse2 sse3 sse4_1 sse4_2 ssse3 vmx bogomips: 17599

These specs were obtained with the bash command inxi. The rest of the RNA-seq pipeline from indexing of the genome, reads mapping and quantification till GSEA was done on an Ubuntu machine with the following system specifications: System: Kernel: 5.15.0-119-generic x86_64 bits: 64 compiler: N/A Desktop: Gnome 3.36.9 wm: gnome-shell dm: GDM3 Distro: Ubuntu 20.04.6 LTS (Focal Fossa) Machine: Type: Desktop System: Dell product: Precision Tower 5810 v: N/A Chassis: type: 7 Mobo: Dell model: 0HHV7N v: A00 UEFI: Dell v: A20 date: 07/26/2017 CPU: Topology: 10-Core model: Intel Xeon E5-2630 v4 bits: 64 type: MT MCP arch: Broadwell rev: 1 L2 cache: 25.0 MiB flags: avx avx2 lm nx pae sse sse2 sse3 sse4_1 sse4_2 ssse3 vmx bogomips: 87798

Addressing not only the software specifications, but also the hardware ones, is important, because a computer hardware architecture can have an impact not only on the time required to execute a certain algorithm, but on the output itself. This is true when it comes to floating-point arithmetic[20]. Floating-point representations have a base β, (which is always assumed to be even) and a precision p. If $\beta = 10$ and $p = 3$, the number 0.1 is represented as 1.00×10^{-1}.

Adapter sequence trimming and QC

The RNA-seq data in fastq format was downloaded from the GEO database[21]. From the downloaded data all the Phred scores for all the bases resulted as leveled to 30 (assigned with the symbol "?"). The real raw SRA data were not available due to unidentified issues. Nonetheless, quality control was completed using trim-galore version 0.6.10 build hdfd78af_0, installed from Bioconda channel. The conda virtual environment in which Trim Galore was launched had Python version 3.8.19, and cutadapt 4.9. The adapter sequence was AGATCG-GAAGAGC. The trimming mode was for single-end reads. The cutoff for removing sequences was 20 bp. The quality Phred score cutoff was set up as 20. Trim Galore uses Cutadapt to identify and remove adapter sequences from the 3' end of reads. By default, it trims the first 13 base pairs of standard Illumina adapters. The tool trims low-quality bases from the 3' end of reads, as mentione before. This is done before adapter trimming to ensure that only high-quality data is retained for downstream analysis[22]. After trimming, reads that are shorter than a specified length can be removed in a length filtering, to prevent issues during alignment and analysis. FastQC was used after trimming the adapter sequences, as it is con-

sidered a best practice. FastQC - which is used together with Cutadapt with Trim Galore acting as a wrapper between the two - performs a series of modular analyses, each focusing on different aspects of the data quality. These analyses include: Per Base Sequence Quality: Assesses the quality scores across all bases at each position in the reads. It helps identify regions with low-quality scores that may need trimming. Per Sequence Quality Scores: Evaluates the distribution of quality scores across all reads to detect any reads with universally low quality. Per Base Sequence Content: Analyzes the proportion of each nucleotide at every position in the reads. Imbalances may indicate biases or contamination. Per Sequence GC Content: Provides a density plot of the GC content across all reads, which can highlight unusual patterns. Overrepresented Sequences: Identifies sequences that occur more frequently than expected, which may indicate contamination or adapter sequences. Adapter Content: Detects the presence of adapter sequences within the reads. K-mer Content: Analyzes the frequency of short nucleotide sequences (k-mers) to identify biases[23].

rRNA sequences removal

After this initial trimming/quality-control step, rRNA sequences removal has been executed with sortmerna version 4.3.6 build h9ee0642_0 installed through bioconda channel. Sort-MeRNA utilizes a seeding strategy that focuses on identifying short regions of similarity between rRNA sequences in a database and the input reads. This approach is akin to the use of Hidden Markov Models (HMMs) but is executed significantly faster[24]. SortMeRNA uses a burst tries data structure coupled with a lookup table to efficiently store and access the rRNA database. This setup enhances the speed and accuracy of the filtering process[24]. The filtering was ran using default parameters, namely seed length as 18, Gumbel lambda as 0.601061, Gumbel K as 0.328323. The other parameters are available in the GitHub[25].

