University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Single-Cell Transcriptomics of Intestinal Epithelial Cells: Insights into the Prediabetic condition

Glenn Ross-Dolan

B.A. (Mod) Microbiology - Trinity College Dublin

Project Thesis in partial fulfilment for the degree of Masters in Bioinformatics and Computational Biology

Supervised by Dr. Silvia Melgar, APC Microbiome Ireland

Table of Contents

List of abbreviations

Acknowledgements

Abstract

Introduction

Type 2 Diabetes Mellitus (T2DM) has emerged as a global health crisis, with its prevalence quadrupling over the past three decades (Zheng et al., 2018). Characterised by insulin resistance and inadequate insulin secretion leading to hyperglycemia, T2DM is often preceded by a condition known as prediabetes. This precursor state is marked by impaired fasting glucose (IFG), impaired glucose tolerance (IGT), or raised HbA1c levels (5.7 – 6.4%), and is closely associated with obesity, diet, sedentary lifestyle, and genetic factors (American Diabetes Association, 2021).

While the systemic effects of T2DM and prediabetes are well-documented, recent evidence has highlighted the role of the intestinal epithelium in the development and progression of prediabetes. The intestinal epithelium, serving as the primary interface between the diet and internal biological systems, performs important functions in nutrient absorption, barrier protection, and hormone secretion. Alterations in the structure and function of this epithelium have been observed in prediabetic individuals, with evidence suggesting a dysregulation of the intestinal barrier and metabolism (Aliluev et al., 2021; Xie et al., 2020). Key areas of interest in studying the intestinal epithelium in the context of prediabetes include changes in the gut microbiota composition, intestinal permeability, inflammation, macronutrient metabolism alterations, and intestinal stem cell (ISC) functions and are reviewed here in detail. Understanding these intricate relationships between the intestinal epithelium and metabolic dysfunction requires advanced research techniques which can capture the complexity of cellular responses at a high resolution.

Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool to address this need. This technology allows for the characterisation of gene expression profiles at the individual cell level, enabling researchers to uncover cellular heterogeneity, identify rare cell populations, and reveal cell-specific, transcriptional responses to physiological changes.

In light of these technological advancements and the growing recognition of the intestinal epithelium's role in metabolic health, this research project aims to employ scRNA-seq analysis to identify genes and signalling pathways within the intestinal epithelium that are characteristic of the prediabetic state. Using a high-fat, high-sugar diet (HFHSD) mouse model to study diet-induced prediabetes, this study seeks to characterise the

transcriptional profiles of individual cell types within the intestinal epithelium in both normal and prediabetic states. By identifying differentially expressed genes and altered signaling pathways associated with prediabetes, we aim to reveal the link between prediabetes-induced changes in the intestinal epithelium and alterations in metabolic function.

The significance of this research lies in its potential to advance our understanding of the molecular basis of prediabetes, particularly in relation to the intestinal epithelium, a significantly underresearched area. By providing a high-resolution map of transcriptional changes in individual cell types, this study aims to reveal novel mechanisms contributing to prediabetes progression in the intestinal epithelium and potentially identify new targets for therapeutic interventions.

Chapter 1: The Intimate Link Between The Intestinal Epithelium And Prediabetes.

1.1 Intesintal Barrier Structure and Function

The intestinal barrier is a dynamic system of several specialised components regulating the absorpotion of nutrients while preventing the entry of harmful substances and microorganisms. Understanding the structure and function of the intestinal barrier is essential for comprehending its role in health and disease, particularly in the context of metabolic disorders such as prediabetes. This section explores the key components of the intestinal barrier, including the gut microbiota, mucus layer, intestinal epithelium, immune barrier as well as factors that can modify barrier function. Furthermore, dysfunctions of the intestinal epithelium in the prediabetic model are discussed in detail.

1.1.1 Microbiota

The intestinal microbiome forms the initial component on the gut barrier consisting of four main phyla Proteobacteria, Bacteroidetes, Actinobacteria, Firmicutes, with Bacteroides and firmicutes totalling to approximately 90% of the total gut microbiota (Rajilić-Stojanović et al., 2007; Vos et al., 2022). The density of these microorganisms increase along the gastrointestnial tract reaching its peak in the colon (McGhee and Fujihashi, 2012). The gut microbiota also plays a crucial role in nutrient metabolism. The microbiota derives its nutrients mostly through carbohydrates in the diet but also lipids, proteins, vitamins and various phenolic compounds. Carbohydrates are primarily used as an energy source by

the microbiota and are primarily metabolised by members of the genus Bacteroides by expressing various carbohydrate digesting enzymes (Jandhyala et al., 2015). Nutrients which are not digestible by the host such as fibers make their way to the colon and are metabolised by the microbiota into a wide array of metabolites such as short-chain fatty acids (SCFAs), a key metabolite which plays an important role in maintaining health and disease displaying roles in inducing reactive oxygen species, altering cell proliferation and function, antinflammatory, antitumorigenic and antimicrobial effect (Tan et al., 2014). SCFAs are also used directly as an energy source by enterocytes in the colon or are transported across the epithelial layer into the blood (Tan et al., 2014). Interactions between the host and the microbiome significantly impacts the maturation of the immune system. Literature indicates that different bacterial species can trigger distinct immune responses, suggesting that microbiota composition significantly influences immunity. The microbiota's impact extends beyond the gut, affecting systemic immune function and influencing disease processes in various organs. Depending on the bacterial species involved, these alterations in the microbiota composition can range from disease promotion to protection and is implicated in the progression of prediabetes and T2DM discussed in more detail later (Kosiewicz et al., 2011).

1.1.2 Mucus Layer

The mucus layer forms the second element of the intestinal barrier, facilitated by a layer of mucins. These are highly O-glycosylated proteins with gel-like properties secreted by goblet cells (Kim and Ho, 2010). This layer is responsible for a unique role in maintaining the intestinal lining from pathogens and mechanical damage (Johansson and Hansson, 2016). The mucus layer of the small intestine is penetrable to microbes however the microbiota are kept at a distance from the intestinal epithelium through antimicrobial metabolites (Johansson and Hansson, 2016). Lysozyme, secretory IgA, and defensins are secreted by Paneth cells are found in the inner mucus layer, providing a mechanism that helps keep bacteria away from the surface of enterocytes (Portincasa et al., 2021; Gibbins et al., 2015). Other antimicrobial molecules, such as REG3G, lypd8, and ZG16, further contribute to bacterial exclusion from the mucosa (Portincasa et al., 2021; Bergström et

al., 2016). Contrastingly, in the large intestine, the mucus layer is stratified into two main layers. The inner layer is impenetrable to the microbiota composed largely of Muc2 multimers whereas the outer layer provides a comfortable penetrable environment for the microorganisms (Johansson and Hansson, 2016). The microbiota influences mucin composition and structure through various mechanisms, including the ratio of Bacteroides and Firmicutes (Wrzosek et al., 2013). Dietary factors, such as fiber content, also impact mucus thickness and the abundance of mucin-degrading bacteria (Desai et al., 2016). Certain bacteria, like Akkermansia muciniphila, play a crucial role in mucus degradation and modulation of inflammatory changes and crosstalk between them and the intestinal epithelium has been shown to modulate obesity, a key risk factor of prediabetes (Everard et al., 2013).

1.1.3 Intestinal Epithelium

The intestinal epithelium forms the main component of the intestinal barrier, consisting of a single layer of cells having roles in nutrient acquisition and thus roles in protection from the external environment. The structure of the epithelium differs between the small and large intestine. The small intestine contains structural protrusions, increasing the surface area for nutrient absorption while the large intestine is flat, reducing the potential for damage from more solid material (Allaire et al., 2018). Furthermore, both the small and large intestine contain invaginations called ‘crypts’, harbouring proliferative intestinal stem cells (ISCs) important in the constant turnover of new intestinal epithelial cells (Allaire et al., 2018). The epithelial layer comprises six primary cell types, with each contributing to these roles. Enterocytes, the most abundant cell type, form the main absorptive surface and play a direct role in the immune response (Snoeck et al., 2005). ISCs as mentioned are involved in regenerating and producing new cells. Goblet cells, responsible for mucus production, contribute to the mucus layer as discussed previously. Tuft cells remain elusive in their functions although exhibit chemosensory functions and secrete effector molecules involved in innate immunity and play an important role in barrier maintenance (Silverman et al., 2024). Enteroendocrine cells are diverse secretory cells which produce hormones regulating nutrient absorption, intestinal barrier function and ISC homeostasis (Nwako and McCauley, 2024). Paneth cells, located in the intestinal crypts produce antimicrobial peptides and proteins to mediate host-microbe interactions and innate immunity, as well as factors that help sustain and modulate the ISCs (Clevers and Bevins, 2013).

The epithelial cells are interconnected by junctional complexes, including tight junctions (Tjs), adherens junctions (Ajs), and desmosomes. TJs are found at the top of the junction, establishes polarity and regulates permeability (Lessey et al., 2022). They comprise over 40 proteins, including claudins, occludin, and zonula occludens proteins (Portincasa et al., 2021). Directly below the TJs are the AJs playing a crucial role in cell-cell adhesion. Below the AJs are the desmosomes which strengthen the adhesion while withstanding mechanical stress (Lessey et al., 2022).

The epithelial barrier allows for both transcellular and paracellular transport of molecules. The paracellular and transcellular routes permit the passage of water, macromolecules, small hydrophilic compounds, lipids, and ions (Lessey et al., 2022). The permeability of this barrier varies along the intestinal tract, with the colonic epithelium being less permeable than the small intestinal epithelium (Portincasa et al., 2021). Various factors can influence junction integrity, including diet, microbiota composition and inflammation which are discussed in detail later. Impaired intestinal barrier function has been observed in animal models of obesity, prediabetes and T2DM, and inflammatory bowel diseases.

The intestinal barrier's integrity and function can be significantly influenced by many factors, most notably, diet, microbiota, physical acitvity, and medication. These modifiers play crucial roles in maintaining or altering the intestinal barrier, potentially having roles in intestinal diseases and dysfunctions such as prediabetes.

1.2 Intestinal Epithelium Alterations In Prediabetes

1.2.1 Microbiota Alterations

Prediabetes is associated with significant alterations in the gut microbiome, which may contribute to the progression towards type 2 diabetes mellitus (T2DM). These changes in microbial composition and diversity are increasingly recognised as potential factors in the development of metabolic disorders. Studies have consistently reported a reduction in microbial diversity and richness in individuals with prediabetes, mirroring observations in patients with established diabetes (Chang et al., 2024). This decreased diversity may

compromise the beneficial functions of the gut microbiome, potentially contributing to metabolic dysregulation.

Several bacterial genera have been found to be differentially abundant in prediabetic individuals compared to those with normal glucose metabolism. Notably, studies have reported lower abundances of Bifidobacterium, Blautia, Clostridium, Faecalibacterium, Mediterraneibacter, Anaerostipes, and Butyricicoccus in prediabetic stool samples (Chang et al., 2024). These bacteria are known to play important roles in maintaining intestinal health, including the production of short-chain fatty acids (SCFAs) and the maintenance of gut barrier integrity.

Particularly noteworthy is the reduced abundance of Akkermansia muciniphila in individuals with prediabetes (Rathi et al., 2023). A. muciniphila has been associated with improved metabolic health, and its depletion may contribute to increased intestinal permeability and metabolic disturbances. Experimental studies have shown that oral administration of A. muciniphila can improve glucose intolerance and insulin resistance in animal models, possibly through Toll-like receptor 2 signaling (Everard et al., 2013; Shin et al., 2014; Plovier et al., 2017; Rathi et al., 2023).

Conversely, some bacterial genera have been found to be more abundant in prediabetic individuals. These include Ruminococcus, Dorea, Streptococcus, Sutterella as well as facultative anaerobes (Rathi et al., 2023; Piccolo et al., 2024). Additionally, increased abundances of Bacteroides, Parabacteroides, Phascolarctobacterium, and Paraprevotella have been observed in prediabetic fecal samples (Chang et al., 2024). The functional implications of these increases are not fully understood however and require further investigation.

It's important to note that while these microbial alterations are consistently observed in prediabetic individuals, the causal relationship between gut dysbiosis and prediabetes development remains to be fully understood. Factors such as diet, which accounts for nearly 60% of gut microbiota composition, play a significant role in shaping the microbial community (Zhang et al., 2010). This underscores the potential for dietary interventions or

supplementation in modulating the gut microbiome and, potentially, in preventing or managing prediabetes. Future research should focus on understanding the functional consequences of these microbial alterations and their specific contributions to the development and progression of prediabetes. Additionally, investigating the potential of microbiome-based interventions, such as targeted probiotic therapies or dietary modifications, may offer new avenues for preventing or managing prediabetes and thus its progression to T2DM.

1.2.2 Intestinal Barrier Permeability

Alterations in intestinal permeability play a crucial role in the pathogenesis of prediabetes and its progression to type 2 diabetes mellitus (T2DM). As previously discussed, the intestinal barrier and its various components are integral in regulating the passage of substances into the body. In prediabetic conditions, several studies have reported increased intestinal permeability, often referred to as "leaky gut". This increased permeability is associated with alterations in the structure and function of tight junctions, critical components of the paracellular barrier (Nascimento et al., 2021).

Olivera et al. have shown that high-fat diet (HFD) intake, often associated with prediabetes, can lead to significant changes in intestinal permeability. In vitro models using Caco-2 cell lines have demonstrated that exposure to intestinal content from the small intestine of mice fed a HFD, can disrupt the tight junction-mediated epithelial barrier in cell culture models (Oliveira et al., 2019). This suggests that an element of the intestinal lumen in HFD conditions may directly impact barrier integrity. In animal models of prediabetes, structural changes in tight junctions have been observed in various segments of the intestine. Notably, these alterations occur early in the development of prediabetes, often preceding major metabolic changes. The duodenum and jejunum appear to be particularly affected, with significant reductions in the junctional content of tight junction proteins (Nascimento et al., 2021). These findings highlight the potential importance of intestinal barrier dysfunction as an early event in the pathogenesis of prediabetes and metabolic disorders associated with high-fat diets. The prevailing theory suggests that alterations in the microbiota via a high-fat diet induces this increase in intestinal permeability and potentially leads to the translocation of bacteria and antigens leading to diabetic like

disturbances such as insulin resistance (de Kort et al., 2011; Matheus et al., 2017). It's important to note that while increased intestinal permeability is consistently observed in prediabetic conditions, the exact mechanisms linking this phenomenon to the development and progression of metabolic disorders are still being clarified. Factors such as diet composition, microbiota alterations, lifestyle factors and genetic susceptibility likely interact in complex ways to influence and progress the metabolic condition.

Future research should focus on further characterising the molecular and cellular changes in the intestinal barrier during the progression from normal glucose tolerance to prediabetes and T2DM. Additionally, investigating potential therapeutic interventions targeting intestinal permeability may offer new strategies for preventing or managing prediabetes and its associated complications.

1.2.3 Inflammation

Inflammation plays a crucial role in the pathogenesis of prediabetes and its progression to T2DM in the larger systemic context. Systemic inflammation in prediabetes is characterised by elevated levels of inflammatory markers and alterations in immune function (Weaver et al., 2021). Studies have reported increased levels of inflammatory proteins in prediabetic patients, including interleukin-6, interleukin-1β, tumor necrosis factor-α, monocyte chemoattractant protein-1, resistin, and C-reactive protein (CRP). The ratio of CRP to albumin is also elevated, indicating a shift towards a pro-inflammatory state (Colloca et al., 2024). Hypoteheses based on protein and gene analysis of pancreatic tissues and isolated islets suggest that inflammation in prediabetes may be initiated by a decrease in CD163+ cells leading to reduced anti-inflammatory protection and thus increased production of pro-inflammatory cytokines and resistin (Weaver et al., 2021). There are significant implications for inflammation in prediabetes. Inflammation during the prediabetic state seems to be a driving force behind pancreatic beta cell dysfunction and insulin resistance, dyslipidemia, and cardiovascular diseases and is a risk factor for peripheral cascular diseases (Saghir et al., 2023). Initiation of inflammation in the prediabetic subject is not fully understood, although research is suggesting that it may begin in the intestines.

