Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

Hanna Sofie Kjemperud

Investigation of single cells in tumor tissue and microRNAs in serum of colorectal cancer patients

Master's thesis in Molecular Medicine
Supervisor: Robin Mjelle
May 2022

NTNU
Norwegian University of Science and Technology

Hanna Sofie Kjemperud

Investigation of single cells in tumor tissue and microRNAs in serum of colorectal cancer patients

Master's thesis in Molecular Medicine
Supervisor: Robin Mjelle
May 2022

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

NTNU
Kunnskap for en bedre verden

Abstract

Colorectal cancer (CRC) is a disease affecting almost 2 million people each year, and about 20% of the patients are diagnosed at advanced stages where treatments have poor effect, due to its lack of symptoms at an early stage. To improve survival, new treatment targets need to be discovered as well as biomarkers for early detection and stratification of subgroups. Understanding the tumor heterogeneity of primary CRC tumor can provide important information about the disease, with the potential for future treatment targets of the cancer at earlier stages, and thus lower CRC mortality rate. Moreover, circulating biomarkers have proven to be usefull with respect to early detection, disease monitoring, and risk predicition.

A protocol for establishing single cell suspension from fresh CRC tumor tissue was developed and optimized, where the single cell suspensions established using the finalized protocol was shown to have cell number, cell appearance, and cell viability compatible with downstream single cell RNA sequencing (scRNA-seq) workflow.

Fresh CRC tissue samples were collected from three stage I-II patients and one stage IV, all who underwent surgery for removal of the primary tumor. Single cell suspensions were then established from the primary tumor of these patients, followed by scRNA-seq. Analysis of the CRC scRNA-seq data revealed clusters of cells with unique expression of cell-type-specific marker genes, in which a total of 18 major cell types were identified: B-cells, CD4+ effector memory T-cells, CD4+ proliferating T-cells, CD8+ effector memory T-cells, dendritic cells/B-cells, fibroblasts, intestinal enterocytes, intestinal epithelial cells (unspecified subgroup), intestinal goblet cells, monocytes, mitochondrial gene-expressing cells, myeloid cells (unspecified subgroup), plasma B-cells, smooth muscle cells, T-cells (unspecified subgroup), unknown cell type, vascular endothelial cells, and vascular smooth muscle cells.

In addition, subtypes of stromal cells (pericytes, cancer-associated fibroblasts, plasma B-cells, crypt-top fibroblasts, myofibroblasts, and lamina propria fibroblasts), endothelial cells (mitochondrial gene-expressing cells, activated tumor-associated endothelial cells (TECs), tip TECs, immature TECs, and proliferative endothelial cells), and intestinal epithelial cells (mitochondrial gene-expressing cells, secretory progenitor 1, crypt base cells/Paneth cells, secretory progenitor 2, enterocytes, plasma B-cells, iron-storing epithelial cells, and tuft-2 cells) were also identified in the CRC tumor tissue. In total, the identified expressed genes and cell type composition of CRC tumor tissue emphasizes the intratumor heterogeneity of CRC.

Lastly, significantly differentially expressed circulating serum microRNAs (miRNAs) between CRC patient groups were identified. A total of 53 significant miRNAs between true positive CRC patients with localized disease and false positive CRC patients, 7 significant miRNAs between true positive CRC patients with metastatic disease and false positive CRC patients, and 30 significant miRNAs between true positive CRC patients with metastatic disease and localized disease were found.

Several miRNAs with biomarker potential were identified. Five miRNAs (miR-142-5p, miR-16-5p, miR-143-3p, miR-126-5p, and miR-16-2-3p) could be markers of CRC, two miRNAs (miR-10a-5p and miR-92b-3p) could be markers of metastatic disease, four miRNAs (miR-122-5p, miR-885-3p, miR-375-3p, and miR-192-5p) could differentiate

patients with different stages of CRC, while two miRNAs (miR-429 and miR-21-5p) were shown to be highly associated with metastatic disease.

Sammendrag

Kolorektal kreft er en sykdom som påvirker nesten 2 millioner mennesker hvert år, og omtrent 20% av pasientene blir diagnostisert ved avanserte kreftstadier hvor behandling har dårlig effekt, som følge av mangel på symptomer ved tidlig stadium. For å bedre overlevelsesraten må nye behandlingsmål og biomarkører oppdages for tidlig påvisning og stratifisering av undergrupper. En forståelse av tumorheterogeniteten i primærtumor av kolorektalkreft kan gi viktig informasjon om sykdommen, og innehar potensial for fremtidige behandlingsmål for kreften på tidlige stadier, noe som dermed kan senke dødelighetsraten for kreftformen. Videre har sirkulerende biomarkører vist seg å være nyttige med hensyn til tidlig påvisning, sykdomsovervåkning og forutsigelse av risiko.

En protokoll for etablering av enkeltcellesuspensjon fra ferskt kolorektalt svulstvev ble utviklet og optimalisert. Enkeltcellesuspensjonene som ble etablert ved bruk av den ferdigstilte protokollen ble vist å ha et celleantall, utseende og andel levedyktige celler kompatibelt med RNA-sekvensering av enkeltceller («scRNA-seq») og videre analyse.

Ferske prøver av kolorektalt svulstvev ble samlet inn fra tre pasienter med stadium I-II og en pasient med stadium IV, der alle pasientene gjennomgikk operasjon for fjerning av primærtumor. Enkeltcellesuspensjoner ble så etablert av primærtumoren fra disse pasientene, etterfulgt av scRNA-seq. En analyse av dataene etter scRNA-seq avslørte klynger av celler med et unikt uttrykk av celle-spesifikke markørgener, hvor totalt 18 hovedcelletyper ble identifisert: B-celler, CD4+ effektor hukommelses T-celler, CD4+ prolifererende T-celler, CD8+ effektor hukommelses T-celler, dendrittiske celler/B-celler, fibroblaster, tarm-enterocytter, tarmepitelceller (uspesifisert undergruppe), tarm-begerceller, monocytter, mitokontrielle gen-uttrykkende celler, myeloide celler (uspesifisert undergruppe), plasma B-celler, glatte muskelceller, T-celler (uspesifisert undergruppe), ukjent celletype, vaskulære endotelceller, og vaskulære glatte muskelceller.

I tillegg ble undergrupper av stromale celler (pericytter, kreft-assosierte fibroblaster, plasma B-celler, krypt-topp fibroblaster, myofibroblaster og lamina propria fibroblaster), endotelceller (mitokondrielle gen-uttrykkende celler, aktiverte kreft-assosierte endotelceller (KECer), tupp KECer, umodne KECer og proliferative endotelceller) og tarmepitelceller (mitokondrielle gen-uttrykkende celler, sekreterende forgjengerceller 1, krypt-bunn celler/Paneth celler, sekreterende forgjengerceller 2, enterocytter, plasma B-celler, jern-lagrende epitelceller og tuft-2 celler) identifisert i kolorektalt svulstvev. Totalt sett understreker de identifiserte uttrykte genene og celletypesammensetningen i kolorektalt svulstvev intratumorheterogeniteten i kolorektalkreft.

Til slutt ble signifikante differensielt uttrykte sirkulerende mikroRNA (miRNA) i serum mellom kolorektal-pasientgrupper identifisert, hvorav 53 var mellom kreftpasienter med lokalisert sykdom og kontrollprøver, 7 var mellom kreftpasienter med spredning og kontrollprøver, og 30 var mellom kreftpasienter med lokalisert sykdom og spredning.

Flere potensielle biomarkør-miRNA-er ble identifisert. Fem miRNA-er (miR-142-5p, miR-16-5p, miR-143-3p, miR-126-5p og miR-16-2-3p) kan fungere som markører for kolorektalkreft, to miRNA-er (miR-10a-5p and miR-92b-3p) kan være markører for metastatisk sykdom, fire miRNA-er (miR-122-5p, miR-885-3p, miR-375-3p og miR-192-5p) kan skille pasienter med forskjellige stadier av kolorektalkreft, mens to miRNA-er (miR-429 og miR-21-5p) ble vist å være sterkt assosiert med metastatisk sykdom.

Acknowledgements

This master's thesis is written as a part of the Molecular Medicine master's degree program at the department of Clinical and Molecular Medicine (IKOM), faculty of Medicine and Health Sciences (MH), at the Norwegian University of Science and Technology (NTNU) in Trondheim.

I would first like to thank my supervisor postdoctoral researcher Robin Mjelle for the opportunity to work on this interesting project within the important field of cancer research. Thank you for your guidance and patience when teaching me laboratory techniques, for help with improving my programming skills, and for giving me valuable feedback and support throughout the master's project.

In addition, I would like to thank Biobank1 for organizing and providing the sample collection of fresh colorectal cancer tissue. I am also grateful to bioengineer Ida Kjølstad Solberg at the Department of Pathology at St. Olav's University hospital for providing operation schedules every week.

I would also like to thank senior engineer Nina-Beate Liabakk for her expertise on flow cytometry analysis, as well as Tone Christensen and the others at the Genomics Core Facility (GCF) at NTNU for their sequencing services. Another thanks to Robin Mjelle at the Bioinformatics Core Facility (BioCore) at NTNU for his technical assistance with processing sequencing data.

A special thanks to my partner, friends, and family for your invaluable support.

Trondheim, May 2022

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACF	Aberrant crypt focus
APC	Adenomatous polyposis coli
CAF	Cancer-associated fibroblast
CBC	Crypt base cell
cDNA	Complementary DNA
CIMP	CpG island methylator phenotype
CIN	Chromosomal instability
CMS	Consensus molecular subtypes
CRC	Colorectal cancer
CSC	Cancer stem cell
CTC	Computed tomographic colonography
CTF	Crypt top fibroblast
DC	Dendritic cell
DCC	Deleted in colorectal cancer
DNA	Deoxyribonucleic acid
EC	Endothelial cell
ERK	Extracellular-signal-regulated kinase
FAP	Familial adenomatous polyposis
GAP	Goblet cell-associated antigen passages
GEM	Gel bead in emulsion
HDI	Human Development Index
HNPCC	Hereditary non-polyposis colorectal cancer
ILC	Innate lymphoid cell
ISC	Intestinal stem cell
ITH	Intratumor heterogeneity
KRAS	KRAS proto-oncogene GTPase
LPF	Lamina propria fibroblast
MAIT cell	Mucosal associated invariant T-cell
MAPK	Mitogen-activated protein kinase
MEK	Mitogen-activated protein kinase kinase
miRNA	MicroRNA
MLH1	MutL homolog 1
MMR	Mismatch repair
mRNA	Messenger RNA
MSH2	MutS homolog 2
MSI	Microsatellite instability
MSS	Microsatellite stable
ncRNA	Non-coding RNA
NGS	Next-generation sequencing

NK cell	Natural killer cell
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
pre-miRNA	Precursor miRNA
pri-miRNA	Primary miRNA
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNAi	RNA interference
RNA-seq	RNA sequencing
SBS	Sequencing-by-synthesis
scRNA-seq	Single cell RNA sequencing
SEER	Surveillance, Epidemiology, and End Results
SMAD2/4	Mothers against decapentaplegic homolog 2/4
TAM	Tumor-associated macrophages
TAN	Tumor-associated neutrophil
TEC	Tumor-associated endothelial cell
TGF-β	Transforming growth factor β
TIL	Tumor-infiltrating lymphocyte
TIM	Tumor-infiltrating mast cell
TME	Tumor microenvironment
TMN	Tumor-node-metastasis
TP53	Tumor protein p53
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique molecular identifier
UTR	Untranslated region
WNN	Weighted nearest neighbor graph

1 Introduction

1.1 Colorectal cancer

The term "colorectal cancer" includes cancer in both the colon and rectum, where the colon and rectum along with the anus make up the large intestine and is a part of the gastrointestinal system of the human body (Figure 1.1) [1].

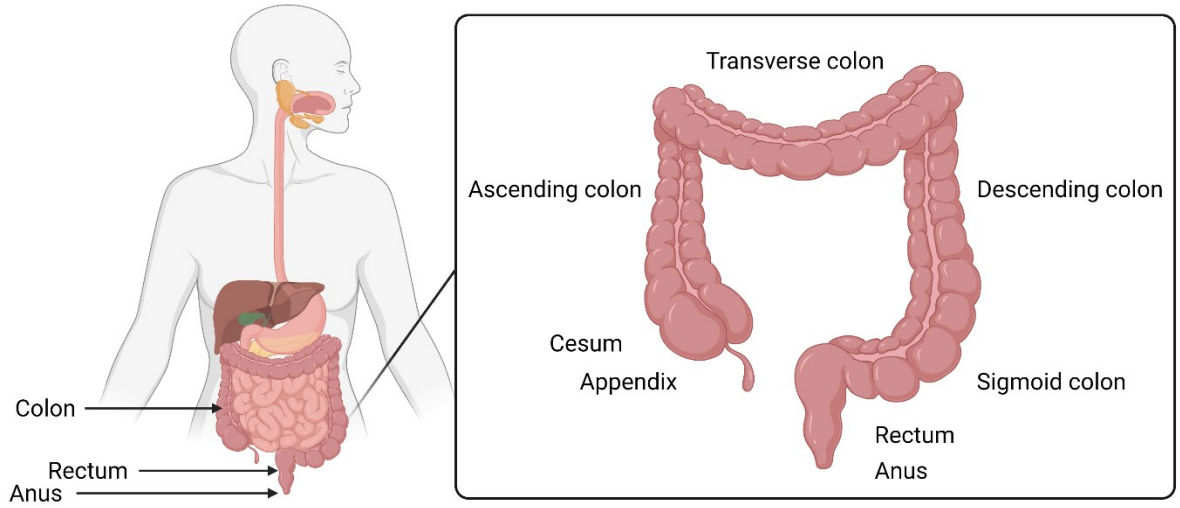


Figure 1.1: The large intestine (colon, rectum, and anus) as part of the human gastrointestinal system. A more detailed structure of the large intestine (box) shows the appendix and cesum lying at the start of the organ, continuing upwards into the ascending colon located on the right side of the abdomen, bending into the transverse colon, and continuing downward into the descending colon at the left abdomen side. The descending colon bends into the sigmoid colon, before opening into the rectum and finally anus. Created with BioRender.com.

1.1.1 Colorectal cancer incidence and risk factors

Colorectal cancer (CRC) is the third most common cancer in the world, with around 1.9 million new diagnosed cases in 2020, which in total accounted for about 10.0% of all new cancer cases that year [2]. In addition, CRC is the second most frequent cause of cancer deaths globally [2]. The disease was responsible for about 935,000 deaths in 2020, which that year was equivalent to 9.4% of all cancer deaths [2]. The CRC incidence rates vary geographically, and a global pattern of increasing disease incidence with a country's increasing Human Development Index (HDI) has been noticed [2]. Thus, CRC can be considered a marker of socioeconomic development [2], and worldwide incidence is predicted to increase to 2.5 million new cases in 2035 due to the continuing progress in developing countries [3].

Modifiable environmental risk factors for CRC include high body fat and obesity, heavy alcohol intake, cigarette smoking, and consumption of red or processed meat [3], which all are associated with socioeconomical development. Epidemiological studies have also shown that male sex and increasing age have strong associations with disease incidence [3], as CRC incidence rates are about 30% higher in men than in women and the majority of cases are diagnosed in patients ages 50 and older [1]. There are also hereditary risk factors for CRC such as a positive CRC family history and inherited cancer susceptibility genes [3].

1.1.2 Staging systems for colorectal cancer

A widely used CRC classification system in clinical settings is the tumor-node-metastasis (TNM) staging system defined by the American Joint Committee on Cancer (AJCC) [1, 4], whereas Surveillance, Epidemiology, and End Results (SEER) is another staging system used for descriptive and statistical analysis of tumor registry data (Table 1.1) [1]. Both staging systems describe the degree of tumor invasion in the body (Figure 1.2).

Table 1.1: Staging systems for colorectal cancer. Two widely used staging systems for colorectal cancer are tumor-node-metastasis (TME) and Surveillance, Epidemiology, and End Results (SEER). The staging systems describe the degree of tumor invasion in the body. Adapted from Ponz de Leon, M. and Di Gregorio, C. [5] and American Cancer Society [1].

Staging system		**Description**
TNM	**SEER**	
(Stage 0)	In situ	Tumor has not yet begun to invade the colorectal wall.
Stage I	Local	Tumor invasion of the submucosa (T1), or further invasion of the muscularis propria or subserosa (T2).
Stage II	Regional	Tumor invasion through the serosa (T3), or further invasion of the abdominal membrane lining and abdominal organs (T4).
Stage III		Metastatic involvement of lymph nodes, either 1-3 nodes (N1) or 3+ nodes (N2).
Stage IV	Distant	Metastatic involvement of liver, lung, or other organs (M1).

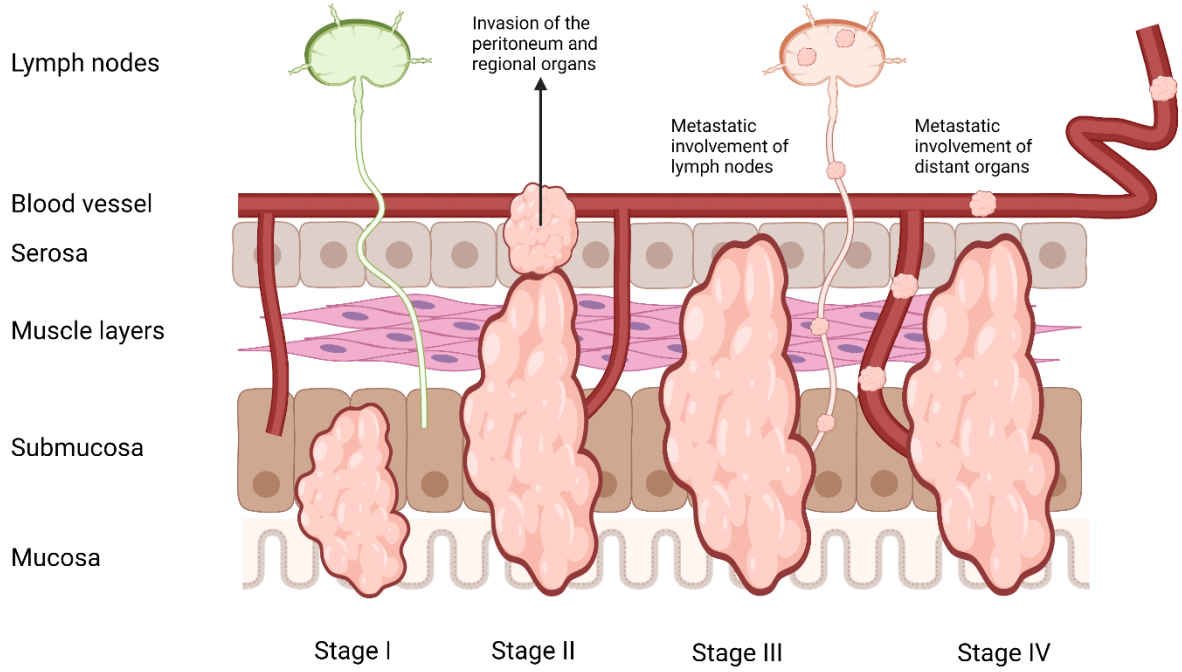


Figure 1.2: A visualization of the degree of tumor invasion in the body as described by the tumor-node-metastasis staging system in terms of colorectal cancer. Stage I includes tumor invasion of the submucosa, stage II includes invasion through the serosa, stage III includes metastatic involvement of lymph nodes, and stage IV includes metastatic involvement of distant organs. Adapted from National Cancer Institute (NCI) [5] and created with BioRender.com.

1.1.3 Colorectal cancer prevention, diagnosis, and treatment

CRC is largely asymptomatic until it reaches an advanced stage, where increasing cancer stages corresponds to more complex disease and thus lower survival rate [1, 6]. In total,

about 40% of CRC patients are diagnosed at an early stage, 40% are diagnosed with regional cancer, and 20% are diagnosed at the late stage with distant metastatic disease [6], where the 5-year survival rate is 90%, 71%, and 13%, respectively [1, 6].

Due to CRCs correlation of mortality with disease stage, preventative measures such as screening programs are important for early diagnosis of the disease [1, 3]. Several detection methods are available in CRC screening programs, like non-invasive stool tests or more invasive endoscopic imaging techniques such as computed tomographic colonography (CTC) or colonoscopy [3, 7]. For people at elevated risk for CRC, such as those with known hereditary risk factors, regular surveillance by colonoscopy is the recommended prevention method [3]. Colonoscopy is also the preferred method for diagnosing CRC [3].

Advancements in pathophysiological understanding have increased the treatment options for both local and advanced disease, focusing on individual treatment plans dependent on tumor-specific molecular features, tumor location, and patient characteristics [1, 3]. Endoscopic treatment or surgical resection is often sufficient to remove early-stage cancers, and in most cases no further treatment is needed [1]. For cancer at the regional stage that has spread to nearby lymph nodes, surgery is usually preceded or followed by chemotherapy to reduce the risk of local recurrence [1, 3]. Advanced CRC with metastatic involvement of other organs typically require surgery, chemotherapy, targeted therapies, and/or immunotherapy, often as palliative treatment to control the cancer, relieve the symptoms, and prolong survival [1, 3].

1.1.4 Molecular basis of colorectal cancer

CRC progression is driven by the continuous acquisition of genetic mutations or epigenetic modifications in both tumor suppressor genes and oncogenes, as described in the well-established adenoma-carcinoma sequence (Figure 1.3) [8, 9]. This unfolds following a CRC molecular pathway such as chromosomal instability (CIN), CpG island methylator phenotype (CIMP), or pure microsatellite instability (MSI) [10], where it is important to note that the pathways not necessarily are mutually exclusive [11].

1.1.4.1 Chromosomal instability (CIN) pathway

The CIN pathway includes inactivation of the tumor suppressor gene adenomatous polyposis coli (APC) affecting the Wnt/β-catenin signaling pathway, followed by activation of KRAS proto-oncogene GTPase (KRAS) affecting the MAPK cascade (also known as the Ras/Raf/MEK/ERK signaling pathway) [9, 10]. Both signaling pathways are involved in cell proliferation, differentiation, and apoptosis [12, 13]. Further on, deleted in colorectal cancer (DCC) and the tumor suppressor genes SMAD2 and SMAD4 are inactivated [9, 10]. Inactivation of DCC hinders cell apoptosis, while inactivation of the SMAD-genes affects the TGF-β signaling pathway controlling cell proliferation, differentiation, and apoptosis [9, 14, 15]. Lastly, inactivation of the tumor suppressor gene TP53 affects the p53 pathway, a pathway ensuring the appropriate responses to cellular stress caused by DNA damage or hyperproliferative signals [9, 10, 16]. TP53 alterations are considered the hallmark of human tumors and is associated with the progression and outcome of sporadic CRC [9].

1.1.4.2 CpG island methylator phenotype (CIMP) pathway

CIMP CRC is characterized by a high density of DNA methylation in promoter CpG islands of tumor suppressor genes [9, 17], where the hypermethylation ultimately blocks the transcription and thus inactivates the gene [9, 18]. The CIMP status in sporadic CRCs can

be classified according to the degree of DNA methylation, often distinguished as the subgroups CIMP-negative, CIMP-low, and CIMP-high [17].

1.1.4.3 Microsatellite instability (MSI) pathway

In the MSI pathway, genetic or epigenetic inactivation of tumor suppressor DNA mismatch repair (MMR) genes occurs [9, 10]. This leads to a dysfunctional MMR system unable to repair the high frequency of replication errors made in microsatellite areas, increasing the mutation rate and potential for malignancy [9, 10]. The MSI status in sporadic CRCs is classified by the number of microsatellite panel markers with altered size, and is often distinguished as the subgroups microsatellite stable (MSS), MSI-low, and MSI-high [11].

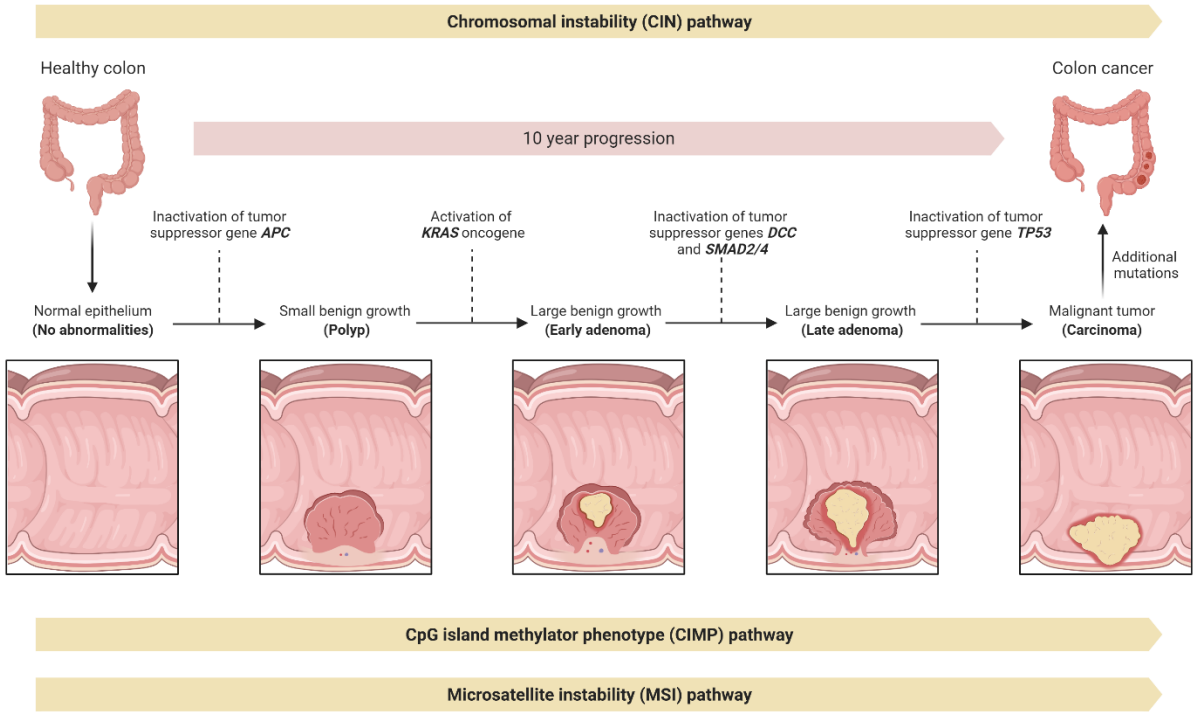


Figure 1.3: The adenoma-carcinoma sequence. It is a well-established multistep genetic model first proposed by Fearon, E.R. and Vogelstein, B. [8] to describe colorectal cancer progression following the chromosomal instability (CIN), CpG island methylator phenotype (CIMP), or microsatellite instability (MSI) pathway. In the CIN pathway, alterations in APC, KRAS, DCC, SMAD2/4, and TP53 drives a healthy colon to develop colon cancer in a 10-year progression span. Abbreviations: APC, adenomatous polyposis coli; KRAS, KRAS proto-oncogene GTPase; DCC, deleted in colorectal cancer; SMAD2/4, Mothers against decapentaplegic homolog 2/4; TP53, Tumor protein p53. Adapted from Nguyen, H.T. and Duong, H.Q. [9] and created with BioRender.com using "The Multi-Hit Model of Colorectal Cancer"-template by Louise De Herdt.

1.1.5 Sporadic and hereditary colorectal cancer

About 70% of all CRC cases occurs sporadically [9]. The cell of origin for sporadic CRC is currently assumed to be an intestinal stem cell located at the base of colonic crypts, that after acquisition of genetic mutations or epigenetic modifications can rapidly expand to occupy the whole crypt and transform it into an aberrant crypt focus (ACF) [3, 19]. In turn, the ACF can evolve into a polyp and eventually progress to CRC over an estimated 10-year period [3]. Regarding CRC molecular pathways, it is estimated that 85% of sporadic CRCs exhibit CIN, 15% displays CIMP, and 15% shows MSI [9-11]. Sporadic CRCs with MSI status are generally not pure, and they often occur in a CIMP-positive context where the MMR gene MLH1 is epigenetic inactivated [9, 10].

The remaining 30% of all CRC cases are due to hereditary risk factors, where roughly 5% are hereditary CRCs and 25% are familial CRC [9, 20]. Hereditary CRCs are associated with specific highly penetrant inherited mutations, such as in the polyposis syndrome familial adenomatous polyposis (FAP) and in the non-polyposis syndrome hereditary non-polyposis colorectal cancer (HNPCC) also known as Lynch syndrome [9]. As hereditary CRCs often occurs 10-15 years earlier than sporadic CRC, people with hereditary CRC syndromes tend to experience disease onset at an earlier age [20, 21]. Familial CRC are likely due to less penetrant but more common single gene mutations, but the entire etiologies are not completely understood [9, 20].

1.1.5.1 Hereditary colorectal cancer syndromes

FAP is a rare hereditary CRC syndrome characterized by the development of hundreds to thousands of colonic adenomas, where affected individuals have a 100% lifetime risk of CRC if left untreated [20]. A less severe form of the disease is attenuated FAP with a lifetime CRC risk of 69% [20]. Both FAP and attenuated FAP result from germline mutations in the tumor suppressing gene APC, where the gene mutation location has been related to severity of the syndrome [20].

Lynch syndrome is the most common hereditary cancer syndrome, accounting for 2-4% of all CRCs [20]. It is a result of germline mutations in MMR genes, primarily MLH1 and MSH2, and are thus characterized by MSI-high status [20]. Affected individuals can develop colonic adenomas with greater frequencies than the general population, and the lifetime CRC risk is estimated to be 50-80% [9, 20].

1.1.6 Proximal and distal colon tumor location

CRC may develop either on the right side or left side of the colon, also called the proximal and distal colon, respectively [22]. The proximal colon includes the cesum, ascending colon, and two thirds of the transverse colon, whereas the distal colon is comprised of the last third of the transverse colon, the descending colon, and sigmoid colon [22]. About 70% of all CRC cases occur in the left-sided distal colon and approximately 10% appears in the proximal colon [22]. Studies have shown that left-sided CRC occurs predominantly in males and at an early age, while right-sided CRC occurs mainly in females and older people [22].

CRC tumors in different parts of the colon exhibit different molecular characteristics and can thus behave differently in terms of disease progression and overall survival, where the difference can be attributed to developmental origin, distinct carcinogenic factors, or a combination of both [22]. Patients with left-sided CRC tend to have tumors with CIN-associated gene mutations like APC, KRAS, and TP53 [11, 22]. These tumors demonstrate polypoid morphology and are thus easier to detect in early stages of carcinogenesis [22]. Right-sided CRC have a flatter morphology that is more difficult to discover, where such tumors tend to be MSI-high [22]. In addition, approximately 30-40% of sporadic proximal CRCs are CIMP-positive, compared to only 3-12% of distal CRCs [11]. A study performed by Mangone, L. et al. confirmed that right-sided CRC has worse survival than left-sided CRC, even when adjusted for screening status [23].

1.2 Intratumor heterogeneity in colorectal cancer

CRC is a heterogeneous disease where the solid tumor consists of many different cell types with distinct genetic and molecular profiles among them (Figure 1.4) [10, 24]. These differences within the same tumor type in patients are referred to as intratumor

heterogeneity (ITH) [24]. CRC can thus include tumors with different biological characteristics and behaviors, making the disease more prone to metastasis, recurrence, and drug resistance due to diverging response to treatment [24].