Genome indexing and reads mapping

After indexing the hg19 fasta file genome from UCSC genome browser, using a GTF EN-SEMBL annotation file from the same portal, the mapping of the reads without rRNA was done. Indexing and mapping were performed using star version 2.5.2b build 0 from bioconda channel. Mapping consists of two phases: seed finding, and then another phase of clustering, stitching and scoring. Seed finding phase is the sequential search for a Maximal Mappable Prefix (MMP)[26]. The splice junctions are detected in a single alignment pass without any a priori knowledge of splice junctions' loci or properties, and without a preliminary contiguous alignment pass needed by the junction database approaches[26]. The MMP in STAR search is implemented through uncompressed suffix arrays (SAs)[27]. Finding MMP is an inherent outcome of the standard binary string search in uncompressed SAs. The MMP search, implemented in STAR, enables finding multiple mismatches and indels. If the MMP search does not reach the end of a read because of the presence of one or more mismatches, the MMPs will serve as anchors in the genome that can be extended, allowing for alignments with mismatches. In some cases, the extension procedure does not yield a good genomic alignment, which allows identification of poly-A tails, library adapter sequences or poor sequencing quality tails[26]. In the second phase of the algorithm, STAR builds alignments of

the entire read sequence by stitching together all the seeds that were aligned to the genome in the first phase. First, the seeds are clustered together by proximity to a selected set of 'anchor' seeds[26]. All the seeds that map within user-defined genomic windows around the anchors are stitched together assuming a local linear transcription model. The stitching is guided by a local alignment scoring scheme, with user-defined scores (penalties) for matches, mismatches, insertions, deletions and splice junction gaps, allowing for a quantitative assessment of the alignment qualities and ranks[26]. STAR was set up for both reads mapping and quantification, in order to have additional control of the quantification step so that reads counts could be compared between STAR and HTseq. The reads average length was 50, so it got diminished of one unit in the STAR parameters, as suggested by the developers. SAM files were also generated during the operations with STAR, as diagnostic tools, and as a best practice. The BAM files generated with STAR and accepted from HTseq as input were of the type "SortedByCoordinate". STAR does not need to know the strandness of the data, but the strandness of the reads and their length are fundamental parameters in both indexing of the genome, and mapping and quantification of the reads. A bash script was used in order to assess the strandness of the reads, parsing "out.tab" files obtained as output from STAR reads mapping and quantification, and another script was used in order to check the average length of the reads by parsing the fastq files with awk text processing tool.

Quantification

The quantification (read counts over transcripts) which was used for downstream analysis was completed with HTseq version 0.6.0[28]. Quantification was ran with the basic default parameters. Again, HTseq parameters had to mirror the ones used for STAR, so a parameter was set up for HTseq to recognize reads on the reverse strand, where the reads selectively map; regarding the key –nonunique parameter, this is by default set as "none", so the read (or read pair) is counted as ambiguous and not counted for any features. Also, if the read (or read pair) aligns to more than one location in the reference, it is scored as alignment_not_unique[29]. Another parameter was set up in order to address the kind of file to parse, in this case BAM sortedByAlignment, as mentioned before.

Downstream analysis

Differential expression analysis on the read counts over transcripts was conducted with pydeseq2[30] version 0.4.4 build pyhdfd78af_0 from bioconda channel. Genes with a total of less than 10 counts were excluded, and after this filtering, genes with baseMean $>= 10$, of which adjusted p value < 0.05 and absolute value of log2FC > 0.5 were selected as differentially expressed genes. Functional gene set enrichment analysis (GSEA) was performed on the whole- transcriptome expression values using gseapy[31] version 1.1.3 build py38he1d39cf_1 from bioconda channel, with a comparison against 'GO Biological Process 2021' gene sets with default parameters. Functional annotation analysis was performed on differentially expressed genes using DAVID v6.8[32,33]. Gene set enrichment analysis is ran on the whole-transcriptome expression values compared to GO annotation analysis (GOA), because it can

capture subtle changes between the samples conditions that GOA could miss. It is done in conjunction with GOA as a best practice in order to get the biological meaning of the data as a whole.