In the context of intestinal inflammation in prediabetes, recent hypotheses are focussed on the relationship between diet, gut microbiota, and intestinal barrier function. One prevalent hypothesis suggests that high-fat diet intake leads to alterations in intestinal microbiota composition. These alterations are thought to increase paracellular permeability and absorption of LPS, dietary antigens, and translocation of bacteria leading to metabolic endotoxemia and low-grade chronic systemic inflammation, which may trigger or exacerbate peripheral insulin resistance (Geurts et al., 2014; Gomes et al., 2017; Nascimento et al., 2021).

However, conflicting evidence exists in the literature regarding the relationship between intestinal permeability, inflammation, and prediabetes development. While some studies report significant increases in intestinal permeability to large molecules, associated with endotoxemia and systemic inflammation, other research has found that increased intestinal TJ permeability in prediabetic mice occurs without significant changes in systemic and intestinal levels of zonulin, TNF-α, and LPS (Nascimento et al., 2021). These discrepancies may be partially explained by differences in prediaebetic models, including variations in animal strains, and diet composition. For instance, studies using diets with very high fat content (e.g., 72% of energy from lipids) have observed significant metabolic and intestinal changes, including metabolic endotoxemia and increased cecal LPS levels, after relatively short exposure periods [ref]. In contrast, studies using more moderate high-fat diets (e.g., 40% of energy from lipids) found that animals became prediabetic after longer periods without significant changes in microbiota composition or luminal LPS levels (Cani et al., 2008; Nascimento et al., 2021). Consistent with this are reports demonstrating that isocaloric diets can have varying diabetogenic and obesogenic effects based on their macronutrient composition, particularly the type of fat (polyunsaturated vs. saturated) and the presence of fructose, rather than just total calorie content (Deol et al., 2015).

Further research is needed to reconcile these conflicting observations and elucidate the specific mechanisms linking intestinal barrier dysfunction to systemic inflammation and metabolic disturbances in prediabetes. Future studies should focus on characterising the temporal relationship between intestinal epithelial alterations, local and systemic inflammatory responses, and metabolic changes in the context of prediabetes

development. Additionally, investigating the role of specific intestinal luminal components and intracellular signaling pathways in regulating TJ structure and function may provide valuable insights into the pathogenesis of prediabetes and potential therapeutic targets.

1.2.4 Macronutrient Metabolism Alterations

The intestinal epithelium undergoes significant changes in macronutrient metabolism during the development of prediabetes, particularly in response to high-fat diets (HFDs) and high-fat high-sugar diets (HFHSDs). These alterations affect lipid, carbohydrate, and amino acid metabolism, contributing to the progression of metabolic dysfunction.

HFDs significantly impact the expression of intestinal genes involved in fatty acid metabolism. A notable example is the Scd1 gene, which converts saturated fatty acids to monounsaturated fatty acids and is upregulated more than tenfold in the jejunum by coconut oil (Martinez-Lomeli et al., 2023). Other affected genes include those involved in linoleic acid and arachidonic acid metabolism such as Cyp2c, Cyp2j, Cyp4a, and Ephx2, which are further associated with pro-inflammatory processes (Martinez-Lomeli et al., 2023). In prediabetic conditions, lipid metabolism pathways are vastly altered. Proteins involved in mitochondrial β-oxidation and peroxisome β-oxidation are upregulated in gut-derived extracellular vesicles (GDEs) from the small intestines of HFD-fed prediabetic mice (Ferreira et al., 2022). This suggests a shift towards fatty acids as a preferred energy source over glucose. Furthermore, a family of acyl-CoA thioesterases (ACOTs) is significantly upregulated, acting as intermediaries in directing fatty acids to either the TCA cycle or storage (Ferreira et al., 2022). Prediabetic mice fed HFHS diets also exhibit increased numbers of enterocytes specialised in carbohydrate and fatty acid absorption suggesting an increase in calorie intake as well as fat accumulation facilitated by an increased expression of the fatty acid binding protein Fabp1 (Aliluev et al., 2021).

Carbohydrate metabolism in the intestinal epithelium is also significantly altered in prediabetes. Enrichment analysis of proteins in GDEs from the small intestines of HFD-fed prediabetic mice show alterations in pyruvate and glycolysis-gluconeogenesis pathways. Key glycolytic enzymes, including hexokinase and phosphofructokinase, show reduced abundance these prediabetic mice subjects (Ferreira et al., 2022). This decrease suggests

changes in sucrose utilisation and lactate production. Pyruvate dehydrogenase A1, part of the pyruvate dehydrogenase complex, is upregulated in HFD prediabetic conditions, a complex important in switching from carbohydrates to lipids as an energy source (Ferreira et al., 2022).

Dietary proteins particularly lysine are used as an energy source by intestinal epithelial cells and are also involved in microbiota composition (van Goudoever et al., 2000; Kar et al., 2017). Ferreira and colleagues noted two interseting findings in their proteomics studies of the prediabetic small intestine in regards to amino acid metabolism. The lysine degradation pathway was observed to be affected, a key mediator in protein biosynthesis such that of cartinine which is involved in fatty acid metabolism. Secondly, nine of the ten proteins involved in arginine and proline metabolism are upregulated in the prediabetic small intestine. Argenine is a well documented insulin secretagogue regulates the release of GLP-1 in the gut having implications in hyperinsulinemia commonly seen in prediabetes (Ferreira et al., 2022).

These alterations in macronutrient metabolism within the intestinal epithelium reflect the complex metabolic changes occurring in prediabetes. The shift in lipid metabolism towards increased fatty acid oxidation, changes in carbohydrate utilisation, and alterations in amino acid metabolism all contribute to the dysregulation of energy homeostasis. Further research should look into the differences in dietary composition on the progression of prediabetes as these finding suggest that HFHSD and HFD may have some discrepancies in carbohydrate metabolism. Furthermore, some lipids are reported to be more diabetogenic than others despite being isocaloric (Deol et al., 2015). Nonetheless, these changes provides insight into the role of the intestinal epithelium in the progression of prediabetes and may offer potential targets for therapeutic interventions.

1.2.5 Intestinal Stem Cell Function Alterations

Overnutrition, a key factor in prediabetes development, is associated with significant alterations in ISC function and proliferation. HFDs and HFHSDs have been shown to stimulate ISC and progenitor cell proliferation, leading to an expansion of the stem cell pool (Pourvali and Monji, 2021). This hyperproliferation is thought to be a key factor in the

increased risk of gastrointestinal cancers observed in individuals with prediabetes and obesity.

The mechanisms underlying this increased ISC proliferation are multifactorial involving several signaling pathways. Some studies suggest that peroxisome proliferator-activated receptor delta (PPAR-δ), a regulator of fatty acid oxidation is implicated in intestinal cancer as it plays a role as a transcriptional target for Wnt/β-catenin signaling cascade (Beyaz and Yilmaz, 2016; Pourvali and Monji, 2021). This activation is thought to increase expression of Wnt target genes, including those involved in cell proliferation and stemness maintenance. There have been some conflicts regarding this hypothesis of PPAR-δ mediated ISC hyperproliferation however. Aliluev and colleagues have demonstrated that in prediabetic HFHS-fed mice, hyperproliferation occurs in the absence of PPAR-δ mediated activation of Wnt/β-catenin signaling cascade. Rather, a combination of qPCR and scRNA-seq data displays an upregulation of PPAR-γ and sterol regulatory element-binding protein 1 (SREBP1)-mediated lipogenesis and insulin-like growth factor 1 receptor (IGF-1)–Akt signalling, which is associated with tumorigenesis as well as increased proliferation (Shao and Espenshade, 2012; Aliluev et al., 2021). Insulin and IGF-1 signaling are elevated in these conditions and have been shown to promote ISC proliferation through the PI3K/Akt pathway. Conversely, adiponectin, which is typically reduced in obesity, has been found to regulate ISC numbers and apoptosis, with its decrease potentially contributing to ISC expansion (Pourvali and Monji, 2021; Colloca et al., 2024).

Interestingly, the literature presents some contradictions regarding the effects of different dietary compositions on ISC function. While HFDs generally promote ISC proliferation, ketogenic diets have been reported to enhance ISC function and self-renewal through activation of the Notch pathway with sugar supplementation attenuating the effects (Pourvali and Monji, 2021). This suggests that the interaction between fats and carbohydrates, rather than fats alone, may be crucial in determining ISC behavior in prediabetic conditions.

The increased proliferation and altered function of ISCs in prediabetes have significant implications for intestinal health and disease risk. The expansion of the stem cell pool and acquisition of stemness properties by progenitor cells may increase susceptibility to oncogenic transformation, potentially explaining the elevated risk of colorectal cancer

observed in individuals with prediabetes and obesity. Furthermore, these alterations in ISC function may contribute to changes in intestinal barrier integrity and nutrient absorption, further exacerbating metabolic dysfunction. It is important to note that while the link between overnutrition, ISC hyperproliferation, and increased cancer risk is well-established, the exact mechanisms and the role of specific dietary components require further investigation. Future research should focus on exploring the complex interactions between diet, obesity, and ISC function to identify potential therapeutic targets for preventing or mitigating the negative effects of prediabetes on intestinal health.

Chapter 2: scRNA-seq and modelling approaches for revealing alterations in the prediabetic disease state.

2.1 Single Cell RNA-sequencing

Multiomics technologies have advanced with major breakthroughs in the last two decades, driven by developments in bioinformatics, computational biology, and multi-omics technologies. The ability to capture large amounts of molecular data through high-throughput technologies provides a new landscape of information in which systems biology can be studied. These approaches collectively enable the comprehensive characterisation of biological systems at multiple levels, including the transcriptome, proteome, metabolome, epigenome, and genome. Multi-omics techniques integrate several methodologies to provide a holistic view of biological processes. The intersections of each of these disciplines are revealing new understandings and mechanisms by which organisms operate, on the multicellular, larger perspective, but also focussed perspectives at the single-cell resolution.

In the context of prediabetes and T2DM research, these multi-omics approaches enable researchers to reveal alterations at multiple biological levels, from genetic predisposition to changes in gene expression, protein function, and metabolic pathways as have been discussed thus far. The application of multi-omics approaches in prediabetes and T2DM research harnesses the potential to reveal novel mechanisms contributing to disease progression, clarify conflicting hypotheses, and uncover potential therapeutic targets. As

these technologies continue to evolve, they promise to provide increasingly detailed and nuanced understandinsg of the molecular basis of metabolic disorders.

Among these techniques, single-cell RNA sequencing (scRNA-seq), has emerged as a leading approach due to its ability to provide high-resolution insights into cellular heterogeneity, gene expression dynamics at the individual cell level and its increasing accessibility to scientists. This technology has proven particularly valuable in studying the diverse cell populations within the intestinal epithelium and their roles in metabolic health and disease (Aliluev et al., 2021; Xie et al., 2020).

Single cell RNA sequncing technologies are employed in this research project in tandem with various bioinformatics and computational biology methods in an attempt to understand the role of the individual cell types of the intestinal epithelium in prediabetes. Here we review the key steps in the analysis process from quality control to cell type annotation to the downstream bioinformatics analysis techniques.

2.1.2 Quality Control

Quality control (QC) is a one of the initial steps in scRNA-seq analysis, aimed at identifying and removing low-quality cells and genes that could skew downstream analyses. The process typically involves evaluating three main aspects of the data: the number of counts per cell or ‘count depth’, the number of genes detected per cell, and the fraction of counts from mitochondrial genes per cell (Carangelo et al., 2022). Cells with exceptionally low count depths, few detected genes, or high fractions of mitochondrial counts (which typically indicates loss of cytoplasmic RNA) are often considered damaged or dead and are removed from the dataset (Ilicic et al., 2016).

However, it's important to note that these quality metrics should be considered in combination with one another, as variation in one parameter may reflect specific biological conditions rather than technical artifacts. For instance, high mitochondrial counts could indicate cells with heavy respiratory activity, while low counts might represent quiescent cells. Therefore, QC thresholds should be set carefully to avoid unintentionally filtering out viable cell populations (Carangelo et al., 2022).

Advanced methods for doublet detection, such as scrublet, have been developed to identify and remove cell multiplets, which can confound downstream analyses (Wolock et al., 2019). These methods can distinguish between embedded doublets and neotypic

doublets (Carangelo et al., 2022), improving the accuracy of cell type identification and are used accordingly in this research project.

2.1.3 Normalisation

Normalisation is essential for making gene expression levels comparable between cells and addressing technical biases introduced during library preparation and sequencing. The most common approach is linear normalisation, which aims to equalise the depth for all cells to a “size factor” such as counts per million (CPM) or counts per ten thousand (CP10K), which scales the count data to obtain relative gene expression abundances. This method assumes that all cells initially contained an equal number of mRNA molecules, and differences in count depth arise solely due to sampling. Following this is typically a variance stabilising transformation of the data such as a log plus one (log1p) transformation. These techniques are commonly used in scRNA-seq analysis and are implemented in the industry standard Scanpy and Seurat programs (Satija et al., 2015; Wolf et al., 2018). One obvious drawback of using these linear normalisation techniques is that they no longer account for count depth as a covariate, some techniques try to account for this via regression based methods although there are conflicting results about how effective it is really is for depth normalisation (Booeshaghi et al., 2022).

Non-linear normalisation methods may be more appropriate for data with large batch effects, particularly for plate-based scRNA-seq data. These methods often employ parametric modeling to correlate biological/techincal sources of variability and correct for both simultaneously (Svensson et al., 2017; Lytal et al., 2020; Carangelo et al., 2022).

2.1.4 Dimensionality Reduction

ScRNA-seq datasets are characteristically of large dimensions consisting often of tens of thousands of cells each with tens of thousands of genes. As a result many data analysis and visualisation approaches, built for lower dimensional data, suffer from the “curse of dimensionality” - a term first coined by Richard Bellman in 1961 but remaining relevant still today (Richard Bellman, 1961). High-dimensional scRNA-seq data is typically reduced to lower dimensions to facilitate visualisation and downstream analyses. A number of techniques can be employed to facilitate this such as feature selection or dimensionality reduction methods which attempt to retain as much of the underlying structure as possible.

Feature selection methods exclude genes which do not explain any variation in the data such as genes which do not vary greatly between cells. These are uninformative genes and fail to explain cell heterogeneity or differences between experimental conditions. Removing these genes allows for more optimised and efficient computational methods (Andrews and Hemberg, 2019; Chen et al., 2019). Feature selection algorithms have been devised for scRNA-seq datasets allowing for the selection of highly variable genes via unsupervised feature selection algorithms such as those seen in Seurat and Scanpy (Satija et al., 2015; Chen et al., 2019).

Dimensionality reduction techniques typically seen in scRNA-seq analyses include Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) (Linderman, 2021). PCA is commonly used as an initial step to capture the main sources of variation in the data. The principle components are defined as the eigenvectors of the covariance matrix and are calculated with the singular value decomposition, a matrix factorisation technique. However, PCA assumes normal linear relationships between features and may not fully capture the complex structure of scRNA-seq data (Chen et al., 2019). Non-linear dimensionality reduction techniques, such t-SNE and UMAP, are frequently employed for visualisation and subsequent clustering. These methods aim to preserve local similarities between cells in the high-dimensional space while projecting them into two or three dimensions at the cost of longer, global similarities. Global structures can however be dramatically improved by initialising the embedding with the first two principle components from a preliminary PCA step (Linderman, 2021). UMAP has gained popularity due to its ability to preserve both local and global structure, faster runtime, and higher reproducibility compared to t-SNE (Becht et al., 2019; Chen et al., 2019).

2.1.5 Clustering and annotation

Clustering is a critical step in scRNA-seq analysis, aiming to group cells with similar transcriptomic profiles. Clustering methods are created on the assumption that cells with similar transcriptomes are of a certain cell type. This approach allows researchers to identify distinct cell populations within heterogeneous samples, revealing cellular diversity and potential new cell types or states. The effectiveness of clustering impacts downstream analyses, including differential gene expression and trajectory inference, making it a crucial component of scRNA-seq data interpretation.