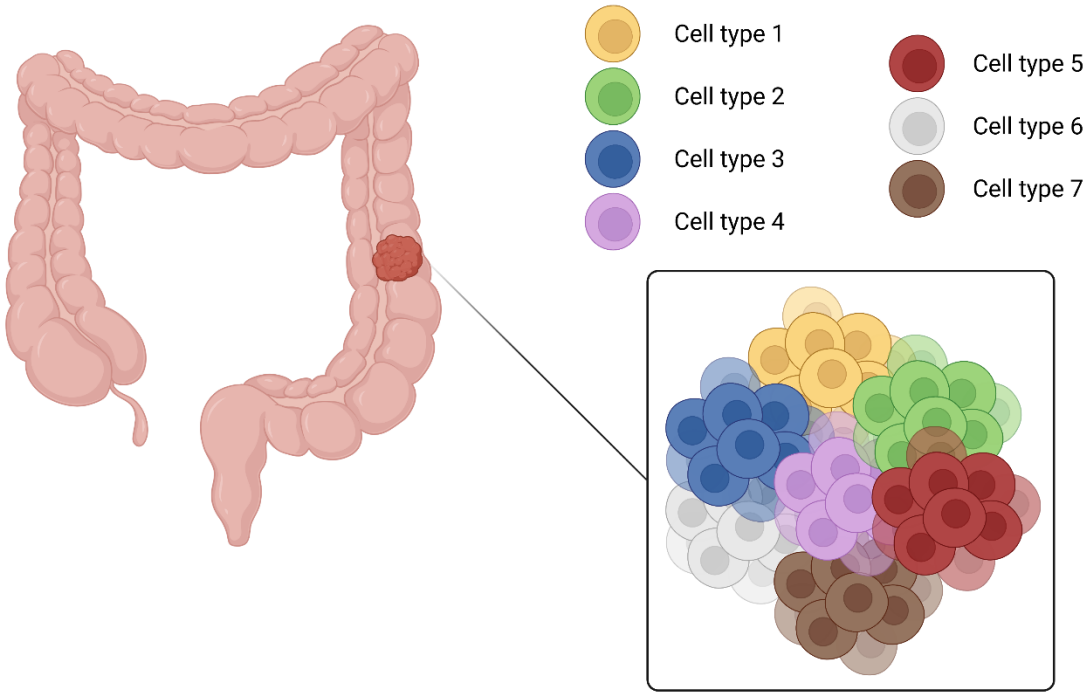


Figure 1.4: Illustration of intratumor heterogeneity in colorectal cancer. Here exemplified with seven distinct cell types. Intratumor heterogeneity is a term used when a solid tumor consists of many different cells showing distinct genetic and molecular profiles among them. Adapted from Zheng, Z. et al. [24] and created with BioRender.com.

ITH can be caused by both genetic and epigenetic variability occurring during CRC molecular pathways, but can also be due to the tumor microenvironment (TME) [25]. TME refers to the cells and their secreted components surrounding a tumor [25]. TME constituents can either promote or suppress tumor formation by interacting with tumor cells during TME signaling pathways, and can thus be considerably involved in CRC progression and metastasis [25].

1.2.1 Relevance of intratumor heterogeneity in cancer prognosis

It appears to be a significant correlation between ITH and cancer prognosis, where increased ITH is associated with a decrease in cancer survival [24]. This trend has been found in various cancers, including CRC [24]. One study found that the 3-year overall survival and progression-free survival in metastatic CRC patients exhibiting low ITH were 66% and 23%, respectively, while patients with high ITH had an overall survival of 18% and progression-free survival of 5% [26]. The exact reason for ITH impact on prognosis is unclear, and there is still a limited amount of research related to the ITH of CRC in general [24].

1.2.2 Identification of intratumor heterogeneity in colorectal cancer

Mapping the ITH in primary CRC tumor is a step on the way to achieve a clearer picture of the internal cell type composition of the tumor and its TME, and it might result in a better understanding of the molecular basis of early biological processes in CRC

development, in addition to determining which factors that can function as better prognostic and predictive markers for the disease [24, 27].

CRC was previously divided into four consensus molecular subtypes (CMS) [28]. This classification system was based on clustering bulk transcriptomic data, not accounting for the relative contribution of each cell type in the tumor tissue [28]. A study using single cell RNA sequencing (scRNA-seq) revealed that the TME-cells had a strong influence on the bulk CMS type, and that clusters of tumor epithelial cells derived from scRNA-seq did not align with CMS [28]. The previous CMS-classes were in fact signatures of the TME and not the intrinsic tumor transcriptome [28].

Identification of cell type composition in primary CRC tumor and TME can thus be achieved by scRNA-seq [29]. ScRNA-seq is a favorable choice for studying cell heterogeneity in complex tissues such as cancer tissue, as the method estimates a distribution of expression levels for each gene across individual cell populations [30, 31]. It is an approach that investigates both the CRC molecular pathway variabilities and the makeup of the TME [27].

1.2.3 Previous research on intratumor heterogeneity in patients with colorectal cancer using single-cell RNA sequencing techniques

A limited number of scRNA-seq studies have been conducted in CRC (Table 1.2). One of the first studies was published in 2017 by Li, H. et al. [32], where scRNA-seq was conducted on primary tumor cells from 11 CRC patients at stages II-IV [29]. Seven distinct cell clusters were obtained and annotated as epithelial cells, fibroblasts, endothelial cells, B-cells, T-cells, mast cells, and myeloid cells [32].

Another early study was performed in 2019 by Dai, W. et al. and involved primary CRC cells from one patient at stage III, where analyses performed on the scRNA-seq data revealed five distinct cell clusters [33]. This was ultimately a clear sign of heterogeneity, where each cluster consisted of specific cell markers with different functions [29].

A third study investigating the overall ITH in CRC was conducted by Mei, Y. et al. in 2021 on primary CRC cells from 12 patients at stage I-IV [34]. This study detected eight distinct cell clusters annotated as T-cells, B-cells, myeloid cells, mucosal associated invariant T-cells (MAIT cells), natural killer (NK) cells, epithelial cells, fibroblasts, and erythrocytes [34].

Another study from 2021 by Khaliq, A.M. et al. was based on profiling primary CRC tissue samples from 16 patients at stages I-IV using scRNA-seq techniques [35]. This resulted in clusters of epithelial cells, fibroblasts, endothelial cells, T-cells, B-cells, and myeloid cells, which again were subclustered to find cell type subsets [35]. TME cells like cancer-associated fibroblast (CAF) subsets, CD4+ subsets, CD8+ subsets, NK cells, innate lymphoid cells (ILCs), monocyte lineage phenotypes, and tumor-associated macrophages (TAMs) were identified [35].

Table 1.2: A summarization of studies on intratumor heterogeneity in patients with colorectal cancer using single cell RNA sequencing techniques. The findings confirm tumor heterogeneity, and highlight different cells present in the colorectal cancer tumor and tumor microenvironment. CRC, colorectal cancer; scRNA-seq, single cell RNA sequencing; RCA, reference component analysis; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; t-SNE, t-stochastic neighbor embedding; MAIT, mucosal-associated invariant T-cells; NK, natural

killer; CAF, cancer-associated fibroblast; ILC, innate lymphoid cell; TAM, tumor-associated macrophage.

Sample	Method	Findings	References
Primary tumor cells from 11 CRC patients at stages II-IV	ScRNA-seq and clustering by RCA	Obtained seven distinct cell clusters (epithelial cells, fibroblasts, endothelial cells, B-cells, T-cells, mast cells, and myeloid cells)	Li, H. et al. (2017) [32]
Primary tumor cells from 1 CRC patient at stage III	ScRNA-seq followed by GO and KEGG pathway analyses	Revealed five distinct cell clusters that all consisted of specific cell markers with different functions	Dai, W. et al. (2019) [33]
Primary tumor cells from 12 CRC patients at stages I-IV	ScRNA-seq and downstream Seurat clustering analysis including t-SNE	Obtained eight distinct cell clusters (T-cells, B-cells, myeloid cells, MAIT cells, NK cells, epithelial cells, fibroblasts, and erythrocytes)	Mei, Y. et al. (2021) [34]
Primary tumor cells from 16 CRC patients at stages I-IV	ScRNA-seq and downstream Seurat clustering analysis including t-SNE	Revealed several distinct cell clusters (epithelial cells, fibroblasts, endothelial cells, T-cells, B-cells, and myeloid cells), where different subsets were identified (CAFs, CD4+, CD8+, NK cells, ILCs, monocyte lineage phenotypes, and TAMs)	Khaliq, A.M. et al. (2021) [35]

Other more current studies using scRNA-seq techniques seem to focus on specific cell type subsets within the CRC tumor by subclustering the cell types of interest. In a study on primary CRC from 2021, Wang, H. et al. found that rare cancer stem cells exist in a dormant state and display plasticity towards cancer epithelial cells which exhibit tumor-initiating features [36]. The same year, Qi, J. et al. found that tumor tissue from CRC patients contains tumor specific ILCs, namely ILC1-like and ILC2 subsets [37], while Domanska, D. et al. did in a study from 2022 analyze macrophages from colonic resections of CRC patients and revealed niche-specific subsets [38].

1.2.4 Cell types in colorectal cancer tumor and tumor microenvironment

Previous research conducted on ITH in CRC using scRNA-seq techniques have revealed many major cell types present in CRC tumors and TME. These findings can be seen in the context of cell types found in normal colorectal tissue and microenvironment, as CRC starts to develop from a healthy tissue.

1.2.4.1 Normal colorectal tissue and microenvironment structure

As mentioned, the colon and rectum do along with the anus make up the large intestine of the human body [1]. The large intestine is lined with a single cell layer of epithelial cells, where this layer at millions of places form tube-like invaginations called colonic crypts into the underlying tissue layer lamina propria (Figure 1.5) [39]. The lamina propria is a layer of connective tissue rich in cells such as fibroblasts, pericytes, endothelial cells, and scattered immune cells [40]. The epithelium do along with the

lamina propria and muscularis mucosae make up the mucosa, where the muscularis mucosae is a thin layer of smooth muscle cells [40]. This means that there is a complex microenvironment composed of diverse cell types surrounding the colonic crypt bases [40].

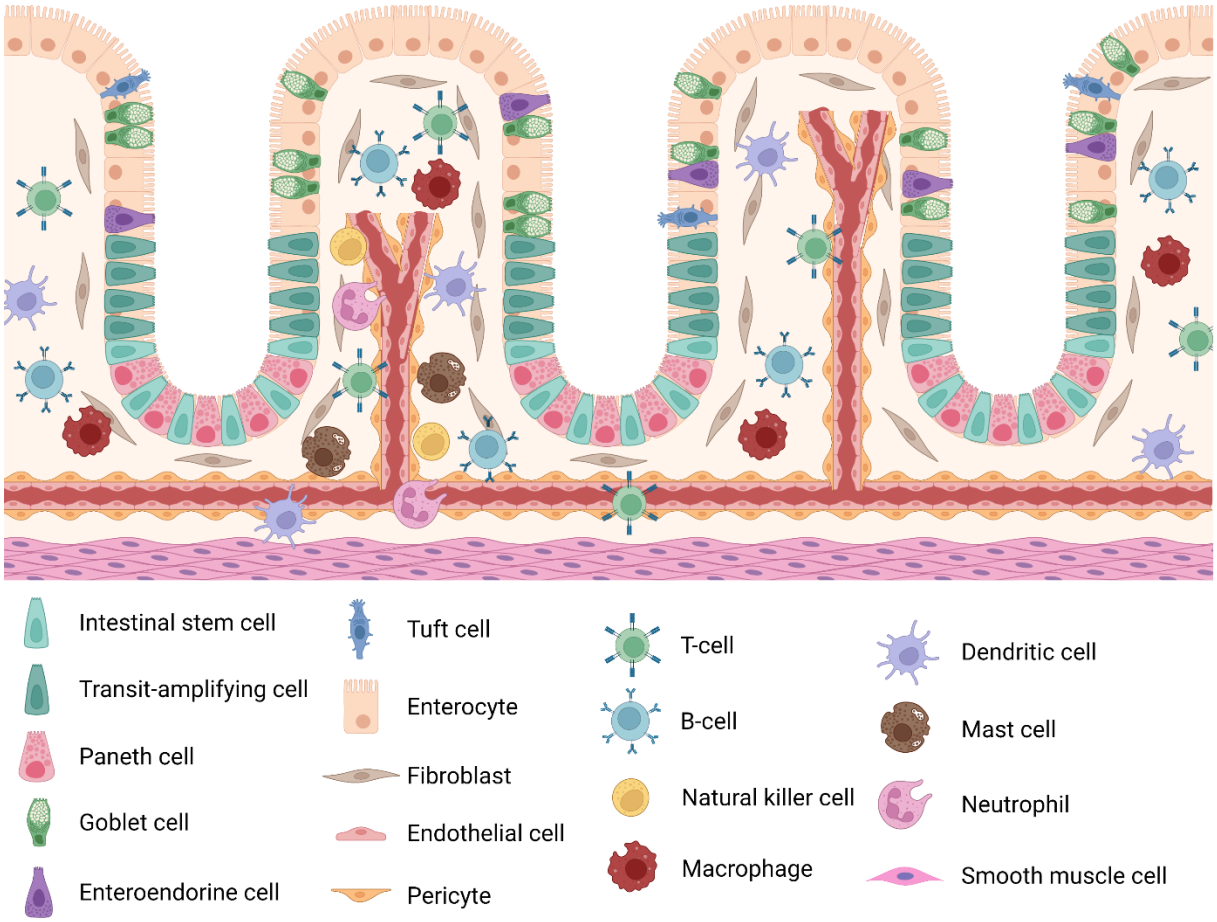


Figure 1.5: An overview of normal colorectal tissue and microenvironment structure. The epithelial cell layer forms colonic crypts into the underlying tissue, where the epithelium along with lamina propria and muscularis mucosae make up the mucosa. Epithelial cells include intestinal stem cells, transit-amplifying cells, Paneth cells, goblet cells, enteroendocrine cells, tuft cells, and enterocytes. Cells in the lamina propria include fibroblasts, endothelial cells, pericytes, T-cells, B-cells, natural killer cells, macrophages, dendritic cells, mast cells, and neutrophils. Muscularis mucosae cells include smooth muscle cells. Adapted from Zhu, G. et al. [40] and created with BioRender.com.

1.2.4.2 Intestinal stem cells can transform into colorectal cancer stem cells

Intestinal stem cells (ISCs) reside at the bottom of intestinal crypts, and are interspersed with a similar number of Paneth cells in the small intestine and Paneth-like cells in the large intestine [40]. This is based on findings in mice, where typical Paneth cells are absent from large intestine crypts, and deep crypt secretory cells instead are intermingled with ISCs and function as the colon equivalent of Paneth cells [40].

The crypt-based ISCs give rise to transit-amplifying cells, which are proliferating progenitors that can differentiate into various intestinal cell types near the top of the crypt and regularly replenish the shedding epithelial cells [40]. Most of the differentiated cells moves upwards, while the Paneth cells or Paneth-like cells move back down to the ISC compartment [40]. There are believed to exist both actively proliferating ISCs and

non- or slowly proliferating reserve ISCs [41]. The active ISCs mediates the normal homeostatic turnover of the intestinal epithelium, whereas reserve ISCs fulfill regenerative tasks specifically after tissue injury or stress [41]. Previously used nomenclature was crypt based columnar cells for active ISCs and +4 cells for reserve ISCs [42].

ISCs in the large intestine displaying tumor-related features due to acquired genetic mutations or epigenetic modifications in tumor suppressor genes and oncogenes [3], possibly with the influence of signals derived from the TME [43], are defined as colorectal cancer stem cells (CSCs) [19]. Mirroring normal ISCs, CSCs can both self-renew and generate all the differentiated cells that comprise the tumor, thus giving rise to heterogenous tumors [19]. Tumor-related characteristics of CSCs are uncontrolled growth, resistance to apoptosis, and increased invasiveness, making the cells play a key role in CRC initiation, invasion and progression, as well as therapy resistance [19].

1.2.4.3 Intestinal epithelial cells

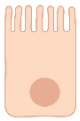
Proliferating transit-amplifying progenitors derived from ISCs can differentiate into various intestinal epithelial cell types such as secretory Paneth cells, goblet cells, enteroendocrine cells, and tuft cells, or adsorptive enterocytes [40]. The various differentiated cell types of the intestinal epithelium are well defined [42], and together they perform several vital physiological functions such as nutrient absorption, energy homeostasis, innate immunity, and tissue regeneration [44]. A disruption of these functions can lead to impaired health conditions such as CRC [44].

Paneth cells are specialized secretory pyramidal-shaped cells possessing dense granules in their cytoplasm, which contains antimicrobial compounds and immunomodulating proteins that function to regulate the composition of the intestinal flora, being important in immunity and host-defense [45].

Goblet cells are cup-shaped cells with the primary function of synthesizing and secreting a protective layer of mucus [46]. Studies have shown that goblet cells can act as antigen importers by taking up antigens and delivering them to antigen-presenting cells in the lamina propria via goblet cell-associated antigen passages (GAPs) [46, 47].

Enteroendocrine cells are rare hormone-producing cells controlling processes related to food intake [48]. The cell type has also been theorized to be involved in intestinal immunity due to their expression of microbial metabolite receptors, secretion of cytokines upon stimulation, and the fact that some of its hormones act directly on immune cells [48].

Tuft cells are flask-shaped chemosensory cells with brush-like microvilli extending from its body [49]. The cells are involved in immune and regulatory metabolic networks, monitoring intestinal content and secreting effector molecules upon stimulation [49]. Tuft cells are especially associated with defense against parasitic infections [49].



Enterocytes are columnar cells with microvilli forming a brush border on the apical surface [50]. They are the most abundant epithelial cell type in the large intestine, with the main function of absorbing nutrients from food passing through the brush border [50]. The cell type functions as a physical

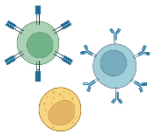
barrier to microbial invasion because of an enzyme surface coat on the brush border preventing the uptake of antigens [50]. Enterocytes also operates as non-professional antigen presenting cells due to their ability to internalize, process, and present antigens directly to T-cells [50].

1.2.4.4 Intestinal microenvironment cells

There is a complex microenvironment composed of diverse cell types surrounding the colonic crypt bases [40]. This microenvironment can influence intestinal regeneration from the stem cell niche by regulating the activity of signaling pathways [51], much in the same way the TME formed by both malignant and non-malignant cells in CRC can regulate signaling pathways to promote CRC progression and metastasis [25].

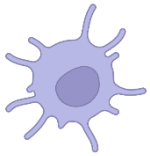
Fibroblasts are flat, spindle-shaped cells producing collagen and ground substance for the extracellular matrix, creating the structural framework for connective tissue [52]. ***Cancer-associated fibroblasts (CAFs)*** can in CRC promote the disease progression via multiple mechanisms, such as regulating signaling pathways involved in tumor progression [25].

Pericytes and endothelial cells are the main components of blood vessels, in which pericytes are mural cells that wrap around the endothelial cells forming the inner lining of the vessel wall [53]. The two cell types can interact with each other, both being involved in angiogenesis and vessel sprouting. ***Tumor-associated endothelial cells (TECs)*** in the TME have been found to produce growth factor receptors to enhance angiogenesis in CRC [25].



Lymphocytes such as T-cells, B-cells and NK cells work together during the immune response to detect and remove antigens or infected cells, and in general control the immune reaction [54, 55]. ***Tumor-infiltrating lymphocytes (TILs)*** can in CRC be involved in tumor immune evasion as well as tumor recognition, destruction, and elimination [25].

Macrophages are cells developed from monocytes, and they are phagocytic antigen presenting cells that can help initiate an immune response [56]. The cells can be classified as M1 or M2 subtypes [25]. ***Tumor-associated macrophages (TAMs)*** are classified as M2 in the TME of CRC, and can promote tumor progression by stimulating angiogenesis and inhibiting immune responses [25].



Dendritic cells (DCs) also process and present antigens to lymphocytes aiming to initiate an immune response [25]. ***Functional defect DCs*** are often present in CRC, as DC maturation is impaired and results in cells with insufficient antigen recognition. Studies on ovarian cancer show that infiltration of the TME by normal mature DCs is correlated with a favorable prognosis [25].

Mast cells can upon stimulation by an antigen or allergen release the contents of its granules to produce local responses characteristic of an allergic reaction [57]. ***Tumor-infiltrating mast cells (TIMs)*** have in CRC been related to promotion of disease progression and a poor

prognosis, although the general consequences of TIMs in a TME have varied based on the type and anatomical site of the tumor [58].

Neutrophils are phagocytic cells with cytoplasmic granules that contains enzymes which can destroy the ingested pathogen [59]. ***Tumor-associated neutrophils (TANs)*** have with accumulating evidence been suggested to support CRC tumor progression, although the cells can also be tumor suppressing due to their defensive function against antigens [25].

Smooth muscle cells in the muscularis mucosae are narrow, spindle-shaped cells which make up the smooth muscle [60]. This muscle tissue is also called involuntary muscle, as it contracts slowly and automatically to control the wall movement of the gastrointestinal tract and help with digestion [61]. Smooth muscle cells are understudied TME partners, but a recent study demonstrated that the cell type produces molecules that are able to modify epithelium behavior and thus affect tumor formation [62].

1.3 MicroRNA in colorectal cancer

The central dogma of molecular biology states that "DNA makes RNA makes proteins", where the genetic information stored in DNA can be transcribed into RNA and further translated into proteins (Figure 1.6) [63]. The gene expression can be regulated by a wide range of mechanisms, and as this expression of genes defines the cell type and function it is fundamental for cellular and organismal life [64].

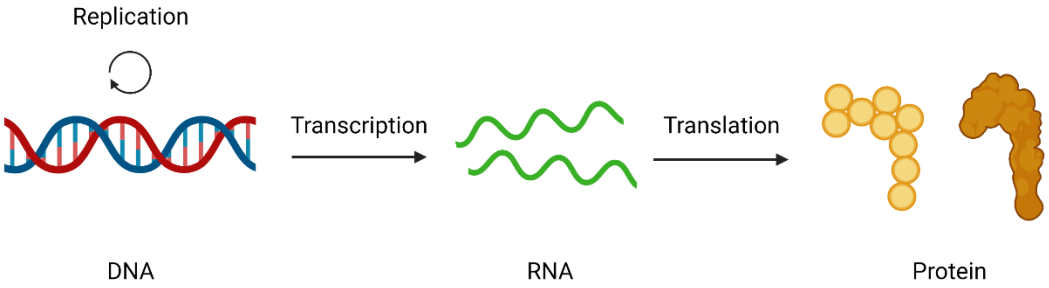


Figure 1.6: The central dogma of molecular biology. DNA is copied in a process known as replication before the genetic information stored in DNA is expressed by transcription into RNA. This is followed by translation into an encoded protein. Created with BioRender.com.

The human cellular transcriptome is a collection of all the RNA transcripts present in a cell, and includes both coding and non-coding molecules [65]. Protein-coding messenger RNA (mRNA) has historically been the most frequently studied RNA species, as non-coding RNA (ncRNA) was thought to be non-functional [65, 66]. It is now clear that ncRNA play multiple structural and regulatory roles in the molecular biology of the cell, and the species includes various subgroups [66].

A subgroup of ncRNA are small ncRNA, which includes microRNA (miRNA) [65]. As the name imply, miRNA are short molecules with a length of about 22 nucleotides that regulate gene expression at the post-transcriptional level by targeting protein-coding mRNA [67]. One miRNA can regulate the expression of many genes, while one gene can be regulated by multiple miRNAs [68]. After its discovery in roundworm in 1993 by Lee,

R.C. et al. [69], miRNA have been found in most living organisms, where some specific molecules have been highly conserved across species [70]. Most mammalian mRNAs are also found to be conserved targets of miRNAs [71], and this might illustrate miRNAs having a widespread importance and role in a broad range of biological processes.

1.3.1 MicroRNA biogenesis and post-transcriptional gene silencing

Production of miRNA starts in the nucleus with gene transcription by RNA polymerase into a long primary miRNA (pri-miRNA) with one or more hairpin structures (Figure 1.7) [67]. The pri-miRNA sequence is cleaved by a microprocessor complex to form precursor miRNA (pre-miRNA), before being exported to the cytoplasm [67]. In the cytoplasm, the hairpin stem of the pre-miRNA is cut by a dicer enzyme to produce mature miRNA in a duplex [67].

One strand of the miRNA duplex can be incorporated into an RNA-induced silencing complex (RISC), while the other strand gets degraded [67, 72]. The retained strand is used as a template by RISC to bind complementary elements mostly located on the 3' untranslated region (UTR) of target mRNA molecules [67]. The interaction between miRNA and mRNA triggers events leading to RNA interference (RNAi) in terms of translation inhibition or mRNA destabilization, and ultimately gene silencing [67, 73].

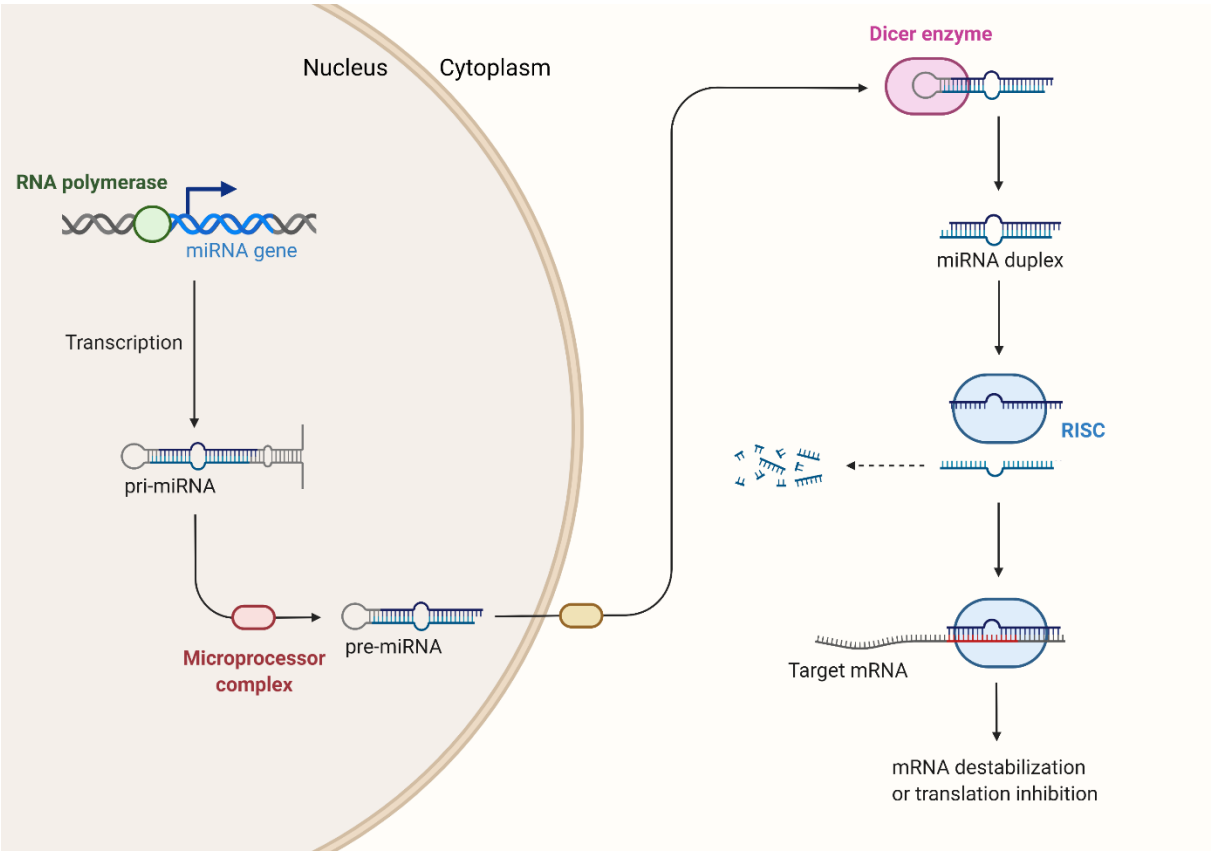


Figure 1.7: Biogenesis and post-transcriptional gene silencing by miRNA. Production of miRNA starts in the nucleus with synthesis of pri-miRNA. The molecule is cleaved into pre-miRNA, transported into the cytoplasm, and cleaved again into a miRNA duplex. One strand of the miRNA duplex is used as a template by RISC to bind target mRNA, where the miRNA-mRNA interaction triggers mRNA destabilization or translation inhibition. Abbreviations: miRNA, microRNA; pri-miRNA, primary microRNA; pre-miRNA, precursor microRNA; RISC, RNA-induced silencing complex; mRNA, messenger RNA. Adapted from Bartel, D.P. [67] and created with BioRender.com.

1.3.2 MicroRNAs are dysregulated in colorectal cancer

Compelling evidence have demonstrated that miRNAs are dysregulated in human cancers [74]. Such dysregulation could be caused by several underlying mechanisms, including amplification or deletion of miRNA genes, abnormal transcriptional control of miRNAs due to dysregulation of key transcription factors, aberrant epigenetic changes, and defects in the miRNA biogenesis machinery [74].

Altered miRNA in CRC tumors is believed to support tumorigenesis by affecting cell proliferation, metastasis, angiogenesis, autophagy, apoptosis, and the radiosensitivity of cancer cells [75]. Different miRNAs may practically function as either tumor suppressor genes or oncogenes under certain circumstances, and their involvement in tumorigenesis could be due to their regulation of only a few specific targets despite their multiple targets [68].

1.3.3 MicroRNAs as biomarkers in colorectal cancer

The current gold standard for diagnosing CRC is colonoscopy, but this procedure is invasive, expensive, and carries patient risk [76]. Newer non-invasive CRC detection methods include biomarker stool-based and blood-based tests, but current tests have relatively poor selectivity and sensitivity and thus produces a high rate of both false positives and false negatives [3, 76]. Therefore, there is a need for a more accurate non-invasive CRC screening procedure [76]. Several biomolecules are being investigated as alternative biomarkers to current screening methods, where a considerable amount of studies have identified miRNAs as a good biomarker candidate for CRC diagnosis, prognosis, and prediction due to their altered expression profiles in cancers [74].

Even though miRNAs are produced in the nucleus and regulate gene expression in the cytoplasm of the cell, they can also be found in the extracellular environment such as serum, possibly originating from passive leakage from apoptotic or damaged cells [72]. The circulating miRNAs can have a role in intercellular communication and affect gene expression in adjacent or distant target cells [72]. Recent studies indicates that the combined signatures of specific circulating miRNAs provide high specificity, sensitivity, and reproducibility in screening of CRC, where these data can be obtained using a non-invasive blood-based test approach [72, 77].

2 Aims of the study

The overall aim of this study was to investigate both tissue heterogeneity of primary CRC tumor and circulating miRNAs in serum of early- and advanced-stage CRC compared to control individuals.

The specific aims related to tissue-investigation were:

- Develop and optimize a protocol for establishing single cell suspension from fresh CRC tissue.
- Analyze and evaluate scRNA-seq data of single cell suspensions to identify expressed genes and cell type composition in CRC tumor.

The specific aim related to blood-investigation was:

- Isolate RNA and conduct miRNA-seq from serum samples of CRC patients with both localized and metastatic disease and control individuals, to identify differentially expressed miRNA between the groups.

3 Methodology

The methodology for this master's thesis can be found in appendices (Appendix A).