4   Results

During the pre-processing with Trim Galore the sequences removed because they became shorter than the length cutoff were between the 0.3-0.5%. The fastqc report signaled warnings regarding the sequence duplication levels, and the per base sequence content, that would be difficult to address without a precise protocol followed by the authors for the PCR and the library preparation. The results of the filtering with SortMeRNA pointed out that between 99,97% and 99,98% of the reads failed the E-value threshold, meaning that they do not contain rRNA. The output of STAR concerning reads mapping and quantification consisted in a range of 16–21 million aligned single-end 50 bp reads per sample, mapped on the reverse strand. This is because the majority of the library preparation protocols are designed to produce stranded reads, most of the time mapping on the reverse strand.

Out of the most interesting DEGs found during the differential expression analysis (DEA), DKK1, a WNT negative regulator[34] was significantly down-regulated. Trefoil factors (TFF1 and TFF2) were also significantly down-regulated. Trefoil factors (TFFs) are a family of small peptides that play crucial roles in maintaining and repairing mucosal surfaces, particularly in the gastrointestinal tract. On the other hand, the upregulation of NLRP6, initiator of the inflammasome complex formation[35], has been confirmed. Another gene confirmed as overexpressed is Defensin, alpha 6 (DEFA6) also known as human alpha defensin 6 (HD6).

One interesting finding is the overexpression of CD36, a membrane protein that belongs to the class B scavenger receptor family. In cancer, CD36 expression promotes tumor development, metastasis, drug resistance[36]. Also known as platelet glycoprotein IV (GPIV), it was demonstrated to be a receptor for thrombospondin-1 (TSP-1), a glycoprotein involved in platelet adhesion and an endogenous inhibitor of angiogenesis[37].
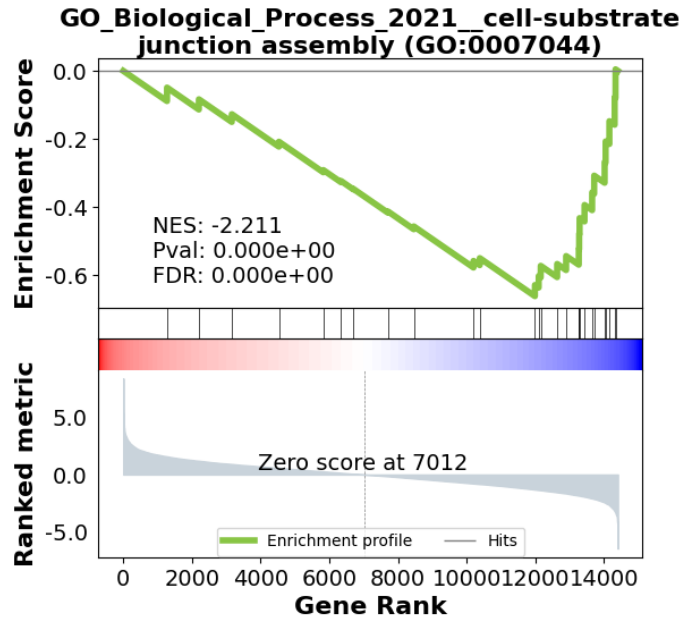
Figure 1: GSEA plot showing the most significant negative normalized enrichment score (NES), considering its absolute value was the highest. The gene set for the term "cell-substrate junction assembly" is more associated with the control group than with the disease group.The red area indicates genes that are more prevalent in the disease group, while the blue area shows those more prevalent in the control group. This means that most of the genes in this set are highly expressed or prevalent in the control samples compared to the disease samples
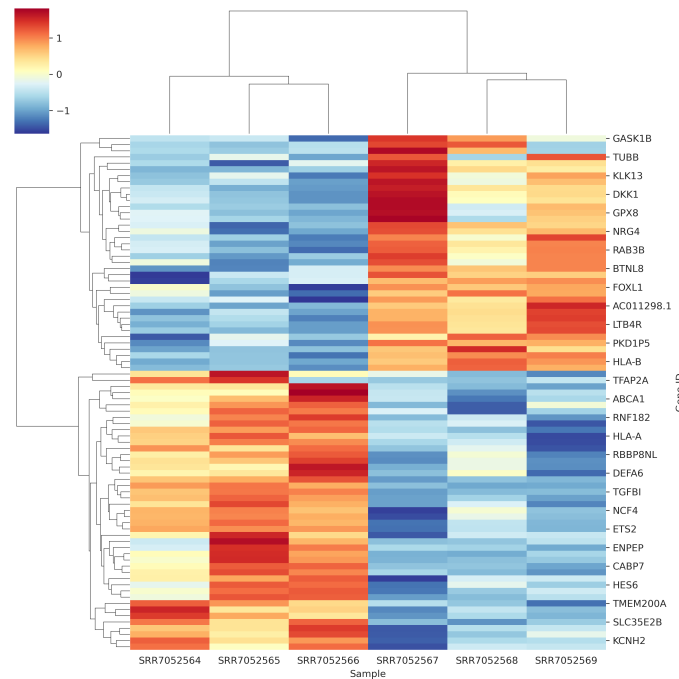