Several clustering algorithms have been applied to scRNA-seq data, each with its own strengths and limitations. K-means clustering, a classical approach, partitions cells into a predefined number of groups based on their gene expression similarities. Hierarchical clustering, another traditional method, constructs a tree-like structure of cell relationships (Peng et al., 2020). However, these methods may struggle with the high dimensionality and sparsity characteristic of scRNA-seq data.

To address these challenges, unsupervised, graph based clustering methods, such as the Louvain and Leiden algorithms, have become popular due to their efficiency and ability to handle large datasets (Blondel et al., 2008; Traag et al., 2019). These methods first construct a graph where cells are nodes and edges represent similarities between cells. The graph is then partitioned to identify communities of cells with similar expression profiles. Leiden algorithm, an improvement over the Louvain algorithm, guarantees that the identified communities are connected, addressing a major limitation of its predecessor (Traag et al., 2019).

Despite their effectiveness, clustering methods face several challenges in the context of scRNA-seq analysis. One primary challenge is determining the appropriate number of clusters or the resolution parameter in graph-based methods (Peng et al., 2020). This choice can significantly impact the resulting cell type classifications and is often not straightforward. Overestimatnig the number of clusters may lead to artificial splitting of cell types, while underestimation can result in the merging of distinct populations.

Another challenge is the need for reclustering to identify rare cell types or more subtle subpopulations (Peng et al., 2020). Initial clustering may reveal broad cell types, but subsequent reclustering of these groups can uncover finer distinctions. However, determining when and how to perform reclustering requires careful consideration and biological knowledge.

Validating clustering results presents an difficult challenge, as clustering is typically based off of the researchers domain knowledge and biological interpretation. Future research should focus on developing more robust and universally applicable methods for cluster number/ resolution determination, rare cell type identification, and clustering result validation in the context of scRNA-seq analysis.

Once clusters have been identified, the next step is to assign biological identities to these cell groups in a process known as annotation. The prevailing method typically involves identifying marker genes that characterise each cluster and matching them to distinct cell types as described in the scientific literature (Cheng et al., 2023). Differential expression analysis between clusters is often used to identify these marker genes, with statistical tests such as the Wilcoxon rank-sum test or t-test employed to rank genes by their difference in expression, techniques which are standard practice in the commonly used Scanpy and Seurat programs.

Automated annotation methods, which compare cluster-specific marker genes to reference datasets or known cell type signatures, can expedite this process (Cheng et al., 2023). However, manual curation by domain experts is often necessary to ensure accurate annotation, particularly for novel or rare cell types.

2.1.6 Differential Gene Expression Analysis

Differential gene expression (DGE) analysis is an important step in RNA sequencing data analysis, enabling the identification of genes that are differentially expressed between experimental conditions or cell types. This analysis provides insights into cellular processes, disease mechanisms, and responses to treatments. This analysis must account for the unique characteristics of scRNA-seq data, including high sparsity, technical noise, and complex experimental designs involving multiple subjects and conditions.

Traditional DGE methods, initially developed for bulk RNA-seq data, have formed the foundation for many current single-cell analytical approaches. Three widely used tools in this domain are DESeq2, edgeR, and limma. DESeq2, developed as a successor to DESeq, employs shrinkage estimators for dispersion and fold change to facilitate more quantitative analysis of comparative RNA-seq data (Love et al., 2014). It introduces features such as automated outlier detection and handling, and hypothesis tests for log-fold changes above or below specified thresholds. edgeR, designed initially for Serial Analysis of Gene Expression (SAGE) data, models count data using an overdispersed Poisson model and applies an empirical Bayes procedure to moderate the degree of overdispersion across genes (Robinson et al., 2010). It has since been adapted for various types of sequencing data, including RNA-seq. Limma, originally created for microarray data analysis, has been extended to handle RNA-seq data through the voom

transformation, which estimates the mean-variance relationship of the log-counts and generates a precision weight for each observation (Ritchie et al., 2015).

While traditional DGE methods share common strategies like information sharing across genes, they face unique challenges with scRNA-seq data, including abundant zero counts, multimodal distributions, and significant cell-to-cell heterogeneity. To address these issues, scRNA-seq-specific tools have been developed, such as SCDE and MAST, which use two-part models for zero counts, and Monocle2, which employs alternative normalization strategies (Wang et al., 2019). Nonparametric methods like SigEMD and EMDomics have also been proposed to handle heterogeneous expression distributions. Comparative studies show that tool performance varies based on dataset characteristics, with some methods exhibiting a higher agreement on highly multimodal data (e.g., DESeq2, EMDomics, Monocle2) (Wang et al., 2019). The varying agreement among these tools in identifying differentially expressed genes demonstrates the complexity of DGE analysis in scRNA-seq data.

DGE analysis in scRNA-seq still faces a number of issues. There is lack of consensus among different tools in identifying differentially expressed genes and their ability to detect biologically relevant genes. Future developments should aim to address multimodality and sparsity to improve the accuracy and reliability of DGE analysis in single-cell RNA-seq data.

2.1.7 Gene Set Enrichment

Gene set enrichment analysis (GSEA) is an important and commonly seen downstream analysis technique seen in scRNA-seq analysis, allowing for the identification of sets of genes that are statistically overrepresented in their results. This method helps highlight the patterns in biological processes and pathways seen in the DGE analysis, providing insights into how experimental conditions affect cellular functions.

Traditional GSEA tools, originally developed for bulk RNA-seq and microarray data, rely on predefined gene sets such as Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto, 2000; The Gene Ontology Consortium et al., 2023). These manually curated gene sets represent known biological processes, molecular functions, and cellular components. Popular tools like DAVID, Enrichr, and clusterProfiler implement various statistical approaches, including

overrepresentation analysis (ORA) and the original GSEA algorithm, to assess the enrichment of these gene sets within a given gene list (Kuleshov et al., 2016; Wu et al., 2021; Sherman et al., 2022).

GSEA methods can produce vastly different results depending on the statistical approach used, making it challenging to interpret findings consistently. The analysis can be computationally intensive, especially for large gene set collections (Geistlinger et al., 2021). Additionally, the increasing number and redundancy of available gene sets can lead to complex, overlapping results that are difficult to interpret.

To address these issues of gene set redundancy, researchers have developed visualisation techniques such as Enrichment Map. This approach organises enriched gene sets into a network-based visualisation, where each gene set is represented as a node, and edges between nodes indicate the overlap of genes between sets. The resulting network is then clustered to group related gene sets, enabling the indentification of major functional themes and allowing for easier interpretation of the enrichment results (Merico et al., 2010).

As scRNA-seq technology continues to advance, GSEA remains an essential tool for extracting biological meaning from transcriptomic data, and will prove to be a valuable technique through the course of this research project. However, the field faces ongoing challenges in adapting these methods to the unique characteristics of single-cell data, such as dropouts, technical noise, and cellular heterogeneity.

Materials and Methods

Experimental Design and Data Generation

Mouse models

This research study used sc-RNA-seq data from a previously published experiment (Aliluev et al., 2021). Briefly, the dataset is composed of scRNA-seq data from small intestinal crypts of control diet and high-fat high-sugar diet obese FVF-enriched mice, as well as villus samaples from ontrol diet and HFHSD C57BL/6N mice. Male mice were randomised into diet groups at 10-12 weeks old and maintained on their respective diets for 11-13 weeks.

Prediabetic Evalutation

Glucose tolerance and insulin secretion were assessed via oral glucose tolerance tests after 12 weeks on the diets. The HFHSD-fed mice developed a prediabetic phenotype characterised by increased hyperinsulinemia, impaired glucose tolerance, and insulin resistance. Insulin resistance and beta cell function were measured using homeostasis model assessment (HOMA-IR and HOMA-β) indices (Aliluev et al., 2021). These assessments confirmed the presence of the prediabetic state in the HFHSD group providing an applicable model for studying the effects of prediabetes on the intestinal epithelium.

Single-cell preparation and RNA-sequencing

Small intestinal crypts and villi were isolated and processed into single-cell suspensions following established protocols as described in the original content (Aliluev et al., 2021). Flow cytometry was used to enrich for viable, single cells, wit an additional enrihment for FVF-positive cells in crypt samples. Single-cell RNA sequencing libraries were prepared using the 10x Genomics Chromium platform and sequenced on an Illumina HiSeq4000 (Aliluev et al., 2021). For comprehensive experiemntal details, including housing conditions, specific dietary compositions, cell isolation protocols, and sequencing paraeters, readers are directed to the original publication (Aliluev et al., 2021).

Upstream Analysis Pipeline

Preprocessing and QC of scRNA-seq data

Raw data was obtained from the Gene Expression Omnibus (GEO) database and cleaned to isolate count data and relevant metadata. Quality control of the scRNA-seq data was performed using the Scanpy package (Wolf et al., 2018) to ensure the integrity and accuracy of the dataset following standard protocols as described previously in the introducion. The quality control steps focused on identifying and mitigating potential confounding factors such as high mitochondrial fractions suggesting cellular stress or cell death. Mitochondrial genes were identified using the prefix "mt-" for mouse genes and similarly prefixes "Rps" and "Rpl" were used for identifying ribosomal genes. The `calculate_qc_metrics()` function from Scanpy was used to compute QC metrics including: total number of genes detected per cell, total count of unique molecular identifiers (UMIs) per cell, percentage of counts from mitochondrial genes, percentage of counts from ribosomal genes, and the percentage of counts attributed to the top 50 most expressed genes. These metrics provide information about the quality and characteristics of individual cells and genes in the dataset, allowing for the identification and removal of low-quality cells or outliers that could skew downstream analyses. The metrics were calculated and stored in the AnnData object for subsequent analysis. Cells were removed with greater than 10% mitochondrial gene content, less than 200 genes by counts as well as cells with less than 500 total counts.

To identify potential doublets, Scrublet (Wolock et al., 2019) was used using a nearest neighbour classifier of observed transcriptomes and simulated doublets. Predicted doublets were subsequently removed from the dataset.

Normalisation and logarithmisation

Counts were normalised per cell by total counts over all genes so that every cell was normalised to the same count depth with raw counts saved for later batch effect correction. Each cell is normalised to a total count equal to the median of total counts for cells before normalisation using the scanpy function normalize_total. The data are then logarithmised using log1p where each count is transformed to the natural log of 1 plus the original count

value allowing for the data to be transformed as well as zero values facilitated by the scanpy function log1p.

Dimensionality Reduction, Batch Effect Correction and Visualisation

Principal Component Analysis (PCA) was performed with a fixed random state for reproducibility and was performed using the scanpy pca function. Batch correction was subsequently applied using single-cell variational inference (scVI) using the scvi-tools package which implements a probabilistic model from scRNA-seq data (Lopez et al., 2018). scVI uses a variational autoencoder with a negative binomial distribution to model gene expression and explicitly accounts for batch effects through a conditional independence property in its generative process (Lopez et al., 2018). An scVI model was instantiated and trained on the cRNA-seq data using the model.train function. Subsequently the trained model was used to generate a latent representation of the data using get_latent_representation capturing biological variability while accounting for batch effects. A nearest neighbor graph was constructed using the scVI-corrected data, and Uniform Manifold Approximation and Projection (UMAP) was applied to visualize the high-dimensional data in 2D space, both with fixed random states for reproducibility. The processed and dimensionality-reduced data was then saved for subsequent clustering and annotation steps.

Clustering and annotation of scRNA-seq data

Clustering and annotation of the scRNA-seq data were performed using a combination of manual curation and algorithmic approaches, guided by marker genes obtained from the original study. The process was implemented in Python once again using the Scanpy package (Wolf et al., 2018). Custom-developed functions were created to allow for reproducibility as well as allowing for readers to follow along the documented code easily and in a transparent manner. The scanpy function leiden marks the beginning of the clustering process, with the resolution parameter set to 0.4, dividing the dataset into 15 initial clusters. Following this is iterative refinement using custom functions for combining clusters, reclustering clusters as well as visualising marker gene expression across the data. Furthermore, a random state was set to allow for reproducibility. Expression levels of marker genes for lymphocytes (Coro1a, Cd52, and Cd37) were visualised in each cluster using violin plots. Marker genes for lymphocytes were obtained from the original pubished

research study (Aliluev et al., 2021). Lymphocyte clusters were subsequently removed from the dataset as intestinal epithelial cells are the focus of the study. Marker gene expression was initially visualised across the dataset using the custom functions plot_gene_heatmap_umap, and plot_combined_gene_heatmap_umap which generate UMAP plots colored by individual and combined gene expression. The recluster_subset function harnesses the leiden scanpy function and is used to select a specific cluster, recluster at a specified resolution and visualise the clustered UMAP before and after reclustering. Furthermore, divide_cluster_by_expression divides a cluster by the expression of marker genes, with before and after results also being visualised through a UMAP. The combine_clusters function works in a similar way where the user specifies a list of clusters, they are subsequently combined, and before/after results visualised. These functions allowed for reclustering of specific cell subsets, division of clusters based on marker gene expression, and combination of similar clusters in a transparent manner where readers can easily follow the logic behind why each step was taken. Throughout this process, clusters were annotated based on their expression of known marker genes. A comprehensive set of marker genes was used, including Lgr5, Olfm4, Axin2, Ascl2, and Slc12a2 for intestinal stem cells; Fabp1, Alpi, Apoa1, Apoa4, and Lct for enterocytes; Muc2, Tff3, Agr2, Spdef, Klf4, Ccl9, and Manf for goblet cells; Dclk1, Trpm5, Gfi1b, and Il25 for tuft cells; and Neurod1, Neurod2, Insm1, Chga, and Chgb for enteroendocrine cells. Progenitor cells were identified by the overlap between intestinal stem cell marker expression with marker genes from mature cell types. Cells which did not appear to express any distinct cell type marker were labelled as ‘Not Annotated’ as there was no information available to assign them a label with confidence. The custom rename_clusters function was created to assign biologically meaningful labels to the identified clusters. Results were compared with the original research papers annotations for comparison. This iterative process of clustering, visualising marker gene expression, and manual annotation was repeated multiple times, each iteration refining the clustering and annotation based on observed gene expression patterns ultimately leading to the identification and characterisation of 12 distinct cell clusters, with five key cell types of particular interest: intestinal stem cells, enterocytes, goblet cells, tuft cells, and enteroendocrine cells. The final annotation provided a comprehensive characterisation of the cellular composition of the intestinal epithelium in the context of the study's dietary conditions and is used in the subsequent sections of the analysis pipeline.

Feature plots

To visualise marker gene expression across cell types, as seen in figure 1c, a custom function plot_combined_gene_heatmaps_umap was developed. This function generates UMAP plots for each cell type based on the combined expression of its associated marker genes. The function takes the annotated data object (adata) and a dictionary of cell types with their corresponding marker genes as input. For each cell type, the average expression of marker genes is calculated per cell, normalized, and visualized on the UMAP. The plots are annotated with the cell type, marker genes, and the range of average expression values. This method allows for simultaneous comparison of marker gene expression patterns across multiple cell types, aiding in the validation of cell type annotations and identification of potential subtypes or states within cell populations.

Marker Gene Heamaps

To visualise the expression patterns of marker genes across the identified cell clusters, heatmaps were generated for both the CD and HFHSD dietary conditions. The proces involved several steps of data preparation and subsequent visualisation with the scanpy package (Wolf et al., 2018). The dataset was first subset to include only cells from the respective dietary conditions (CD and HFHSD). To focus on diet-related effects, and align with the approach described in the original study, cell cycle genes were removed from the analysis. These genes were identified using the GO term for cell cycle, GO:0007049, and excluded prior to marker gene identification. These methods are consistent with the methods of Aliluev and colleagues (Aliluev et al., 2021). Marker genes were then identified using a wilcoxon rank-sum test via Scanpy’s rank_genes_groups function. Genes with a score greater than 5 were selected, and a maximum of 19 genes per cluster. This setting allowed for a manageable number of features for visualisation while still capturing the most significant markers. The data was then subset to include only the identified marker genes. Cell type raitios were calculated and normalised to account for differences in cluster sizes. The expression data were subsequently scaled using scanpys scale function, with a maximum value of 4 and zero-centering enabled to enhance the visualisation of expression patterns across different genes and clusters. The final heatmaps were created using scanpy’s pl.heatmap function. The function was parameterised to use the list of

identified marker genes (var_names) and cluster annotations (groupby='leiden15'). A diverging colour map ("RdBu_r") was used, ranging fro blue (low expression) to red (high expression). The axes were swapped, displaying genes on the y-axis, and cell types on the x-axis. Gene labels were hidden as there were too many to display without visual clutter.