4 Results

4.1 Validation of protocol for establishing single cell suspension from fresh colorectal cancer tissue

A protocol for establishing single cell suspension from fresh CRC tissue was finalized and evaluated in terms of cell number by cell counting, cell appearance by microscopy, and cell viability by flow cytometry. Established single cell suspensions had an average cell count of about 3 million cells/mL, where most cells were intact and single with some cell clumping and debris (Supplementary Figure 1). Cell viability was estimated to be 68.8% (Supplementary Table 1), where cell viability over 60% also was confirmed by the Genomics Core Facility (GCF) at the Norwegian University of Science and Technology (NTNU).

4.2 Identified expressed genes and cell type composition in colorectal cancer tumor tissue

The biological material used in this part of the study were CRC single cell suspensions established from fresh CRC tissues by using a protocol described in this thesis (Appendix A). ScRNA-seq was performed on 4 samples (Table 4.1) followed by a downstream analysis workflow using Seurat in R [78]. A total number of 23,440 cells were obtained, where separate cluster analyses of each sample found that CRC tissue could be classified into 11-13 clusters (Supplementary Figure 2) with respect to their mRNA transcriptomes. The different clusters were manually annotated based on their top 10 expressed cell-type-specific marker genes (Supplementary Table 2-Supplementary Table 9), ultimately detecting major cell types in each sample (Figure 4.1).

Table 4.1: Patient sample characteristics. Three samples from men with stage I-II colorectal cancer and one sample from a man with stage IV cancer were used in this study. 50% of the patients had right-sided cancer (sample 554 and 559) and 50% had left-sided cancer (sample 556 and sample 569).

Sample ID	Gender	CEA [µg/L]	CRP [mg/L]	Colorectal cancer staging by TNM			Tumor location
				Tumor (T)	*Node (N)*	*Metastasis (M)*	
554	Male	9,6	25	pT4a	N2b (9/14)	Peritoneal metastasis (M1)	Right-sided proximal colon (ileocecal valve)
556	Male	2,6	<5	cT3-T4	N0	M0	Left-sided distal colon (sigmoid colon)
559	Male	3	<5	pT3	N0	M0	Right-sided proximal colon (ileocecal valve)
569	Male	26,7	<5	pT4a	N1b (2/20)	M0	Left-sided distal colon (sigmoid colon)

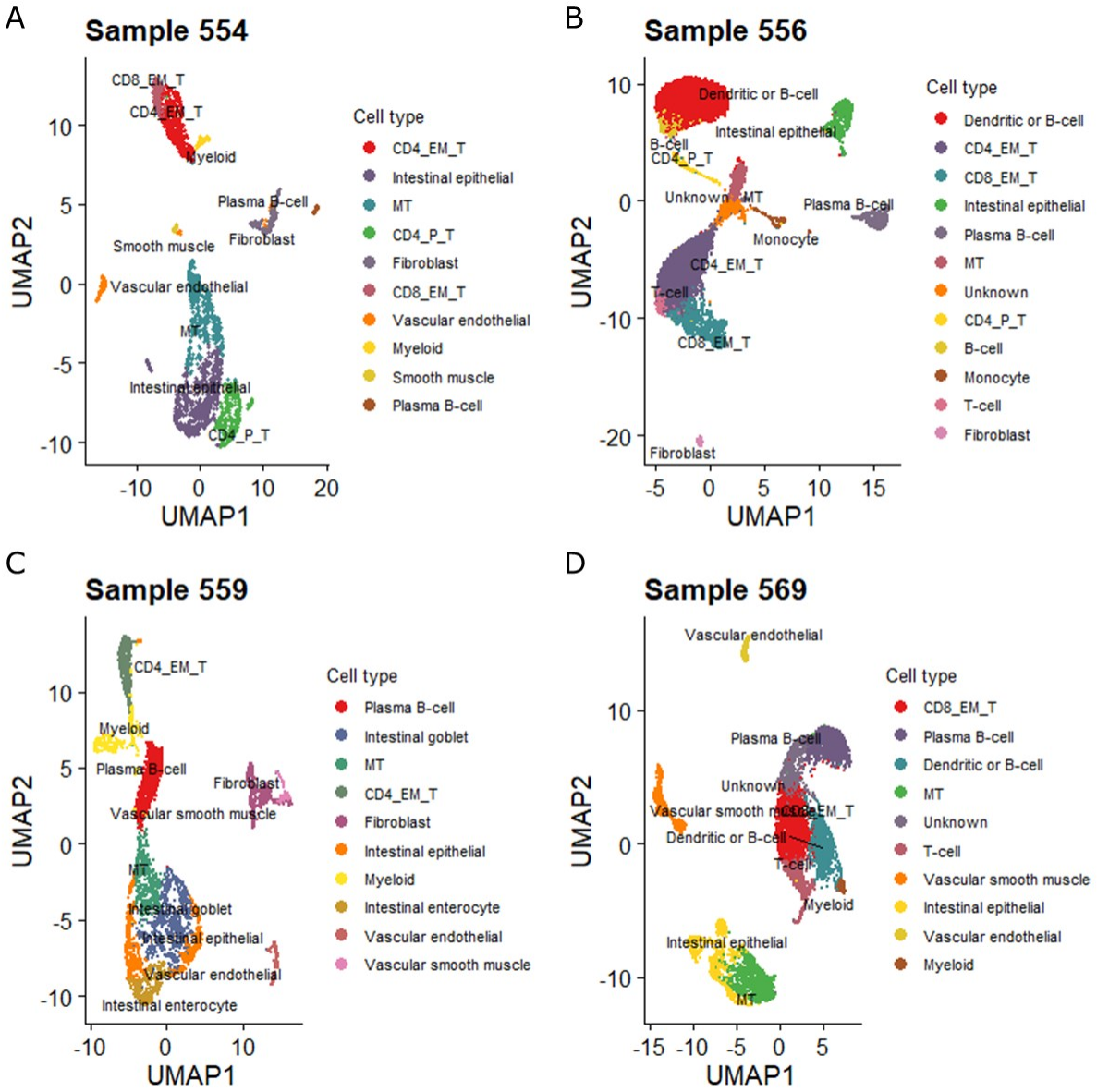


Figure 4.1: UMAP plot representation of cell types present in different samples of CRC tissue. (A) The major cell types found in sample 554. Abbreviations listed in (D). **(B)** The major cell types found in sample 556. Abbreviations listed in (D). **(C)** The major cell types found in sample 559. Abbreviations listed in (D). **(D)** The major cell types found in sample 569. Abbreviations: CD4_EM_T, CD4+ effector memory T-cell; MT, mitochondrial; CD4_P_T, CD4+ proliferating T-cell; CD8_EM_T, CD8+ effector memory cell.

After performing cell-type annotation for each individual sample, the 4 samples were integrated using the integration procedure in Seurat (Appendix A). A new cluster analysis was performed in the integrated data, revealing a total of 18 unique clusters (Figure 4.2-A), with contribution of varying cell numbers from each sample (Table 4.2). The top 10 cell-type-specific marker genes for each automatically annotated cluster were identified (Supplementary Table 10). A heatmap confirmed unique cell-type-specific gene expression for the 18 major identified cell types B-cells, CD4+ effector memory T-cells, CD4+ proliferating T-cells, CD8+ effector memory T-cells, dendritic cells/B-cells, fibroblasts, intestinal enterocytes, intestinal epithelial cells (unspecified subgroup), intestinal goblet cells, monocytes, mitochondrial gene-expressing cells, myeloid cells (unspecified subgroup), plasma B-cells, smooth muscle cells, T-cells (unspecified

subgroup), unknown cell type, vascular endothelial cells, and vascular smooth muscle cells (Figure 4.2-B).

Next, the specific cell types that had similar gene expression of the marker genes were grouped into 6 rough cell types: stromal cells, endothelial cells, intestinal epithelial cells, and three different subtypes of immune cells (Figure 4.2-C). A group of unknown cells were identified that showed similarity in gene expression with both the intestinal epithelial cells and different subtypes of immune cells, clustering together in the middle of these clusters. Mitochondrial gene-expressing cells displayed similarity with intestinal epithelial cells. The rough classification showed that immune cells separated clearly from stromal, endothelial, and epithelial cells. Further, stromal cells and endothelial cells were clearly separated from each other and from other types of cells.

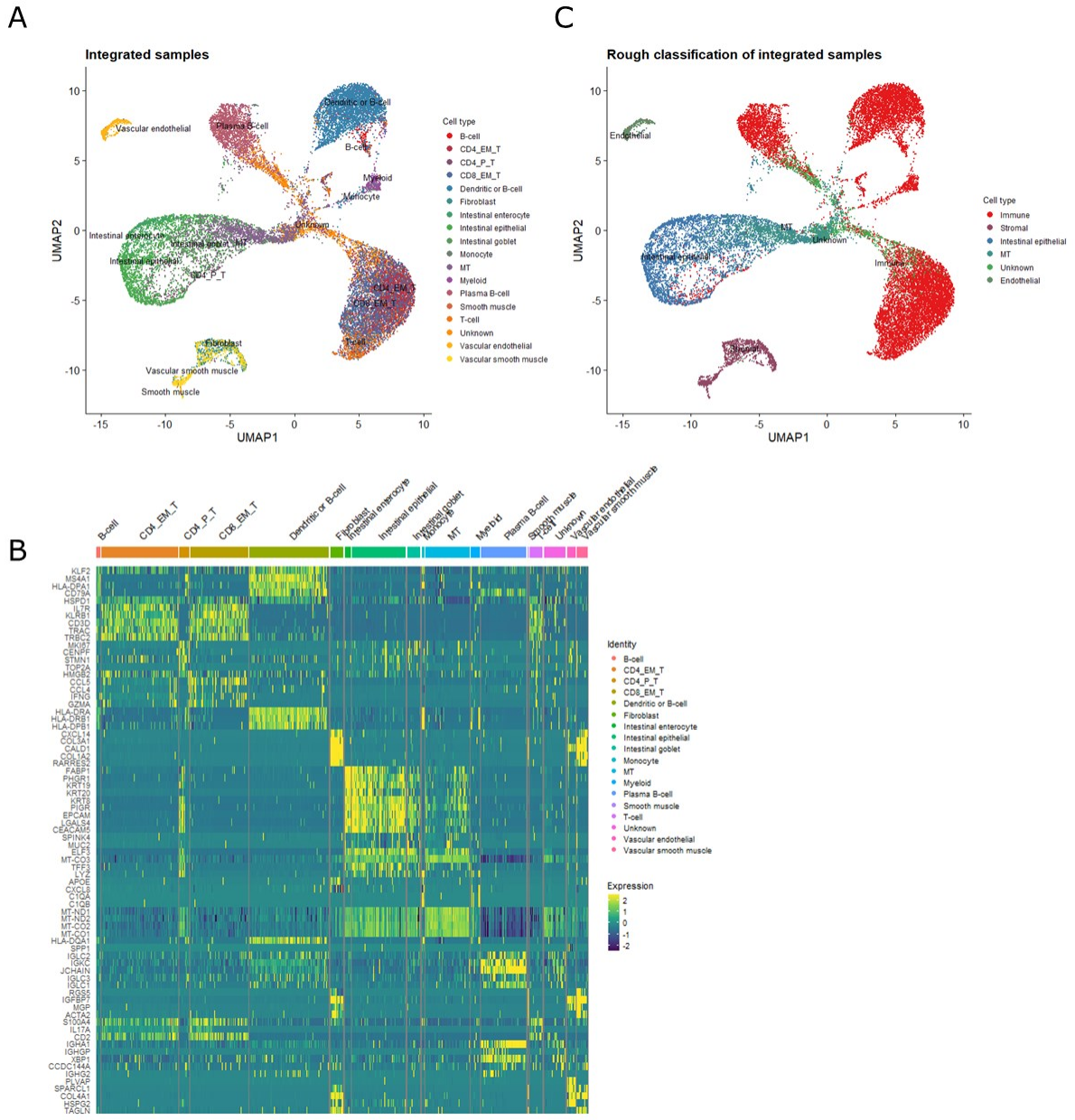


Figure 4.2: Portrayals of cell types present in CRC tissue. (A) UMAP plot representation of CRC tissue with 18 distinct cell types. Abbreviations listed in (C). **(B)** Heatmap showing the top 5 cell-type-specific marker genes of each cluster in CRC tissue. Abbreviations listed in (C). **(C)** UMAP plot representation of CRC tissue, where the major cell types have been classified into rough cell

type groups. Abbreviations: CD4_EM_T, CD4+ effector memory T-cell; CD4_P_T, CD4+ proliferating T-cell; CD8_EM_T, CD8+ effector memory T-cell; MT, mitochondrial.

Table 4.2: Number of cells in every major cell type for each sample. Intestinal epithelial cells (unspecified subgroup), plasma B-cells, mitochondrial gene-expressing cells, and at least one type of effector memory T-cell were the cell types found in all four samples. Abbreviations: MT, mitochondrial; CD4_EM_T, CD4+ effector memory T-cell; CD8_EM_T, CD8+ effector memory T-cell; CD4_P_T, CD4+ proliferating T-cell.

	Number of cells				
Cell type cluster name	*Sample 554*	*Sample 556*	*Sample 559*	*Sample 569*	*Total number of cells*
Intestinal epithelial	668	494	633	889	*2684*
Plasma B-cell	37	409	817	1002	*2265*
MT	498	407	447	828	*2180*
CD4_EM_T	447	2927	441		*3815*
CD8_EM_T	161	972		1745	*2878*
Fibroblast	210	75	388		*673*
Vascular endothelial	104		152	174	*430*
Myeloid	57		321	78	*456*
CD4_P_T	295	198			*493*
Dendritic or B-cell		3063		884	*3947*
T-cell		143		498	*641*
Unknown		387		705	*1092*
Vascular smooth muscle			103	446	*549*
Smooth muscle	41				*41*
B-cell		185			*185*
Monocyte		150			*150*
Intestinal enterocyte			302		*302*
Intestinal goblet cell			659		*659*
Total number of cells	**2518**	**9410**	**4263**	**7249**	***23440***

The rough cell type groups (stromal cells, endothelial cells, and intestinal epithelial cells) were further subclustered and analyzed. The subclusters were manually annotated based on their top 10 expressed cell-type-specific marker genes (Supplementary Table 11-Supplementary Table 16), where some specific marker genes ultimately determined the assignment of cluster cell type names (Table 4.3).

Table 4.3: Cell-type-specific marker genes ultimately determining stromal, endothelial, and intestinal epithelial subset cell type names. Abbreviations: CAF, cancer-associated fibroblast; CTF, crypt-top fibroblast; LPF, lamina propria fibroblast, MT, mitochondrial; TEC, tumor-associated endothelial cell; EC, endothelial cell; CBC, crypt base cell.

Cell type group	**Annotated cell type cluster name**	**Cell-type-specific marker gene**	**References**
Stromal cells	Pericyte	RGS5	PanglaoDB [79] Elmentaite, R. et al. (2021) [80] Dasgupta, S. et al. (2021) [81]
		MCAM	PanglaoDB [79] Elmentaite, R. et al. (2021) [80] Kotsiliti, E. (2022)[82]
		NOTCH3	PanglaoDB [79] Elmentaite, R. et al. (2021) [80] Tefft, J.B. et al. (2022) [83]
	CAF	MMP3	Bigaeva, E. et al. (2020) [84] Uhlitz, F. et al. (2020) [85]
		MMP11	Bigaeva, E. et al. (2020) [84] Uhlitz, F. et al. (2020) [85]
		COL1A1	Uhlitz, F. et al. (2020) [85]
		COL1A2	Uhlitz, F. et al. (2020) [85]
	Plasma B-cell	IG-genes	PanglaoDB [79]
	CTF	PDGFRA	Brügger, M.D. et al. (2020) [86]

		BMP5	Brügger, M.D. et al. (2020) [86]
	Myofibroblast	HHIP	Elmentaite, R. et al. (2021) [80]
		MYH11	Bigaeva, E. et al. (2020) [84]
		NPNT	Elmentaite, R. et al. (2021) [80]
		ACTG2	Bigaeva, E. et al. (2020) [84]
	LPF	CCL13	Bigaeva, E. et al. (2020) [84]
		CCL11	PanglaoDB [79] Bigaeva, E. et al. (2020) [84]
		ADAMDEC1	Bigaeva, E. et al. (2020) [84]
		CCL2	Bigaeva, E. et al. (2020) [84]
		APOE	Bigaeva, E. et al. (2020) [84]
Endothelial cells	Mitochondrial gene-expressing cells	MT-genes	HPA [87]
	Activated TEC	CCN2 (CTFG)	Liu, S.C. et al. (2014) [88]
		CPE	Goveia, J. et al. (2020) [89]
		CLU	Goveia, J. et al. (2020) [89]
		CCL14	Goveia, J. et al. (2020) [89]
		HLA-DRB1	Goveia, J. et al. (2020) [89]
		HLA-DRA	Goveia, J. et al. (2020) [89]
		HLA-DPA1	Goveia, J. et al. (2020) [89]
	Tip TEC	SPARC	Goveia, J. et al. (2020) [89]
		CD34	Siemerink, M.J. et al. (2012) [90]
		ANGPT2	Goveia, J. et al. (2020) [89] Zarkada, G. et al. (2021) [91]
	Immature TEC	HSPG2	Goveia, J. et al. (2020) [89]
		JAG1	Goveia, J. et al. (2020) [89]
	Proliferative EC	MKI67	Uxa, S. et al. (2021) [92]
		NUSAP1	Han, G. et al. (2018) [93]
		HMGB2	Kalucka, J. et al. (2020) [94]
		STMN1	Kalucka, J. et al. (2020) [94]
		TUBA1B	Kalucka, J. et al. (2020) [94]
Intestinal epithelial cells	Mitochondrial gene-expressing cell	MT-genes	HPA [87]
	Secretory progenitor 1	SOX4	Fazilaty, H. et al. (2021) [95] Gracz, A.D. et al. (2018) [96]
		EPHB3	Sancho, R. et al. (2015) [97]
		FCGBP	Habowski, A.N. et al. (2020) [98]
	CBC/Paneth cell	LYZ	PanglaoDB [79] Nakanishi, Y. et al. (2016) [99]
		OLFM4	PanglaoDB [79] van der Flier, L.G. et al. (2009) [100]
	Secretory progenitor 2	PLA2G2A	Rajagopal, J. et al. (2021) [101]
		DMBT1	Rajagopal, J. et al. (2021) [101]
		C15orf48	Rajagopal, J. et al. (2021) [101]
		PIGR	Rajagopal, J. et al. (2021) [101]
	Enterocyte	FABP1	PanglaoDB [79]
		KRT20	PanglaoDB [79]
		SLC26A3	PanglaoDB [79]
	Plasma B-cell	IG-genes	PanglaoDB [79]
	Iron-storing epithelial cell	FTH1	Xu, M. et al. (2020) [102] Xu, S. et al. (2021) [103]
		FTL	Xu, M. et al. (2020) [102] Xu, S. et al. (2021) [103]
	Tuft-2 cell	SH2D6	Xiong, Z. et al. (2022) [104]
		HPGDS	Xiong, Z. et al. (2022) [104]
		SIPB	Xiong, Z. et al. (2022) [104]

Subtypes of the cell groups were identified based on the mentioned marker genes (Figure 4.3), with contribution of varying cell numbers from each sample (Table 4.4). Stromal cells were clustered into 6 cell types (pericytes, cancer-associated fibroblasts (CAFs), plasma B-cells, crypt-top fibroblasts (CTFs), myofibroblasts, and lamina propria fibroblasts (LPFs)) with pericytes being the largest cluster with 303 cells and LPFs the smallest with 105 cells.

Endothelial cells were clustered into 5 cell types (mitochondrial gene-expressing cells, activated tumor-associated endothelial cells (TECs), tip TECs, immature TECs, and proliferative endothelial cells). Here a clear separation of immature TECs and the other clusters were identified, with immature TECs showing more similarities to mitochondrial gene-expressing cells than the other endothelial subtypes. The largest subcluster were

constituted of 127 mitochondrial gene-expressing cells and the smallest cluster included 18 proliferative ECs.

Intestinal epithelial cells were clustered into 8 cell types (mitochondrial gene-expressing cells, secretory progenitor 1, crypt base cells (CBCs)/Paneth cells, secretory progenitor 2, enterocytes, plasma B-cells, iron-storing epithelial cells, and tuft-2 cells), with mitochondrial gene-expressing cells making up the largest cluster of 985 cells, followed by 719 secretory progenitor 1 cells and 665 CBC/Paneth cells. Tuft-2 cells comprised the smallest cluster with 44 cells and was also the only cell type clearly separated from the other intestinal epithelial subtypes.

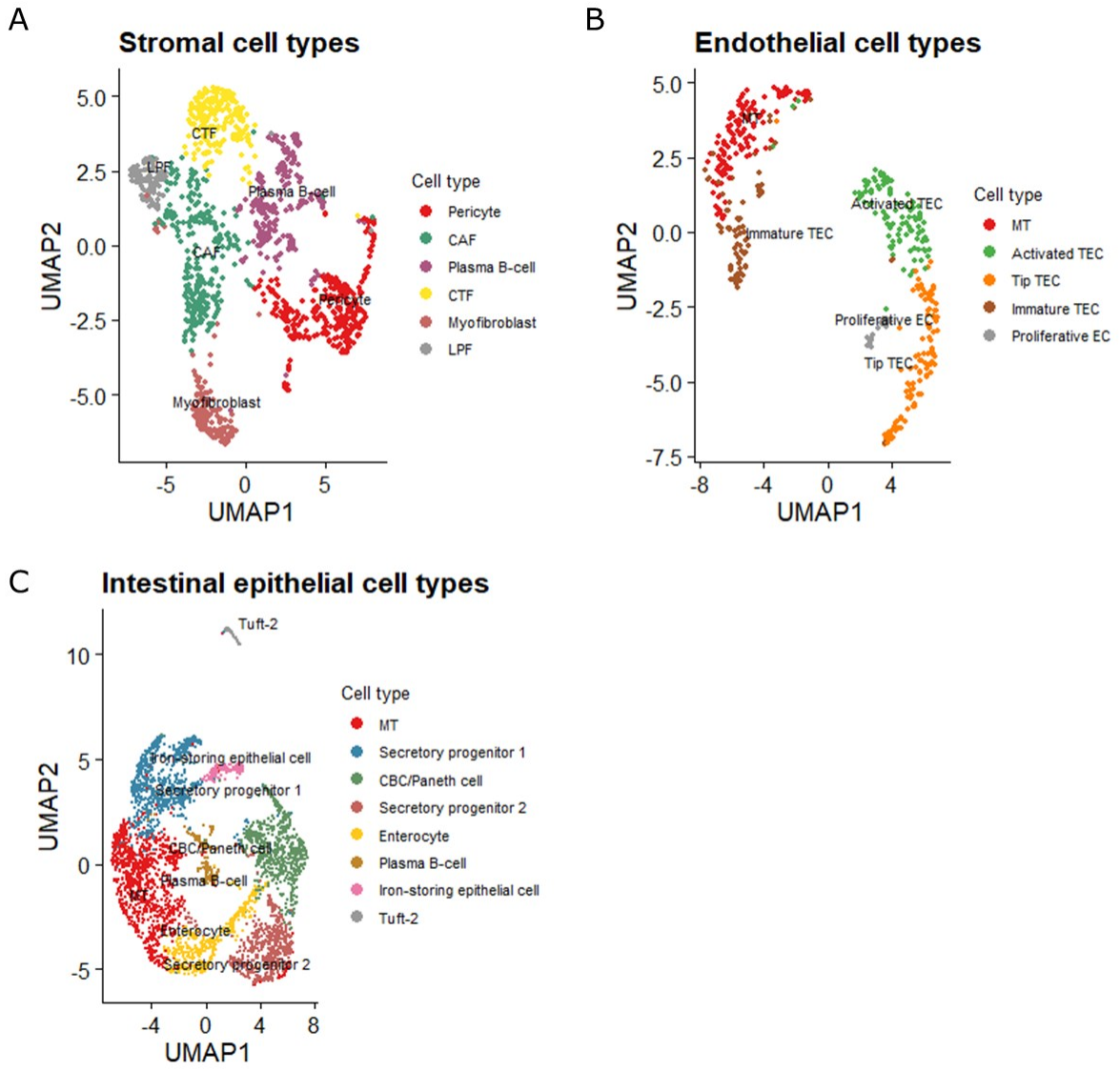


Figure 4.3: Subset cell types of rough cell type groups present in CRC tissue. (A) UMAP plot representation of the subclusters present within the stromal cell type group. Abbreviations listed in (C). **(B)** UMAP plot representation of the subclusters present within the endothelial cell type group. Abbreviations listed in (C). **(C)** UMAP plot representation of the subclusters present within the intestinal epithelial cell type group. Abbreviations: CAF, cancer-associated fibroblast; CTF, crypt-top fibroblast; LPF, lamina propria fibroblast; MT, mitochondrial; TEC, tumor-associated endothelial cell; EC, endothelial cell; CBC, crypt base cell.

Table 4.4: Number of cells in every subcluster of each rough cell type group. Each sample had varying contribution of cell number in each cell type subset cluster. Abbreviations: CAF,

cancer-associated fibroblast; CTF, crypt-top fibroblast; LPF, lamina propria fibroblast; TEC, tumor-associated endothelial cell; EC, endothelial cell; CBC, crypt base cell.

Cell type group	**Cell type cluster name**	**Number of cells**				***Total number of cells***
		Sample 554	*Sample 556*	*Sample 559*	*Sample 569*	
Stromal cells	Pericyte	43	24	118	118	*303*
	CAF	92	34	89	67	*282*
	Plasma B-cell	3	4	118	104	*229*
	CTF	45	2	73	94	*214*
	Myofibroblast	42	10	45	33	*130*
	LPF	26	1	48	30	*105*
	Total number of cells	**251**	**75**	**491**	**446**	***1263***
Endothelial cells	Mitochondrial gene-expressing cells	32	44	0	51	*127*
	Activated TEC	28	25	0	50	*103*
	Tip TEC	30	41	0	32	*103*
	Immature TEC	12	31	0	36	*79*
	Proliferative EC	2	11	0	5	*18*
	Total number of cells	**104**	**152**	**0**	**174**	***430***
Intestinal epithelial cells	Mitochondrial gene-expressing cells	158	185	528	114	*985*
	Secretory progenitor 1	23	79	496	121	*719*
	CBC/Paneth cell	157	41	240	227	*665*
	Secretory progenitor 2	174	70	147	160	*551*
	Enterocyte	125	60	129	65	*379*
	Plasma B-cell	3	51	46	68	*168*
	Iron-storing epithelial cell	2	0	7	125	*134*
	Tuft-2 cell	26	8	1	9	*44*
	Total number of cells	**668**	**494**	**1594**	**889**	***3645***
Grand total number of cells		*1023*	*721*	*2085*	*1509*	*5338*

Heatmaps confirmed unique cell-type-specific gene expression for the 6 identified subtypes in stromal cell type group, the 5 identified subtypes in epithelial cell type group, and the 8 identified subtypes in intestinal epithelial cell type group (Figure 4.4). In the stromal cell type group, CTFs, myofibroblasts, and LPFs ultimately showed the most distinct cell-type-specific gene expression profiles with all their top 5 marker genes being uniquely and highly expressed. Plasma B-cells had a less distinct expression profile than the other stromal cells.

Proliferative ECs exhibited the most defined gene expression profiles among the endothelial subtypes, followed by activated TECs. Both tip TECs and immature TECs showed similarity in gene expression, but a higher expression of the genes was found in the tip TECs.

The subtypes of endothelial cells showed expression of many of the same genes, apart from plasma B-cells and tuft-2 cells. Tuft-2 cells did not express any of the top 5 cell-type-specific marker genes found in the other subtypes. The secretory progenitor 2 cells had a more distinct profile than the other secretory progenitor cell type, at the same time as secretory progenitor 1 cells expressed more MT-genes.

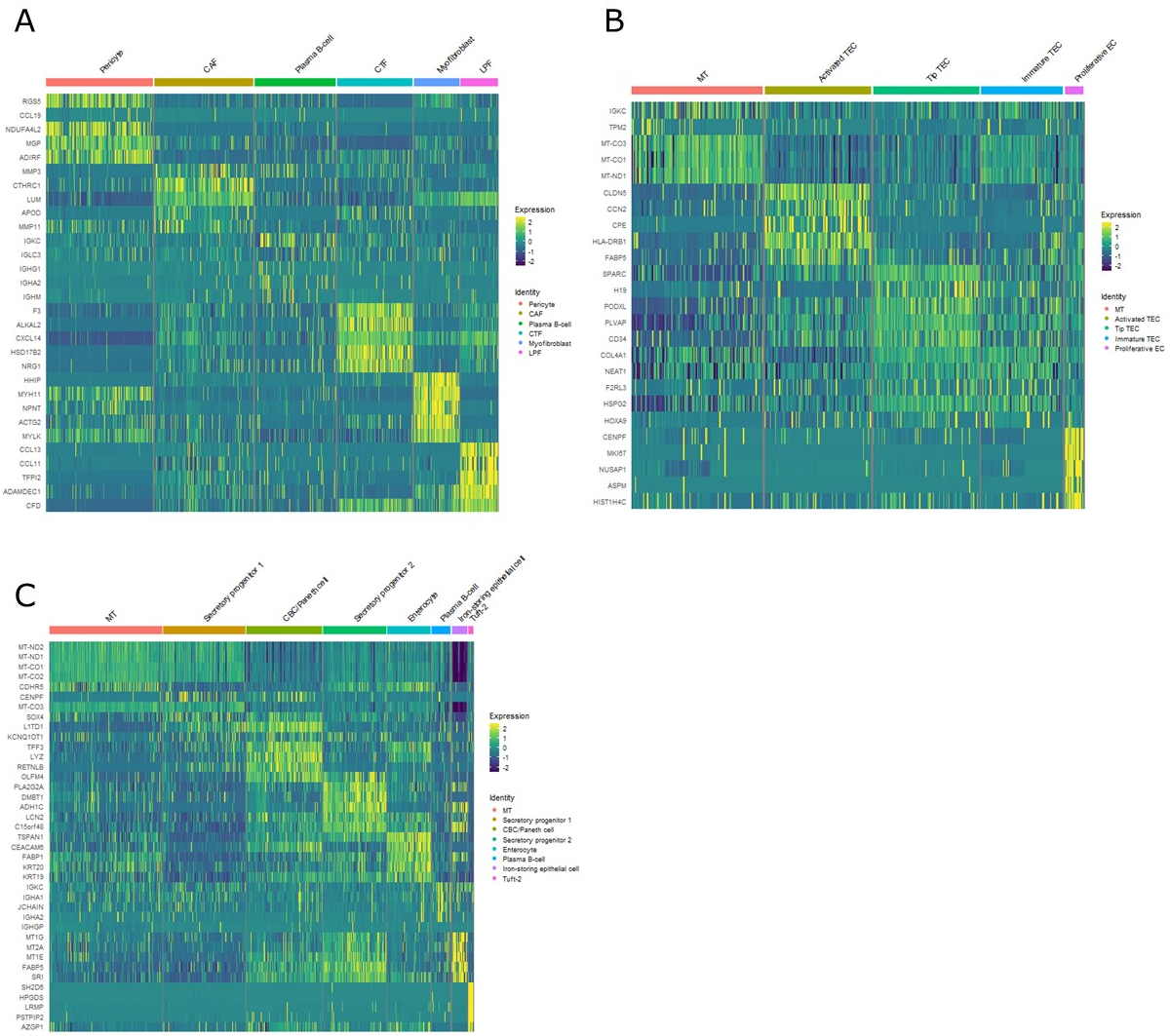


Figure 4.4: Heatmaps showing the top 5 cell-type-specific marker genes of each subset in the rough cell type groups. (A) Heatmap over the identified clusters in the stromal cell type group. Abbreviations listed in (C). **(B)** Heatmap over the identified clusters in the endothelial cell type group. Abbreviations listed in (C). **(C)** Heatmap over the identified clusters in the intestinal epithelial cell type group. Abbreviations: CAF, cancer-associated fibroblast; CTF, crypt-top fibroblast; LPF, lamina propria fibroblast; MT, mitochondrial; TEC, tumor-associated endothelial cell; EC, endothelial cell; CBC, crypt base cell.