Figure 2: Heatmap representing RNA-seq expression values for the genes differentially expressed in organoids from patients with active celiac disease (CD, n = 3) compared to non-celiac controls (NC, n = 3). A color code from blue to red indicates low and high expression levels, respectively.

Interesting is also the overexpression of TGFBI with a high level of significance, and it has been suggested that the role of TGF-β in Th17 differentiation is the suppression of Th1 and Th2 differentiation (i.e. suppression of the production of IFN-γ and IL-4), since these cyokines strongly inhibit Th17 differentiation, but according to a report, RORγt, the master regulator of Th17, was induced by TGF-β alone even in the absence of IL-6[38].
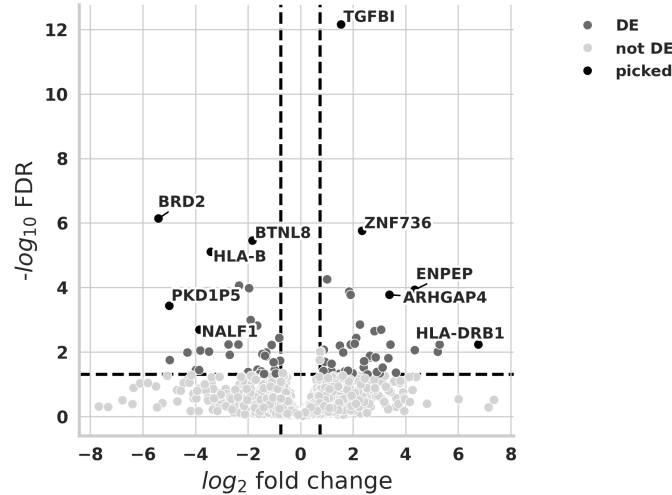

Figure 3: Volcano plot showing fold change (X axis in log_2 scale) and statistical significance (FDR, Y axis in $-log_{10}$ scale) for differentially expressed genes (RNA-seq) in active celiac (CD, n = 3) compared to non-celiac (n = 3) organoids

Another suspect and overexpressed gene, visible in the volcano plot, is HLA-DRB1. HLA-DRB1 gene alleles are often linked to these celiac-associated HLA-DQ aplotypes such as HLA-DQ2 and HLA-DQ8. In a report on tens of Lybian children suffering both CD and type 1 diabetes mellitus, 70% carried the HLA-DQ2 linkage with HLA-DRB1 alleles[39]. Human leukocyte antigen (HLA) region is located on chromosome 6. The region consists of genes encoding molecules responsible for regulating immune response. HLA molecules are cell surface–bound glycoproteins classified into three classes. HLA class II is expressed on the surface of antigen-presenting cells (including macrophages, B cells and dendritic cells) and is essential in order to display peptides to T-helper CD4+ cells, inducing their activation. HLA class II antigens are mainly encoded by DR, DQ[40].