Downstream Analysis Pipeline

DGE analysis

To identify genes which were differentiall expressed between the CD and HFHSD conditions across different cell types, a differential gene expression (DGE) analysis was performed. Three methods were explored, Scanpy's in-built function, DESeq2, and limma-voom. After careful consideration, DESeq2 was selected as the primary method for the analysis due to its robust performance with count data. Scanpy's rank_gene_groups function was used with the wilcoxon rank-sum test parameter to perform the DGE analysis. This method was implemented for each identified cluster, with genes filtered to remove ambient genes and those expressed in less than 10% of cells in either diet condition. Files were exported as CSV files for each cluster for further downstream analysis.

DGE analysis was also experimented using the limma-voom method. This approach involved creating a design matrix accounting for diet condition, batch effects as well as the number of genes expressed per cell. The limma package was then used to fit linear models and perform empirica Bayes moderation.

For the final analysis, DESeq2 was employed which is specifically designed for RNA-seq count data and provides a sophisticated approach to differential expression analysis. The DESeq2 analysis was performed separately for each cell type identified in the clustering step. The process involved several key steps of data preparation, filtering, running the DESeq2 algorithm, results processing, and output generation.

For each of the DGE analysis methods, ambient genes were filtered out of the data. These represent RNA molecules that were free-floating in the cell suspension solution which were incorporated with the single cell droplets. The list of genes were identified by Aliluev and colleagues and subsequently published in the research paper associated with these data. In brief the genes were identified by examining empty droplets (UMI's with a total count

less than 200) based off the assumption that the only RNA present in these droplets must be from free floating RNA. Genes expressed in more than 1% of empty droplets were considered. Furthermore, genes showing a high fold change in expression between diet conditions in empty droplets were noted as potentially generating false positives in the DGE analysis. The list of ambient genes identified are: Defa20, Ang4, Gm14850, Gm7861, Defa22, Gm15308, Itln1, Zg16, Defa17, Lyz1, Gm15284, Defa21, Fcgbp, Agr2, Gm15308, Clps, Spink4, Gm14851, AY761184, Defa24, Tff3. Finally, genes which were expressed in less than 10% of cells were filtered out.

The DESeq2 analysis was performed by creating a DESeqDataSet object for each cell type in R studio using the diet condition as the main factor of interest in the design formula. The analysis was run using the DESeq function estimating size factors, dispersions, and fits a negative binomial generalised linear model. Results were extracted and saved as CSV files for each individual cell type.

Volcano plots were generated for each cell type, with the log2 fold change plotted on the x-axis and the adjusted p-value on the y-axis in the form $-\log_{10}P$. The log2 fold change cut off was set to 0.5 and the adjusted p-value cutoff set to 0.05. The most significantly differentially expressed genes are labelled with their respective gene symbols on the plot.

Gene Ontology Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) was employed to attach biological meaning to the differentially expressed genes using the R package clusterProfiler (Wu et al., 2021). The gseGO function was employed to conduct GSEA against the GO database. Initially, the Generally Applicable Gene-set Enrichment (GAGE) R package was evaluated as an alternative method for GSEA. However, due to the presence of outdated Gene Ontology terms in its database, it was not selected for the final analysis.

Gene symbols were converted to Entrez ID format using the org.Mm.eg.db annotation package. Duplicate genes were removed, and only genes with an adjusted p-value < 0.05 were retained for the analysis. A named vector of log-fold change values was created and sorted descendingly to sere as input fo the gseGO function. The analysis was performed separately for each class of GO terms (Biological Process, Cellular Component, and Molecular Function) using the mouse genome as the reference organism. The gseGO

function was parameterised with nPermSimple = 10000 to ensure statistical robustness at a slight inconvenience of increased computational time.

To address the challenge of redundancy in GO terms, a custom clustering approach was developed. This method aimed to aggregate similar GO terms, facilitating a more easily interperatable representation of enrichment results. The clustering proess began by computing pairwise Jaccard similarities between GO terms based on their associated gene lists. A distance matrix was constructed using the complement of these similarity scores (1 – similarity).

Hierarchical clustering was applied to this distance matrix using the average linkage method. The resulting dendrogram was cut at a modifiable height of 0.9 to define clusters, balancing specificity and interpretability. This process effectively grouped GO terms with substantial gene set overlap, allowing for a more streamlined analysis of functional enrichment patterns.

For visualisation of these results. A custom plotting function was developed using ggplot2 R package which integrates the enrichment results with the clustering results. The plot displays GO terms as points with the x-axis represnting the absolute Normalised Enrichment Score (NES) and the y axis represnting individual GO terms. The size of each point corresponds to the number of genes associated with the term, while colour coding denotes cluster membership. To enhance the readability, GO terms were faceted by ontology and regulation (Upregulation or Downregulation). P-value significance was indicated using asterisks, and cluster numbers were superimposed on each point for easy reference.

To further explore the relationship between enriched GO terms, the R studio package aPEAR was used for generating an enrichment map. This approach provided a network-based visualisation of the enrichment results, offering additional insights into the broader biological functions enriched from the differentially expressed genes. The enrichmentNetwork function from aPEAR was used with parameters set to detect clusters of size 3 or greater. Edge cutoff values were set at 0.6 and 0.1 for outer and inner connections respectively. These parameters offered the best balancing of the connectivity between and within clusters. This network representation complements the dotplot visualisation by highlighting the interconnectedness of enriched biological processes and functions.

KEGG Enrichment Analysis

To further understand the biological pathways affected by differential gene expression between diet conditions, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed. The analysis is composed of three main components, enrichment analysis using the GAGE bioconductor package, visualisation of enrichment results via custom developed density plots, and the creation of pathway graphs using the pathview Bioconductor package.

For the KEGG enrichment analysis, GAGE was employed for GSEA analysis against the KEGG databse. Gene symbols were first mapped to their corresponding Entrez IDs using the org.Mm.eg.db annotation package. A foldchanges vector was created using the log2 fold change values, with Entrez IDs as names. The KEGG pathway sets for mouse (kegg.sets.mm) were used, focussing on signalling and metabolic pathways (sigmet.idx.mm). After the analysis, the results were separated into upregulated and downregulated pathways. Upregulated/downregulated pathways were extracted from keggres$greater / keggres$less, respectively. In both cases, pathways with a p-value < 0.05 were considered significantly enriched. The pathway IDs were then processed to extract the KEGG-specific identifier.

To visualise the KEGG enrichment results, a custom function was developed to create density plots of log2 fold change distributions for each enriched pathway for each cell type. This function integrates both the enrichment results as well as the associated gene expression data from the differential gene expression analysis. KEGG pathways were categorised into major and subcateogries based on information obtained from the KEGG website, which was parsed into a usable format with the parse_kegg_categories function. For each pathway, a gene ratio was calculated representing the proportion of significantly differentially expressed genes. P-values are formatted in scientific notation, with significance levels indicated using asterisks. The density plots were created using the ggplot2 and ggridges packages. Each plot represents an enriched pathway with the x-axis showing log2 fold change values and the y-axis displaying individual pathways. The fill colour for each density plot represents the gene ratio, while the outline colour indicates the major pathway category. This visualisation approach allowed for the simultaneous representation of pathway enrichment, gene expression information and pathway cetegorisation.

For a more detailed view of gene expression changes within specific pathways, the pathview Bioconductor package was used to create KEGG pathway graphs. For each significantly enriched pathway, a graph was generated using the pathview function. The function parameters were set to highlight upregulated genes in green, and downregulated genes in red, with colour intesity corresponding to the fold change magnitude. This provided a visual represntation of how differentially expressed genes fit into the context of known biological pathways.

To further investigate the gene-level changes within pathways of iterest, another custom function was developed allowing for the visualisation of the expression values of individual genes across multiple cell types. This function, named ‘pathway_heatmap’, uses the pheatmap Bioconductor package to create the framework of the plots. In the plots, columns represent cell types, rows represent genes, and the colour represents the logFC value of the gene allowing for the simultaneous visualisation of gene expression changes in a pathway across multiple cell types.

The function accepts a list of differential expression results for various cell types and either a KEGG pathway ID or a custom gene set. It then constructs a matrix of log2 fold change values and adjusted p-values for the genes of interest across all provided cell types. To enhance clarity and reduce visual clutter, genes with entirely NA values or zeroes across all cell types were optionally removed from the visualisation.

A key feature of the function is the ability to order genes on the sum of their significant log2FC values. This ranking helps to highlight the genes with the most substantial and consistent changes across cell types. The significance of gene expression changes is visually represented using asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$) overlaid on the heatmap cells. These visualisations complement the pathway enrichment analysis and density plots by providing a detailed view of gene-level changes within pathways of interest, allowing for a more nuanced interpretation of the transcriptomic changes induced by the HFHSD condition.

Results

Single-Cell RNA-Sequencing Reveals Cell Type Heterogeneity and Diet-Induced Alterations in the Intestinal Epithelium

Single-cell RNA-sequencing analysis of small intestinal tissue displayed the variation the transcriptomes of the cells in the dataset and revealed distinct cell populations across CD and HFHSD conditions. A total of 27,419 cells passed quality control criteria, with 13,363 cells from CD and 14,056 cells from HFHSD conditions. The average number of genes detected per cell was 3,566 for CD and 3,490 for HFHSD, while the mean UMI count per cell was 18,647 for CD and 20,385 for HFHSD. Unsupervised Leiden clustering and annotation based on known marker genes identified 12 distinct cell clusters (Figure 1a). These clusters represented six main cell types, ISCs, enterocytes, goblet cells, tuft cells, paneth cells, and enteroendocrine cells. Additionally, progenitor populations were identified for all main cells types excluding ISCs. Cells which did not express known epithelial cell markers were labelled as ‘Not Annotated’. UMAP visualisation of cells from CD and HFHSD conditions (Figure 1b) showed overall similarity in cluster distribution, indicating sucessful batch correction. Notable differences were observed in the density in the Paneth cell cluster which were more dense in the HFHSD condition. This cluster of Paneth cells were reported to have sampling issues however, and were thus removed from further analyses. Feature pots displaying the average expression of cell type-specific marker genes (Fig 1c) confirmed the identity of the annotated clusters. Each cell type exhibited a distinct pattern of marker gene expression. ISCs exhibited a high expression of *Lgr5*, *Olfm4*, *Axin2*, *Ascl2*, and *Slc12a2* (max average expression: 2.22), enterocytes displayed high expression of *Fabp1*, *Alpi*, *Apoa1*, *Apoa4*, and *Lct* (max average expression: 4.27), Goblet cells displayed high expression of *Muc2*, *Tff3*, *Agr2*, *Spdef*, *Klf4*, *Ccl9* (max average expression: 4.43), and *Manf*, Paneth cells displayed high expression of *Lyz1*, *Mmp7*, *Defa17*, *Defa22*, and *Anf4* (max average expression: 6.10). Enteroendocrine cells displayed high expression of *Neurod1*, *Neurod2*, *Insm1*, *Chga*, and *Chgb* (max average expression: 4.33), and tuft cells displayed high expression of *Dclk1*, *Trpm5*, *Gfi1b*, and *Il25* (max average expression: 2.42). Quantification of cell type proportions between CD and HFHSD conditions (Fig 1d) revealed changes in several cell populations. The ISC population was lower in the HFHSD condition, while enterocyte progenitors showed a

higher proportion, indicating increased cell turnover of ISCs in the HFHSD condition. Conversely, both tuft progenitors and tuft cells appeared in lower proportions in the HFHSD condition. These trends were consistent with previously published findings using this dataset (Aliluev et al., 2021). Enterocyte progenitors were the largest cluster with a proportion of approximately 0.3 in CD and 0.375 in HFHSD. Second to this cluster were ISCs and goblet proportions. ISCs exhibited a proportion of approximately 0.23 in the CD group and 0.13 in the HFHSD group. Goblet cells accounted for approximately a protion of 0.2. Heatmaps of marker gene expression for CD and HFHSD conditions (Figure 1e) illustrated distinct transcriptional profiles for each of the 12 identified clusters. The heatmaps revealed both shared and diet-specific gene expression patterns across cell types, with certain genes showing differential expression between CD and HFHSD conditions within the same cell type. These data exhibit the diverse cellular heterogeneity within the intestinal epithelium as well as some HFHSD induced alterations in cell type proportions and expression profiles.

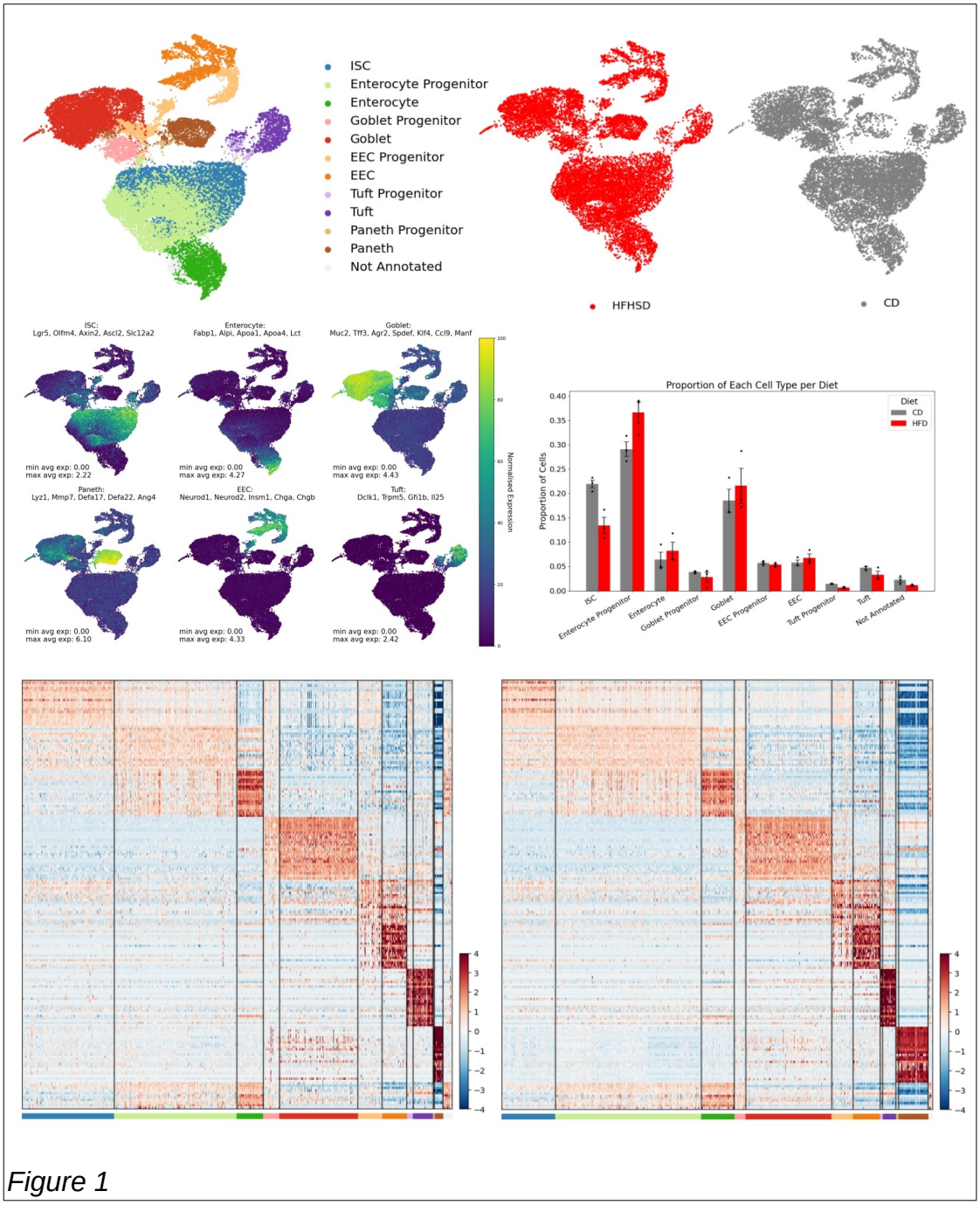

ISC
Enterocyte Progenitor
Enterocyte
Goblet Progenitor
Goblet
EEC Progenitor
EEC
Tuft Progenitor
Tuft
Paneth Progenitor
Paneth
Not Annotated
HFHSD
CD
ISC:
Lgr5, Olfm4, Axin2, Ascl2, Slc12a2
Enterocyte:
Fabp1, Alpi, Apoa1, Apoa4, Lct
Goblet:
Muc2, Tff3, Agr2, Spdef, Klf4, Ccl9, Manf
Paneth:
Lyz1, Mmp7, Defa17, Defa22, Ang4
EEC:
Neurod1, Neurod2, Insm1, Chga, Chgb
Tuft:
Dclk1, Trpm5, Gfi1b, Il25
Normalised Expression
Proportion of Each Cell Type per Diet
Proportion of Cells
Diet
CD
HFD