4.3 Identified differentially expressed circulating miRNAs between colorectal cancer patient groups

The biological material used in this part of the study consisted of serum samples from CRC patients. The patients were separated into three main groups, which included false positive CRC patients (hereafter referred to as healthy or false positives) (n=21), true positive CRC patients with localized disease (n=16), and true positive CRC patients with metastatic disease (n=16). The false positives represent individuals seeking medical consultation with CRC symptoms that were characterized as healthy. Small RNA-seq were performed on 47 samples with high quality of all samples (Supplementary Figure 3). Several classes of RNAs were detected (Supplementary Figure 4), where the sequencing data revealed a clear enrichment of RNA fragments with 22 nucleotides in length, corresponding to miRNAs (Supplementary Figure 5).

A principal component analysis (PCA) of mature miRNA expression found that the localized CRC samples were different from the metastatic CRC samples (Figure 4.5-A). The local CRC samples had slightly more variation in the PCA than the metastatic samples. The PCA analysis also revealed a similarity between healthy samples and the two groups of CRC samples, where the healthy samples formed a cluster along PC2 roughly in the middle of the two CRC sample groups.

A differential expression analysis was performed to detect differentially expressed miRNA between the patient groups. 53 significant miRNAs between local CRC and healthy individuals, 7 significant miRNAs between true positive CRC patients with metastatic disease and healthy individuals, and 30 significant miRNAs between true positive CRC patients with metastatic disease and localized disease were identified (Supplementary Table 17-Supplementary Table 19) and visualized by volcano plots (Figure 4.6). In general, different sets of significant miRNAs were observed between the three comparisons (Figure 4.5-B), where no miRNAs were significant across all comparisons.

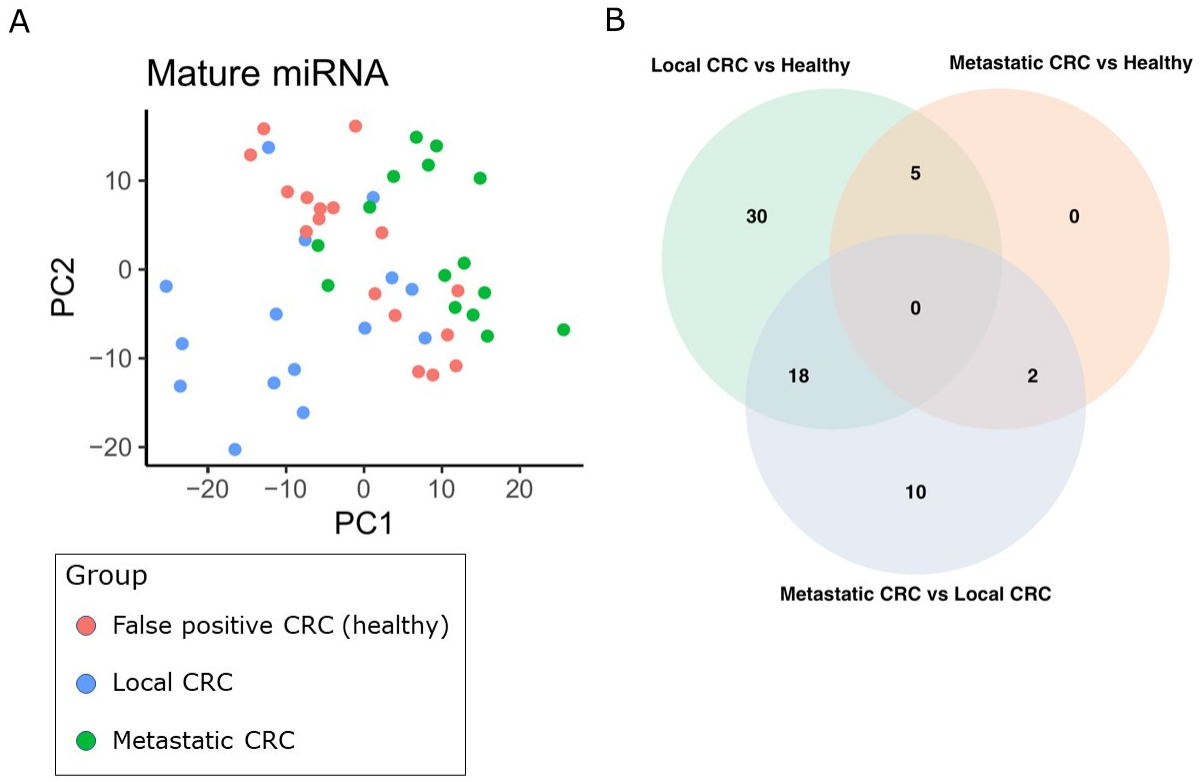


Figure 4.5: PCA plot and Venn diagram for serum miRNA in CRC patient groups. (A) PCA plot of mature miRNA expression (cpm, log2) in serum of CRC patient groups. Dots represent samples and are colored according to patient group. Patients with low miRNA expression (less than 1 cpm in more than 50% of samples) were excluded from the PCA. Abbreviations: PC, principal component; PCA, principal component analysis. **(B)** Number of differentially expressed miRNAs for each comparison. The Venn diagram includes both up- and down-regulated miRNAs.

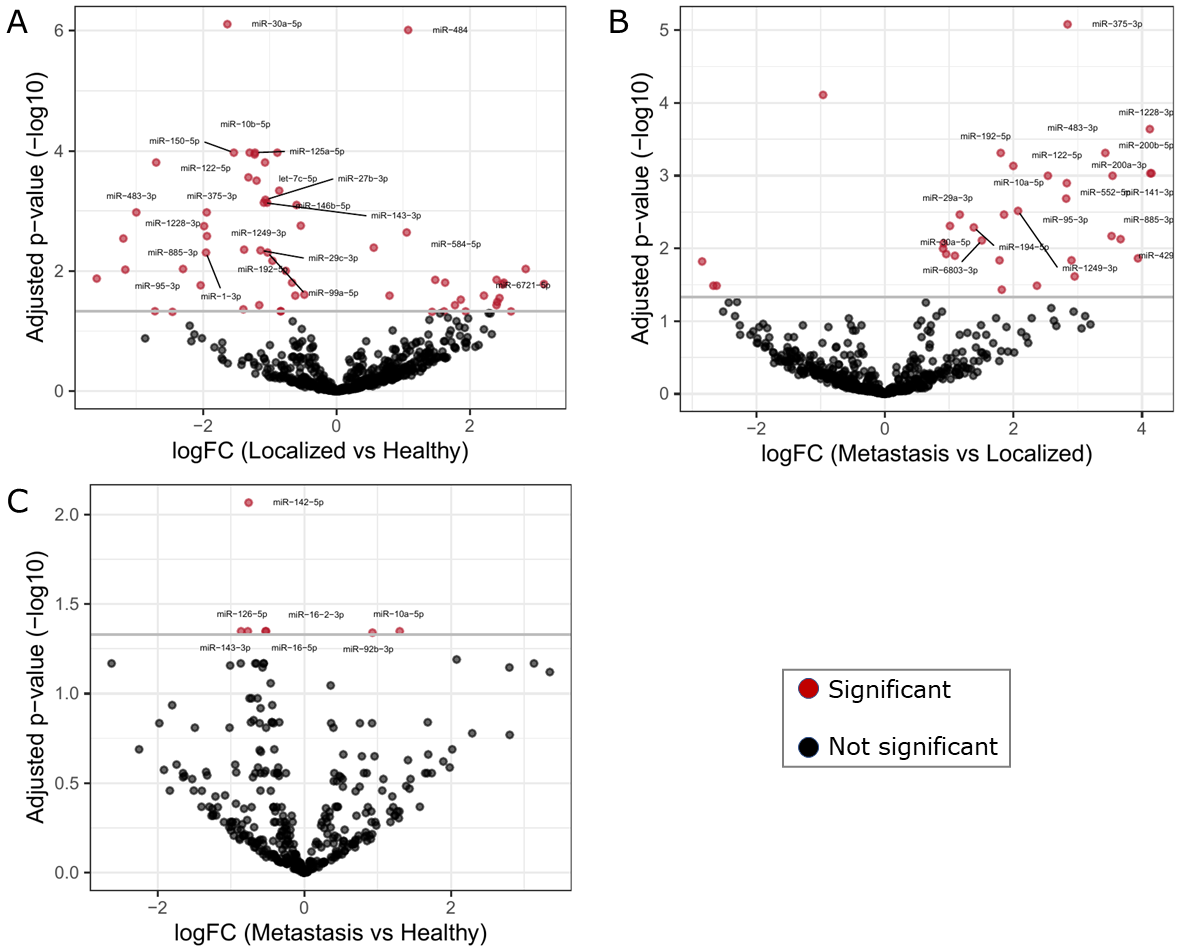


Figure 4.6: Differentially expressed miRNAs for the three comparisons of CRC patient groups. The x-axes of the volcano plots show the log2 fold change between the comparison indicated in the x-axis text. The y-axes show the inverted Benjamini-Hochberg adjusted p-values.

Unsupervised clustering of all samples in the three patient groups with respect to their differentially expressed miRNAs also demonstrated that the localized CRC samples were distinct from the metastatic CRC samples (Figure 4.7). The samples in the two patient groups were clustered distinctly into two separate groupings, except for two localized CRC samples being included in the group with the metastatic CRC samples. The heatmap also revealed a similarity in miRNA expression between the healthy samples and the two groups of CRC samples, mostly the metastatic ones.

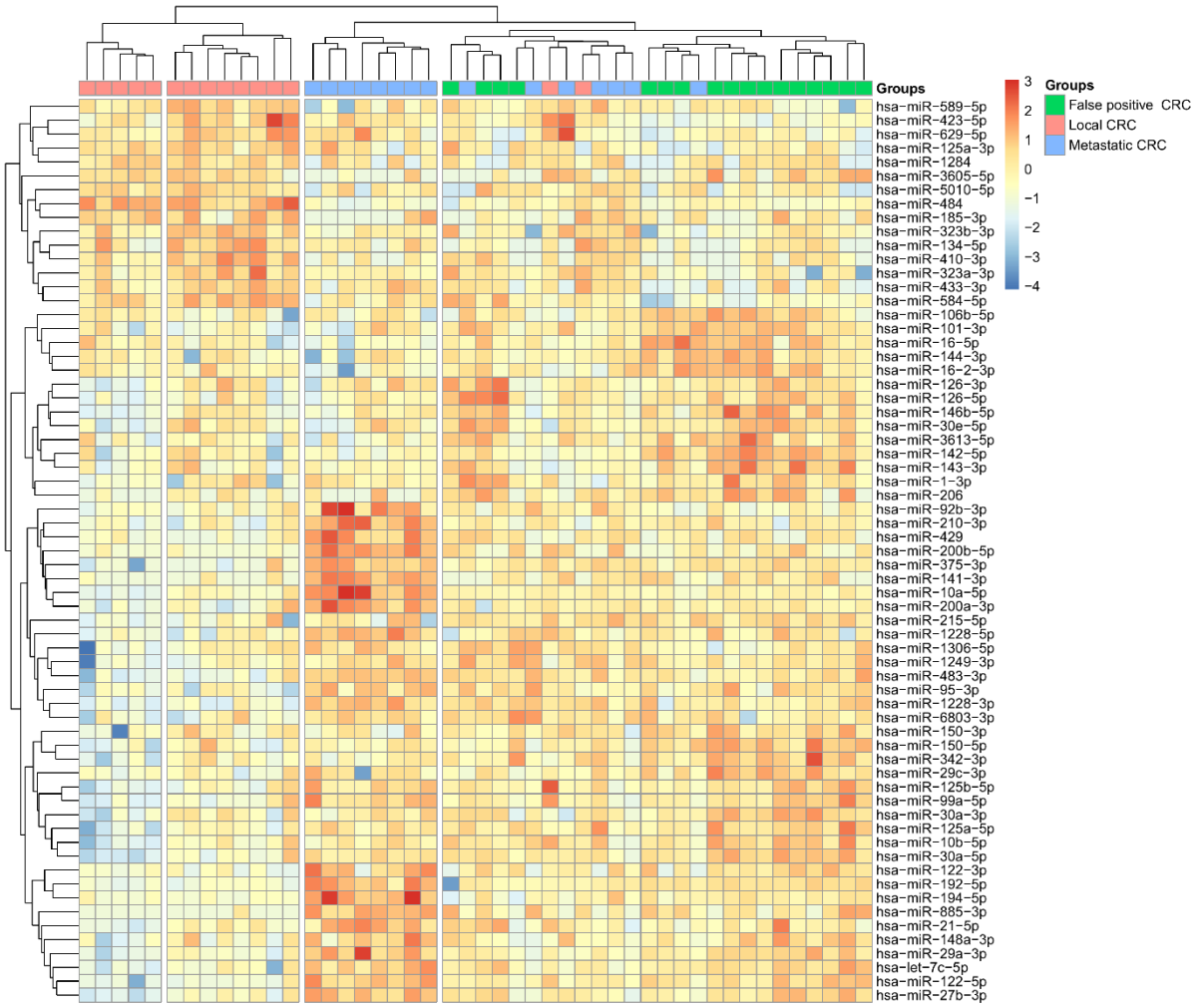


Figure 4.7: Heatmap of miRNA expression in CRC patient groups. Unsupervised clustering of samples with respect to the differentially expressed miRNAs. The heatmap includes miRNAs that are significant in at least one of the three comparisons. Red color indicates high expression and blue color indicates low expression. Groups are indicated at the top.

A selection of differentially expressed miRNAs identified in the comparisons of all three patient groups were further investigated (Table 4.5). Five miRNAs (miR-142-5p, miR-16-5p, miR-143-3p, miR-126-5p, and miR-16-2-3p) were downregulated in both stages of CRC compared to healthy subjects, suggesting these miRNAs could be an indicator of the disease. Two miRNAs (miR-10a-5p and miR-92b-3p) were upregulated in CRC patients with metastasis compared to both healthy subjects and to patients with localized CRC, indicating that these miRNAs could be a marker of metastatic disease. Four miRNAs (miR-122-5p, miR-885-3p, miR-375-3p, and miR-192-5p) were downregulated in CRC patients with localized disease compared to healthy subjects but upregulated in patients with CRC metastasis compared to CRC patients with localized disease. These miRNAs could participate in the differentiation between the early and late stages of CRC. The last two miRNAs (miR-429 and miR-21-5p) were only found differentially expressed in the comparison between metastatic and localized disease, where they were upregulated and thus associated with metastatic CRC.

Table 4.5: A selection of differentially expressed miRNAs between CRC patient groups. The first seven miRNAs were differentially expressed between true positive CRC patients with metastatic disease and false positive CRC patients (healthy). The next four miRNAs were found in comparisons of true positive CRC patients with localized disease and healthy subjects, and true

positive CRC patients with metastatic and localized disease. The last two miRNAs were only found differentially expressed in the comparison between metastatic and localized disease.

miRNA	**Local vs. healthy**	**Metastatic vs. local**	**Metastatic vs. healthy**
miR-142-5p	Downregulated	-	Downregulated
miR-16-5p	Downregulated	-	Downregulated
miR-143-3p	Downregulated	-	Downregulated
miR-10a-5p	-	Upregulated	Upregulated
miR-126-5p	Downregulated	-	Downregulated
miR-16-2-3p	Downregulated	-	Downregulated
miR-92b-3p	-	Upregulated	Upregulated
miR-122-5p	Downregulated	Upregulated	-
miR-885-3p	Downregulated	Upregulated	-
miR-375-3p	Downregulated	Upregulated	-
miR-192-5p	Downregulated	Upregulated	-
miR-429	-	Upregulated	-
miR-21-5p	-	Upregulated	-

5 Discussion on methodology and results

5.1 Investigating tissue of colorectal cancer patients at the single cell-level

An overall aim of this study was to investigate tissue of CRC patients at the single cell-level, as this can provide a better understanding of expressed genes and cell type composition in the disease. ScRNA-seq is a rapidly increasing analysis tool used to gain knowledge of heterogeneous complex samples such as cancer tissue, but the method requires highly viable single cell suspensions to provide good-quality data, as well as the knowledge of novel bioinformatic methods to analyze the data output. In order to investigate CRC tissue at the single cell-level, there was a need to both set up a good protocol for establishing single cell suspension from fresh CRC tissue and implement a functional computing method and annotation approach for processing and analyzing scRNA-seq data, before expressed genes and cell type composition could be identified.

5.1.1 Development and optimization of a protocol for establishing single cell suspension from fresh colorectal cancer tissue

5.1.1.1 Protocol evaluation by microscopy, cell counting, and flow cytometry

A protocol for establishing single cell suspension from fresh colorectal cancer tissue was adapted and developed, where the single cell suspensions were used in the downstream scRNA-seq. A high-quality single cell suspension for scRNA-seq is characterized by cell viability of at least 70%, as stressed or dead cells can lyse and release ambient RNA [105, 106]. Ambient RNA can contribute to cross-contamination and increased background noise, something that will compromise the quality of single cell data [105]. The mean cell viability of the single cell suspensions was in this study estimated to be 68.8% using flow cytometry (Supplementary Table 1) and confirmed to be over 60% by GCF at NTNU, and this result was assessed to be sufficient.

A low level of cell debris and clumping also characterizes a high-quality single cell suspension [106], and the cell appearance of the single cell solutions were evaluated by microscopy (Supplementary Figure 1). Most cells looked intact and single, with some cases of cell clumping (often referred to as duplicates) and debris. As the cases of cell clumping and debris did not surpass what to be expected when establishing a single cell suspension, the appearance of cells were assessed as good. The cell number of the single cell solutions were evaluated by cell counting, where an average result of 3,000,000 cells/mL met the desired criteria of a cell stock concentration of about 700,000-1,200,000 cells/mL to achieve the targeted number of recovered cells [105, 106]. Based on the total protocol evaluation, it was concluded that the developed and optimized protocol was satisfactory for establishing a single cell suspension from CRC tissue to be used further in downstream scRNA-seq analysis.

5.1.1.2 Using a protocol for pancreatic tissue as basis for colorectal tissue

Protocols on establishing viable single cell suspension from solid tissue can vary significantly between studies and can also be unique for different tissue types. In this thesis, a protocol for converting fresh pancreatic tissue into single cell suspension by Bernard, V. et al. [107] (Appendix C) was used as basis. Minimal changes were made to the original protocol, and protocol evaluation by cell counting, microscopy, and flow

cytometry, showed that it worked well with colorectal tissue too. The protocol is believed to have worked on another tissue type than described in the publication as it is a general protocol in terms of following a best practice workflow for preparing a single cell suspension from solid tissue, as presented by Reichard, A. and Asosingh, K. [108] (Appendix D), and not containing any tissue-specific steps or special enzymes.

5.1.1.3 Evaluation of flow cytometry results on cell viability

As presented in the mean flow cytometry results (Supplementary Table 1), the different fluorochromes used in this study gave quite divergent results, ranging from 42.0% live cells (calcein green) to 74.3% live cells (live/dead far red) and 90.1% live cells (PI). The varying numbers are most likely the result of differences in staining method. Calcein green marks live cells, as the nonfluorescent dye is converted to a green-fluorescent calcein during intracellular processes in live cells [109]. Live/dead far red discriminates between live and dead cells, as the dye reacts with free amines on the cell surface of live cells and free amines on both the cell surface and interior of dead cells with compromised membranes [110]. PI marks dead cells, as the dye can penetrate the cell membrane of dead or dying cells and insert itself between the bases of the cell's DNA [111].

Looking at the individual results from which the mean is calculated (Appendix E), there was a small variation in the live/dead far red and PI results with a maximum variation of 6.7% and 5.2%, respectively. For calcein green, the maximum variation between the individual results was 37.9% and indicates that this fluorochrome gives the most inaccurate results of the three. It could therefore be a possibility to not include the calcein green results when calculating a total percentage of live cells in the samples, changing the total mean value of live cells from 68.8% to 82.2%. Nevertheless, as none of the mean flow cytometry results for the three fluorochromes were consistent with each other, all the fluorochromes are presented and used in the result.

5.1.2 Implementation of a functional computing method and annotation approach for processing and analyzing single cell RNA sequencing data

5.1.2.1 Choice of single cell RNA sequencing processing tool

It has been developed an overwhelming number of methods to use in the computing processing and analysis of scRNA-seq data, making it difficult to select the ideal method to implement in a specific study. The single cell genomics R toolkit Seurat was used in this thesis because of its functionality with common clustering pipelines and popularity within the single-cell field.

5.1.2.2 Adjustment of parameters to create cluster graphs

As other methods for analyzing scRNA-seq data, Seurat comes with default parameter settings [112]. These parameters can be altered to fit a specific study's use and need, but it is important to be aware that changes to some of the parameters can have a significant effect on cell clustering [112]. In this study, the parameters of most commands (`NormalizeData`, `FindVariableFeatures`, `ScaleData`, `RunPCA`, `FindNeighbors`, and `RunUMAP`) were kept at their default setting (Appendix F).

The resolution parameter in the command `FindClusters` were customized when creating cluster graphs for each sample (554, 556, 559, and 569) and for the rough cell type groups of the integrated subsets (stromal cells, endothelial cells, and intestinal epithelial

cells). The resolution parameter sets the detail level of the downstream clustering, where increased resolution values results in a greater number of clusters [113]. The optimal resolution is believed to be 0.4-1.2 for scRNA-seq datasets around 3,000 cells and increases for lager datasets [113].

To create the cluster graphs for each sample (554, 556, 559, and 569), the resolution parameter was eventually set to 0.5. This resulted in sample 554, 556, and 569 being separated into a total of 13 clusters and sample 559 being separated into 11 clusters (Supplementary Figure 2). Even though the sample cell numbers variated from 2518 to 9410 (Table 4.2), the resolution parameter was not adjusted up for the larger datasets. This was because it was only necessary to get an overview of the major cell types in the individual samples, as they later were to be integrated and further subclustered to reveal distinct cell type subsets.

Cluster graphs for each of the rough cell type groups (stromal cells, endothelial cells, and intestinal epithelial cells) were created by setting the resolution parameter to 0.2, 0.5, and 0.2, leading to the cell type groups being separated into a total of 6, 5, and 8 clusters, respectively (Figure 4.3). The number of cells were 1263 in stromal subsets, 430 in endothelial subsets, and 3645 in intestinal epithelial subsets (Table 4.4). Based on the cell numbers, the resolution parameter should theoretically be put lower for endothelial subsets and higher for intestinal epithelial subsets. When decreasing the endothelial resolution parameter to 0.2, the only main difference occurring was the merging of one cluster into another. As the "disappearing" cluster seemed to stand out from the others, the resolution was put at a higher number where it could be further investigated. An increase of the intestinal epithelial resolution parameter to 0.4 resulted in an additional 3 clusters, but as the new cluster locations were in the area where it was believed to only exist mitochondrial gene-expressing cells, the resolution was put at a lower number to remove these extra clusters.

5.1.2.3 Choice of cell type annotation approach

Cell type annotation can be conducted automatically or manually [114]. Automated cell type annotation is an efficient way to label clusters, where the general principle is based on using a computer algorithm to match the gene expression of a cluster to the gene expression of a known cell type, and thus assign the cluster that respective label [114]. In manual cell type annotation, this matching is done manually with a wider set of resources, and the approach can therefore be more slow, labor-intensive, and subjective [114]. As not all cell types are known, have well-characterized gene expression and/or are found in specific resources, inaccurate or incomplete cluster labeling can occur using an automated approach [114], which was the main reason for choosing a manual cell type annotation approach in this thesis.

Cell type annotation can also be reference-based or marker-based [114]. In a reference-based approach, a gene expression pattern formed by all cell-type-specific marker genes in a cluster is compared to the gene expression pattern of known cell types [114]. The specific resources used for such an approach are existing manually annotated scRNA-seq reference data [114]. A marker-based annotation approach involves matching a cell-type-specific marker gene in a cluster to an identical marker gene in a known cell type, where the resource used for this approach usually is a cell-type-specific marker gene database that can be supplemented with findings in literature [114].

When manually annotating the cluster graphs of each sample (554, 556, 559, and 569), a reference-based approach was mainly used because an overall gene expression pattern

similarity was sufficient to identify major cell types, which was the main goal for that part of the study. For manual annotation of the clusters of the rough cell type groups (stromal cells, endothelial cells, and intestinal epithelial cells), where the idea was to reveal distinct cell type subsets and/or rare cell types, a marker-based approach was used because of access to a larger number of cell-type-specific marker genes in different literature.

5.1.2.4 Choice of annotation resources

The resources used for cell type annotation in this thesis was the Human Protein Atlas (HPA) [87], Azimuth [115], and PanglaoDB [79], in addition to supplementary literature. Both HPA, Azimuth, and PanglaoDB were all thought to be solid web-based resources, as HPA contains scRNA-seq reference data from 25 human tissues (including intestinal tissue) and peripheral blood mononuclear cells (PBMCs) [87], while Azimuth includes reference data sets from different human tissues such as pancreas, fetal development, lung, and kidney, in addition to PBMC [115]. PanglaoDB contains 1368 scRNA-seq datasets which can be filtered to include tumor/cancer samples and cell lines from only human species, in addition to providing a cell type marker database for 178 cell types in 29 tissues [79].

5.1.2.5 General considerations taken during cell type annotation

Ideally, each cluster should uniquely express canonical cell-type-specific marker genes of one cell type [114]. However, a cluster often express marker genes of more than one cell type [114]. In such cases, statistical values were used to see which cell-type-specific marker genes were most uniquely expressed in the cluster. These marker genes were "weighted" when assigning a cell type label. Statistical values were calculated for all top 10 cell-specific marker genes for each cluster and included the percentage of cells where the gene is detected in the cluster (pct.1), the percentage of cells where the gene is detected on average in the other clusters (pct.2), and the average log2 fold change (avg_log2FC) [116]. It is recommended to "weight" markers with a high pct.1 value and large differences in pct.1 and pct.2, and larger fold changes [116]. In cases where the cluster contained cells expressing the same amount of marker genes for multiple cell types, it was annotated as "unknown".

Some clusters showed expression of mitochondrial genes. Mitochondrial genes are expressed in most cells and are cell-type specific, and a high expression of mitochondrial genes among other cell-type-specific genes within a cluster can indicate poor sample quality [117]. This is because lysed cells with intact mitochondria can be registered during scRNA-seq analysis, and thus increase the fraction of mitochondrial transcripts detected within a cluster [117]. In this study, mitochondrial genes were in a few cases found among the top 10 cell-type-specific marker genes of a cluster, but it did not significantly affect the annotation of cluster cell types. In addition, each sample (554, 556, 559, and 569) contained a single or couple of clusters showing upregulation of only mitochondrial genes and no other cell-type-specific marker genes. These clusters were annotated as mitochondrial gene-expressing cells and were not thought to be an indication of poor sample quality, but rather a representation of a cell population of dead or dying cells.

5.1.3 Identified expressed genes and cell type composition in colorectal cancer tumor tissue

5.1.3.1 Identification of 18 major cell types in colorectal cancer tumor tissue

ScRNA-seq were performed on 4 CRC samples followed by a downstream clustering workflow, leading to the detection of a variety of cell types in each of the samples (Figure 4.1). After integrating the manually annotated sample datasets, 18 major cell types in CRC tissue were identified (Table 4.2).

Intestinal epithelial cells (unspecified subgroup), plasma B-cells, mitochondrial gene-expressing cells, and at least one type of effector memory T-cell were the cell types found in all four samples. These cells reflect the CRC tissue well, with intestinal epithelial cells lining the colorectal tissue, tumor-infiltrating lymphocytes exerting an immune response against the tumor cells, and mitochondrial gene-expressing cells representing a population of dead or dying cells. Additional cell types found in the samples also corresponded to known colorectal tissue and microenvironment structure. The differences in identified cell types and cell numbers emphasizes the ITH of CRC, where the solid tumors of different patients can consist of many different cell types.

A rough classification of the major cell types emphasized the similar gene expression within stromal cells, endothelial cells, intestinal epithelial cells, and three different subtypes of immune cells (Figure 4.2-C). Unknown cells indicated to show similarity in gene expression with both the intestinal epithelial cells and different subtypes of immune cells, clustering together in the middle of these clusters. The expression of cell-type-specific genes for many different cell types was why the cells were annotated as "unknown" in the first place, and this cell population could represent cells in the middle of a dynamic process such as stem cell differentiation. For future work, a trajectory inference analysis could be performed using Seurat to allocate the cells to lineages and then order them based on pseudotimes within these lineages [118].

5.1.3.2 Cell group subtypes identified in this study

A selection of cell type groups (stromal cells, endothelial cells, and intestinal epithelial cells) was further subclustered to identify cell subtypes in CRC tissue (Figure 4.3). Of the top 10 expressed cell-type-specific marker genes for the different subclusters, some specific marker genes ultimately determined the assignment of cluster cell type names (Table 4.3). The gene expression profiles of the identified subtypes were visualized in a heatmap (Figure 4.4).

Stromal cells: Cancer-associated fibroblasts, crypt-top fibroblasts, myofibroblasts, lamina propria fibroblasts, and plasma B-cells

The stromal subtypes identified in this study were pericytes, CAFs, CTFs, myofibroblasts, and LPFs, in addition to plasma B-cells. The finding of immune plasma B-cells among stromal subsets was unexpected, but is not considered accurate, as the cell type did not exhibit a defined gene expression profile as the other stromal cells and had several mitochondrial genes included in the clusters top 10 cell-type-specific markers (Supplementary Table 11).

The other cell types identified among the stromal cells were different subtypes of fibroblasts, apart from pericytes. Pericytes have been shown to differentiate into stromal myofibroblasts under pathological conditions [119], something which can explain the shared expression of marker genes between pericytes and myofibroblasts. Some cell-

type-specific markers of pericytes were RGS5 and MCAM. RGS5 is a signature molecule of tumor-associated pericytes [81], whereas analyses of human CRC tissues have shown increased MCAM expression in pericytes during tumorigenesis [82].

The other stromal subgroups seem to exhibit a defined gene expression profile unique for the individual cell types, with some shared genes. LPFs are distributed throughout the lamina propria and are involved in the structural organization of the extracellular matrix [84]. Myofibroblasts are also distributed throughout the lamina propria, but are specialized LPFs with contractile activity [84]. CTFs are located at the top of the crypt in close proximity to epithelial cells, and incudes differentiation in the nearby epithelial cells by secretion of Bmp ligands [86], such as the cell-type-specific BMP5 marker. CAFs are a fibroblast subtype exclusive in CRC tissue, secreting a variety of active factors to regulate tumor development and metastasis [120].