5   Discussion

TFF1 contributes to the protection of the gastric mucosa by enhancing the stability and viscosity of the mucous layer, which acts as a barrier against gastric acid and pathogens[41]. TFF2, also known as spasmolytic polypeptide, is predominantly found in the stomach, secreted by the mucous neck cells. TFF2 plays a significant role in protecting the gastric lining and aiding in its repair after injury. It is involved in maintaining the integrity of the mucosal barrier and supports epithelial restitution[42].TFF2 is expressed in immune cells and has been implicated in modulating immune responses, particularly in the context of gastrointestinal

inflammation[43]; in addition to that, it forms complexes with mucin MUC6, contributing to the stabilization of the inner layer of gastric mucus, which is crucial for protecting the stomach lining from digestive enzymes and pathogens[43]. Defensin, alpha 6 (DEFA6) also known as human alpha defensin 6 (HD6) affords protection against invasion by enteric bacterial pathogens by self-assembly to form fibrils and nanonets that surround and entangle bacteria[44]. NLRP6, initiator of the inflammasome complex formation is overexpressed, as mentioned before. In celiac disease, the NLRP6 inflammasome is expressed in intestinal epithelial cells. This expression is associated with the production of caspase-1 and biologically active interleukin-18, which can further induce interferon-γ in intraepithelial lymphocytes[45], potentially exacerbating inflammation and tissue damage. Intestinal epithelial cells (IECs) produce inflammasome cytokines in response to noxious agents and selected microorganisms. The inflammasome cytokine IL-18, also known as IFN-γ-inducing factor, is most likely the inducer. It is expressed at very high levels in IECs in active CD[45]. PTPRK transcription is directly regulated by TGF-β through binding of Smad3/4 to the proximal promoter region of the PTPRK gene, so TGF-β induce PTPRK gene expression[46]. And one of the roles of PTPRK (Protein Tyrosine Phosphatase Receptor Type K), a nodal phosphatase is the regulation of EGFR activation in the proliferation of the enterocytes from CD biopsies and organoids. The overexpression of PTPRK in CD organoids is able to reduce pEGFR, pERK and proliferation. The activation of the EGFR/ERK pathway is linked to increased proliferation of enterocytes[47]. This pathway is crucial in CD, where enterocyte proliferation is dysregulated, leading to crypt hyperplasia and villous atrophy. TGF-βs signal via a pair of transmembrane serine/threonine protein kinases, known as the type I (TGF-βR1 or ALK5) and type II receptors (TGF-βR2). Binding of one of these cytokines to a TGF-βR2 dimer triggers the recruitment and subsequent transphosphorylation of TGF-βR1 which, in turn, activates the downstream canonical signaling cascade via the receptor-regulated SMADs (R-SMADs). Upon receptor activation, phosphorylation of SMADs causes their dissociation from SARA (anchor for receptor activation) and formation of SMAD2/3–SMAD4 complexes[48]. BMP subfamily receptors phosphorylate SMAD1, SMAD5 and SMAD8, while TGF-β subfamily receptors mainly phosphorylate SMAD2 and SMAD3. Upon phosphorylation, R-SMADs dissociate from the receptor and form a heterotrimeric transcriptional complex with SMAD4. Heterotrimers of SMAD2, SMAD3 and SMAD4 are formed and they translocate and accumulate in the nucleus where they bind DNA directly but with low affinity and specificity or indirectly by interacting with other DNA-binding protein. In order to contain the expression levels, SKI factor inhibits the expression of TGF-β target genes by binding Smad Binding Elements (SBEs) complexed with SMAD4 and by recruiting other repressors and histone deacetylases (HDA)[49]. Bi-allelic LOF (loss of function) variants in TGFB1 were found in patients with inflammatory bowel disease (IBD), ultimately resulting in decreased downstream SMAD2/3 signaling, causing immunological abnormalities and fatal infectious complications[50]. So SMAD2/3 signaling could have a role in containing the inflammation process in CD, especially considering how this complex leads PTPRK transcription. On the other hand though, TGF-β is responsible for Th17 differentiation and and consequent activation of the Th17-driven immune response in CD[51]. Meanwhile, the overexpression of TGF-β contrasts the effects of the overexpression of NLRP6, responsible for the production of IFN-γ, which inhibit Th17 differentiation. But the net effect in this contrast is likely the Th17 differentiation, with the related immune processes. The downexpression of