Figure 1

High-Fat High-Sugar Diet Alters ISC Function

Differential gene expression analysis of intestinal stem cells revealed a total of 337 differentially expressed genes between the CD and HFHSD conditions (Figure 2). The top 10 highest log2 fold change significantly DEGs were *Fabp1*, *Mgll*, *H2-Aa*, *Creb3l3*, *Etv4*, *Slc16a12*, *H2-Ab1*, *H2-Eb1*, *Cst6*, and *Scd2*. The top 10 lowest log2 fold change significantly DEGs were *Sepp1*, *Sis*, *S100g*, *Mtus2*, *Anpep*, *Emp2*, *AA467197*, *Lzts3*, *Prr15*, *Zfp467*. Trends were observed in genes relating to both fat metabolism and cell cycle regulation. Notably, genes involved in fatty acid metabolism and transport showed consistent upregulation. Among these, *Fabp1* stood out with one of the highest positive log2 fold change (log2FC) of approximately 2.7, while *Fabp2*, a key intracellular protein involved in the uptake and transport of long-chain fatty acids, also displayed significantly increased expression albeit with a lower log2FC of approximately 0.6. Similarly, *Hmgcs2*, a mitochondrial enzyme essential for the first steps of ketogenesis, exhibited higher expression levels under the HFHSD with a log2FC of around 1. Upregulation was also seen in *Acot1*, an enzyme that facilitates the conversion of acyl-CoAs into fatty acids and CoA, and *Scd2*, which plays a role in fatty acid biosynthesis both highly significant with log2FC values around 1. The enzyme encoded by *Mgll*, responsible for breaking down monoacylglycerols into free fatty acids, also showed increased expression with a log2FC over 2.5.

In addition to changes in metabolism-related genes, several genes associated with cell cycle regulation and proliferation were upregulated. *Hsp90aa1*, a molecular chaperone known to be involved in cell cycle control, showed a marked increase in expression. *Nme1*, involved in nucleoside triphosphate synthesis. In contrast, *Zfas1*, a non-coding RNA linked to cellular differentiation, was downregulated, as was *Sepp1*, a protein that plays a role in redox balance and epithelial cell proliferation. Additionally, the tumor suppressor *Pdcd4* showed significant downregulation, along with *Ypel3*, a gene associated with the regulation of cellular senescence.

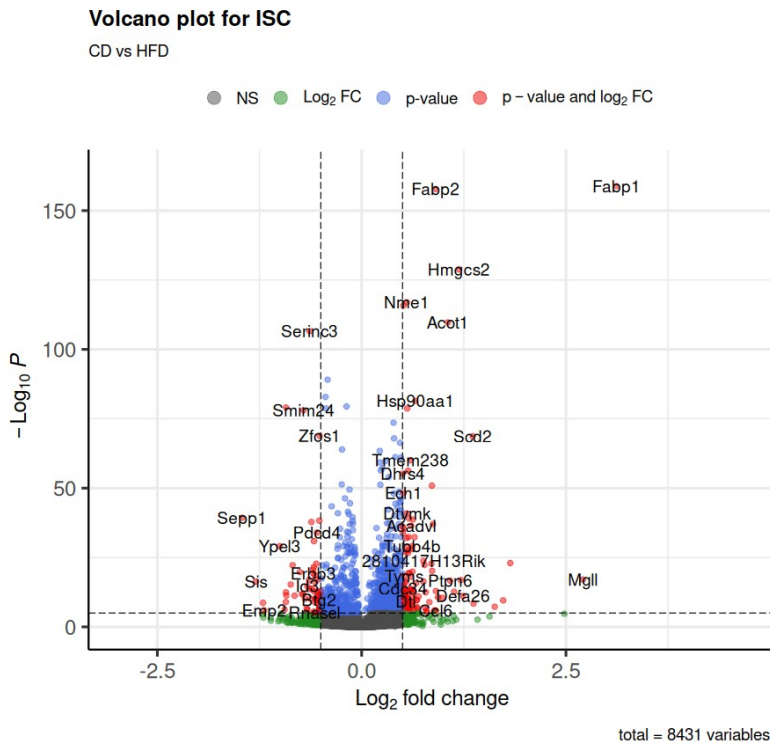


Figure 2

Gene Ontology Enrichment Analysis

GO enrichment analysis of DEGs in ISCs exposed to a HFHSD revealed significant alterations across CC, MF, and BP categories (Figure 3). The analysis revealed 16 distinct clusters of enriched GO terms and demonstrated a trend in the upregulation of cell cycle-related processes consistent with findings described from volcano plot analysis. As the top 50 GO terms from each class were selected by adjusted p-value, all terms are highly significant.

Cluster 3 in the BP category displayed enrichment for terms associated with chromosome separation, mitotic cell cycle, and cell cycle regulation. This pattern was mirrored in the CC section, where the enrichment of terms in the same cluster were related to chromosomes and mitotic spindles. Complementing these findings, cluster 1 showed enrichment of DNA replication GO terms in the BP category, corresponding to replication fork-related terms in the CC group. The MF category further supported this trend, with upregulation of terms associated with helicase activity, ATP hydrolysis acting on DNA, and single-strand DNA binding. Each of these GO terms were positively enriched with NESs between 1.5 and 2.5.

Mitochondrial function also appeared significantly enriched, as evidenced by cluster 6. This cluster showed upregulation of mitochondrial translation and gene expression in the BP category, while the CC section indicated increased expression of mitochondrial ribosome-related terms.

RNA processing mechanisms were positively. Cluster 5 revealed upregulation of spliceosome and ribonucleoprotein complex-related terms, particularly in protein-RNA complex assembly. Similarly, cluster 2 showed increased ribonucleoprotein complex biogenesis, with corresponding upregulation of snoRNA binding in the MF category.

Cluster 12 showed decreased DNA-binding transcription factor activity specific to RNA polymerase II.

Cluster 4 indicated downregulation in genes related to adherens junctions in the CC category, with associated decreases in transmembrane protein kinase activity and cell adhesion molecule binding in the MF category. All related GO terms were found in the -1.5 to -2 NES range.

Overall this GO enrichment analysis of ISCs under prediabetic HFHSD conditions, revealed trends in the broad upregulation of processes related to cell cycle progression, DNA replication, mitochondrial function, and RNA processing. Moreover, it showed downregulation in specific transcription factor activities and cell adhesion processes.

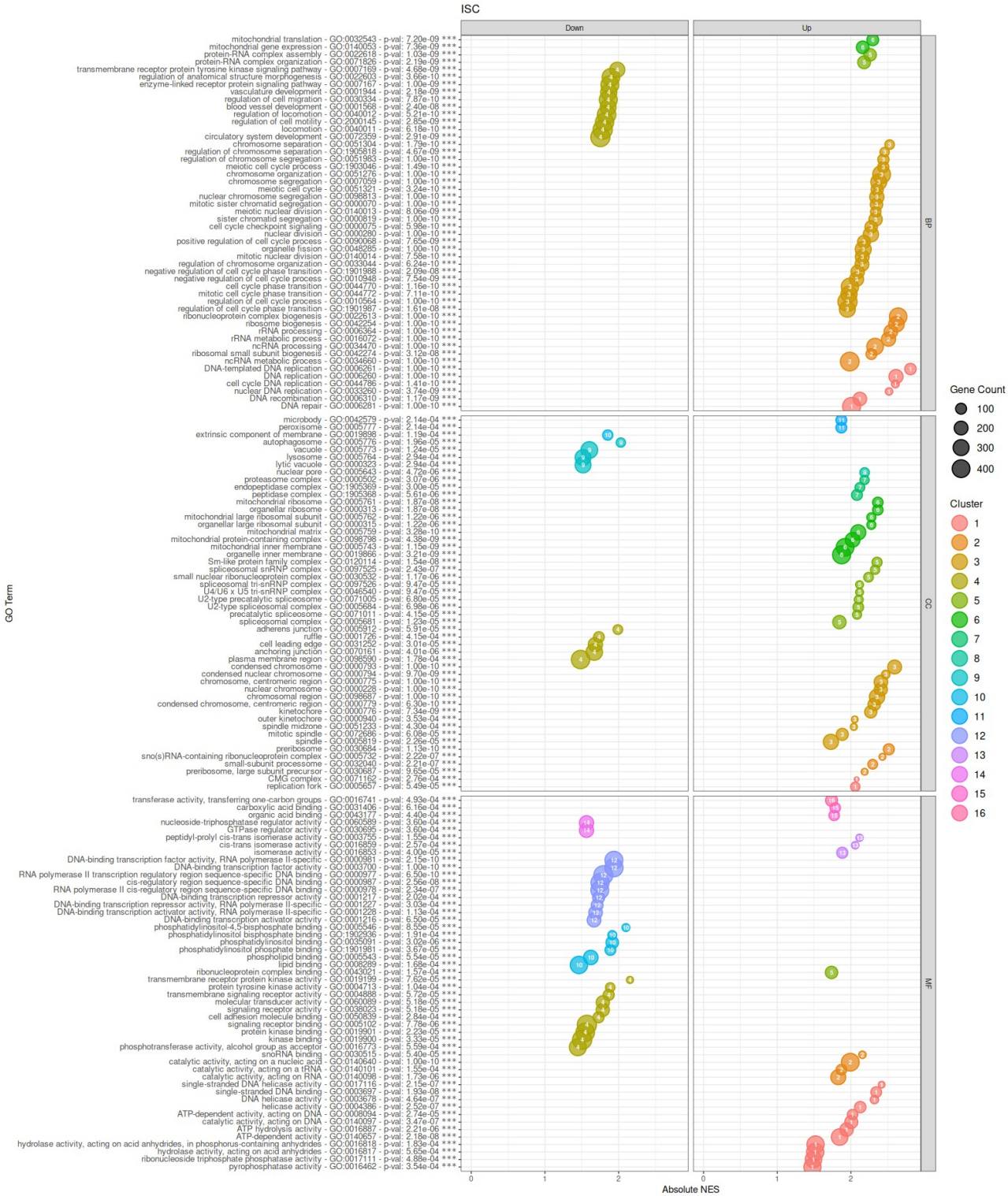


Figure 3

The Gene Ontology enrichment network analysis (Figure 4) further corroborated and expanded upon the findings observed in the dotplot visualisation. This network

representation integrated Cellular Component, Biological Process, and Molecular Function categories, revealing 33 distinct clusters of functionally related GO terms.

A prominent cluster centered around the "regulation of mitotic sister chromatid segregation" GO term emerged, exhibiting positive Normalised Enrichment Scores (NES). This cluster encompassed various cell cycle-related processes, reinforcing the earlier observation of cell cycle upregulation in ISCs under HFHSD. Similarly, a separate cluster with most edges leading to the "double-strand break repair via homologous recombination" GO term also displayed positive NES values, further emphasising the enrichment of DNA replication and repair processes.

Conversely, a cluster associated with "brush border" and "regulation of developmental process" GO terms showed negative NES values. This cluster is consistent with the previous findings relating to the downregulation of adherens junction-related GO terms, providing additional evidence for altered cell adhesion and developmental processes in the ISCs.

The enrichment map also highlighted several other functional clusters. Notably, terms related to "mitochondrial translation" and "ribosomal large subunit biogenesis" formed distinct clusters with positive NES values, aligning with the earlier findings of upregulated mitochondrial and protein synthesis processes. Additionally, clusters associated with "fatty acid binding" and "carboxylic acid metabolic process" were observed, indicating significant alterations in lipid metabolism pathways.

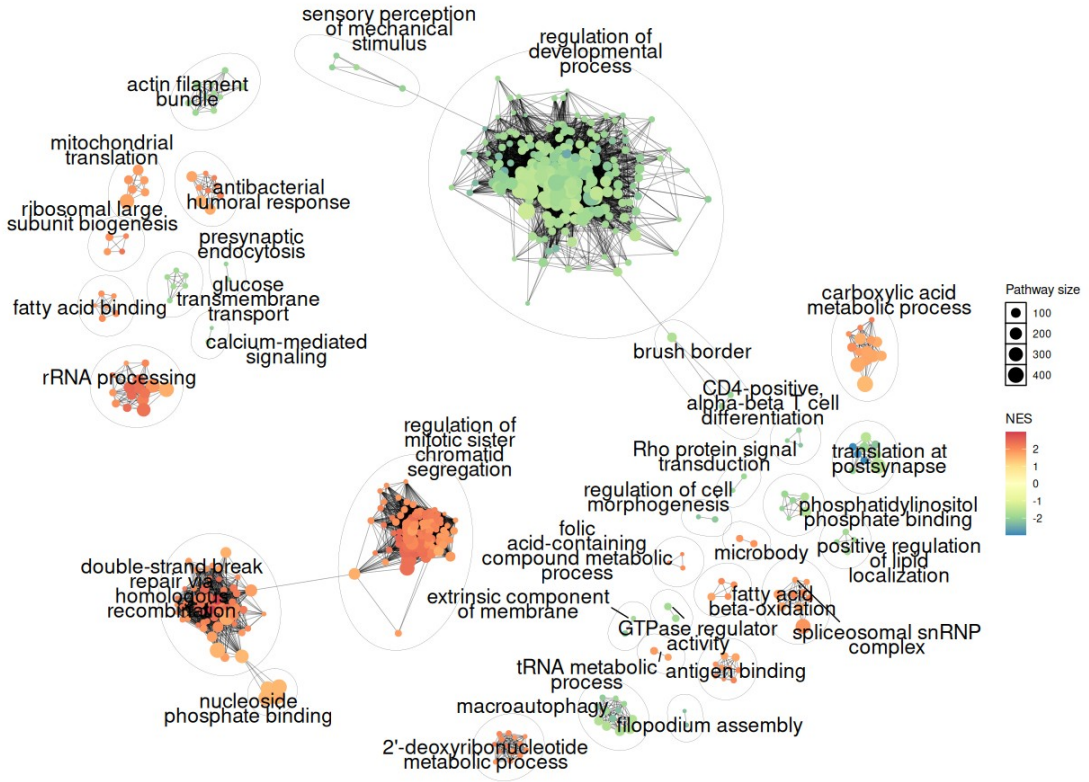


Figure 4

Building upon the GO enrichment analysis, KEGG pathway enrichment analysis of ISCs further revealed the functional alterations associated with the prediabetic state (Figure 5). The analysis found 23 significant enrichment of pathways. 5 of these pathways were related to genetic information processing, each with a spread of DEGs in the upregulated region, with DNA replication showing high significance (p = 3.57e-04) and a high gene ratio of approximately 0.9. This was complemented by enrichment of related pathways including ribosome biogenesis, mismatch repair, and spliceosome. 3 pathways were enriched in the cellular pathways category with the cell cycle pathway most significantly enriched (p = 6.33e-03), and with most significant DEGs upregulated (gene ratio approximately 0.5), reinforcing the earlier observations of cell cycle upregulation. Notably, the cell adhesion molecules pathway was enriched (p = 1.29e-02) with a predominance of downregulated genes. The analysis also highlighted alterations in the endocrine system, with significant enrichment of the PPAR signaling pathway (p = 1.24e-06), with most of the significant DEGs in the positive log2 fold change region. Furthemore, the fatty acid metabolism pathway was significantly enriched (p = 2.08e-02) with many of its genes

upregulated, coinciding with the alterations revealed in the DEG and GO analyses. Interestingly, several immune-related pathways showed enrichment, including the Staphylococcus aureus infection pathway (p = 1.23e-07) and the antigen processing and presentation pathway (p = 8.69e-03), with many of their respective significant DEGs in displayed in the positive og2 fold change region.