Endothelial cells: Mitochondrial gene-expressing cells, activated tumor-associated endothelial cells, tip tumor-associated endothelial cells, immature tumor-associated endothelial cells, and proliferative endothelial cells

The endothelial subtypes identified in this study were mitochondrial gene-expressing cells, activated TECs, tip TECs, immature TECs, and proliferative endothelial cells. The mitochondrial gene-expressing cells are thought to be a specific cell population of dead or dying cells, most likely endothelial cells that go through apoptosis as part of vessel remodeling during angiogenesis [121].

Tumor angiogenesis typically involves the formation of new blood vessels from pre-existing vessels in a process called vessel sprouting [122]. During vessel sprouting, tip endothelial cells navigate the sprout at the forefront, while proliferating stalk cells elongate the sprout [94]. In this study, the proliferative endothelial cells exhibited the most defined gene expression profiles among the endothelial subgroups, expressing the same cell-type-specific markers as found in proliferative endothelial cells of healthy murine liver and spleen tissues (HMGB2, STMN1, TUBA1B) [94].

Tip TECs and immature TECs both shared some gene expression patterns, but the tip cells were ultimately annotated based on the known endothelial tip marker gene CD34 [90] and cell-type-specific marker genes also found in endothelial tip cells in lung tumor (SPARC and ANGPT2) [89].

Activated TECs were also annotated based on findings in lung tumor, where CPE, CLU, and CCL14 were genes expressed in activated post-capillary vein TECs in lung tumor tissue. The activated TECs also expressed several genes involved in antigen presentation (HLA-DRB1, HLA-DRA, and HLA-DPA1), which supports their known role as non-professional antigen-presenting cells [123].

Intestinal epithelial cells: Mitochondrial gene-expressing cells, secretory progenitor 1, crypt-base-cells/Paneth cells, secretory progenitor 2, enterocytes, plasma B-cells, iron-storing epithelial cells, and tuft-2 cells

The intestinal epithelial subgroups identified in this study were mitochondrial gene-expressing cells, secretory progenitor 1, CBCs/Paneth cells, secretory progenitor 2, enterocytes, plasma B-cells, iron-storing epithelial cells, and tuft-2 cells. As enterocytes are the most abundant epithelial cell type in the large intestine [50], it was expected to be identified in large quantity in the CRC samples, although the cell type only amounted 379 cells of a total of 3645 intestinal epithelial cells (Table 4.4).

The cluster of intestinal mitochondrial gene-expressing cells mostly showed expression for mitochondrial genes, but some of the cluster's top 10 cell-type-specific marker genes also included some marker genes for intestinal epithelial cell types (Supplementary Table 13). The cluster was therefore assessed as a population of dead or dying intestinal epithelial cells.

Based on the low difference in pct.1 and pct.2 values of the expressed genes (Supplementary Table 13), most of the intestinal epithelial subgroups express many of the same genes. Some clear exceptions were plasma B-cells and tuft-2 cells. The finding of immune plasma B-cells among intestinal epithelial subsets was unexpected, but increasing evidence suggests that immunoglobulin can be produced by cancer cells such as intestinal epithelial cells [124]. This might imply that the cluster annotated as plasma B-cells could be immunoglobulin-producing epithelial cancer cells.

A relatively low number of tuft-2 cells were identified, where the cells had a gene expression pattern quite unique for that cluster, enriched for immune-related genes [104] and a couple CRC-related genes (CRIP1 and RASSF6). CRIP1 is shown overexpressed in CRC tissues and suppress apoptosis [125], while RASSF6 have been demonstrated to act as a tumor suppressor in CRC cells [126]. Only a few studies have examined tuft cells in humans and their relation to gut disease so far [49], and a more specific function of tuft-2 cells in CRC were therefore not found.

One cluster was annotated as CBCs/Paneth cells based on the expression of LYZ and OLFM4. LYZ is a well-established marker of Paneth cells, encoding the enzyme lysozyme found in the granules of the cell [99], while OLFM4 is a gene shown to be highly expressed in crypt base cells in the human small intestine and colon [100]. The cluster cell type could also have been annotated as Paneth-like cells or deep crypt secretory cells, which are other names for the colon equivalent of Paneth cells [40].

Two different types of secretory progenitor cells were identified and were in this study annotated as secretory progenitor 1 and secretory progenitor 2. Secretory progenitor 1 did not express a particularly unique cell-type-specific marker gene pattern and shared several genes with CBC/Paneth cells in addition to expressing many genes related to CRC stem cells. Annotation of the cluster as a progenitor cell type was ultimately done based on expression of FCGBP, a gene activated in colonic stem cell's transition to the progenitor stage [98]. Secretory progenitor 2 showed a more distinct gene expression profile than secretory progenitor 1 at the same time as expressing many genes related to CRC stem cells. This cluster was ultimately annotated as a progenitor cell type based on PLA2G2A-expression, a marker gene for transit-amplifying cells [101].

Both secretory progenitors 1 and 2 are most likely transit-amplifying cells originating from crypt-based CRC stem cells, which further can differentiate into a tumor intestinal epithelial cell. The differences in gene expression profiles can reflect the progenitor cells being at different stages in the CRC stem cell's transition to the progenitor stage or further transition into differentiated cells. As secretory progenitor 1 expressed more mitochondrial genes than secretory progenitor 2, it could also mean that the cells in that cluster are in somewhat of a more "worse shape".

The last subgroup identified were iron-storing epithelial cells, annotated after their expression of FTH1 and FTL. These two genes together make up the main intracellular iron storage protein ferritin [103], and it has been shown that increased ferritin expression limits ferroptosis, which is a novel form of regulated cell death [127]. Dietary

iron can be absorbed as ferritin, while excess cellular iron can be stored in in intestinal epithelial enterocytes [128, 129]. The iron-storing intestinal epithelial cells were not further annotated as enterocytes, as the rest of the top 10 cell-type-specific expressed marker genes also implied that the cell type could be annotated as enteroendocrine cells (Supplementary Table 16).

5.1.4 Future remarks

As mentioned, manual cell type annotation can be slow and labor-intensive. The manual annotation of roughly 700 genes in this thesis were time-consuming, unfortunately not leaving any time to investigate the scRNA-seq results further. For future work it would be interesting to perform a trajectory inference analysis on the unknown cell types to confirm if this cell population represents cells in the middle of a dynamic process such as stem cell differentiation. In addition, the prognostic values for some of the top 10 cell-type-specific marker genes could be analyzed to detect key genes contributing to CRC progression.

The influence of patient cohort characteristics (Table 4.1) could also be a basis for future research. One sample (sample 554) were shown to have peritoneal metastasis, and differences between this sample and the others could have been investigated. In addition, 50% of the samples were collected from the left-sided distal colon (sample 556 and sample 569), while the other 50% were retrieved from the right-sided proximal colon (sample 554 and sample 559). ITH differences between distal and proximal tumor location could be examined.

5.2 Investigating blood of colorectal cancer patients at the microRNA-level

Another overall aim of this study was to investigate circulating miRNA in serum of CRC patients, where the focus was on finding differentially expressed miRNAs between false positive CRC patients (healthy), true positive CRC patients with localized disease, and true positive CRC patients with metastatic disease. A considerable amount of studies have identified miRNA as a good biomarker candidate for CRC diagnosis, prognosis, and prediction due to their altered expression profiles in cancer [74]. Identification of differentially expressed serum miRNA between cancer patient groups could therefore potentially reveal novel miRNA biomarkers for CRC screening.

5.2.1 Evaluation of fragment length for small RNA reads

Small RNA reads produced after small RNA-seq were trimmed to filter out other RNA species than miRNA during data processing. As seen in the fragment length distribution of the trimmed small RNA reads (Supplementary Figure 5), the final length of the RNA molecules was mainly found to be at around 22 nucleotides, confirming that miRNAs were present. Both shorter and longer fragments were also found. The shorter fragments are theorized to include degraded RNA, while the longer fragments were found to be other small RNAs such as snoRNAs, tRNAs, and other classes of RNAs (Supplementary Figure 4).

5.2.2 MicroRNA expression profiles in colorectal cancer patient groups

PCA of mature miRNA expression found that the localized CRC samples were different from the metastatic CRC samples, and the analysis also revealed a similarity between healthy samples and the two groups of CRC samples (Figure 4.5-A). It was not expected

that healthy subjects had expression profiles similar to both local and metastatic CRC. One would expect that the expression profiles of both cancer stages would be more like each other, and the profile of healthy subjects to be different, as miRNAs has been shown to be dysregulated in cancer and thus make subjects exert different miRNA profiles. A possible explanation for the pattern occurring in the PCA plot could be the choice of control subjects. Control subjects used were false CRC patients, meaning they were first believed to have CRC then did not. These individuals could potentially have had other underlying conditions or diseases with similar biomarker changes as for CRC. Since we expect that many of the miRNA-changes in blood are related to changes in the immune system, other conditions could lead to the same miRNAs being altered. The subjects could therefore show a degree of similarity to other cancer patients. However, we would still expect the control group to resemble more the local CRC group, and not the metastatic group.

5.2.3 Differentially expressed microRNA between the patient groups

Differentially expressed miRNA between patient groups were detected by conducting a differential expression analysis, ultimately distinguishing 53 significant miRNAs between true positive CRC patients with localized disease and false positive CRC patients, 7 significant miRNAs between true positive CRC patients with metastatic disease and false positive CRC patients, and 30 significant miRNAs between true positive CRC patients with metastatic disease and localized disease (Supplementary Table 17-Supplementary Table 19). Of these, no miRNAs were differentially expressed between all three groups.

The lowest amount of differentially expressed miRNAs was found between true positive CRC patients with metastatic disease and healthy subjects, where five miRNAs (miR-142-5p, miR-16-5p, miR-143-3p, miR-126-5p, and miR-16-2-3p) were downregulated and two miRNAs were upregulated (miR-10a-5p and miR-92b-3p) (Table 4.5). The five downregulated miRNAs were also found downregulated in the comparison between true positive CRC patients with localized disease and healthy patients, suggesting the miRNAs could indicate CRC. The two upregulated miRNAs were also found upregulated in the comparison between true positive CRC patients with metastatic and localized disease, which possibly could make them markers of metastatic CRC. It should be noted that when not removing lowly expressed miRNAs, metastatic CRC had the highest number of differentially expressed miRNAs compared to healthy individuals (data not shown), indicating that deeper sequencing could have revealed more robust differentially expressed miRNAs for the metastatic group.

A selection of the 18 differentially expressed miRNAs identified in the comparisons of true positive CRC patients with localized disease and healthy patients, and true positive CRC patients with metastatic and localized disease (miR-122-5p, miR-885-3p, miR-375-3p, and miR-192-5p) were further investigated (Table 4.5). The selected miRNAs were all downregulated in the localized versus healthy comparison and upregulated in the metastatic versus localized comparison, potentially being able to differentiate patients with different stage of CRC.

Of the 10 miRNAs only found differentially expressed in the comparison between metastatic and localized disease, two selected miRNAs (miR-429 and miR-21-5p) were upregulated (Table 4.5). The miRNAs have previously been described as highly associated with tumor size, distant metastasis, and poor prognosis in CRC [130-132], which coincides with the findings in this study of significant upregulation of the miRNAs in metastatic CRC compared to localized disease.

6 Conclusion

In this study, a protocol for establishing single cell suspension from fresh CRC tumor tissue was developed and optimized. The established single cell suspensions were shown to have cell number, cell appearance, and cell viability compatible with downstream scRNA-seq workflow.

ScRNA-seq of primary CRC tumor tissue revealed clusters of cells with unique expression of cell-type-specific marker genes, in which a total of 18 major cell types were identified: B-cells, CD4+ effector memory T-cells, CD4+ proliferating T-cells, CD8+ effector memory T-cells, dendritic cells/B-cells, fibroblasts, intestinal enterocytes, intestinal epithelial cells (unspecified subgroup), intestinal goblet cells, monocytes, mitochondrial gene-expressing cells, myeloid cells (unspecified subgroup), plasma B-cells, smooth muscle cells, T-cells (unspecified subgroup), unknown cell type, vascular endothelial cells, and vascular smooth muscle cells.

A rough classification of the major cell types demonstrated a clear separation of immune cells, stromal cells, endothelial cells, and intestinal epithelial cells. An unknown cell type was shown to have similar gene expression with both immune- and intestinal epithelial cells. Further investigations could be performed to identify the unknown cell type and find out if they represent a dynamic cell population such as differentiating stem cells.

Subtypes of stromal cells (pericytes, CAFs, plasma B-cells, CTFs, myofibroblasts, and LPFs), endothelial cells (mitochondrial gene-expressing cells, activated TECs, tip TECs, immature TECs, and proliferative ECs), and intestinal epithelial cells (mitochondrial gene-expressing cells, secretory progenitor 1, CBCs/Paneth cells, secretory progenitor 2, enterocytes, plasma B-cells, iron-storing epithelial cells, and tuft-2 cells) were also identified in the CRC tumor tissue. The findings of stromal plasma B-cells were suggested not to be accurate; the endothelial mitochondrial gene-expressing cells were proposed to be dead or dying endothelial cells; intestinal epithelial plasma B-cells were implied to be immunoglobulin-producing epithelial cancer cells; and both secretory progenitors 1 and 2 were suggested to be differentiating transit-amplifying cells originating from crypt-based CRC stem cells. In total, the identified expressed genes and cell type composition of CRC tumor tissue emphasizes the ITH of CRC.

Lastly, significantly differentially expressed circulating miRNAs between CRC patient groups were identified. A total of 53 significant miRNAs between true positive CRC patients with localized disease and false positive CRC patients, 7 significant miRNAs between true positive CRC patients with metastatic disease and false positive CRC patients, and 30 significant miRNAs between true positive CRC patients with metastatic disease and localized disease were found.

It was demonstrated that the localized CRC samples were distinct from the metastatic CRC samples in terms of miRNA expression. Five miRNAs (miR-142-5p, miR-16-5p, miR-143-3p, miR-126-5p, and miR-16-2-3p) were downregulated in both stages of CRC, suggesting these miRNAs could be an indicator of the disease. Two miRNAs (miR-10a-5p and miR-92b-3p) were upregulated in CRC patients with metastasis, indicating that these miRNAs could be a marker of metastatic disease. Four miRNAs (miR-122-5p, miR-885-3p, miR-375-3p, and miR-192-5p) were downregulated in localized CRC but upregulated in patients with CRC metastasis compared to patients with localized disease, suggesting these miRNAs could participate in the differentiation between the early and late stages of

CRC. Two miRNAs (miR-429 and miR-21-5p) were only found differentially expressed in the comparison between metastatic and localized disease, where they were upregulated and thus associated with metastatic CRC.

References

1. American Cancer Society, Colorectal Cancer Facts & Figures 2020-2022, American Cancer Society, Atlanta, 2020.
2. Sung, H. et al. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71 (3), 209-249.
3. Dekker, E. et al. (2019) Colorectal cancer. The Lancet 394 (10207), 1467-1480.
4. Migliore, L. et al. (2011) Genetics, cytogenetics, and epigenetics of colorectal cancer. J Biomed Biotechnol 2011, 792362.
5. National Cancer Institute (NCI) Colon Cancer Treatment. https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq, (accessed 29.11.2021).
6. World Cancer Research Fund (WCRF) Colorectal cancer. https://www.wcrf.org/dietandcancer/colorectal-cancer/, (accessed 29.11.2021).
7. Schreuders, E.H. et al. (2015) Colorectal cancer screening: a global overview of existing programmes. Gut 64 (10), 1637-1649.
8. Fearon, E.R. and Vogelstein, B. (1990) A genetic model for colorectal tumorigenesis. Cell 61 (5), 759-67.
9. Nguyen, H.T. and Duong, H.Q. (2018) The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy (Review). Oncol Lett 16 (1), 9-18.
10. Worthley, D.L. and Leggett, B.A. (2010) Colorectal cancer: molecular features and clinical opportunities. Clin Biochem Rev 31 (2), 31-8.
11. Yamagishi, H. et al. (2016) Molecular pathogenesis of sporadic colorectal cancers. Chin J Cancer 35, 4-4.
12. Hiremath, I.S. et al. (2021) The multidimensional role of the Wnt/β-catenin signaling pathway in human malignancies. J Cell Physiol n/a (n/a).
13. Plotnikov, A. et al. (2011) The MAPK cascades: Signaling components, nuclear roles and mechanisms of nuclear translocation. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1813 (9), 1619-1633.
14. Sabbadini, F. et al. (2021) The Multifaceted Role of TGF-β in Gastrointestinal Tumors. Cancers (Basel) 13 (16).
15. Mehlen, P. and Fearon, E.R. (2004) Role of the Dependence Receptor DCC in Colorectal Cancer Pathogenesis. J Clin Oncol 22 (16), 3420-3428.
16. Liebl, M.C. and Hofmann, T.G. (2021) The Role of p53 Signaling in Colorectal Cancer. Cancers (Basel) 13 (9).
17. Kim, M.S. et al. (2010) DNA methylation markers in colorectal cancer. Cancer Metastasis Rev 29 (1), 181-206.
18. Moore, L.D. et al. (2013) DNA Methylation and Its Basic Function. Neuropsychopharmacology 38 (1), 23-38.
19. Hervieu, C. et al. (2021) The Role of Cancer Stem Cells in Colorectal Cancer: From the Basics to Novel Clinical Trials. Cancers (Basel) 13 (5), 1092.
20. Jasperson, K.W. et al. (2010) Hereditary and familial colon cancer. Gastroenterology 138 (6), 2044-2058.
21. Lavik, L.A.S. and Sjursen, W. (2009) Molekylærgenetiske analyser ved utredning av arvelig kolorektal cancer. Bioingeiøren.
22. Baran, B. et al. (2018) Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature. Gastroenterology research 11 (4), 264-273.
23. Mangone, L. et al. (2021) Colon cancer survival differs from right side to left side and lymph node harvest number matter. BMC Public Health 21 (1), 906.
24. Zheng, Z. et al. (2020) Intratumor heterogeneity: A new perspective on colorectal cancer research. Cancer Medicine 9 (20), 7637-7645.
25. Li, J. et al. (2022) Tumor Microenvironment Shapes Colorectal Cancer Progression, Metastasis, and Treatment Responses. Frontiers in Medicine 9.

26. Sveen, A. et al. (2016) Intra-patient Inter-metastatic Genetic Heterogeneity in Colorectal Cancer as a Key Determinant of Survival after Curative Liver Resection. PLoS Genet 12 (7), e1006225.
27. Buikhuisen, J.Y. et al. (2020) Exploring and modelling colon cancer inter-tumour heterogeneity: opportunities and challenges. Oncogenesis 9 (7), 66.
28. Chowdhury, S. et al. (2021) Implications of Intratumor Heterogeneity on Consensus Molecular Subtype (CMS) in Colorectal Cancer. Cancers (Basel) 13 (19), 4923.
29. Tieng, F.Y.F. et al. (2020) Single Cell Transcriptome in Colorectal Cancer—Current Updates on Its Application in Metastasis, Chemoresistance and the Roles of Circulating Tumor Cells. Front Pharmacol 11.
30. Chen, G. et al. (2019) Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Frontiers in Genetics 10 (317).
31. Wellcome Sanger Institute Analysis of single cell RNA-seq data. https://www.singlecellcourse.org/, (accessed 09.12.2021).
32. Li, H. et al. (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet 49 (5), 708-718.
33. Dai, W. et al. (2019) Single-cell transcriptional profiling reveals the heterogenicity in colorectal cancer. Medicine 98 (34).
34. Mei, Y. et al. (2021) Single-cell analyses reveal suppressive tumor microenvironment of human colorectal cancer. Clinical and translational medicine 11 (6), e422-e422.
35. Khaliq, A.M. et al. (2021) Redefining tumor classification and clinical stratification through a colorectal cancer single-cell atlas. bioRxiv, 2021.02.02.429256.
36. Wang, H. et al. (2021) Colorectal Cancer Stem Cell States Uncovered by Simultaneous Single-Cell Analysis of Transcriptome and Telomeres. Advanced science (Weinheim, Baden-Wurttemberg, Germany) 8 (8), 2004320-2004320.
37. Qi, J. et al. (2021) Single-cell transcriptomic landscape reveals tumor specific innate lymphoid cells associated with colorectal cancer progression. Cell Reports Medicine 2 (8), 100353.
38. Domanska, D. et al. (2022) Single-cell transcriptomic analysis of human colonic macrophages reveals niche-specific subsets. J Exp Med 219 (3).
39. Humphries, A. and Wright, N.A. (2008) Colonic crypt organization and tumorigenesis. Nature Reviews Cancer 8 (6), 415-424.
40. Zhu, G. et al. (2021) The cellular niche for intestinal stem cells: a team effort. Cell Regeneration 10 (1), 1.
41. Bankaitis, E.D. et al. (2018) Reserve Stem Cells in Intestinal Homeostasis and Injury. Gastroenterology 155 (5), 1348-1361.
42. Barker, N. et al. (2012) Identifying the Stem Cell of the Intestinal Crypt: Strategies and Pitfalls. Cell Stem Cell 11 (4), 452-460.
43. van der Heijden, M. and Vermeulen, L. (2019) Stem cells in homeostasis and cancer of the gut. Mol Cancer 18 (1), 66.
44. Shanahan, M.T. et al. (2021) Multiomic analysis defines the first microRNA atlas across all small intestinal epithelial lineages and reveals novel markers of almost all major cell types. American Journal of Physiology-Gastrointestinal and Liver Physiology 321 (6), G668-G681.
45. Lueschow, S.R. and McElroy, S.J. (2020) The Paneth Cell: The Curator and Defender of the Immature Small Intestine. Front Immunol 11.
46. Dao DPD and Le PH (2022) Histology, Goblet Cells. https://www.ncbi.nlm.nih.gov/books/NBK553208/?report=classic, (accessed 25.04.2022).
47. Kulkarni, D.H. and Newberry, R.D. (2019) Intestinal Macromolecular Transport Supporting Adaptive Immunity. Cell Mol Gastroenterol Hepatol 7 (4), 729-737.
48. Beumer, J. et al. (2020) Enteroendocrine Dynamics – New Tools Reveal Hormonal Plasticity in the Gut. Endocr Rev 41 (5).
49. Hendel, S.K. et al. (2022) Tuft Cells and Their Role in Intestinal Diseases. Front Immunol 13.
50. Snoeck, V. et al. (2005) The role of enterocytes in the intestinal barrier function and antigen uptake. Microbes and Infection 7 (7), 997-1004.

51. Hageman, J.H. et al. (2020) Intestinal Regeneration: Regulation by the Microenvironment. Dev Cell 54 (4), 435-446.
52. The Editors of Encyclopaedia Britannica (2018) fibroblast. https://www.britannica.com/science/fibroblast, (accessed 26.04.2022).
53. Ramirez, M. et al. (2019) Pericytes in the Gut. In Pericyte Biology in Different Organs (Birbrair, A. ed), pp. 73-100, Springer International Publishing.
54. Humphrey, J.H. and Perdue, S.S. (2020) immune system. https://www.britannica.com/science/immune-system, (accessed 27.04.2022).
55. The Editors of Encyclopaedia Britannica (2020) T cell. https://www.britannica.com/science/T-cell, (accessed 27.04.2022).
56. Britannica, T.E.o.E. (2020) macrophage. https://www.britannica.com/science/macrophage, (accessed 27.04.2022).
57. The Editors of Encyclopaedia Britannica (2018) mast cell. https://www.britannica.com/science/mast-cell, (accessed 27.04.2022).
58. Komi, D.E.A. and Redegeld, F.A. (2020) Role of Mast Cells in Shaping the Tumor Microenvironment. Clin Rev Allergy Immunol 58 (3), 313-325.
59. The Editors of Encyclopaedia Britannica (2018) neutrophil. https://www.britannica.com/science/neutrophil, (accessed 27.04.2022).
60. The Editors of Encyclopaedia Britannica (2022) smooth muscle. https://www.britannica.com/science/smooth-muscle, (accessed 27.04.2022).
61. Hafen, B.B. and Burns, B. (2021) Physiology, Smooth Muscle. https://www.ncbi.nlm.nih.gov/books/NBK526125/, (accessed 27.04.2022).
62. Martín-Alonso, M. et al. (2021) Smooth muscle-specific MMP17 (MT4-MMP) regulates the intestinal stem cell niche and regeneration after damage. Nature Communications 12 (1), 6741.
63. Crick, F. (1970) Central dogma of molecular biology. Nature 227 (5258), 561-3.
64. O'Connor, C. et al., Essentials of Cell Biology, NPG Education, 2010.
65. Kukurba, K.R. and Montgomery, S.B. (2015) RNA Sequencing and Analysis. Cold Spring Harbor protocols 2015 (11), 951-969.
66. Pertea, M. (2012) The human transcriptome: an unfinished story. Genes 3 (3), 344-360.
67. Bartel, D.P. (2018) Metazoan MicroRNAs. Cell 173 (1), 20-51.
68. Xu, P. et al. (2020) A Systematic Way to Infer the Regulation Relations of miRNAs on Target Genes and Critical miRNAs in Cancers. Frontiers in Genetics 11.
69. Lee, R.C. et al. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75 (5), 843-54.
70. O'Brien, J. et al. (2018) Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front Endocrinol (Lausanne) 9.
71. Friedman, R.C. et al. (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 19 (1), 92-105.
72. Gmerek, L. et al. (2019) MicroRNA regulation in colorectal cancer tissue and serum. PLoS One 14 (8), e0222013-e0222013.
73. Hammond, S.M. et al. (2001) Post-transcriptional gene silencing by double-stranded RNA. Nature Reviews Genetics 2 (2), 110-119.
74. Peng, Y. and Croce, C.M. (2016) The role of MicroRNAs in human cancer. Signal Transduction and Targeted Therapy 1 (1), 15004.
75. Zhang, N. et al. (2021) The role of miRNAs in colorectal cancer progression and chemoradiotherapy. Biomed Pharmacother 134, 111099.
76. Alves Martins, B.A. et al. (2019) Biomarkers in Colorectal Cancer: The Role of Translational Proteomics Research. Front Oncol 9.
77. Ahadi, A. (2020) The significance of microRNA deregulation in colorectal cancer development and the clinical uses as a diagnostic and prognostic biomarker and therapeutic agent. Non-coding RNA Research 5 (3), 125-134.
78. Hao, Y. et al. (2021) Integrated analysis of multimodal single-cell data. Cell 184 (13), 3573-3587.e29.
79. Franzén, O. et al. (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database 2019.

80. Elmentaite, R. et al. (2021) Cells of the human intestinal tract mapped across space and time. Nature 597 (7875), 250-255.
81. Dasgupta, S. et al. (2021) RGS5–TGFβ–Smad2/3 axis switches pro- to anti-apoptotic signaling in tumor-residing pericytes, assisting tumor growth. Cell Death Differ 28 (11), 3052-3076.
82. Kotsiliti, E. (2022) Origin of CAFs in colorectal cancer. Nature Reviews Gastroenterology & Hepatology 19 (2), 79-79.
83. Tefft, J.B. et al. (2022) Notch1 and Notch3 coordinate for pericyte-induced stabilization of vasculature. American Journal of Physiology-Cell Physiology 322 (2), C185-C196.
84. Bigaeva, E. et al. (2020) Understanding human gut diseases at single-cell resolution. Hum Mol Genet 29 (R1), R51-R58.
85. Uhlitz, F. et al. (2020) Single cell analysis of colorectal cancer identifies mitogen-activated protein kinase as a key driver of tumor cell plasticity. bioRxiv, 2020.01.10.901579.
86. Brügger, M.D. et al. (2020) Distinct populations of crypt-associated fibroblasts act as signaling hubs to control colon homeostasis. PLoS Biol 18 (12), e3001032.
87. Karlsson, M. et al. (2021) A single-cell type transcriptomics map of human tissues. Science Advances 7 (31).
88. Liu, S.C. et al. (2014) CTGF increases vascular endothelial growth factor-dependent angiogenesis in human synovial fibroblasts by increasing miR-210 expression. Cell Death Dis 5 (10), e1485-e1485.
89. Goveia, J. et al. (2020) An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. Cancer Cell 37 (1), 21-36.e13.
90. Siemerink, M.J. et al. (2012) CD34 marks angiogenic tip cells in human vascular endothelial cell cultures. Angiogenesis 15 (1), 151-163.
91. Zarkada, G. et al. (2021) Specialized endothelial tip cells guide neuroretina vascularization and blood-retina-barrier formation. Dev Cell 56 (15), 2237-2251.e6.
92. Uxa, S. et al. (2021) Ki-67 gene expression. Cell Death Differ 28 (12), 3357-3370.
93. Han, G. et al. (2018) NUSAP1 gene silencing inhibits cell proliferation, migration and invasion through inhibiting DNMT1 gene expression in human colorectal cancer. Exp Cell Res 367 (2), 216-221.
94. Kalucka, J. et al. (2020) Single-Cell Transcriptome Atlas of Murine Endothelial Cells. Cell 180 (4), 764-779.e20.
95. Fazilaty, H. et al. (2021) Tracing colonic embryonic transcriptional profiles and their reactivation upon intestinal damage. Cell Rep 36 (5), 109484.
96. Gracz, A.D. et al. (2018) Sox4 Promotes Atoh1-Independent Intestinal Secretory Differentiation Toward Tuft and Enteroendocrine Fates. Gastroenterology 155 (5), 1508-1523.e10.
97. Sancho, R. et al. (2015) Stem cell and progenitor fate in the mammalian intestine: Notch and lateral inhibition in homeostasis and disease. EMBO reports 16 (5), 571-581.
98. Habowski, A.N. et al. (2020) Transcriptomic and proteomic signatures of stemness and differentiation in the colon crypt. Communications Biology 3 (1), 453.
99. Nakanishi, Y. et al. (2016) Control of Paneth Cell Fate, Intestinal Inflammation, and Tumorigenesis by PKCλ/ι. Cell Rep 16 (12), 3297-3310.
100. van der Flier, L.G. et al. (2009) OLFM4 Is a Robust Marker for Stem Cells in Human Intestine and Marks a Subset of Colorectal Cancer Cells. Gastroenterology 137 (1), 15-17.
101. Rajagopal, J. et al., Modulation of epithelial cell differentiation, maintenance and/or function through T cell action, and markers and methods of use thereof, The Broad Institute, Inc., Massachusetts Institute of Technology, The General Hospital Corporation, United States, 2021.
102. Xu, M. et al. (2020) Ferroptosis involves in intestinal epithelial cell death in ulcerative colitis. Cell Death Dis 11 (2), 86.
103. Xu, S. et al. (2021) The emerging role of ferroptosis in intestinal disease. Cell Death Dis 12 (4), 289.