both TFF1 and TFF2 is pointing to a compromised gastric mucus, which does not protect the stomach anymore. This could lead to a disbiosis that extends itself to the small intestine area, generating overall abdominal pain in CD patients, caused by digestive enzymes attacking the stomach, and by bacteria attacking both the stomach and the small intestine. The overepression of DEFA6 could also point out the struggle in the organism to contain the proliferation of opportunistic bacteria, that find favourable conditions at the expense of other groups of bacteria. The overexpression of CD36 with its negative impact on angiogenesis through the binding with thrombospondin-1 (TSP-1), could justify the typical emaciated aspect of CD patients, if it is summed up with crypt hyperplasia and villous blunting. This scenario imply a generally diminished absorption of nutrients. The other main culprit behind CD is likely to be HLA-DRB1, with its overexpression. Considering how highly polymorphic this locus can be, with 2690 distinct alleles, encoding 1899 proteins[40], and being its alleles associated with CD risk aplotypes, this gene could indirectly contribute to both the disbiosis and the autoimmunity. There could be multiple reasons behind this hypothesis: shared epitope (SEP) hypothesis, CD4+ T cell dependent somatic hypermutation. SE is a 30 years old hypothesis of a pathogenic role of three amino acid sequences (70QRRAA74, 70RRRAA74 or 70QKRAA74) located at positions 70–74, i.e., within the HVR3 of the DRβ1 chain. It was hypothesized that the presence of these SE sequences allows the presentation of self-antigens to T lymphocytes, and thus plays a key role in the development of Rheumatoid Arthritis (RA)[52]. Based on this hypothesis, there could be other sequences, that conversely allow presentation of self-antigens to T lymphocytes, playing a role in CD. In addition, the production of anti-citrullinated protein antibodies (ACPA) is driven by CD4+ T cell dependent somatic hypermutation. This process occurs during B-cell activation, and in the case of ACPA, this hypermutation process may also result in the formation of glycosylation sites that enhance the pathogenicity of these antibodies in RA[40]. The same mechanism could happen for CD in the case of HLA-DRB1 presenting gluten-derived peptides to T cells, thus exacerbating the autoimmune response. So there could be more parallelisms between RA and CD than we could expect. Regarding the disbiosis contribution and the role of HLA-DRB1 in the immune response to bacterial infections, according to a report, the association of poststreptococcal reactive arthritis (ReA) with HLA-DRBl*O1, but not with HLA-B27, would gest that its pathogenesis may be more similar to that of rheumatic fever than to that of ReA associated with enteric pathogens[53]. But the main point is that there must be a deeper connection between gut diseases like CD and RA or even ReA. According to the latest estimates, the prevalence rate of CD in RA patients could approximately 3%, which is three times the rate in the healthy population[54], suggesting a possible shared autoimmune mechanism.

6   References

[1]   Seungil Kim et al. "Comparison of cell and organoid-level analysis of patient-derived 3D organoids to evaluate tumor cell growth dynamics and drug response". In: *SLAS DISCOVERY: Advancing the Science of Drug Discovery* 25.7 (2020), pp. 744–754.

[2]   Anna Chiara Piscaglia et al. "Circulating hematopoietic stem cells and putative intestinal stem cells in coeliac disease". In: *Journal of translational medicine* 13 (2015), pp. 1–9.

[3] DM Shalimar et al. "Mechanism of villous atrophy in celiac disease: role of apoptosis and epithelial regeneration". In: *Archives of Pathology and Laboratory Medicine* 137.9 (2013), pp. 1262–1269.

[4] Maria Vittoria Barone et al. "Pivotal role of inflammation in celiac disease". In: *International Journal of Molecular Sciences* 23.13 (2022), p. 7177.

[5] Monia Porpora et al. "Inflammation is present, persistent and more sensitive to proinflammatory triggers in celiac disease enterocytes". In: *International Journal of Molecular Sciences* 23.4 (2022), p. 1973.

[6] Anne M Minihane et al. "Low-grade inflammation, diet composition and health: current research evidence and its translation". In: *British Journal of Nutrition* 114.7 (2015), pp. 999–1012.

[7] Ludvig M Sollid and Bana Jabri. "Triggers and drivers of autoimmunity: lessons from coeliac disease". In: *Nature Reviews Immunology* 13.4 (2013), pp. 294–302.

[8] Maureen M Leonard et al. "Celiac disease and nonceliac gluten sensitivity: a review". In: *Jama* 318.7 (2017), pp. 647–656.