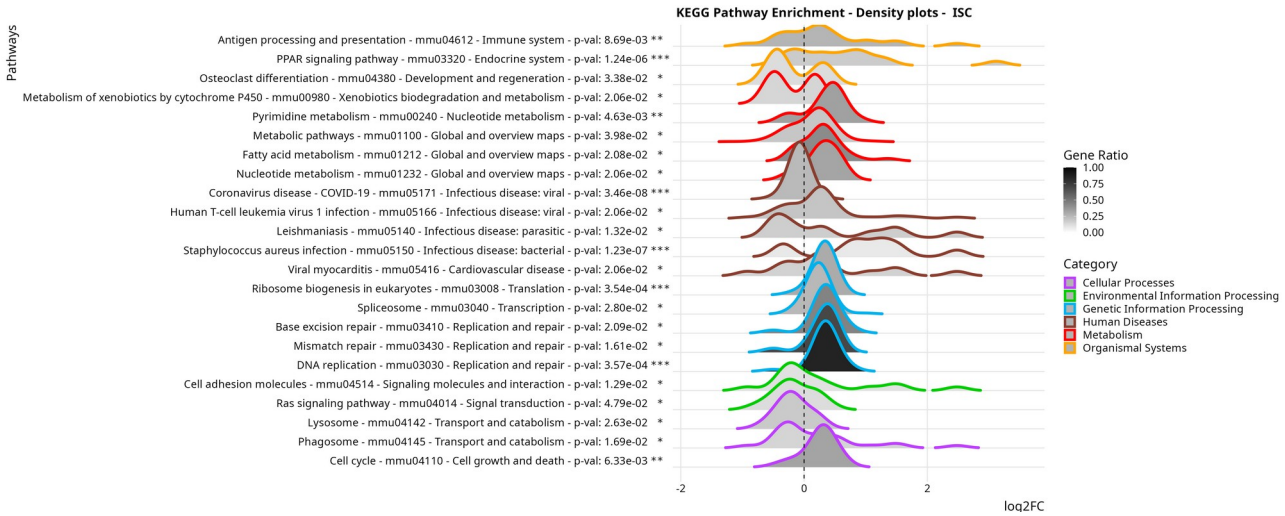


Figure 5

Further examination of the cell cycle pathway through KEGG graphs and heatmaps of log 2 fold change values of DEGs provided additional insights into the mechanisms of cell cycle alterations in the ISCs under prediabetic conditions (Figure 6). The heatmap analysis revealed that approximately 52% of genes involved in the cell cycle were reported significantly differentially expressed in the epithelium with ISCs and enterocyte progenitors being the most affected cell types within the pathway. ISCs displayed widespread upregulation of cell cycle-related genes, with a slightly attenuated signal observed in enterocyte progenitors. This trend builds on the current evidence of the rapid proliferation of ISCs followed by differentiation into enterocyte cell types, consistent with the cell type proportions seen in Figure 1d. The KEGG graph visualisation of the cell cycle pathway highlighted significant upregulation of key cyclins, including Cyclin D, E, A, and B, as well as Cyclin-dependent kinase 1 (CDK1). The graph also indicated increased expression of CDK45, with the regulatory signals predominantly tending towards DNA replication upregulation. These observations at the pathway level corroborate and extend the earlier findings of cell cycle upregulation in ISCs, providing a more detailed view of the specific cell cycle components and mechanisms affected in the prediabetic intestinal epithelium.

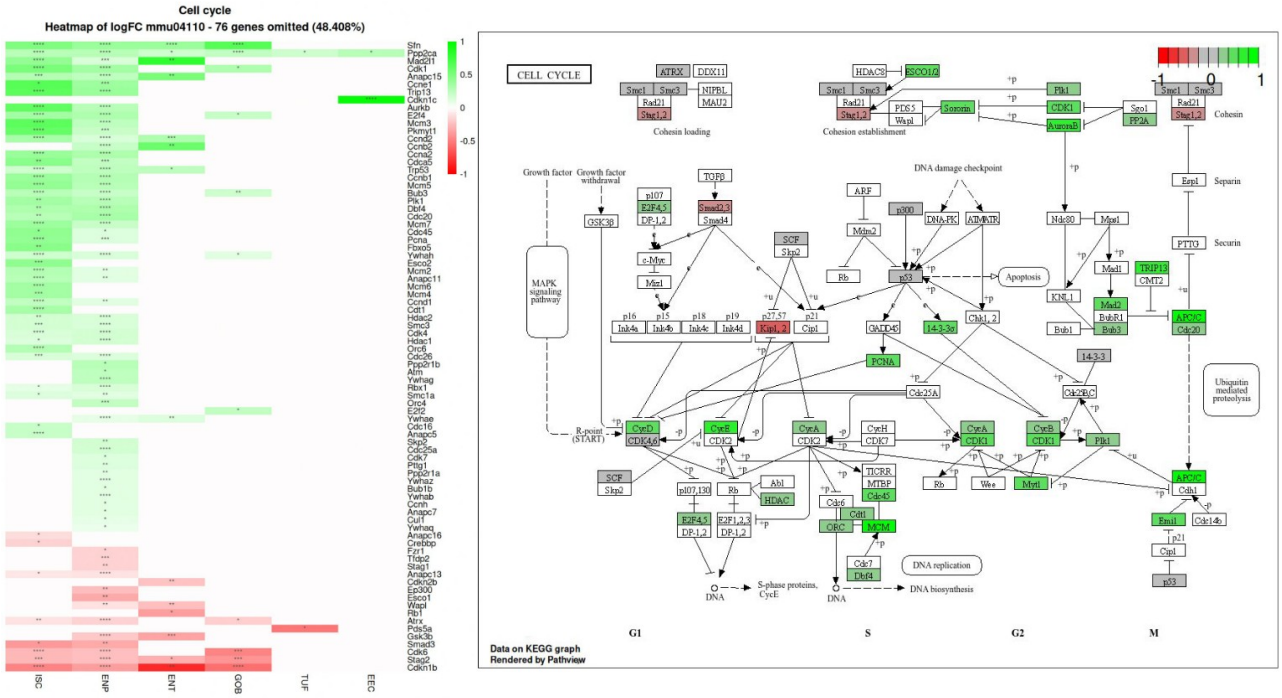


Figure 6

High-Fat High-Sugar Diet Alters Enterocyte Progenitor Function

Differential gene expression analysis of enterocyte progenitors in the intestinal epithelium revealed a total of 339 significantly differentially expressed genes between the CD and HFHSD conditions (Figure 7). The top 10 highest log2 fold change significantly DEGs were *Cyp4a10*, *Fabp1*, *Mgll*, *Creb3l3*, *Acot1*, *Me1*, *Hmgcs2*, *Akr1b8*, *Acaa1b*, *Acacb*. The top 10 lowest log2 fold change significantly DEGs were *Sepp1*, *Sis*, *Gngt2*, *Sh3bgr*, *Mfsd4*, *S100g*, *Sectm1b*, *AA467197*, *Abi3*, *Tnfsf10*. Trends were observed in genes relating to in nutrient and energy utilisation pathways between CD and prediabetic HFHSD conditions, which were closely related to lipid metabolism processes. Genes involved in fatty acid uptake and intracellular transport showed marked upregulation. *Fabp1* and *Fabp2*, encoding fatty acid binding proteins crucial for fatty acid absorption in enterocytes, exhibited log2 fold changes approximately 3.5 and 0.7 respectively, similar to what was seen in the ISC population. Hmgcs2, a key enzyme in ketogenesis, also showed significant upregulation with a log2 fold change value of approximately 2.

The prediabetic model further induced upregulation of genes associated with mitochondrial fatty acid metabolism and beta-oxidation. *Ech1* (enoyl-CoA hydratase), *Acaa2* (acetyl-CoA

acyltransferase 2), *Etfb* (electron transfer flavoprotein beta subunit), and *Acad1* (acyl-CoA dehydrogenase) all showed increased expression, all related in theircapacity for fatty acid breakdown. *Me1*, encoding malic enzyme 1, which generates NADPH for fatty acid biosynthesis, was also upregulated. Additionally, *Cyp4a10*, involved in fatty acid omega-oxidation, showed increased expression, suggesting activation of alternative lipid metabolism pathways.

Conversely, several genes related to carbohydrate metabolism and cellular stress responses were downregulated in the HFHSD condition. *Aldob*, encoding fructose-bisphosphate aldolase B, a key enzyme in glycolysis and fructose metabolism, showed reduced expression. *Oat*, encoding ornithine aminotransferase, a mitochondrial enzyme involved in amino acid metabolism, was also downregulated.

Genes involved in cellular stress responses and redox regulation also showed decreased expression. *Sepp1*, which encodes selenoprotein P, an extracellular antioxidant, and Txnip, a regulator of cellular redox state and glucose metabolism, both exhibited downregulation.

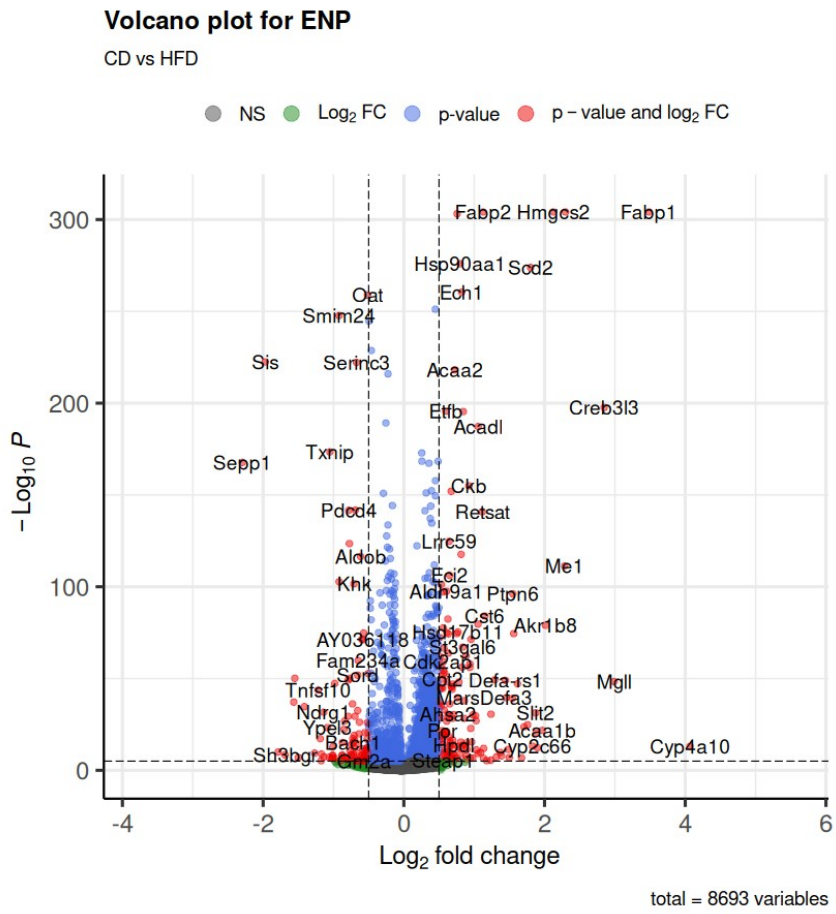


Figure 7

GO enrichment analysis of DEGs in enterocyte progenitors revealed extensive alterations across BP, CC, and MF categories in response to the prediabetic HFHSD model (Figure 8). The analysis uncovered 22 distinct clusters of enriched GO terms with trends in x y and z consistent with findings of the volcano plot. As seen in the previous GO enrichment analysis of ISCs, in this analysis of enterocyte progenitors, the top 50 GO terms from each class were selected by adjusted p-value, and thus all terms are highly significant.

Cluster 3 emerged as a focal point of upregulation, predominantly in lipid metabolism processes. This cluster showed enrichment in fatty acid beta-oxidation, lipid catabolism, fatty acid metabolism, and monocarboxylic acid transport. Notably, long-chain fatty acid metabolic process and fatty acid oxidation terms displayed high enrichment scores. These processes were primarily associated with the peroxisome cellular component. The cluster also revealed upregulation of molecular functions crucial for lipid processing, including

carboxylic ester hydrolase activity, fatty acyl-CoA hydrolase activity, NADP binding, and various hydrolase and oxidoreductase activities.

In contrast, cluster 5 exhibited consistent downregulation across all three GO categories, suggesting a coordinated change in cell adhesion and signaling. Biological processes affected included transmembrane receptor protein tyrosine kinase signaling pathway, positive regulation of cell migration, cell junction organisation, and actin filament-based processes. The analysis revealed downregulation of cellular components including adherens junctions, focal adhesions, and the actin cytoskeleton. Corresponding molecular functions, such as protein kinase activity, transmembrane signaling receptor activity, and cell adhesion molecule binding, were also downregulated.

Several other smaller clusters also displayed HFHSD induced alterations in enterocyte progenitor functions. Cluster 22 displayed a downregulation of calcium-dependent protein binding molecular functions, affected calcium-mediated cellular processes. Cluster 21 indicated reduced cysteine-type endopeptidase regulator activity involved in apoptotic processes. Cluster 20 revealed downregulation of beta-catenin binding molecular function. Additionally, Cluster 16 demonstrated a set of downregulated GO terms associated with DNA-binding transcription repressor activity, including specific RNA polymerase II-related transcription factor activities.

