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.