[9] Jorma Isola et al. "The oral transglutaminase 2 inhibitor ZED1227 accumulates in the villous enterocytes in celiac disease patients during gluten challenge and drug treatment". In: *International journal of molecular sciences* 24.13 (2023), p. 10815.

[10] Steffen Husby et al. "European Society for Pediatric Gastroenterology, Hepatology, and Nutrition guidelines for the diagnosis of coeliac disease". In: *Journal of pediatric gastroenterology and nutrition* 54.1 (2012), pp. 136–160.

[11] Rasmus Iversen et al. "Evidence that pathogenic transglutaminase 2 in celiac disease derives from enterocytes". In: *Gastroenterology* 159.2 (2020), pp. 788–790.

[12] Nobuyo Maeda et al. "Duplication within the haptoglobin Hp 2 gene". In: *Nature* 309.5964 (1984), pp. 131–135.

[13] Karen M Lammers et al. "Gliadin induces an increase in intestinal permeability and zonulin release by binding to the chemokine receptor CXCR3". In: *Gastroenterology* 135.1 (2008), pp. 194–204.

[14] Sandro Drago et al. "Gliadin, zonulin and gut permeability: Effects on celiac and non-celiac intestinal mucosa and intestinal cell lines". In: *Scandinavian journal of gastroenterology* 41.4 (2006), pp. 408–419.

[15] Justin Hollon et al. "Effect of gliadin on permeability of intestinal biopsy explants from celiac disease patients and patients with non-celiac gluten sensitivity". In: *Nutrients* 7.3 (2015), pp. 1565–1576.

[16] Alessio Fasano. "Zonulin and its regulation of intestinal barrier function: the biological door to inflammation, autoimmunity, and cancer". In: *Physiological reviews* (2011).

[17] Karen M Lammers et al. "Identification of a novel immunomodulatory gliadin peptide that causes interleukin-8 release in a chemokine receptor CXCR3-dependent manner only in patients with coeliac disease". In: *Immunology* 132.3 (2011), pp. 432–440.

[18] Antonio Di Sabatino et al. "Epithelium derived interleukin 15 regulates intraepithelial lymphocyte Th1 cytokine production, cytotoxicity, and survival in coeliac disease". In: *Gut* 55.4 (2006), pp. 469–477.

[19] Gloria Serena et al. "The role of gluten in celiac disease and type 1 diabetes". In: *Nutrients* 7.9 (2015), pp. 7143–7162.

[20] David Goldberg. "What every computer scientist should know about floating-point arithmetic". In: *ACM computing surveys (CSUR)* 23.1 (1991), pp. 5–48.

[21] URL: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA451302&o=acc_s%3Aa.

[22] URL: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/trim_galore_User_Guide_v0.4.1.pdf.

[23] URL: https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf.

[24] Evguenia Kopylova, Laurent Noé, and Hélène Touzet. "SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data". In: *Bioinformatics* 28.24 (2012), pp. 3211–3217.

[25] URL: https://github.com/DavideBiondi/fmp_masteromics_thesis.

[26] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[27] Udi Manber and Gene Myers. "Suffix arrays: a new method for on-line string searches". In: *siam Journal on Computing* 22.5 (1993), pp. 935–948.

[28] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python framework to work with high-throughput sequencing data". In: *bioinformatics* 31.2 (2015), pp. 166–169.

[29] URL: https://htseq.readthedocs.io/en/latest/htseqcount.html.

[30] Boris Muzellec et al. "PyDESeq2: a python package for bulk RNA-seq differential expression analysis". In: *Bioinformatics* 39.9 (2023), btad547.

[31] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. "GSEApy: a comprehensive package for performing gene set enrichment analysis in Python". In: *Bioinformatics* 39.1 (2023), btac757.

[32] Da Wei Huang et al. "The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists". In: *Genome biology* 8 (2007), pp. 1–16.

[33] Da Wei Huang et al. "DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists". In: *Nucleic acids research* 35.suppl_2 (2007), W169–W175.

[34] Paolo Fedi et al. "Isolation and biochemical characterization of the human Dkk-1 homologue, a novel inhibitor of mammalian Wnt signaling". In: *Journal of biological Chemistry* 274.27 (1999), pp. 19465–19472.

[35] Maayan Levy et al. "NLRP6: a multifaceted innate immune sensor". In: *Trends in Immunology* 38.4 (2017), pp. 248–260.