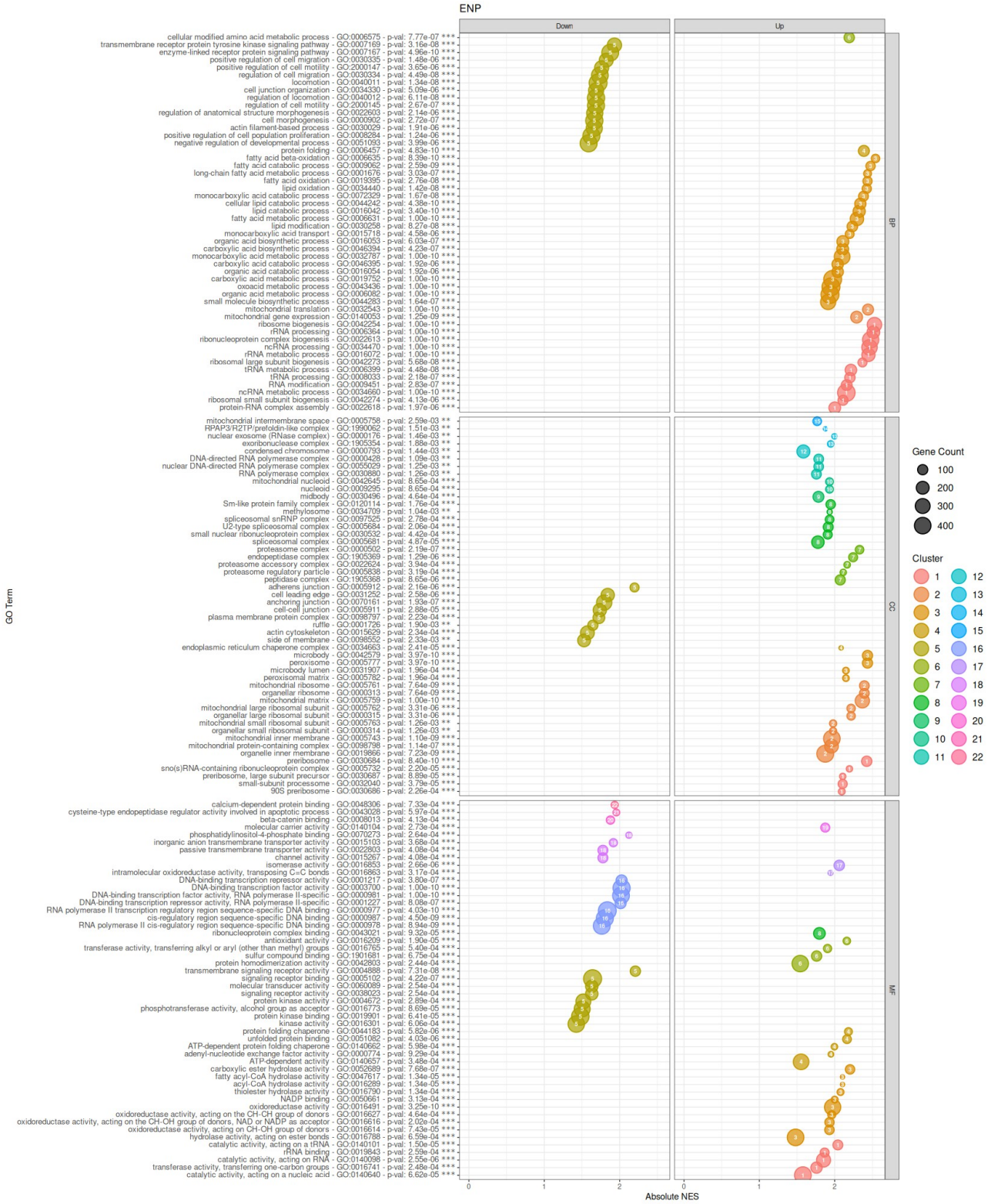

ENP
Down
Up
BP
CC
MF
GO Term
Absolute NES
Gene Count
100
200
300
400
Cluster
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Figure 8

KEGG pathway enrichment analysis of enterocyte progenitors revealed 32 significantly enriched pathways across various categories (Figure 9). The figure displays 32 significantly enriched KEGG pathways from various categories such as cellular processes, Environmental information processing, genetic information processing, human diseases, metabolism, and organismal systems. Notably, the peroxisome pathway from the cellular processes category is enriched, showing a wide distribution of upregulated DEGs with a gene ratio of approximately 0.5. The Notch signaling pathway is also significantly enriched, exhibiting the highest gene ratio in its category and a p-value of 2.78e-02.

In the genetic information processing category, the proteasome pathway and protein processing in the endoplasmic reticulum show significant enrichment, with high gene ratios (approximately 0.9 for the proteasome and approximately 0.6 for protein processing in the ER). The proteasome pathway has a p-value of 5.47e-03, while the ER protein processing pathway has a p-value of 2.20e-02.

In the metabolism category, 10 pathways are enriched, all showing a predominance of upregulated DEGs. Of these, six belong to the lipid metabolism subcategory, each displaying high gene ratios. Retinol metabolism is particularly significant, with a p-value of 1.64e-04. In the amino acid metabolism subcategory, glutathione metabolism is enriched with a p-value of 3.52e-02, alongside the valine, leucine, and isoleucine degradation pathway, which has a p-value of 1.14e-02.

Additionally, 10 enriched pathways fall under the organismal systems category, with two in the digestive system, six in the endocrine system, and the remaining two in the sensory and immune system subcategories. The fat digestion and absorption pathway is highly significant (p=7.00e-04) and displays a majority of genes in the upregulated region, while the carbohydrate digestion and absorption pathway (p=1.18e-02) predominantly shows downregulated genes. Within the endocrine system, the PPAR signaling pathway is the most significant, with an extremely low p-value of 1.48e-08.

These trends display vast alterations in metabolism in the prediabetic enterocyte progenitor population with widespread upregulation of lipid metabolism and digestion pathways. Furthermore, upregulatory trends are apparent in the endocrine system pathways, and protein folding, sorting and degradation pathways. Moreover, the prediabetic enterocyte population display downregulatory trends in notch signalling, and carbohydrate digestion.

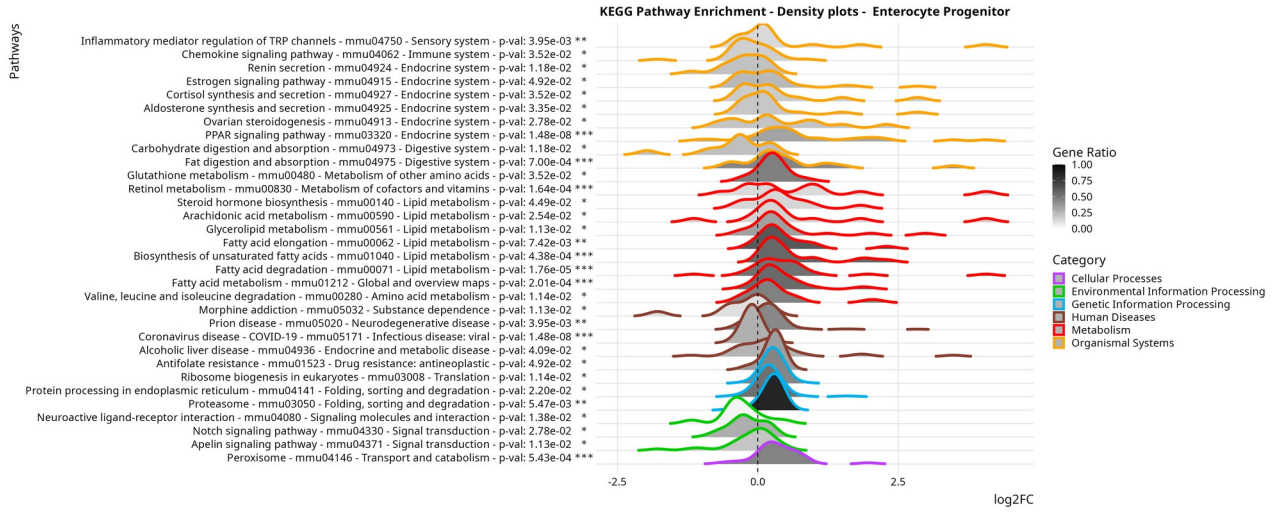


Figure 9

Further examination of the tight junction pathway through KEGG graphs and heatmaps of log 2 fold change values of DEGs provided additional insights into the mechanisms and components altered under the prediabetic condition (Figure 10).

In enterocyte progenitors, several claudin genes showed notable downregulation, particularly *Cldn15*, *Cldn4*, and *Cldn3*. The mature enterocyte population exhibited downregulation of *Cldn23*. Genes related to actin cytoskeleton also displayed significant downregulation in both enterocyte and enterocyte progenitor populations, with *Actb*, *Actg1*, and *Actn4* showing the most pronounced decreases.

Conversely, genes encoding alpha-tubulin proteins showed upregulatory trends across multiple cell types. *Tuba4a* and *Tuba1c* were significantly upregulated in most cell populations examined. *Tuba1b* showed specific upregulation in the ISC and enterocyte progenitor populations.

The KEGG graph visualisation of the tight junction pathway (Figure 10) provided further insights into the altered gene expression patterns. It highlighted the upregulation of PP2A, known to interact with occludin. The graph also illustrated downregulation of claudins in the tight junctional space. Furthermore, the visualisation indicated decreased expression in genes associated with cell polarity and paracellular permeability mechanisms within the tight junction complex. Collectively, these results demonstrate substantial changes in the expression of tight junction-related genes in the intestinal epithelium under HFHSD

conditions, with distinct patterns observed across different cell types and specific pathway components.

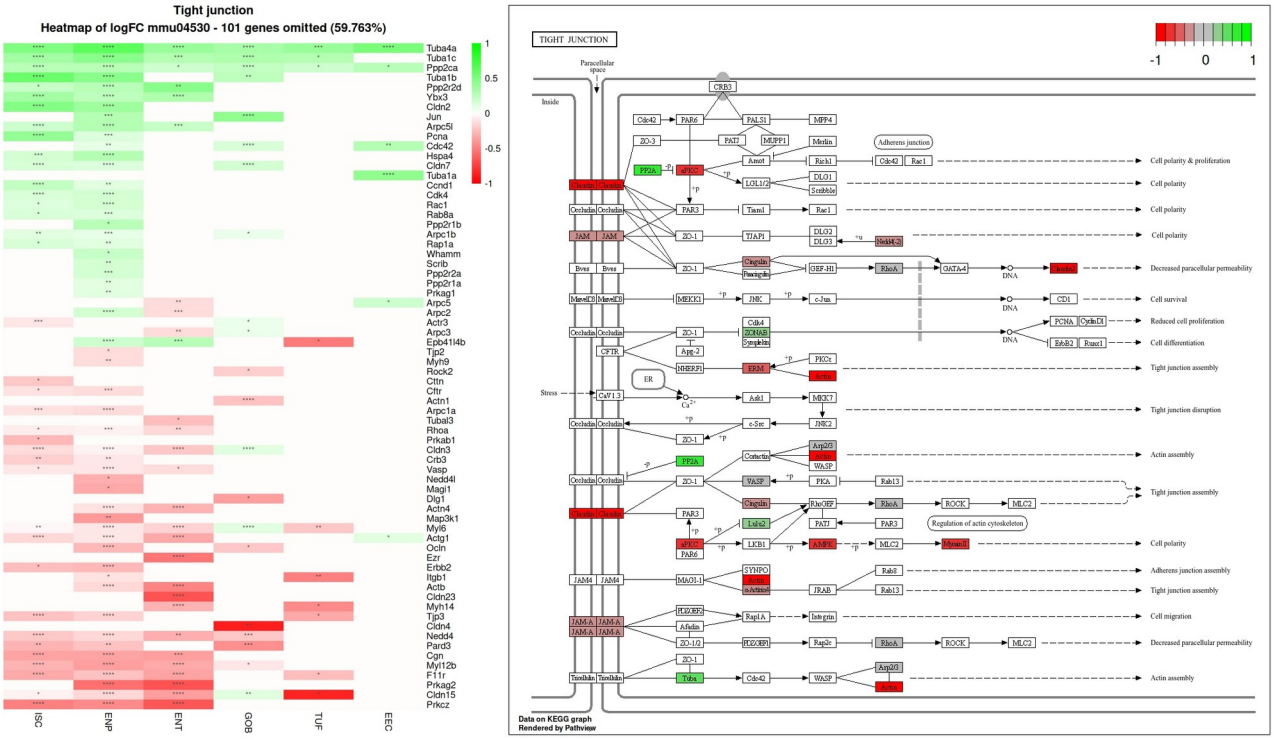


Figure 10

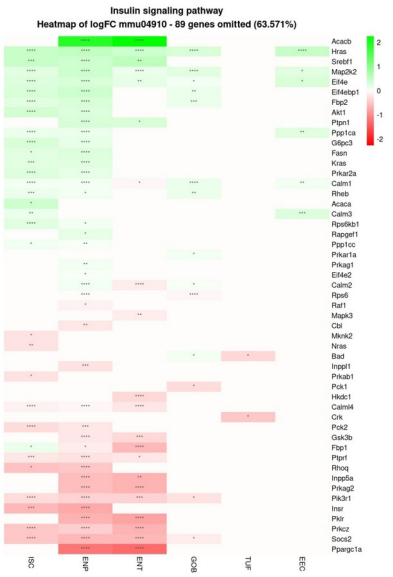
Analysis of the insulin signaling pathway, carbohydrate digestion and absorption, and fatty acid degradation pathways revealed significant alterations in gene expression patterns across different cell types in response to the HFHSD condition (Figure 11).
In the insulin signaling pathway, notable changes were observed in genes related to fatty acid biosynthesis and glycolysis/gluconeogenesis. The AMPK protein showed downregulation, while the Acetyl-CoA carboxylase (ACC) protein exhibited upregulation. Sreb-1c gene expression was increased in ISCs, enterocyte progenitors, and enterocytes, coinciding with decreased expression of the pyruvate kinase (PYK) protein. The *Hras* gene, involved in proliferation and differentiation pathways, showed upregulation in these cell types.

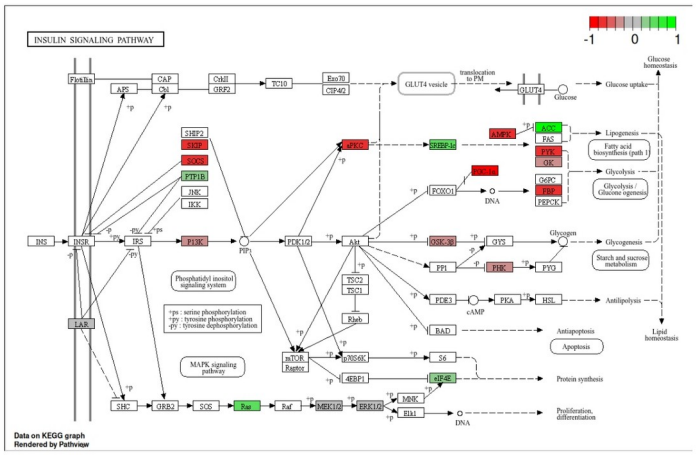
The carbohydrate digestion and absorption pathway analysis revealed significant changes, particularly in enterocytes. The *Sis* gene was markedly downregulated across all cell types except enteroendocrine cells (EECs). The *Slc2a2* gene, encoding the GLUT2 protein, a

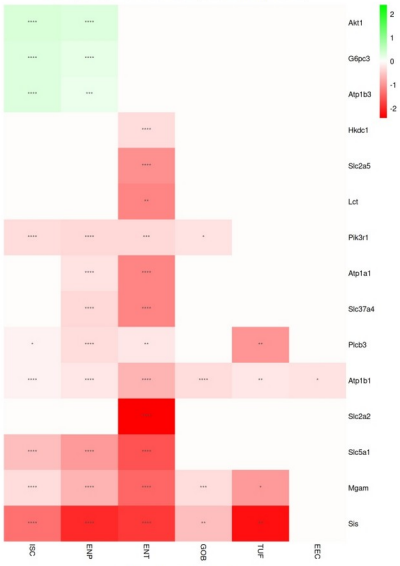
glucose/fructose transporter, showed substantial downregulation in the heatmap and KEGG graph visualisations.

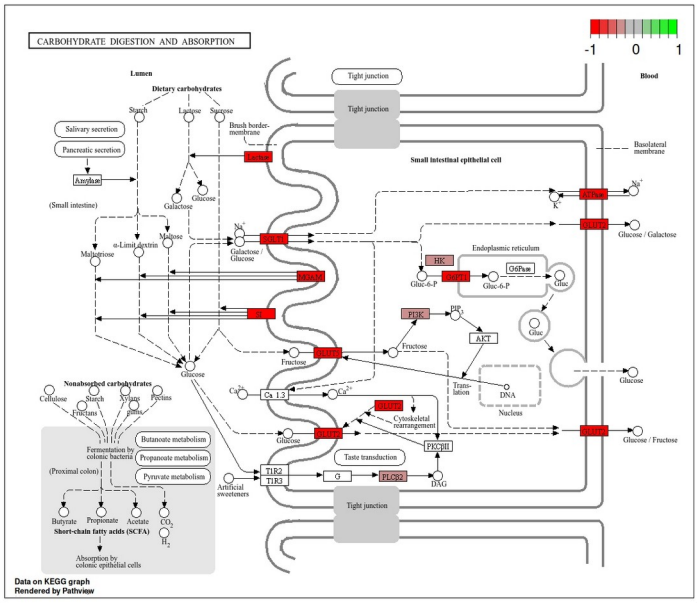
Examination of the fatty acid degradation pathway in enterocytes revealed widespread changes in enzyme expression. Upregulated enzymes included long-chain fatty acid-CoA ligase (EC:6.2.1.3), carnitine O-palmitoyltransferase 1 and 2 (CPT1, CPT2), fatty acyl-CoA oxidase (EC:1.3.3.6), long-chain acyl-CoA dehydrogenase (EC:1.3.8.8), very-long-chain acyl-CoA dehydrogenase (EC:1.3.8.9), short-chain acyl-CoA dehydrogenase (EC:1.3.8.1), acetyl-CoA C-acyltransferase (EC:2.3.1.16), alcohol dehydrogenase (EC:1.1.1.1), aldehyde dehydrogenase (NAD+) (EC:1.2.1.3), and long-chain fatty acid omega-monooxygenase (EC:1.14.14.80). Conversely, Delta3-Delta2-enoyl-CoA isomerase (EC:5.3.3.8) and alkane 1-monooxygenase (EC:1.1.1.35) showed downregulation.