104. Xiong, Z. et al. (2022) Intestinal Tuft-2 cells exert antimicrobial immunity via sensing bacterial metabolite N-undecanoylglycine. Immunity.
105. 10x Genomics Sample Preparation Tips for Single Cell Gene Expression. https://pages.10xgenomics.com/rs/446-PBO-704/images/10x_LIT037_Sample_Prep_Tips_Single_Cell_digital.pdf, (accessed 04.05.2022).
106. 10x Genomics Questions & Answers - Single Cell Gene Expression. https://kb.10xgenomics.com/hc/en-us/categories/360000149952-Single-Cell-3-Gene-Expression, (accessed 04.05.2022).
107. Bernard, V. et al. (2019) Single-Cell Transcriptomics of Pancreatic Cancer Precursors Demonstrates Epithelial and Microenvironmental Heterogeneity as an Early Event in Neoplastic Progression. Clinical cancer research : an official journal of the American Association for Cancer Research 25 (7), 2194-2205.
108. Reichard, A. and Asosingh, K. (2019) Best Practices for Preparing a Single Cell Suspension from Solid Tissues for Flow Cytometry. Cytometry. Part A : the journal of the International Society for Analytical Cytology 95 (2), 219-226.
109. Thermo Fisher Scientific CellTrace™ Calcein Green, AM. https://www.thermofisher.com/order/catalog/product/C34852?SID=srch-srp-C34852, (accessed 07.12.2021).
110. Thermo Fisher Scientific LIVE/DEAD™ Fixable Far Red Dead Cell Stain Kit. https://www.thermofisher.com/order/catalog/product/L34974?SID=srch-srp-L34974, (accessed 07.12.2021).
111. Thermo Fisher Scientific Propidium Iodide. https://www.thermofisher.com/order/catalog/product/P1304MP?SID=srch-srp-P1304MP, (accessed 07.12.2021).
112. Schneider, I. et al. (2021) Use of "default" parameter settings when analyzing single cell RNA sequencing data using Seurat: a biologist's perspective. J Transl Genet Genom (Progresses on the Application of Single - Cell Sequencing in the Human Diseases Research).
113. Satija Lab (2022) Seurat - Guided Clustering Tutorial. https://satijalab.org/seurat/articles/pbmc3k_tutorial.html, (accessed 21.01.2022).
114. Clarke, Z.A. et al. (2021) Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat Protoc 16 (6), 2749-2764.
115. Azimuth - App for reference-based single-cell analysis. https://azimuth.hubmapconsortium.org/, (accessed 04.05.2022).
116. Harvard Chan Bioinformatics Core (HBC) Lessons in scRNA-seq. https://github.com/hbctraining/scRNA-seq/tree/master/lessons, (accessed 21.01.2022).
117. 10x Genomics Why do I see a high level of mitochondrial gene expression? https://kb.10xgenomics.com/hc/en-us/articles/360001086611-Why-do-I-see-a-high-level-of-mitochondrial-gene-expression, (accessed 21.02.2022).
118. Van den Berge, K. et al. (2020) Trajectory-based differential expression analysis for single-cell sequencing data. Nature Communications 11 (1), 1201.
119. Hosaka, K. et al. (2016) Pericyte-fibroblast transition promotes tumor growth and metastasis. Proceedings of the National Academy of Sciences 113 (38), E5618-E5627.
120. Ping, Q. et al. (2021) Cancer-associated fibroblasts: overview, progress, challenges, and directions. Cancer Gene Ther 28 (9), 984-999.
121. Watson, E.C. et al. (2017) Endothelial cell apoptosis in angiogenesis and vessel regression. Cell Mol Life Sci 74 (24), 4387-4403.
122. Sun, W. (2012) Angiogenesis in metastatic colorectal cancer and the benefits of targeted therapy. J Hematol Oncol 5 (1), 63.
123. Nagl, L. et al. (2020) Tumor Endothelial Cells (TECs) as Potential Immune Directors of the Tumor Microenvironment – New Findings and Future Perspectives. Frontiers in Cell and Developmental Biology 8.
124. Cui, M. et al. (2021) Immunoglobulin Expression in Cancer Cells and Its Critical Roles in Tumorigenesis. Front Immunol 12.

125. Zhang, L. et al. (2019) Cysteine-rich intestinal protein 1 suppresses apoptosis and chemosensitivity to 5-fluorouracil in colorectal cancer through ubiquitin-mediated Fas degradation. J Exp Clin Cancer Res 38 (1), 120.
126. Chen, E. et al. (2016) Decreased level of RASSF6 in sporadic colorectal cancer and its anti-tumor effects both in vitro and in vivo. Oncotarget 7 (15), 19813-19823.
127. Hou, W. et al. (2016) Autophagy promotes ferroptosis by degradation of ferritin. Autophagy 12 (8), 1425-1428.
128. Bonilla, D.A. et al. (2022) A Bioinformatics-Assisted Review on Iron Metabolism and Immune System to Identify Potential Biomarkers of Exercise Stress-Induced Immunosuppression. Biomedicines 10 (3), 724.
129. Pandrangi, S.L. et al. (2022) Role of dietary iron revisited: in metabolism, ferroptosis and pathophysiology of cancer. Am J Cancer Res 12 (3), 974-985.
130. Toiyama, Y. et al. (2013) Serum miR-21 as a Diagnostic and Prognostic Biomarker in Colorectal Cancer. JNCI: Journal of the National Cancer Institute 105 (12), 849-859.
131. Xing, X.-L. et al. (2020) MicroRNA-Related Prognosis Biomarkers from High-Throughput Sequencing Data of Colorectal Cancer. BioMed Research International 2020, 7905380.
132. Li, J. et al. (2013) MiR-429 is an independent prognostic factor in colorectal cancer and exerts its anti-apoptotic function by targeting SOX2. Cancer Lett 329 (1), 84-90.
133. The Norwegian National Research Ethics Committees (REC) (2014) Regionale komiteer for medisinsk og helsefaglig forskningsetikk (REK). https://www.forskningsetikk.no/om-oss/komiteer-og-utvalg/rek/, (accessed 01.12.2021).
134. Lov om medisinsk og helsefaglig forskning (helseforskningsloven), Lovdata, 2009.
135. The Norwegian National Research Ethics Committees (REC) (2015) Human biological material. https://www.forskningsetikk.no/en/resources/the-research-ethics-library/research-on-human-biological-material/human-biological-material/, (accessed 01.12.2012).
136. The Norwegian National Research Ethics Committees (REC) (2020) The Health Research Act. https://www.forskningsetikk.no/en/resources/the-research-ethics-library/legal-statutes-and-guidelines/the-health-research-act/, (accessed 01.12.2021).
137. Sayers, E.W. et al. (2022) Database resources of the national center for biotechnology information. Nucleic Acids Res 50 (D1), D20-D26.
138. Genomics Core Facility (GCF). https://www.ntnu.edu/mh/gcf, (accessed 03.05.2022).
139. Bioinformatics (BioCore). https://www.ntnu.edu/mh/biocore, (accessed 03.05.2022).
140. R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021.
141. RStudio Team, RStudio: Integrated Development Environment for R, RStudio, PBC, Boston, MA, 2021.
142. Mjelle, R. et al. (2019) Comprehensive transcriptomic analyses of tissue, serum, and serum exosomes from hepatocellular carcinoma patients. BMC Cancer 19 (1), 1007.
143. Mjelle, R. et al. (2017) Identification of metastasis-associated microRNAs in serum from rectal cancer patients. Oncotarget 8 (52), 90077-90089.
144. McKinnon, K.M. (2018) Flow Cytometry: An Overview. Curr Protoc Immunol 120, 5.1.1-5.1.11.
145. Adan, A. et al. (2017) Flow cytometry: basic principles and applications. Crit Rev Biotechnol 37 (2), 163-176.
146. Illumina Main Steps in Next-Generation Sequencing. https://www.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-workflow.html, (accessed 08.12.2021).
147. National Human Genome Reserach Institute (NHGRI) Transcriptome Fact Sheet. https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet, (accessed 08.12.2021).
148. See, P. et al. (2018) A Single-Cell Sequencing Guide for Immunologists. Front Immunol 9.

149. Salomon, R. et al. (2019) Droplet-based single cell RNAseq tools: A practical guide. Lab on a Chip 19.
150. 10X Genomics (2021) Inside Chromium Next GEM Technology. https://pages.10xgenomics.com/rs/446-PBO-704/images/10x_BR025_Chromium-Brochure_Letter_Digital.pdf, (accessed 18.01.2022).
151. Illumina A targeted method for both small RNA profiling and discovery applications. https://www.illumina.com/techniques/sequencing/rna-sequencing/small-rna-seq.html, (accessed 04.05.2022).
152. New England Biolabs Protocol for use with NEBNext Small RNA Library Prep Set for Illumina. https://international.neb.com/protocols/2018/03/27/protocol-for-use-with-nebnext-small-rna-library-prep-set-for-illumina-e7300-e7580-e7560-e7330, (accessed 04.05.2022).
153. Benesova, S. et al. (2021) Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis. Diagnostics (Basel, Switzerland) 11 (6), 964.
154. Schneider, I. et al. (2021) Use of "default" parameter settings when analyzing single cell RNA sequencing data using Seurat: a biologist's perspective. Journal of Translational Genetics and Genomics 5 (1), 37-49.
155. Nugent, R. and Meila, M. (2010) An Overview of Clustering Applied to Molecular Biology. In Statistical Methods in Molecular Biology (Bang, H. et al. eds), pp. 369-404, Humana Press.
156. Butler, A. et al. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36 (5), 411-420.
157. Programmathically (2022) Principal Components Analysis Explained for Dummies. https://programmathically.com/principal-components-analysis-explained-for-dummies/, (accessed 26.01.2022).
158. McInnes, L. and Healy, J. (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.
159. Coenen, A. and Pearce, A. Understanding UMAP. https://pair-code.github.io/understanding-umap/, (accessed).
160. Qiu, Y. et al. (2021) Identification of cell-type-specific marker genes from co-expression patterns in tissue samples. Bioinformatics 37 (19), 3228-3234.
161. Satija Lab (2022) Introduction to scRNA-seq integration. https://satijalab.org/seurat/articles/integration_introduction.html, (accessed 23.03.2022).
162. Potla, P. et al. (2021) A bioinformatics approach to microRNA-sequencing analysis. Osteoarthritis and Cartilage Open 3 (1), 100131.

Appendices

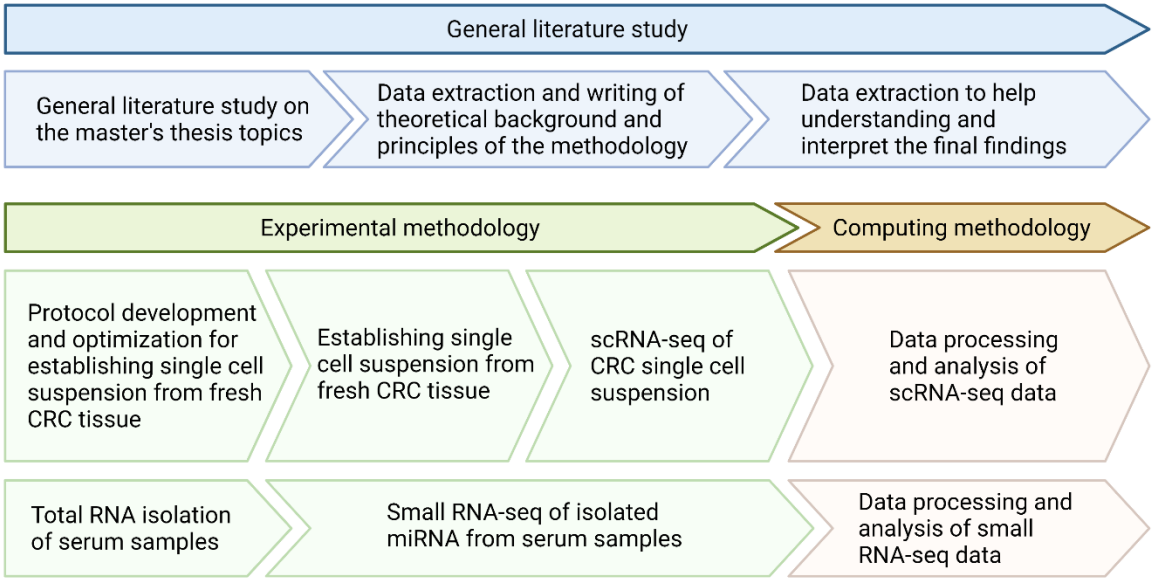
A Methodology

A.1 Master's thesis preparation

Before writing the master's thesis, preparations were done in terms of putting up a planned workflow and ensuring good research ethics were followed during the project.

A.1.1 The master's thesis workflow

A workflow of this master's thesis was put up to achieve the aims of the study in a good and precise manner, consisting of the main parts literature study, experimental methodology, and computing methodology (Appendix Figure 1).



Appendix Figure 1: An illustration of the master's thesis workflow. The workflow was put up to achieve the aims of the study in a good and precise manner. Created with Biorender.com.

A.1.2 Considerations of research ethics

All medical and health research in Norway involving human biological material needs an approval from the Regional Committees for Medical and Health Research Ethics (REC) before project start. REC is an organ ensuring research projects are in line with Norway's legislation, and assesses if the project is ethically justifiable to carry out [133]. The Health Research Act is a part of Norwegian medical and health research legislation, where the purpose of the Act is to promote good and ethically sound research by facilitating the protection of the individual's self-determination, integrity, and privacy [134-136]. This is in practice done by collecting a voluntary, informed consent from the patient before obtaining human biological material [135].

Prior to conducting the master's thesis' experimental methods, a REC-approval was obtained by the master's thesis' supervisor Robin Mjelle for the project "Pre-diagnostic biomarkers for colorectal cancer" (ref. no 2016/534) and a more recent project change related to this master's thesis (ref. no. 30022) (Appendix B). This included ensuring a consent-based sample collection from the CRC patients involved in this study, something that was done by Biobank1 during tissue sample collection.

A.2 General literature study

A general literature study was conducted both prior to and during writing of the master's thesis with the main purposes of (I) understanding the basics of relevant topics such as CRC, ITH, and miRNA, (II) gaining knowledge of experimental and computing methods used for providing relevant research results in terms of the master's thesis' aims, and (III) help understanding and interpret the final findings. Several databases were used to find information (e.g., MEDLINE with PubMed as search interface [137]), and the literature considered relevant in terms of purpose (I) and (II) were used as a basis for writing the included theoretical background and methodology principles.

A.3 Experimental and computing methodology

A.3.1 Applied experimental materials and computing software

Different technical equipment and instruments (Appendix Table 1), reagents (Appendix Table 2), and commercial kits (Appendix Table 3) were used in the experimental part of this study, and various software (Appendix Table 4) were used in the computing part.

Appendix Table 1: Technical equipment and instruments used in the experimental part of this study.

Product name	Manufacturer
Countess™ Automated Cell Counter	Invitrogen™
Moxi Z Mini Automated Cell Counter	ORFLO
AE30 Binocular Inverted Microscope	Motic®
FACSCanto™ Flow Cytometer	BD Biosciences
New Brunswick™ Innova® 44 Incubator Shaker	Eppendorf®
Centrifuge 5810 R	Eppendorf®
CryoTube™ Vials	Thermo Scientific
GFL-1083 Shaking Water Bath	GFL
Protein LoBind Safe-Lock Tube	Eppendorf®
2100 Bioanalyzer	Agilent Technologies
Labchip GX	Caliper Life Sciences
BluePippin	Sage Science
NextSeq 500 sequencing system	Illumina

Appendix Table 2: Reagents used in the experimental part of this study.

Product name	Art.nr/Catalogue nr.	Manufacturer
DMEM (Dulbecco's Modified Eagle's Medium) high glucose	D6429	Sigma-Aldrich
HEPES	H4034	Sigma-Aldrich
BSA (Bovine Serum Albumin)	A2153	Sigma-Aldrich
PBS (Phosphate buffered saline)	-	Pre-made in lab
Liberase™ TH Research Grade	5401151001	Roche
Accutase® solution	A6964	Sigma-Aldrich
RPMI (Roswell Park Memorial Institute Medium) 1640	R8758	Sigma-Aldrich
FCS/FBS (fetal calf serum/fetal bovine serum)	-	Pre-made in lab

DMSO (Dimethyl sulfoxide)	D8418	Sigma-Aldrich

Appendix Table 3: Commercial kits used in the experimental part of this study.

Product name	Art.nr/Catalogue nr.	Manufacturer
miRNeasy Serum/Plasma Kit	217184	QIAGEN
NEXTflex sRNA-seq kit v3	5132-05	Bioo Scientific
QIAquick PCR Purification Kit	28104	QIAGEN
High sensitivity DNA kit	5067-4626	Agilent Technologies
KAPA Library Quantification Kit	07960140001	Roche

Appendix Table 4: Software used in the computing part of this study.

Software name	Version
Seurat	4.1.0
R	4.1.2
RStudio	2021.09.1
bcl2fastq2 conversion software (Illumina)	2.20.0422
cutadapt	3.7
bowtie2	2.4.5
htseq-count	2.0

A.3.2 Protocol development and optimization before establishing single cell suspension from fresh colorectal cancer tissue

A.3.2.1 Collection of fresh colorectal cancer tissue

The sample collection of fresh CRC tissue was organized and provided by Biobank1 in collaboration with the Department of Pathology at St. Olav's University hospital. These samples were used both for protocol development and optimization, and for establishing single cell suspension to be used in scRNA-seq analysis.

A.3.2.2 Protocol development and optimization

A protocol for converting fresh pancreatic tissue into single cell suspension by Bernard, V. et al. [107] (Appendix C) and a workflow for preparing a single cell suspension from solid tissue by Reichard, A. and Asosingh, K. [108] (Appendix D), in addition to standard protocols for freezing and thawing intact cells, were used as basis for development and optimization in terms of establishing single cell suspension from fresh CRC tissue.

Cell specifications such as number of cells was estimated using Countess™ Automated Cell Counter (Invitrogen™) and Moxi Z Mini Automated Cell Counter (ORFLO), appearance of cells was evaluated using AE30 Binocular Inverted Microscope (Motic®), and flow cytometry analysis using FACSCanto™ Flow Cytometer (BD Biosciences) and fluorochromes Calcein Green, Live/Dead Far Red, and Propidium Iodide (PI) was performed for validating changes made to the protocols regarding identifying live and dead cells (Appendix E), ensuring cell viability for further analyzes. Cell viability was also confirmed by the Genomics Core Facility (GCF) at the Norwegian University of Science and Technology (NTNU) [138].

A.3.2.3 Establishing single cell suspension from fresh colorectal cancer tissue

Single cell suspension from fresh CRC tissue was established by following the developed and optimized protocol; CRC tissue (approximately 1 cm^2) was transported to the laboratory on ice in DMEM, HEPES (25 mM), and BSA (1%) in a conical tube (15 mL) after surgical resection. The tissue was then rinsed with PBS to remove blood and unwanted material. Liberase™ TM Research Grade (Roche, 5 mg/mL) and Accutase® solution (Sigma-Aldrich, 2 mL) was mixed, and 1 mL of the solution was transferred to a petri dish followed by the rinsed tissue. The tissue was then minced with a sterile surgical scalpel to 0.5 to 1.0 mm fragments, and the mixture was transferred to a conical tube (15 mL) containing 1 mL of the Liberase-Accutase solution. Warm tissue digestion was done by incubating the tissue fragments at the orbital shaker New Brunswick™ Innova® 44 Incubator Shaker (Eppendorf®, 37°C, 250 RPM, 20 min), gently pipetting the solution every 10 minutes. After the digestion period, the tissue slurry was filtrated through a 70 µm cell strainer followed by a 40 µm cell strainer. The single cell suspension was transferred to a new conical tube (15 mL) and centrifuged using Centrifuge 5810 R (Eppendorf®, 4°C, 400 RCF, 5 min). The supernatant was discharged, and the cell pellet was resuspended in 1 mL cold RPMI with 20% FCS for downstream cell counting, microscopy and freezing of cells.

A.3.2.4 Freezing single cell suspension

1 mL of DMSO (20%), RPMI (40%), and FCS (40%) was added to the single cell suspension dropwise, and the solution was mixed slowly before being transferred to CryoTube™ Vials (Thermo Scientific). The single cell suspension was then put in a freezer (-80°C) for about 48 hours before being transferred and stored at liquid nitrogen.

A.3.2.5 Thawing single cell suspension

A cryotube containing single cell suspension was collected from the liquid nitrogen tank and thawed at a GFL-1083 laboratory water bath (GFL, 37°C, 10 min). 1 mL RPMI with 20% FCS was transferred to a conical tube (50 mL), followed by the single cell suspension. A double volume of RPMI with 20% FCS relative to its volume in the conical tube was then added a total of 5 times (1, 2, 4, 8, and 16 mL), with a pause of 1 min in between each round of adding. The solution was then centrifuged using Centrifuge 5810 R (New Brunswick Scientific, 22°C, 300 g, acceleration 9, brake 5, 5 min). The supernatant except 1 mL was discharged, and 10 mL PBS was added. The solution was centrifuged again under the same conditions, and the supernatant was discharged. The cell pellet was resuspended in 1 mL PBS, and the solution was filtrated through a 40 µm cell strainer before being transferred to Protein LoBind Safe-Lock Tubes (Eppendorf®) for downstream analysis.

A.3.3 Single cell RNA sequencing and downstream data analysis

A.3.3.1 Single cell RNA sequencing of colorectal cancer single cell suspension

ScRNA-seq was performed on CRC single cell suspensions (n=4) by the Genomics Core Facility (GCF) at the Norwegian University of Science and Technology (NTNU) [138] by following a standard protocol (principle described in more detail in chapter A.4.2.1).

A.3.3.2 Processing and analyzing single cell RNA sequencing data

Pre-processing of raw scRNA-seq output files was conducted by the master's thesis' supervisor Robin Mjelle at the Bioinformatics Core Facility (BioCore) at NTNU [139] (principle described in more detail in chapter A.4.3.1). Further data processing and

analysis was performed using Seurat V4.1.0 [78] in R V4.1.2 [140] with RStudio V2021.09.1 [141], which in short involved writing scripts for creating cluster graphs for each sample, integrating the sample data, and subcluster the integrated data (Appendix F).

A.3.3.2.1 Creating cluster graphs for each sample and manually annotating the clusters
The pre-processed scRNA-seq data was first normalized, multiplied by a scale factor of 10,000, and log-transformed using `NormalizeData`, before identifying the 2,000 most variable genes by `FindVariableFeatures`. Uninteresting sources of variation were regressed out using `ScaleData`. Linear dimensional reduction of the data was then performed running a principal component analysis (PCA) using `RunPCA`. A weighted nearest neighbor (WNN) algorithm was employed to create a multidimensional cluster graph with `FindNeighbors`, and the resolution was adjusted with `FindClusters`. Visualization of the cluster graph was done through Uniform Manifold Approximation and Projection (UMAP) for non-linear dimension reduction using `RunUMAP`.

The top 10 cell-type-specific marker genes for each cluster were found with `FindAllMarkers`, only looking at genes found in a minimum of 25% of the clusters and showing at least 0.25-fold (log-scale) difference between the clusters. Manual cell type annotation was performed by comparing the gene expression pattern formed by the identified cluster cell-type-specific marker genes to the gene expression pattern of known cell types found in existing manually annotated scRNA-seq reference data in the Human Protein Atlas (HPA) [87], Azimuth [115], and PanglaoDB [79].

A.3.3.2.2 Integration of sample data to perform a joint analysis of all samples
Integration of the datasets from all samples were performed by first identifying cells from each data set within the same clusters ("anchors") with `FindIntegrationAnchors` and then combining the results using `IntegrateData`. An integrated analysis of the data assay was then performed by normalizing, scaling, and running PCA followed by UMAP to visualize the integrated clusters.

A.3.3.2.3 Subclustering of the integrated data
Higher-level cell type groups of interest in the integrated data were clustered again using `NormalizeData`, `FindVariableFeatures`, `ScaleData`, `RunPCA`, `FindNeighbors`, `FindClusters`, and `RunUMAP`. The subclusters were manually annotated by comparing the top 10 cell-type-specific marker genes to marker genes of known cell types found in existing manually annotated scRNA-seq reference data in the Human Protein Atlas (HPA) [87], in the cell type marker database-part of PanglaoDB [79], and in additional literature.

A.3.4 Small RNA sequencing of isolated RNA from serum samples of colorectal cancer patients

A.3.4.1 Collection of serum samples from CRC patients
Serum samples from false positive CRC patients (n=21), true positive CRC patients with localized disease (n=16), and true positive CRC patients with metastatic disease (n=16) were provided by Biobank1.

A.3.4.2 Total RNA isolation of serum
Total RNA was isolated from thawed serum samples (200 µL) using miRNeasy Serum/Plasma Kit (QIAGEN). In brief, cell lysate preparation was done by adding QIAzol

Lysis Reagent (5 times volume of the sample, 1 ml) to the plasma sample and vortexed until lysed. Phenol (equal volume to the starting sample, 200 µL) was then added and vortexed additionally. The lysate was separated into aqueous and organic phase by centrifugation, and RNA was extracted from the upper aqueous phase. 100% ethanol (1.5 volumes to the aqueous phase, approximately 900 µL of ethanol to 600 µL aqueous phase) was added and mixed by pipetting. The sample was loaded onto a spin column provided by the kit. RNA bound to the spin column membrane, and contaminants were washed away using Buffer RWT (700 µL), Buffer RPE (500 µL), and 80% ethanol (500 µL). The column was finally transferred to a new elution tube, where RNA was eluted using RNase-Free Water (14 µL). Isolated RNA was stored at -80°C.

A.3.4.3 Small RNA sequencing of isolated RNA from serum samples

Small RNA-seq of isolated RNA from serum samples of CRC patients was performed by GCF at NTNU [138] by following a standard protocol; Assessment of RNA quality and relative size were conducted by measuring the samples using Eukaryote Total RNA Pico assay on the 2100 Bioanalyzer (Agilent Technologies). Isolated miRNA from serum generally has lengths of about 22 nucleotides, and the presence of this peak were checked on the bioanalyzer trace for quality control (Supplementary Figure 3). RIN values were not considered as ribosomal RNA are degraded in serum and plasma.

Small RNA-seq of 47 samples/libraries were performed using the NEXTflex sRNA-seq kit v3 (Bioo Scientific). The adapter-dimer reduction technology incorporated into this kit allows low input library preparation. Reducing ligation-associated bias involves the use of adapters with randomized bases at the ligation junctions, resulting in greatly decreased bias in comparison to standard protocols. In brief, 10.5 µl total RNA extracted from 200 µl serum, was used as a template for 3' 4N and 5' 4N adenylated adapter ligation, followed by reverse transcription-first strand synthesis. In the first ligation step, 10 calibrator RNAs were mixed with the RNA to control for technical variation during the data analysis. The sequences of the calibrators are previously descried by Mjelle, R. et al. [142]. By applying these products as a template for second-strand synthesis, double-stranded cDNA was prepared by PCR amplification (22 cycles). Fragments/libraries were run on a Labchip GX (Caliper Life Sciences), for quality control and quantitation. Individual libraries were normalized to 25 nM and pooled. The library pool was purified with the QIAquick PCR Purification Kit (QIAGEN) according to instructions.

Automated size selection was performed using the BluePippin (Sage Science), with a range of 135-165 bp to select the ~ 152 bp fragment. Following size selection, the pool was evaluated on the 2100 Bioanalyzer (Agilent Technologies) using the High Sensitivity DNA kit (Agilent Technologies). The pool of libraries was quantified with the KAPA Library Quantification Kit (Roche). Libraries were normalized to 2.6 pM subjected to clustering. Single read sequencing was performed for 51 cycles on NextSeq 500 (Illumina) high output flow cell, according to the manufacturer's instructions. Sequence reads were demultiplexed and converted from BCL to fastq file format using bcl2fastq2 conversion software V2.20.0422 (Illumina).

A.3.4.4 Processing and analyzing small RNA sequencing data

The raw small RNA-seq data was processed by the Bioinformatics Core Facility (BioCore) at NTNU [139] by following a protocol previously described by Mjelle, R. et al. [143]. Specifically, the adapters and the random nucleotides at both ends were removed using cutadapt (v.3.7) followed by alignment to the human genome (hg38) using bowtie2 (v2.4.5). The aligned reads were counted using htseq-count (v2.0) with the

corresponding GFF files from miRBase. An expression matrix for mature miRNAs were constructed from the htseq-count output by combining the individual expression data for each sample and used for statistical analyses in R.

Differentially expressed small RNAs between groups were detected using the Limma-Voom procedure in R. The Limma-Voom procedure for the comparisons is shown below:

```
Smallrna.exp.dge <- DGEList(Smallrna.mature,group =
Smallrna.ss.order$Sample_Group)

# Smallrna.mature is the data frame with the expression values

# Smallrna.ss.order is the samplesheet with the groups

Smallrna.exp.dge <- calcNormFactors(Smallrna.exp.dge)

keep <- rowSums(edgeR::cpm(Smallrna.exp.dge)>1) >=
dim(Smallrna.exp.dge)[2]/2

Smallrna.exp.dge <- Smallrna.exp.dge[keep,]

Smallrna.exp.dge <- calcNormFactors(Smallrna.exp.dge, method="TMM")

Smallrna.exp.dge$samples$norm.factors <-
Smallrna.calibrator.dge$samples$norm.factors

des <- model.matrix(~0+Smallrna.ss.order$Sample_Group)

Smallrna.ss.order$ID2==colnames(Smallrna.exp.dge)

colnames(des) <- c("Falsepositive","LocalCRC","MetastaticCRC")

v <- voom(Smallrna.exp.dge,design = des,plot = T)

fit <- lmFit(v, design=des)

contrasts <- makeContrasts(Local_vs_False=LocalCRC-Falsepositive,

                           Met_vs_False=MetastaticCRC-Falsepositive,

                           Met_cs_Local=MetastaticCRC-LocalCRC,

                           levels=des)

fit2 <- contrasts.fit(fit, contrasts=contrasts)

fit2 <- eBayes(fit2)

colSums(decideTests(fit2)!=0)

Local_vs_False.toptable <-
topTable(fit2,coef="Local_vs_False",sort.by="P",adjust.method="BH",n=Inf)

Met_vs_False.toptable <-
topTable(fit2,coef="Met_vs_False",sort.by="P",adjust.method="BH",n=Inf)

Met_cs_Local.toptable <-
topTable(fit2,coef="Met_cs_Local",sort.by="P",adjust.method="BH",n=Inf)
```

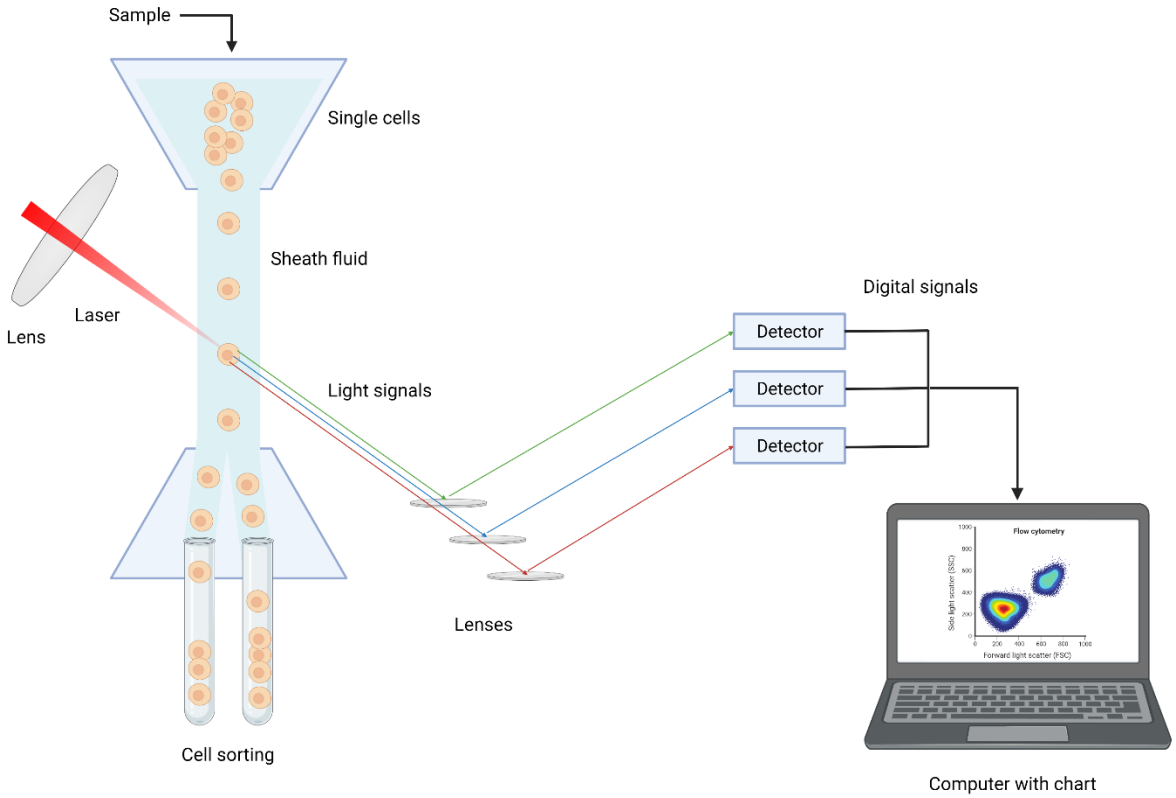
All plots for the sequencing data were generated in R using the libraries ggplot2, pheatmap and VennDiagram.