[36] Sandra L Guerrero-Rodríguez et al. "Role of CD36 in cancer progression, stemness, and targeting". In: *Frontiers in Cell and Developmental Biology* 10 (2022), p. 1079076.

[37] William W Feng, Hannah T Zuppe, and Manabu Kurokawa. "The role of CD36 in cancer progression and its value as a therapeutic target". In: *Cells* 12.12 (2023), p. 1605.

[38] Kenji Ichiyama et al. "Foxp3 Inhibits RORγt-mediated IL-17A mRNA Transcription through Direct Interaction with RORγt*". In: *Journal of Biological Chemistry* 283.25 (2008), pp. 17003–17008.

[39] Millad Ghawil et al. "HLA-DQ types of celiac disease in Libyan children with type 1 diabetes mellitus". In: *European journal of gastroenterology & hepatology* 24.1 (2012), pp. 59–63.

[40] Tomasz Wysocki, Marzena Olesińska, and Agnieszka Paradowska-Gorycka. "Current understanding of an emerging role of HLA-DRB1 gene in rheumatoid arthritis–from research to clinical practice". In: *Cells* 9.5 (2020), p. 1127.

[41] Nayara Braga Emidio et al. "Structure, function, and therapeutic potential of the trefoil factor family in the gastrointestinal tract". In: *ACS Pharmacology & Translational Science* 3.4 (2020), pp. 583–597.

[42] Eitaro Aihara, Kristen A Engevik, and Marshall H Montrose. "Trefoil factor peptides and gastrointestinal function". In: *Annual review of physiology* 79.1 (2017), pp. 357–380.

[43] Werner Hoffmann. "Trefoil factor family (Tff) peptides and their links to inflammation: A re-evaluation and new medical perspectives". In: *International journal of molecular sciences* 22.9 (2021), p. 4909.

[44] Hiutung Chu et al. "Human α-defensin 6 promotes mucosal innate immunity through self-assembled peptide nanonets". In: *Science* 337.6093 (2012), pp. 477–481.

[45] Grzegorz Pietz et al. "Immunopathology of childhood celiac disease—Key role of intestinal epithelial cells". In: *PLoS One* 12.9 (2017), e0185025.

[46] Yiru Xu et al. "Notch and TGF-β pathways cooperatively regulate receptor protein tyrosine phosphatase-κ (PTPRK) gene expression in human primary keratinocytes". In: *Molecular biology of the cell* 26.6 (2015), pp. 1199–1206.

[47] Merlin Nanayakkara et al. "PTPRK, an EGFR phosphatase, is decreased in CeD biopsies and intestinal organoids". In: *Cells* 12.1 (2022), p. 115.

[48] Tomoo Tsukazaki et al. "SARA, a FYVE domain protein that recruits Smad2 to the TGFβ receptor". In: *Cell* 95.6 (1998), pp. 779–791.

[49] Angeles C Tecalco-Cruz et al. "Transcriptional cofactors Ski and SnoN are major regulators of the TGF-β/Smad signaling pathway in health and disease". In: *Signal transduction and targeted therapy* 3.1 (2018), p. 15.

[50] Daniel Kotlarz et al. "Human TGF-β1 deficiency causes severe inflammatory bowel disease and encephalopathy". In: *Nature genetics* 50.3 (2018), pp. 344–348.

[51] G Serena et al. "Proinflammatory cytokine interferon-γ and microbiome-derived metabolites dictate epigenetic switch between forkhead box protein 3 isoforms in coeliac disease". In: *Clinical & Experimental Immunology* 187.3 (2017), pp. 490–506.

[52] Peter K Gregersen, Jack Silver, Robert J Winchester, et al. "The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis". In: *Arthritis & Rheumatism* 30.11 (1987), pp. 1205–1213.

[53] Stanford T Shulman and Elia M Ayoub. "Poststreptococcal reactive arthritis". In: *Current opinion in rheumatology* 14.5 (2002), pp. 562–565.

[54] Elham Elhami et al. "Prevalence of celiac disease in Iranian patients with rheumatologic disorders". In: *Gastroenterology and Hepatology From bed to Bench* 11.3 (2018), p. 239.