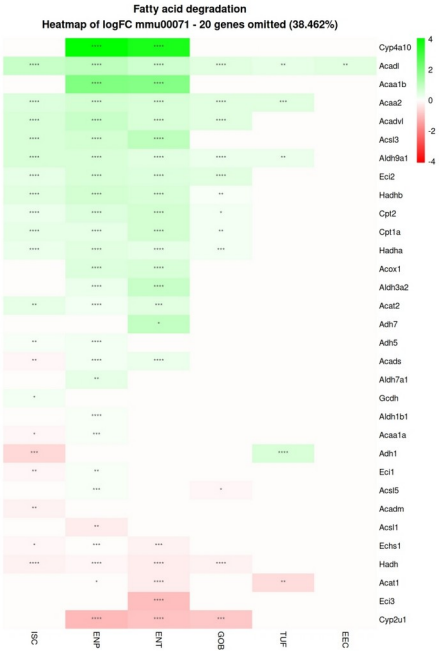
These results collectively demonstrate substantial alterations in gene expression patterns related to insulin signaling, carbohydrate metabolism, and fatty acid degradation in the intestinal epithelium under HFHSD conditions, with distinct changes observed across different cell types and specific pathway components.


Insulin signaling pathway
Heatmap of logFC mmu04910 - 89 genes omitted (63.571%)
Acacb
Hras
Srebf1
Map2k2
Eif4e
Eif4ebp1
Fbp2
Akt1
Ptpn1
Ppp1ca
G6pc3
Fasn
Kras
Prkar2a
Calm1
Rheb
Acaca
Calm3
Rpsl6kb1
Rapgef1
Ppp1cc
Prkar1a
Prkag1
Eif4e2
Calm2
Rps6
Raf1
Mapk3
Cbl
Mknk2
Nras
Bad
Inppl1
Prkab1
Pck1
Hkdc1
Calml4
Grk
Pck2
Gsk3b
Fbp1
Ptprf
Rhoq
Inpp5a
Prkag2
Pik3r1
Insr
Pkiv
Prkcz
Socs2
Ppargc1a
ISC
ENP
ENT
GOB
TUF
EEC


INSULIN SIGNALING PATHWAY
Data on KEGG graph
Rendered by Pathview


Carbohydrate digestion and absorption
Heatmap of logFC mmu04973 - 34 genes omitted (69.388%)
Akt1
G6pc3
Atp1b3
Hkdc1
Slc2a5
Lct
Pik3r1
Atp1a1
Slc37a4
Plcb3
Atp1b1
Slc2a2
Slc5a1
Mgam
Sis
ISC
ENP
ENT
GOB
TUF
EEC


CARBOHYDRATE DIGESTION AND ABSORPTION
Data on KEGG graph
Rendered by Pathview


Fatty acid degradation
Heatmap of logFC mmu00071 - 20 genes omitted (38.462%)
Cyp4a10
Acadl
Acaa1b
Acaa2
Acadvl
Acsl3
Aldh9a1
Eci2
Hadhb
Cpt2
Cpt1a
Hadha
Acox1
Aldh3a2
Acat2
Adh7
Adh5
Acads
Aldh7a1
Gcdh
Aldh1b1
Acaa1a
Adh1
Eci1
Acsl5
Acadm
Acsl1
Echs1
Hadh
Acat1
Eci3
Cyp2u1
ISC
ENP
ENT
GOB
TUF
EEC

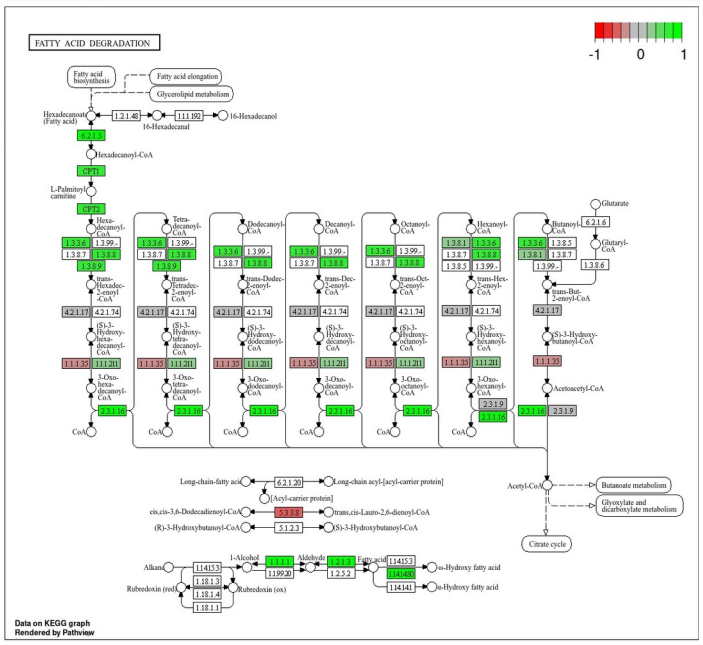

FATTY ACID DEGRADATION
Data on KEGG graph
Rendered by Pathview

 Figure 11

Alterations in the Endoplasmic Reticulum and Proteasome

The KEGG graph for the enterocyte progenitors highlights widespread upregulation across the protein processing pathway in the endoplasmic reticulum (ER), with notable enhancements in the ubiquitin ligase complex, ER-associated degradation (ERAD), and protein recognition by luminal chaperones (Figure 12). Within the ubiquitin ligase complex, proteins such as Hsp40, Hsp70, and CHIP are significantly upregulated. CHIP, known for its dual role as a co-chaperone and E3 ubiquitin ligase, is crucial for the degradation of chaperone-bound misfolded proteins, indicating increased proteostasis in response to ER stress. In ERAD and chaperone recognition, key upregulated proteins include Hsp70, Hsp90, Hsp40, and nucleotide-exchange factor (NEF), encoded by *Bag1*, which facilitates the release of ADP from HSP70 proteins, thus promoting protein folding and release. Additional upregulation is seen in Derlin, encoded by *Derl2*, which plays a role in the degradation of misfolded glycoproteins, and DOA1, encoded by *Plaa*, which is involved in ubiquitin-mediated protein degradation.

The accompanying heatmap reinforces this upregulation, showing increased expression levels across key genes involved in protein processing within the ER pathway. Genes such as *Hspa1a*, *Hspa1b*, *Hsp90aa1*, *Hsph1* exhibit marked upregulation, particularly in the enterocyte progenitors and enterocytes populations.

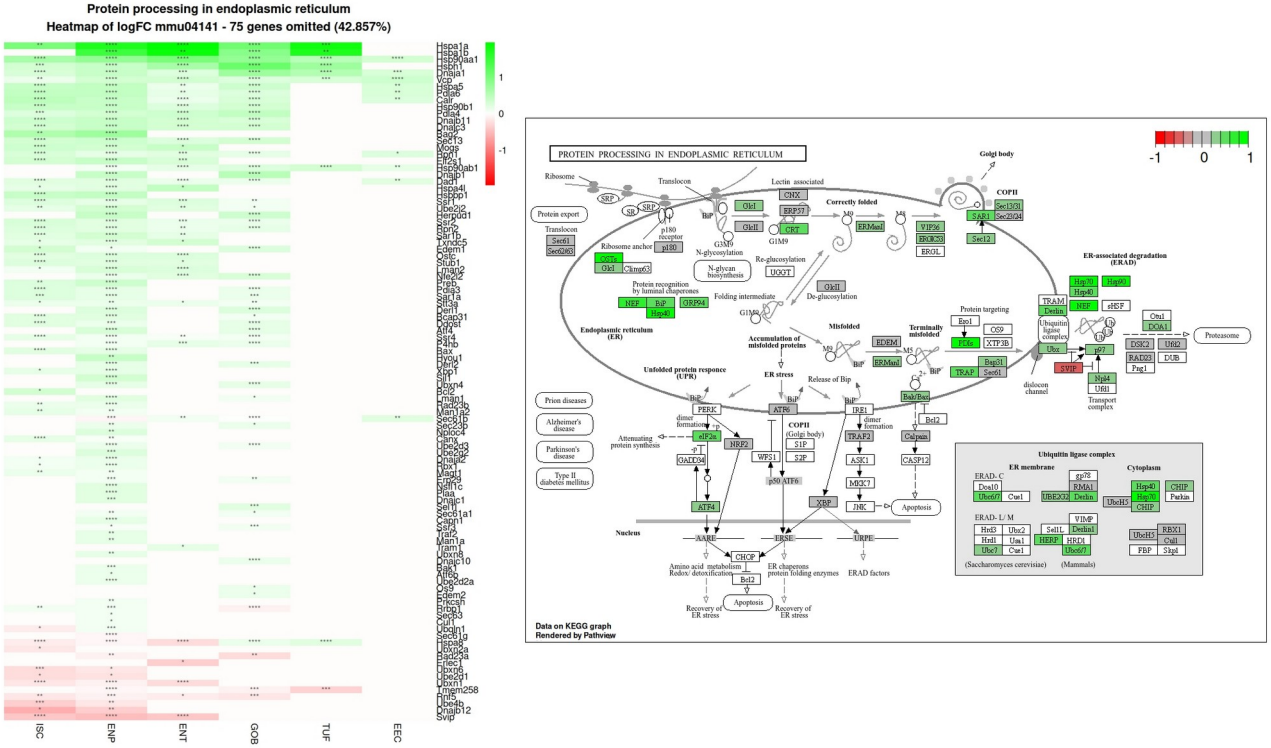


Figure 12

Following on from the ER response in protein processing, the proteasome pathway also exhibits extensive upregulation across all cell types, particularly in the 20S core particle, which involves both the alpha and beta ring subunits (Figure 13). This upregulation is mirrored in the 19S regulatory particle, especially in the PA700 lid and PA700 base subunits, indicating a broad activation of proteasome-mediated protein degradation via the HFHSD condition. Contrastingly, there is a slight downregulation observed in the immunoproteasome subunit, specifically the β5i subunit encoded by Psmb8, within the enterocyte progenitors and enterocytes populations.

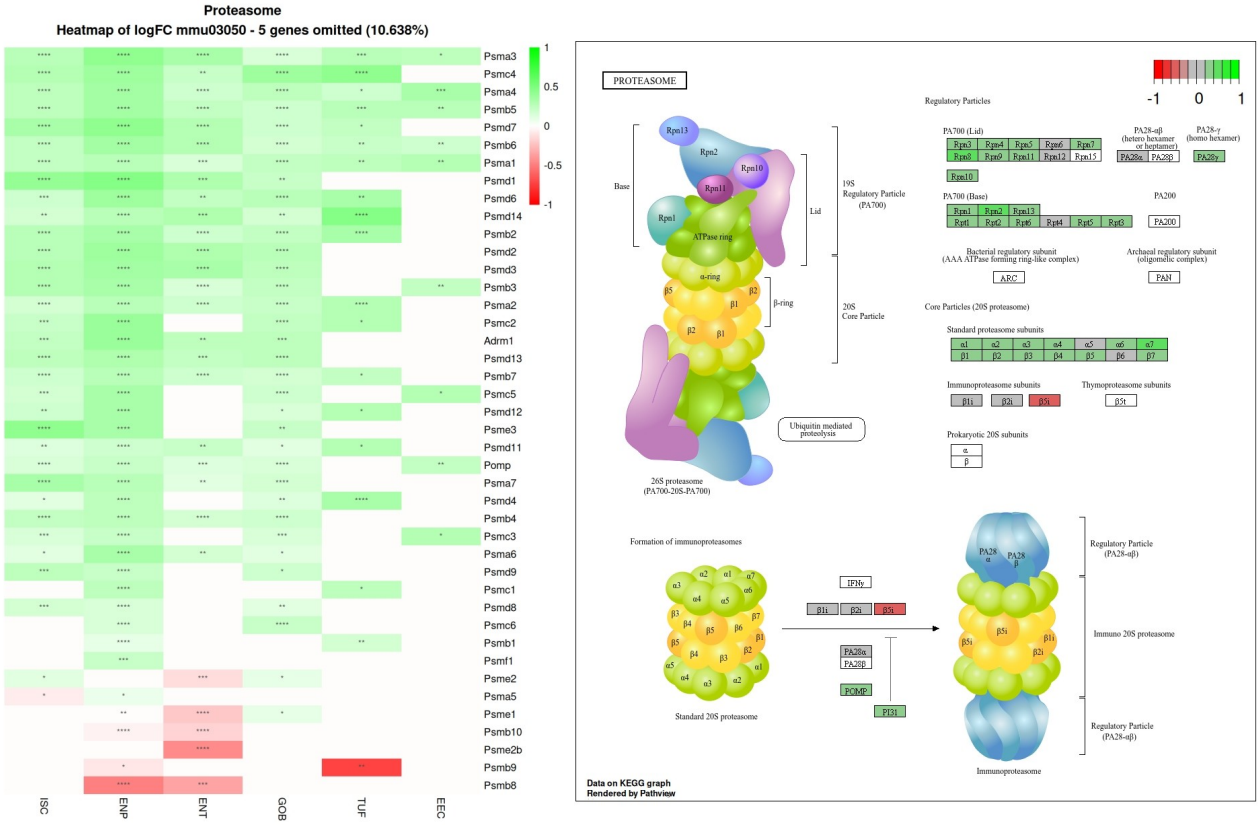

Proteasome
Heatmap of logFC mmu03050 - 5 genes omitted (10.638%)
Psma3
Psmc4
Psma4
Psmb5
Psmd7
Psmb6
Psma1
Psmd1
Psmd6
Psmd14
Psmb2
Psmd2
Psmd3
Psmb3
Psma2
Psmc2
Adrm1
Psmd13
Psmb7
Psmc5
Psmd12
Psme3
Psmd11
Pomp
Psma7
Psmd4
Psmb4
Psmc3
Psma6
Psmd9
Psmc1
Psmd8
Psmc6
Psmb1
Psmf1
Psme2
Psma5
Psme1
Psmb10
Psme2b
Psmb9
Psmb8
ISC
ENP
ENT
GOB
TUF
EEC
1
0.5
0
-0.5
-1
PROTEASOME
Regulatory Particles
Rpn13
Rpn2
Rpn10
Rpn11
Rpn1
Base
ATPase ring
α-ring
β-ring
Lid
19S Regulatory Particle (PA700)
20S Core Particle
Ubiquitin mediated proteolysis
26S proteasome (PA700-20S-PA700)
PA700 (Lid)
PA700 (Base)
PA28-αβ (hetero hexamer or heptamer)
PA28-γ (homo hexamer)
PA200
Bacterial regulatory subunit (AAA ATPase forming ring-like complex)
Archaeal regulatory subunit (oligomeric complex)
ARC
PAN
Core Particles (20S proteasome)
Standard proteasome subunits
Immunoproteasome subunits
Thymoproteasome subunits
Prokaryotic 20S subunits
Formation of immunoproteasomes
IFNγ
PA28α
PA28β
POMP
PI[3]
Standard 20S proteasome
Immunoproteasome
Regulatory Particle (PA28-αβ)
Immuno 20S proteasome
Regulatory Particle (PA28-αβ)
Data on KEGG graph
Rendered by Pathview

Figure 13