A.4 Principles of the experimental and computing methodology

A.4.1 Principle of flow cytometry

Flow cytometry is a technology that provides multi-parametric analysis of cells or cell populations based on their light scattering or fluorescent characteristics (Appendix Figure 2) [144]. A flow analysis is conducted by suspending a single cell solution in sheath fluid and pressurizing it to make a coaxial flow where cells align in a single file fashion in the core of the sheath fluid stream [145]. The stream is then directed into lasers, which generates cell-specific light scatter and fluorescent signals that are detected by a computer which displays the data as charts [144, 145].

The light scattering is directly related to morphological properties of the cell, while fluorescence emission is proportional to the amount of fluorochrome bound to it [145]. Fluorochromes are fluorescent probes used to stain components in a cell [145], where examples of such probes are calcein green, live/dead far red, and propidium iodide (PI). Some flow cytometers are sorting, meaning they can perform a fluorescence activated cell sorting (FACS) analysis, which includes an extra step where a heterogenous sample is physically sorted into separate populations for further analysis [145].

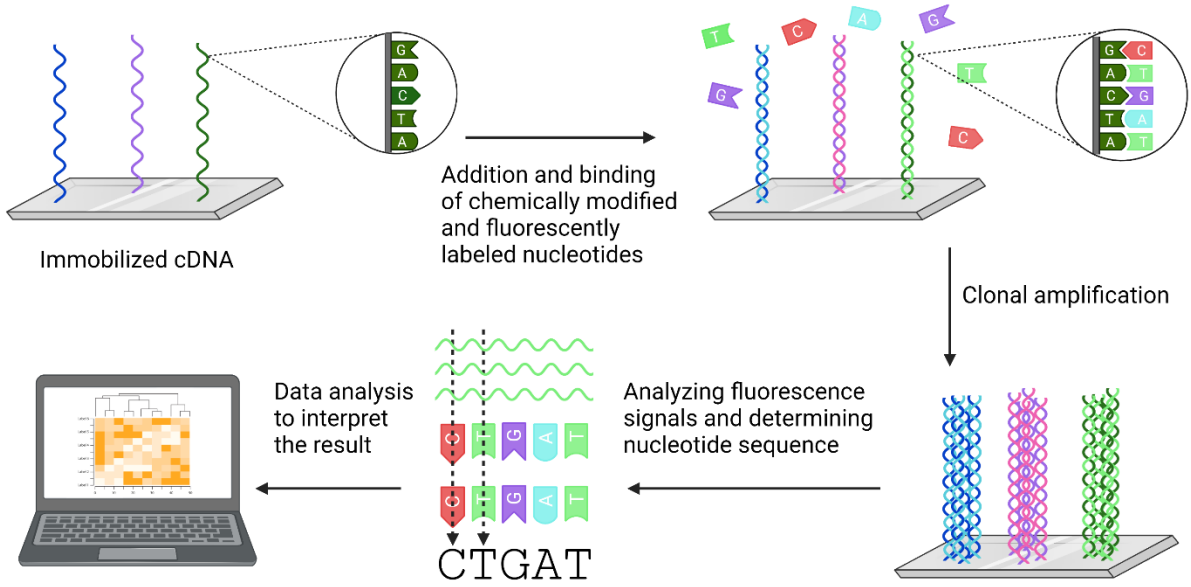


Appendix Figure 2: The main components and underlying working principle of a sorting flow cytometer. A single cell solution sample is suspended in a sheath fluid and pressurized to make a coaxial flow, where lasers generate cell-specific light scatter and fluorescent signals related to cell morphology and bound fluorochrome, respectively. Adapted from Adan, A. et al. [145] and created with BioRender.com.

A.4.2 Principle of RNA sequencing techniques such as single cell RNA sequencing and small RNA sequencing

RNA-seq is a technique that uses high-throughput massive parallel sequencing/next-generation sequencing (NGS) methods to provide insight to into the transcriptome of a

sample by determining the nucleotide sequence in millions of sequence clusters in parallel, indicating e.g. gene expression [65]. A widely used NGS platform is provided by Illumina and use a sequencing-by-synthesis (SBS) approach (Appendix Figure 3) [65]. Here, chemically modified and fluorescently labeled nucleotides bind to immobilized cDNA fragments through natural complementarity [146]. Clusters of the same strand are then created by clonal amplification to ensure detectable, relative fluorescent signals, making it possible for instrument software to identify nucleotides and thus sequence the molecule [65, 146].

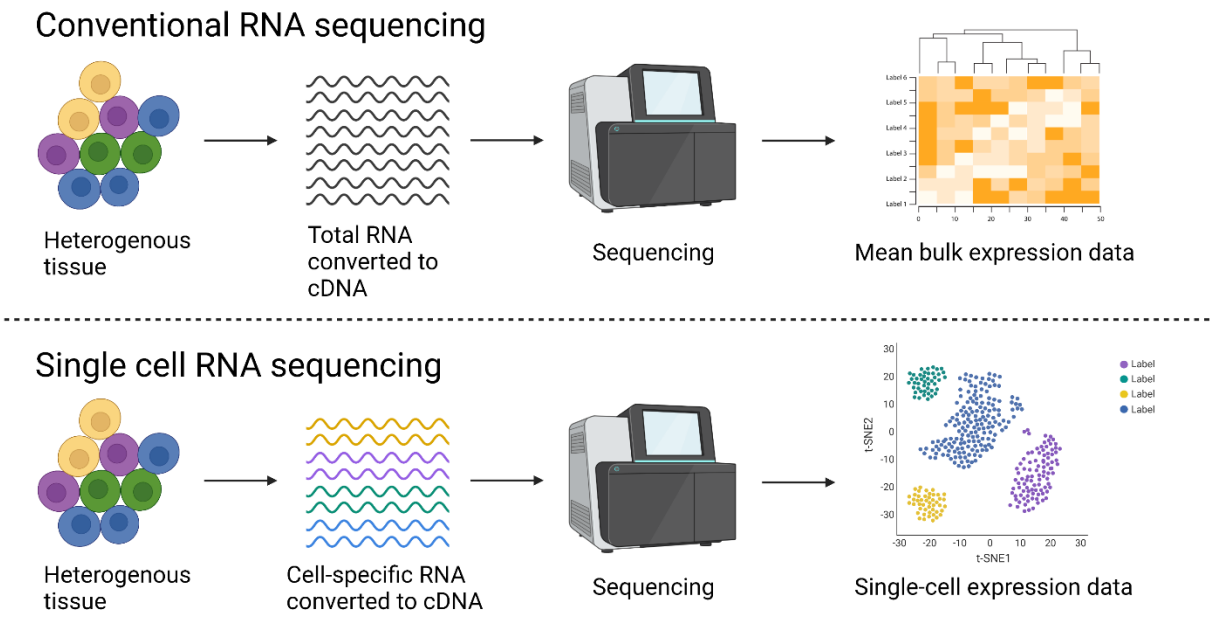


Appendix Figure 3: Sample sequencing using a next-generation sequencing (NGS) method. Chemically modified and fluorescently labeled nucleotides binds to immobilized sample cDNA fragments before clusters of the same strand are created by clonal amplification to provide detectable fluorescent signals that can be analyzed. Created with BioRender.com.

Sequencing-ready libraries must be created before performing NGS [65], and different types of RNA-seq tends to differ in terms of library preparation. Something that is possible in most library preparation protocols are multiplexing and the use of "spike-ins". Multiplexing is a process where multiple libraries can be pooled together to save resources, where unique barcode sequences are added to each library in advance to distinguish between them during data analysis [146]."Spike-ins" are positive controls that can be added to sequencing-ready libraries at different concentrations, working as a quality control tool for separating technical from biological variation and thus further improve the accuracy of e.g. gene expression levels [65].

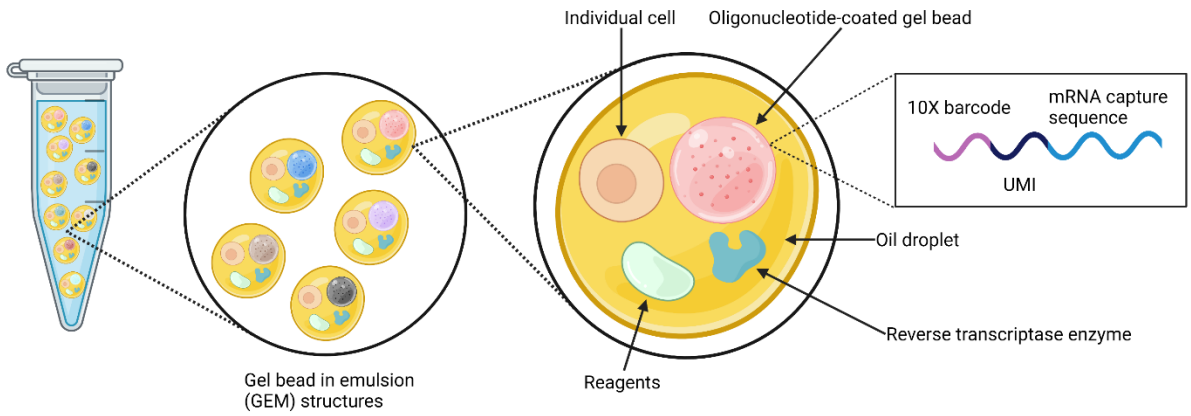
A.4.2.1 Principle of single cell RNA sequencing

Different cells show different patterns of gene expression, which reflects the cells different properties, functionalities, and behaviors [147]. Conventional RNA-seq only indicates the average expression level for each gene across a large bulk of sample cells [30], while the more novel method scRNA-seq estimates a distribution of expression levels for each gene across a cell population (Appendix Figure 4) [31]. This ultimately makes scRNA-seq data more complex and challenging to analyze compared to conventional RNA-seq [30].



Appendix Figure 4: A broad comparison of conventional RNA sequencing (RNA-seq) and single cell RNA sequencing (scRNA-seq). Conventional RNA-seq only indicates the average expression level for each gene across a large bulk of sample cells, while scRNA-seq estimates a distribution of expression levels for each gene across a cell population. Adapted from Wellcome Sanger Institute [31] and created with BioRender.com.

Sequencing-ready libraries for scRNA-seq can be prepared using the popular droplet-based 10X Genomics Chromium scRNA-seq platform [148], which requires the sample being in the form of single cell suspension [149]. The sample cells are individually encapsulated by Chromium into an oil droplet along with a gel bead, reagents, and reverse transcriptase enzymes, forming gel bead in emulsion (GEM) structures (Appendix Figure 5) [150]. Each gel bead is coated with oligonucleotides that consists of a 10X cell barcode to distinguish the different cells, a unique molecular identifier (UMI) to differentiate molecules within the same cell, and specific sequences to capture the 3' end of the cells' mRNA molecules [148, 150].



Appendix Figure 5: Gel bead in emulsion (GEM) structures. A GEM consist of oil droplet-encapsulated cells, oligonucleotide-coated gel beads, reagents, and reverse transcriptase enzyme. Adapted from 10X Genomics [150] and created with BioRender.com.

The GEM reagents ultimately cause a reaction which results in dissolved gel beads and cell lysis, ending in the cells' mRNA molecules being captured and barcoded [150].

Further on in the GEM structures, reverse transcription is performed by the enzymes to generate cDNA from the mRNA template [149]. Barcoded cDNA fragments for all the cells are then pooled and amplified by PCR, and further downstream NGS are conducted [149, 150].

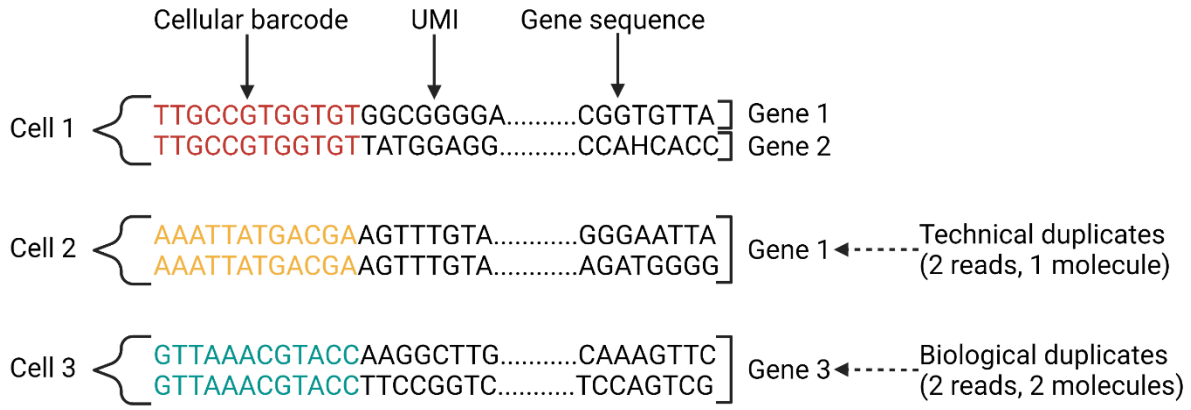
A.4.2.2 Principle of small RNA sequencing

Small RNA-seq is a technique used for isolating and sequencing small RNA species such as miRNAs [151]. Sequencing-ready libraries for small RNA-seq can be prepared following protocols compatible with popular Illumina sequencing platform tools, which requires the sample being in the form of total RNA [151, 152]. Isolation of miRNAs can be accomplished using two-adaptor ligation-based methods to extend their length and introduce primer-binding sites for reverse transcription and subsequent amplification [153]. In such methods, two adaptors are sequentially ligated to the 5' and 3' ends of the RNA molecules [153]. The adaptors contain unique sequences used to distinguish the different RNA species and thus quantify small RNAs like miRNA, and to discriminate RNA molecules deriving from different samples [153]. After adaptor ligation, reverse transcriptase is performed on the ligated fragments to generate cDNA, followed by PCR amplification and NGS [153].

A.4.3 Principle of processing and analyzing data from different RNA sequencing techniques such as single cell RNA sequencing and small RNA sequencing

A.4.3.1 Principle of processing and analyzing single cell RNA sequencing data

ScRNA-seq data consists of several different strings of nucleotide sequences, which all includes a cellular barcode, a UMI, and a gene-derived mRNA sequence (Appendix Figure 6) [116]. The workflow of processing this data is often divided into pre-processing and downstream analysis. During pre-processing, a cell x gene matrix of UMI counts is generated (Appendix Figure 7) [154]. Technical duplicates sharing the same barcode and UMI is collapsed into a single string for counting, whereas biological duplicates originating from the same cell but from different mRNA molecules are counted as separate strings [116]. The count matrix essentially shows the counts for mRNA molecules originating from a gene across all cells [116].



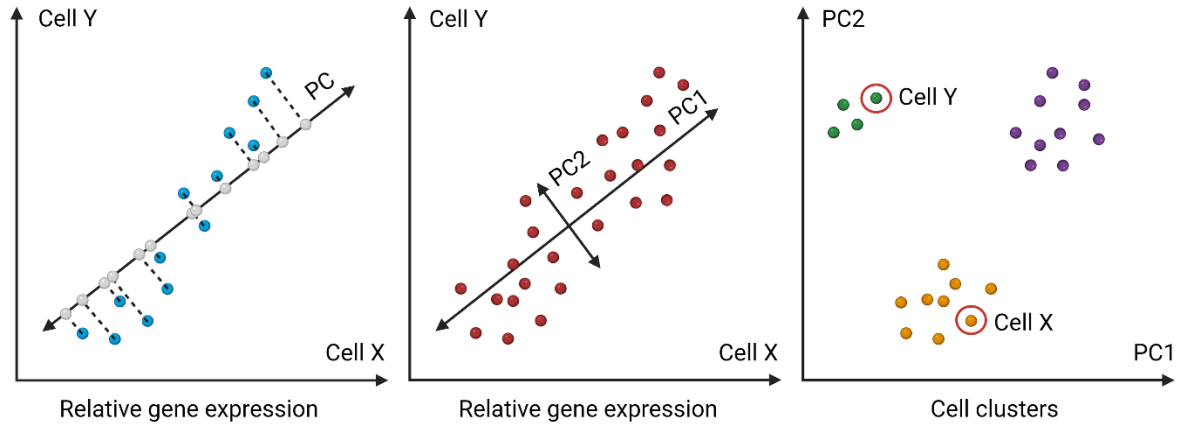
Appendix Figure 6: A visualization of single cell RNA sequencing (scRNA-seq) data. ScRNA-seq data consists of several nucleotide sequence strings which all includes a cellular barcode, unique molecular identifier (UMI), and gene sequence. Adapted from Harvard Chan Bioinformatics Core (HBC) [116] and created with BioRender.com.

	Cell 1	Cell 2	Cell 3	...
Gene 1	1	1	0	
Gene 2	1	0	0	
Gene 3	0	0	2	
...				

Appendix Figure 7: Unique molecular identifier (UMI) count matrix generated from single cell RNA sequencing (scRNA-seq) raw data. The UMI count matrix includes different cells for each column and different genes for each row. Adapted from Harvard Chan Bioinformatics Core (HBC) [116] and created with BioRender.com.

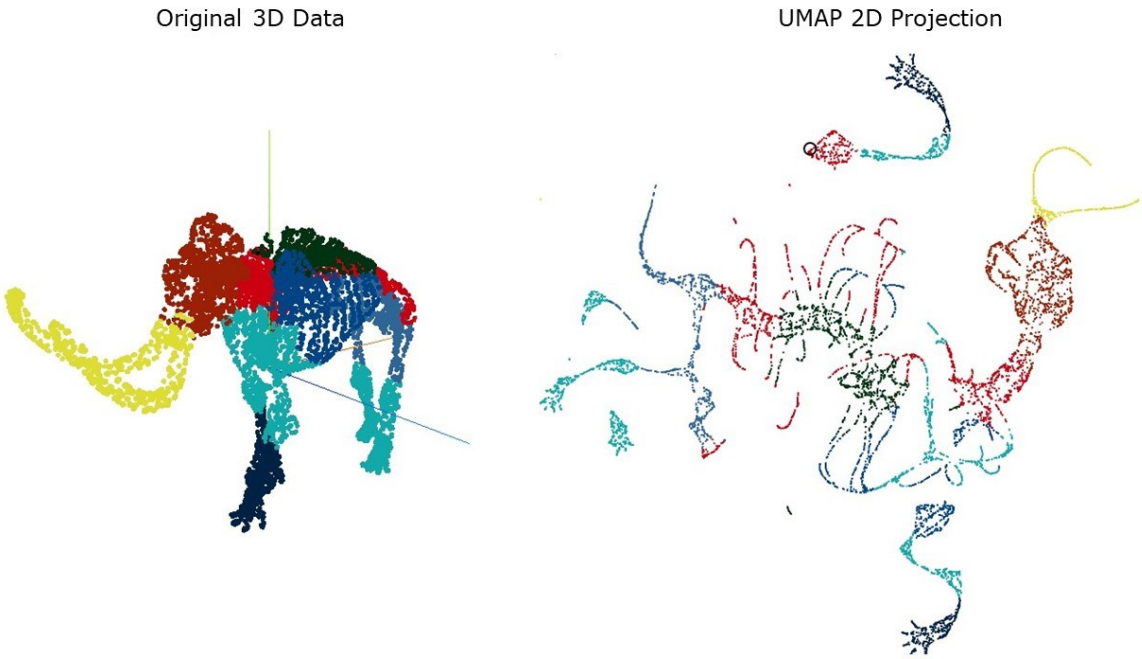
The count matrix can further be processed consistent with a downstream clustering analysis workflow [113]. Clustering is the process of distributing observations similar to each other into their own separate clusters [155]. In biological experiments, cells with similar gene expression patterns can be clustered and thus correspond to different cell types [156]. Several clustering methods have been developed [155], e.g. graph-based clustering as implemented in the single cell genomics toolkit Seurat [78]. In short, this includes data normalization and scaling, principal component analysis (PCA), and constructing a weighted nearest neighbor graph (WNN) to use as input in Uniform Manifold Approximation and Projection (UMAP) for cluster visualization [116].

Linear dimension-reducing PCA can be performed on the most variable genes by projecting data points onto principal components (PCs) and giving the cells PC scores [116]. PCs are mathematical optimized vectors constructed to maximize the variance between data points as much as possible along that line [157]. PCs are orthogonal to each other and thus independent, meaning variance explained in the different PCs do not overlap and therefore represent information as efficiently as possible [157]. In a data set of *n* cells there would be *n* PCs, where PC1 represents the largest variation, PC2 the second largest variation not covered by PC1, and so forth [116, 157]. A general PC score plot will have axes that maximize the variance, and cells with similar gene expression will therefore cluster together (Appendix Figure 8) [157].



Appendix Figure 8: A visualization of dimension reducing principal component analysis (PCA). PCA includes data projection onto a principal component (PC) (left), orthogonal PCs (middle), and PC score plot (right). Adapted from Programmathically [157] and created with BioRender.com.

A selection of significant PCs, where clusters of cells with similar gene expression are presented, can be used for constructing a multidimensional WNN graph [116]. This is done by employing an unsupervised algorithm to assign a cell the cluster most common among its k nearest neighbors, where nearer neighbors are weighted and thus contribute more than the more distant ones [78, 116]. This allows for more detailed clustering, where the resolution can be adjusted to partition the cells into an even greater number of clusters [116]. The WNN cell clusters can be visualized using UMAP, a manifold learning technique for dimension reduction which uses mathematical algorithms to describe nonlinear relationships within a data set and visualize it in a two-dimensional space [116, 158]. Simplified, UMAP uses a graph layout algorithm to arrange the multidimensional WNN graph data in a low-dimensional graph the best way possible (Appendix Figure 9) [159].



Appendix Figure 9: Simplified example of how Uniform Manifold Approximation and Projection (UMAP) arranges multidimensional graph data in a low-dimensional graph. A 3D-figure of a Wooly mammoth (left) is projected by UMAP in 2D (right). The figure is created by Coenen, A. and Pearce, A. [159] for Google People + AI research (PAIR) based on open-source tools with Apache License 2.0.

Continuing a scRNA-seq data analysis, cell-type-specific marker genes are identified and used for cluster verification and annotation of the visualized cell clusters [113]. Cell-type-specific marker genes can be defined as genes which are highly expressed primarily in a single cell type [160], where such genes are identified by comparing the expressed genes of a single cluster to the other clusters [113]. The cell-type-specific marker genes of a cluster can ultimately define cellular identity [160], where clusters reflecting known markers or marker combinations as described in literature or databases are verified on those grounds and can thus be annotated.

To summarize the standard workflow principle and analysis of scRNA-seq data, the data is pre-processed into a UMI count matrix and normalized, highly variable features are identified and scaled, linear dimension-reduction by PCA is performed, cells are clustered in more detail by employing a WNN algorithm to create a multidimensional WNN graph, non-linear dimension-reduction by UMAP is performed to arrange clusters in 2D, and cell-

type-specific marker genes are found before cell type identities are assigned to the different clusters based on marker gene information in literature or databases. In addition, it is possible to perform a joint analysis of several scRNA-seq datasets with the goal of identifying common cell types across the sets [161]. This can be done by integration, where cross-dataset pairs of cells in a matched biological state ("anchors") are used both to correct for technical differences between datasets and to perform comparative scRNA-seq analysis [161].

A.4.3.2 Principle of processing and analyzing small RNA sequencing data

The processing and analysis of small RNA-seq data to investigate miRNA expression does in brief consist of quality assessment, UMI analysis and filtering, reference-based alignment, and creation of a UMI count matrix [162]. UMI analysis and filtering involves mapping the structure of the entire read, which includes two adaptors, a reverse transcription primer, and an UMI [162]. The reads are then retained and trimmed to remove the reads that are either too short (<18 bp) or too long (>30 bp) [162]. Reference-based alignment to different mature miRNA-databases can then be performed to reveal which specific miRNAs are present in the sample, before finally collapsing the UMI reads to generate a count matrix representing the counts for each miRNA that were present in the original biological sample prior to amplification [162].

B REC approval

Forskningsprosjekt

Pre-diagnostiske biomarkører for kolorektal kreft.

Vitenskapelig tittel:

Pre-diagnostic biomarkers for Colorectal Cancer

Prosjektbeskrivelse:
Tykktarms- og endetarmskreft er blant de krefttypene med høyest forekomst i Norge. i 2013 ble 2781 personer diagnostisert med denne krefttypen i Norge. Kreft som oppdages tidlig kan lettere behandles og øker sannsynligheten for overlevelse betraktelig. Tykktarms- og endetarmskreft diagnostiseres idag hovedsalig ved hjelp av koloskopi ved å se etter ondartede polypper i tarmen. Dette er en ressurskrevende metode som medfører ubehag for pasienten. En blod-basert test vil styrke diagnostiseringen av sykdommen og gjøre det lettere å foreta en bred screening av risikogrupper. MikroRNA er en lovende gruppe molekyler som er vist å kunne predikere ulike typer kreft ved å måle nivået i blod. Ved å bruke blod fra HUNT biobank ønsker vi å se på endringen i mikroRNA 1-3 år før pasienten blir diagnostisert for tykktarms- og endetarmskreft.
(Prosjektleders prosjektbeskrivelse)

Ref. nr.: 2016/534 Prosjektstart: 01.02.2016 Prosjektslutt: 31.12.2019

Behandlingsstatus: Pågående
Prosjektleder: Robin Mjelle
Forskningsansvarlig(e): NTNU, Institutt for samfunnsmedisin
Norges teknisk-naturvitenskapelige universitet
Norges teknisk-naturvitenskapelige universitet
Initiativtaker: Bidragsforskning
Finansieringskilder:

Prosjektet er tildelt 250 000 kroner fra St Olavs Hospital.

Forskningsdata: Registerdata, Humant biologisk materiale
Utvalg: Pasienter/klienter, Kontrollgruppe(r)

Materiale fra biobank:
HUNT

Behandlet i REK

Dato	REK
22.04.2016	REK midt
21.10.2016	REK midt
22.09.2017	REK midt
27.10.2017	REK midt
05.12.2018	REK midt
20.03.2019	REK midt

REK
REGIONALE KOMITEER FOR MEDISINSK OG HELSEFAGLIG FORSKNINGSETIKK

Region:	Saksbehandler:	Telefon:	Vår dato:	Vår referanse:
REK midt	Ramunas Kazakauskas		22.02.2021	30022
			Deres referanse:	

Robin Mjelle

30022 Pre-diagnostiske biomarkører for kolorektal kreft.

Forskningsansvarlig: Norges teknisk-naturvitenskapelige universitet

Søker: Robin Mjelle

REKs vurdering

Du sendte en søknad om prosjektendring den 02.02.2021. Søknaden ble behandlet av leder for REK midt på fullmakt, med hjemmel i helseforskningsloven § 11 og forskrift om behandling av etikk og redelighet i forskning § 10.

Du søker om "å gjøre enkeltcelle-sekvensring av RNA fra tarmbiopsier. Bakgrunn for endringene er at vi ønsker å se på endringer i RNA-nivå i enkeltceller mellom ulike pasientgrupper i vevsmateriale i tillegg til blod som vi allerede har godkjenning til å gjøre. Disse analysene vil gi et mer komplett bilde av endringer i immunceller hos tarmkreftpasienter og hvordan endringene påvirker selve kreftcellene i tarmen. Prøvemateriale vil bli innsamlet i samarbeid med Biobank1 og er samtykkebasert. Analysene vil bli gjennomført ved NTNU sine laboratorier ved St.Olavs Hospital. RNA-data som blir produsert fra sekvenseringen blir lagret ved HUNT-cloud som er godkjent for sikker lagring av sensitive data. Undersøkelsene gir ikke genetisk informasjon som kan være handlingsutløsende."

Vi har vurdert søknad om prosjektendring, og har ingen forskningsetiske innvendinger mot endringen av prosjektet. Endringen vil ikke føre til handlingsutløsende funn. Hensynet til deltakernes velferd og integritet er fremdeles godt ivaretatt. Vi minner om at prosjektet må gjennomføres i henhold til tidligere vedtak i saken.

Vedtak

Godkjent

Med vennlig hilsen

Vibeke Videm
Dr. med.
Leder, REK midt

Ramunas Kazakauskas
Rådgiver

C Protocol for converting fresh pancreatic tissue into single cell suspension

The following bullet points list a protocol for converting fresh pancreatic tissue into single cell suspension as described by Bernard, V. et al. [107], and was used as a basis for development and optimization in terms of establishing single cell suspension from fresh CRC tissue.

- Transport tissue (approximately 1 cm^2) to the laboratory on ice after surgical resection in DMEM, high-glucose, GlutaMAX™ Supplement, HEPES (Thermo Fisher, 10564011) in 1% bovine serum albumin (Thermo Fisher, B14) in a 15-mL conical tube.

- Rinse tissue with PBS to remove blood and other unwanted material.

- Transfer tissue to a 35×12 mm Petri Dish (Thermo Fisher, #150318), and minced with sterile surgical scalpel to 0.5-1.0 mm fragments in approximately 1 mL of the media.

- Digest tissue using Liberase TH Research Grade and Accutase solution for PDAC tissues (Sigma-Aldrich, A6964). For warm digestion with Liberase TH Research Grade, tissue fragments are incubated to a final concentration of 10 mg/mL and placed on an incubated orbital shaker at 37°C, 225 RPM for 20 minutes and gently pipetted every 10 minutes.

- A second digestion is performed by incubating the sample in sterile-filtered Accutase solution on a shaker at 37°C, 225 RPM for 30 minutes, with gentle pipetting every 10 minutes.

- At the end of the digestion period, the fragments (tissue slurry) were gently pipetted and washed to maximize the release of single cells.

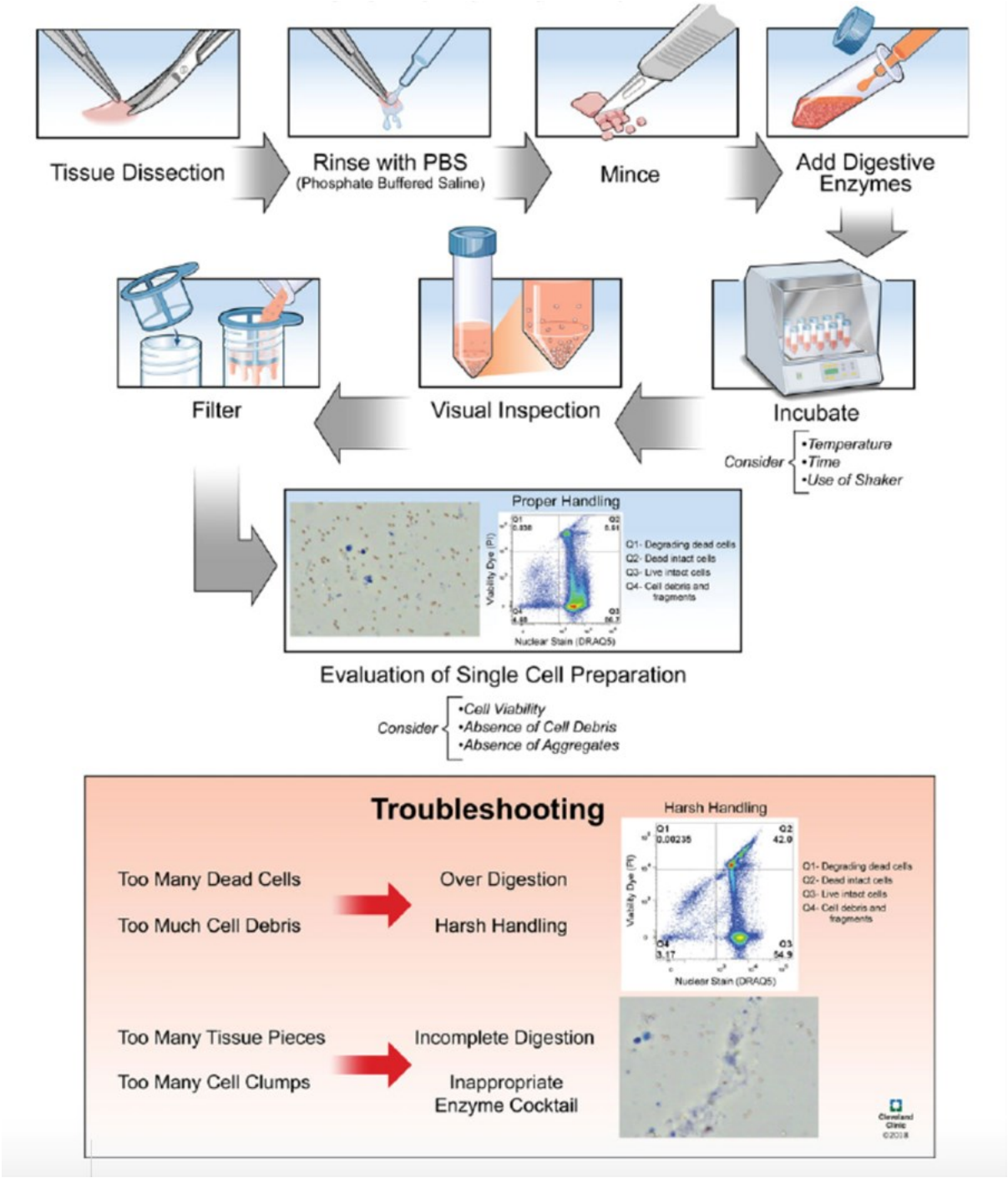
- The tissue slurry is passed through a 100-µm cell strainer followed by a 35-µm cell strainer.

- The single cell suspension is transferred to a new 15-mL conical tube and centrifuged for 5 minutes at 400 RCF at 4°C.

- The supernatant is discarded, and the cell pellet is resuspended in 400 µL of PBS for downstream cell viability analysis and cell counting.

D General workflow for preparing single cell suspension from solid tissue

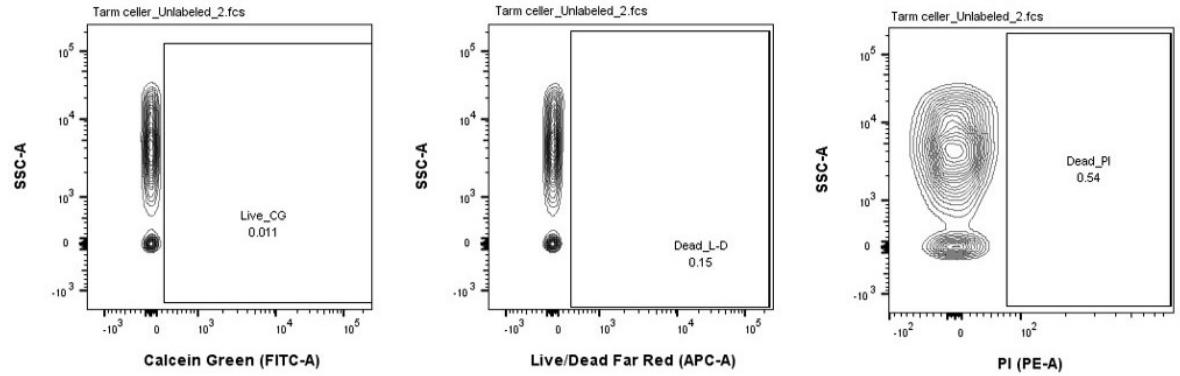
A general workflow for preparing single cell suspension from solid tissue by Reichard, A. and Asosingh, K [108] (Appendix Figure 10) was used as a basis for development and optimization in terms of establishing single cell suspension from fresh CRC tissue.

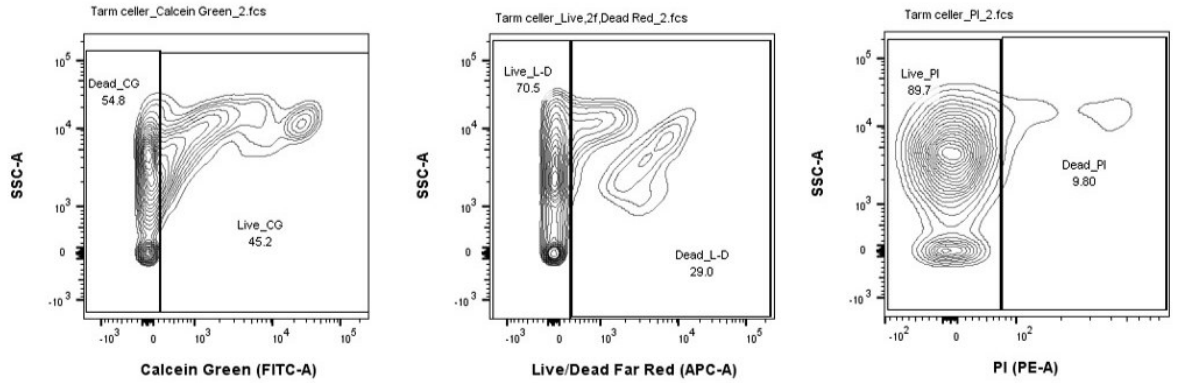


Appendix Figure 10: A general workflow for preparing single cell suspension from solid tissue. This workflow was presented by Reichard, A. and Asosingh, K. [108].

E Flow cytometry validation of single cell suspension protocol

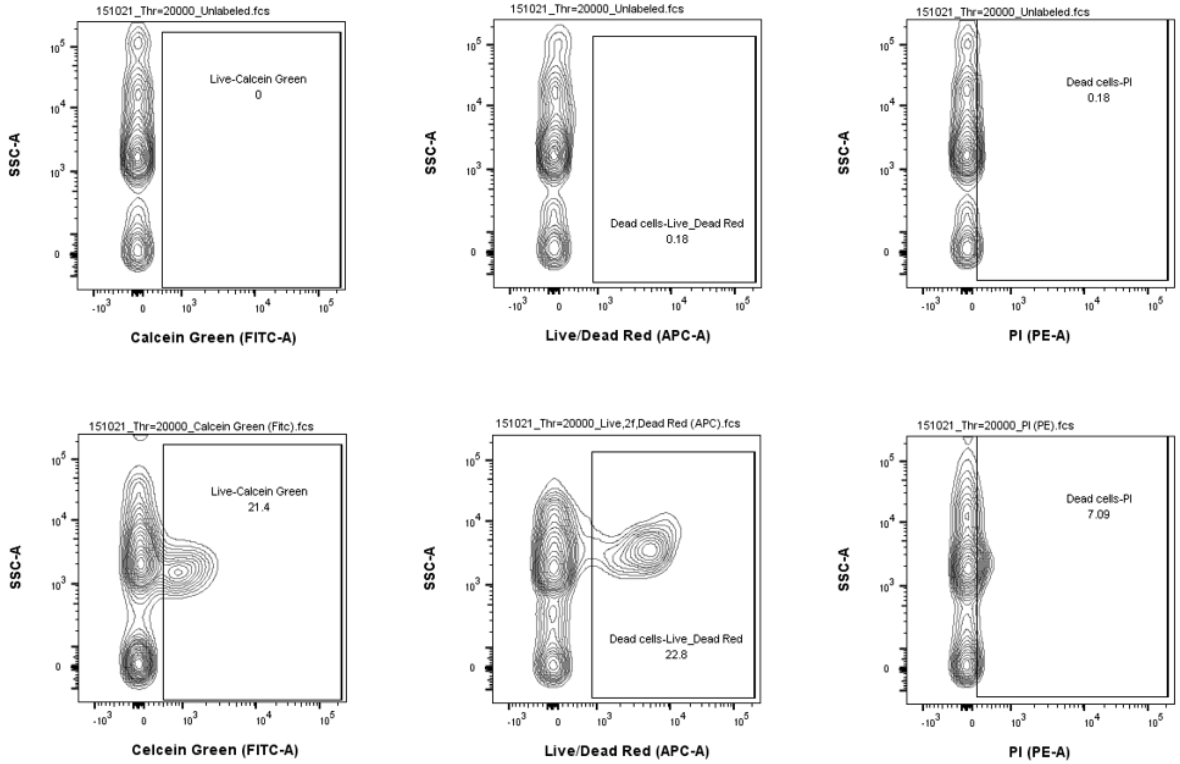
E.1 Results 31 August 2021





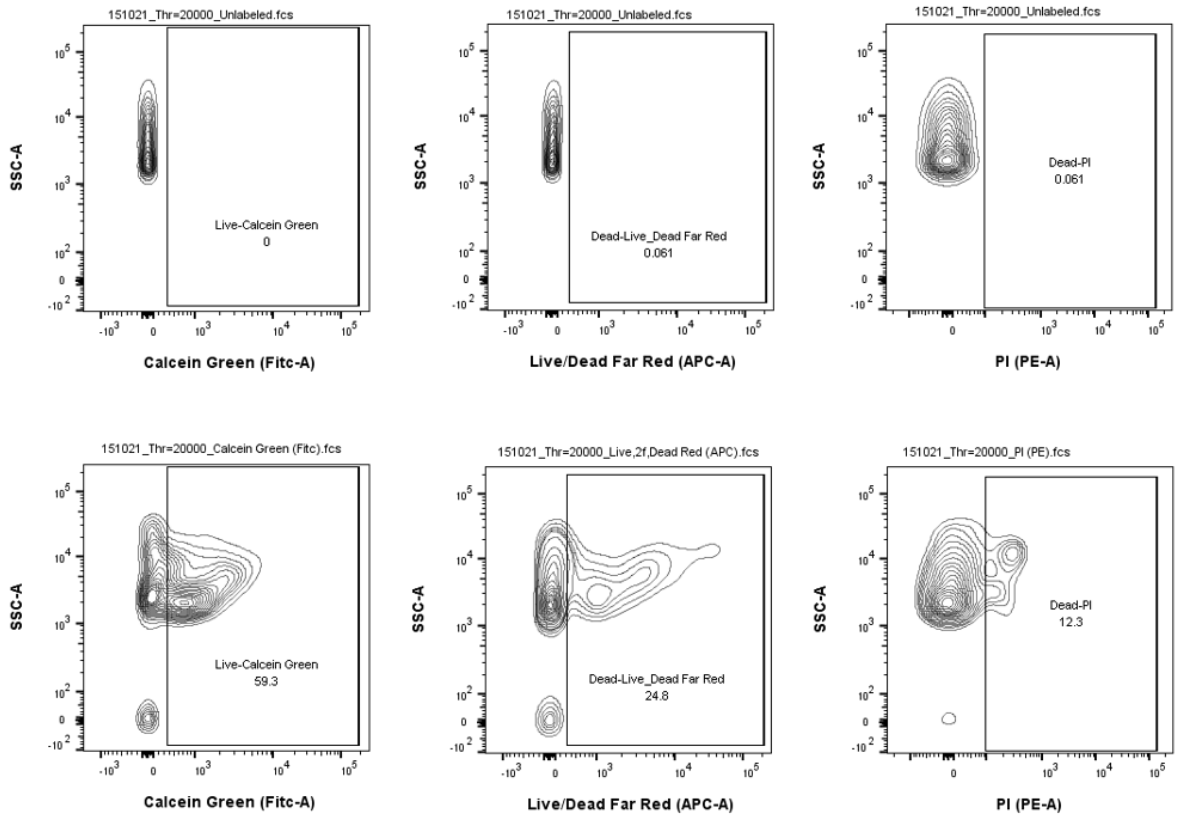
Fluorochrome	**Marks**	**Live cells [%]**	**Dead cells [%]**
Calcein green	Live cells	45.2	54.8
Live/dead far red	Dead cells	70.5	29.0
PI	Dead cells	89.7	9.8

E.2 Results 15 October 2021



Fluorochrome	**Marks**	**Live cells [%]**	**Dead cells [%]**
Calcein green	Live cells	21.4	78.6
Live/dead far red	Dead cells	77.2	22.8
PI	Dead cells	92.9	7.1

E.3 Results 20 October 2021



Fluorochrome	**Marks**	**Live cells [%]**	**Dead cells [%]**
Calcein green	Live cells	59.3	40.7
Live/dead far red	Dead cells	75.2	24.8
PI	Dead cells	87.7	12.3

F R scripts for processing and analyzing the single cell RNA sequencing data

F.1 Creating cluster graphs for each sample and manually annotating the clusters (shown for sample 554)

```
# Making packages available
library(dplyr) # For data manipulation
library(Seurat) # Single cell genomics toolkit
library(scater) # Single cell genomics toolkit
library(patchwork) # For combining complex data plots
library(ggplot2) # For graph visualization
library(RColorBrewer) # For graph color theme

# Creating clusters
filtered_554 <- Read10X_h5(file = "554_filtered_feature_bc_matrix.h5")
# Data is read and UMI count matrix is returned
filtered_554.obj <- CreateSeuratObject(counts = filtered_554,project =
"554",min.cells = 3,min.features = 200,names.field=1) # Creating a
Seurat object, a container for both data and analysis results
filtered_554.obj <- NormalizeData(filtered_554.obj,normalization.method
= "LogNormalize",scale.factor = 10000) # Normalizing data to account
for differences in sequencing depth
filtered_554.obj <- FindVariableFeatures(filtered_554.obj,
selection.method = "vst", nfeatures = 2000) # Identifying the 2000 most
variable genes (high and low expressed)
filtered_554.obj <- ScaleData(filtered_554.obj) # Scaling data to
reflect both high and low expressed genes
filtered_554.obj <- RunPCA(filtered_554.obj) # Performing linear
dimensional reduction by PCA
filtered_554.obj <- FindNeighbors(filtered_554.obj, dims = 1:10) #
Using WNN algorithm to creating a clustering graph
filtered_554.obj <- FindClusters(filtered_554.obj, resolution = 0.5) #
Determining the clusters for various resolutions (1 = many clusters,
0.5 = fewer clusters)
filtered_554.obj <- RunUMAP(filtered_554.obj, dims = 1:10) # Performing
non-linear dimension reduction by UMAP, enabling clustering
visualization

# Adding a color theme to cluster visualization
nb.cols <- 13 # Number of wanted colors in palette (equals number of
clusters)
mycolors <- colorRampPalette(brewer.pal(9, "Set1"))(nb.cols) # Classic
color palette Set1 with 9 colors, which is expanded

#Cluster visualization and aesthetic
DimPlot(filtered_554.obj, label = T, label.size = 3.5) +
  ggtitle('Sample 554') +
  scale_color_manual(values=mycolors) +
  labs(y = "UMAP2", x = "UMAP1", color = "Cluster") +
  theme(legend.text = element_text(size = 8), legend.title =
element_text(size = 8))

# Identifying cell-type-specific marker genes for every cluster
compared to all remaining cells
filtered_554.obj.markers <- FindAllMarkers(filtered_554.obj, only.pos =
TRUE, min.pct = 0.25, logfc.threshold = 0.25)
```

```
filtered_554.obj.markers <- filtered_554.obj.markers %>%
group_by(cluster) %>% slice_max(n = 10, order_by = avg_log2FC)
filtered_554.obj.markers <- as.data.frame(filtered_554.obj.markers)

# Saving the list of cell-type-specific marker genes (NB! Rename for
new lists)
write.table(filtered_554.obj.markers,
file="filtered_554.obj.markers.res0.5_n10.csv", quote = F, col.names =
NA)

# Naming cluster cell types based on cell-type-specific marker genes
filtered_554.obj.labels<- c("CD4_EM_T",
                            "Intestinal epithelial",
                            "MT",
                            "CD4_P_T",
                            "Intestinal epithelial",
                            "Fibroblast",
                            "MT",
                            "CD8_EM_T",
                            "Vascular endothelial",
                            "Myeloid",
                            "Smooth muscle",
                            "Plasma B-cell",
                            "Intestinal epithelial"
)

# Renaming clusters to new annotated name
filtered_554.obj.new <- filtered_554.obj # Making copy of object to not
overwrite old file
names(filtered_554.obj.labels) <- levels(filtered_554.obj.new)
filtered_554.obj.new <- RenameIdents(filtered_554.obj.new,
filtered_554.obj.labels) # Changing cluster names to manually defined
names
filtered_554.obj.new@meta.data$Annotated <-
Idents(filtered_554.obj.new)

# Adding a color theme to cluster visualization
nb.cols <- 13 # Number of wanted colors in palette (equals number of
clusters)
mycolors <- colorRampPalette(brewer.pal(9, "Set1"))(nb.cols) # Classic
color palette Set1 with 9 colors, which is expanded

# Cluster visualization and aesthetic
DimPlot(filtered_554.obj.new,label = T, label.size = 3, repel = T) +
  ggtitle('Sample 554') +
  scale_color_manual(values=mycolors) +
  labs(y = "UMAP2", x = "UMAP1", color = "Cell type") +
  theme(legend.text = element_text(size = 8), legend.title =
element_text(size = 10))

# Saving Seurat objects including annotated cluster names
library(SeuratDisk)
library(SeuratData)
SaveH5Seurat(filtered_554.obj.new, filename = "annotated_554_obj")

# Creating table with cell numbers in each cluster
filtered_554.obj.new_table <-
table(filtered_554.obj.new@meta.data$Annotated)
```

```
write.table(filtered_554.obj.new_table,
file="cellnumber_in_cluster_554.csv", quote = T, col.names = T,
row.names = F)
```

F.2 Integration of sample data and joint analysis of all samples

```
# Making packages available
library(dplyr) # For data manipulation
library(Seurat) # Single cell genomics toolkit
library(scater) # Single cell genomics toolkit
library(patchwork) # For combining complex data plots
library(ggplot2) # For graph visualization
library(RColorBrewer) # For graph color theme

# Finding saved annotated cell data in the sample's Seurat objects
filtered_554.obj.new@meta.data$Annotated <-
Idents(filtered_554.obj.new)
filtered_556.obj.new@meta.data$Annotated <-
Idents(filtered_556.obj.new)
filtered_559.obj.new@meta.data$Annotated <-
Idents(filtered_559.obj.new)
filtered_569.obj.new@meta.data$Annotated <-
Idents(filtered_569.obj.new)

# Defining the Seurat objects to integrate
objects.to.integrate <- (list(filtered_554.obj.new,
                              filtered_556.obj.new,
                              filtered_559.obj.new,
                              filtered_569.obj.new))

# Performing integration
anchors <- FindIntegrationAnchors(object.list = objects.to.integrate,
anchor.features = 2000) # Identifying anchors (cells from each data set
within the same clusters)
integrated.samples <- IntegrateData(anchorset = anchors) # Creating an
integrated data assay based on the anchors

# If running into an error about not enough memory, increase the memory
limit
memory.limit()
memory.limit(24000)

# Sets active assay from RNA to integrated, to use the integrated data
in downstream analysis
DefaultAssay(integrated.samples) <- "integrated"

# Performing integrated analysis and creating new clusters for the
integrated samples
integrated.samples <- NormalizeData(integrated.samples,
normalization.method = "LogNormalize", scale.factor = 10000) #
Normalizing data to account for differences in sequencing depth
integrated.samples <- ScaleData(integrated.samples) # Scaling data to
reflect both high and low expressed genes
integrated.samples <- RunPCA(integrated.samples) # Performing linear
dimensional reduction by PCA
integrated.samples <- RunUMAP(integrated.samples, dims = 1:10) #
Performing non-linear dimension reduction by UMAP, enabling clustering
visualization
```

```
# Adding a color theme to cluster visualization
nb.cols <- 30 # Number of wanted colors in palette (equals number of
clusters)
mycolors <- colorRampPalette(brewer.pal(9, "Set1"))(nb.cols) # Classic
color palette Set1 with 9 colors, which is expanded

# Cluster visualization and aesthetic (had to minimize the legend in
order to get enough space for labels not to overlap)
DimPlot(integrated.samples, label = T, label.size = 3.5, repel = T) +
  ggtitle('Integrated samples') +
  scale_color_manual(values=mycolors) +
  labs(y = "UMAP2", x = "UMAP1", color = "Cell type") +
  theme(legend.text = element_text(size = 8), legend.title =
element_text(size = 8))

# Identifying cell-type-specific marker genes for every cluster
compared to all remaining cells
integrated.samples.markers <- FindAllMarkers(integrated.samples,
only.pos = TRUE, min.pct = 0.25, logfc.threshold = 0.25)
integrated.samples.markers <- integrated.samples.markers %>%
group_by(cluster) %>% slice_max(n = 10, order_by = avg_log2FC)
integrated.samples.markers <- as.data.frame(integrated.samples.markers)

# Saving the list of cell-type-specific marker genes (NB! Rename for
new lists)
write.table(integrated.samples.markers,
file="integrated.samples.markers.n10.csv", col.names = NA)

# Creating heatmap displaying the top 5 cell-type-specific marker genes
for each cluster
top5 <- integrated.samples.markers %>% group_by(cluster) %>% top_n(n =
5, wt = avg_log2FC)
DoHeatmap(integrated.samples, features = top5$gene, size = 3) +
  theme(text = element_text(size = 8), legend.key.size = unit(0.35,
'cm'),) +
  scale_fill_viridis_c()

### Creating rougher plot ###

# Manually annotating integrated cluster graph for a rougher plot
integrated.samples.labels<- c("Immune",
                      "Immune",
                      "Immune",
                      "Immune",
                      "Stromal",
                      "Intestinal epithelial",
                      "Intestinal epithelial",
                      "Intestinal epithelial",
                      "Immune",
                      "MT",
                      "Immune",
                      "Immune",
                      "Stromal",
                      "Immune",
                      "Unknown",
                      "Endothelial",
                      "Stromal"
```

```
)

# Renaming clusters to new annotated name
integrated.samples.new <- integrated.samples # Making copy of object to
not overwrite old file
names(integrated.samples.labels) <- levels(integrated.samples.new)
integrated.samples.new <- RenameIdents(integrated.samples.new,
integrated.samples.labels) # Changing cluster names to manually defined
names
integrated.samples.new@meta.data$Annotated <-
Idents(integrated.samples.new)

# Adding a color theme to cluster visualization
nb.cols <- 18 # Number of wanted colors in palette (equals number of
clusters)
mycolors <- colorRampPalette(brewer.pal(9, "Set1"))(nb.cols) # Classic
color palette Set1 with 9 colors, which is expanded

# Cluster visualization and aesthetic
DimPlot(integrated.samples.new,label = T, label.size = 3.5, repel = T)
+
  ggtitle('Rough classification of integrated samples') +
  scale_color_manual(values=mycolors) +
  labs(y = "UMAP2", x = "UMAP1", color = "Cell type") +
  theme(legend.text = element_text(size = 9), legend.title =
element_text(size = 10))
```

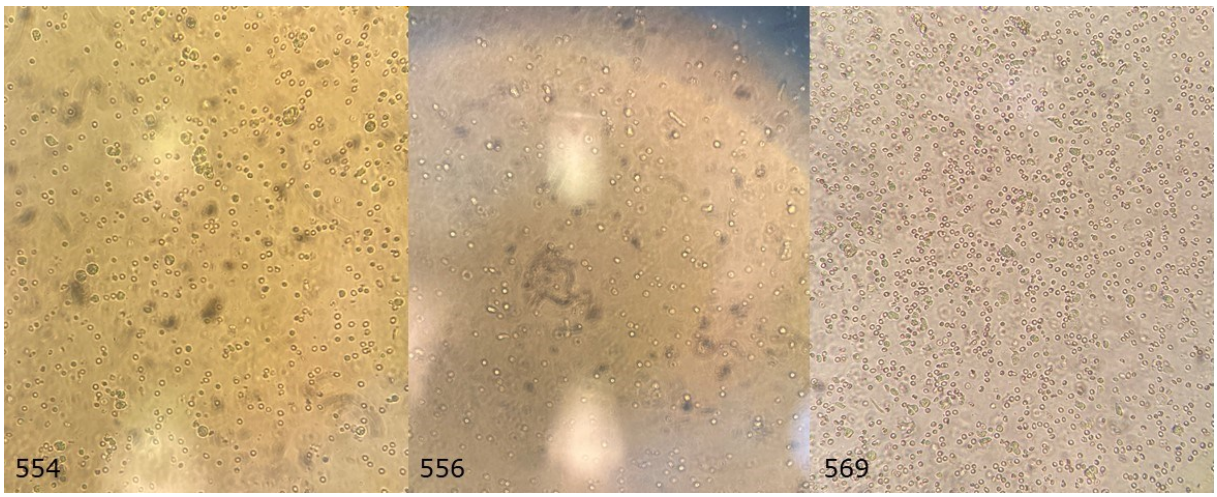
F.3 Subclustering the integrated data (shown for stromal subset)

```
# Making packages available
library(dplyr) # For data manipulation
library(Seurat) # Single cell genomics toolkit
library(scater) # Single cell genomics toolkit
library(patchwork) # For combining complex data plots
library(ggplot2) # For graph visualization
library(RColorBrewer) # For graph color theme

# Creating a subset variable from the integrated clustering graph
integrated.samples.stromal <- subset (x = integrated.samples, idents =
c("Fibroblast", "Smooth muscle", "Vascular smooth muscle"))

# If running into an error about not enough memory, increase the memory
limit
memory.limit()
memory.limit(24000)

# Creating clusters
integrated.samples.stromal <- NormalizeData(integrated.samples.stromal,
normalization.method = "LogNormalize", scale.factor = 10000) #
Normalizing data to account for differences in sequencing depth
integrated.samples.stromal <-
FindVariableFeatures(integrated.samples.stromal, selection.method =
"vst", nfeatures = 2000) # Identifying the 2000 most variable genes
(high and low expressed)
integrated.samples.stromal <- ScaleData(integrated.samples.stromal) #
Scaling data to reflect both high and low expressed genes
```

```
integrated.samples.stromal <- RunPCA(integrated.samples.stromal,
features = VariableFeatures(object = integrated.samples.stromal)) #
Performing linear dimensional reduction by PCA
integrated.samples.stromal <- FindNeighbors(integrated.samples.stromal,
dims = 1:10) # Using WNN algorithm to creating a clustering graph
integrated.samples.stromal <- FindClusters(integrated.samples.stromal,
resolution = 0.2) # Determining the clusters for various resolutions (1
= many clusters, 0.5 = fewer clusters)
integrated.samples.stromal <- RunUMAP(integrated.samples.stromal, dims
= 1:10) # Performing non-linear dimension reduction by UMAP, enabling
clustering visualization
DimPlot(integrated.samples.stromal, label = T)

# Identifying cell-type-specific marker genes for every cluster
compared to all remaining cells
integrated.samples.stromal.markers <-
FindAllMarkers(integrated.samples.stromal, only.pos = TRUE, min.pct =
0.25, logfc.threshold = 0.25)
integrated.samples.stromal.markers <-
integrated.samples.stromal.markers %>% group_by(cluster) %>%
slice_max(n = 10, order_by = avg_log2FC)
integrated.samples.stromal.markers <-
as.data.frame(integrated.samples.stromal.markers)

# Saving the list of cell-type-specific marker genes (NB! Rename for
new lists)
write.table(integrated.samples.stromal.markers,
file="integrated.samples.markers.stromal.res0.2_n10.csv", quote = F,
col.names = NA)

# Naming cluster cell types based on cell-type-specific marker genes
integrated.samples.stromal.labels<- c("Pericyte",
                              "CAF",
                              "Plasma B-cell",
                              "CTF",
                              "Myofibroblast",
                              "LPF"
)

# Renaming clusters to new annotated name
integrated.samples.stromal.new <- integrated.samples.stromal # Making
copy of object to not overwrite old file
names(integrated.samples.stromal.labels) <-
levels(integrated.samples.stromal.new)
integrated.samples.stromal.new <-
RenameIdents(integrated.samples.stromal.new,
integrated.samples.stromal.labels) # Changing cluster names to manually
defined names
integrated.samples.stromal.new@meta.data$Annotated <-
Idents(integrated.samples.stromal.new)

# Adding a color theme to cluster visualization
nb.cols <- 6 # Number of wanted colors in palette (equals number of
clusters)
mycolors <- colorRampPalette(brewer.pal(9, "Set1"))(nb.cols) # Classic
color palette Set1 with 9 colors, which is expanded
```

```
# Cluster visualization and aesthetic (had to minimize the legend in
order to get enough space for labels not to overlap)
DimPlot(integrated.samples.stromal.new, label = T, label.size = 3,
repel = T) +
  ggtitle('Stromal cell types') +
  scale_color_manual(values=mycolors) +
  labs(y = "UMAP2", x = "UMAP1", color = "Cell type") +
  theme(legend.text = element_text(size = 8), legend.title =
element_text(size = 10))

# Creating heatmap displaying the top 5 cell-type-specific marker genes
for each cluster
top5 <- integrated.samples.stromal.markers %>% group_by(cluster) %>%
top_n(n = 5, wt = avg_log2FC)
DoHeatmap(integrated.samples.stromal.new, features = top5$gene, size =
3) +
  theme(text = element_text(size = 8), legend.key.size = unit(0.35,
'cm'),) +
  scale_fill_viridis_c()

### Creating tables ###

# Creating table with cell numbers in each cluster
integrated.samples.stromal_table <-
table(integrated.samples.stromal.new@meta.data$Annotated)
write.table(integrated.samples.stromal_table,
file="integrated.samples.cellnumber.stromal.csv", quote = T, col.names
= T, row.names = F)

# Creating table with patient-specific cell numbers in each cluster
integrated.samples.stromal.patient_table <-
table(integrated.samples.stromal.new@meta.data$orig.ident,
integrated.samples.stromal.new@meta.data$Annotated)
write.table(integrated.samples.stromal.patient_table,
file="integrated.samples.cellnumber.patient.stromal.csv", quote = T,
col.names = T, row.names = F)
integrated.samples.stromal.patient_table

### Saving file ###

# Saving Seurat objects including annotated cluster names
library(SeuratDisk)
library(SeuratData)
SaveH5Seurat(integrated.samples.stromal.new, filename =
"subset_stromal")
```

G Supplementary section: Validation of protocol for establishing single cell suspension from fresh colorectal cancer tissue

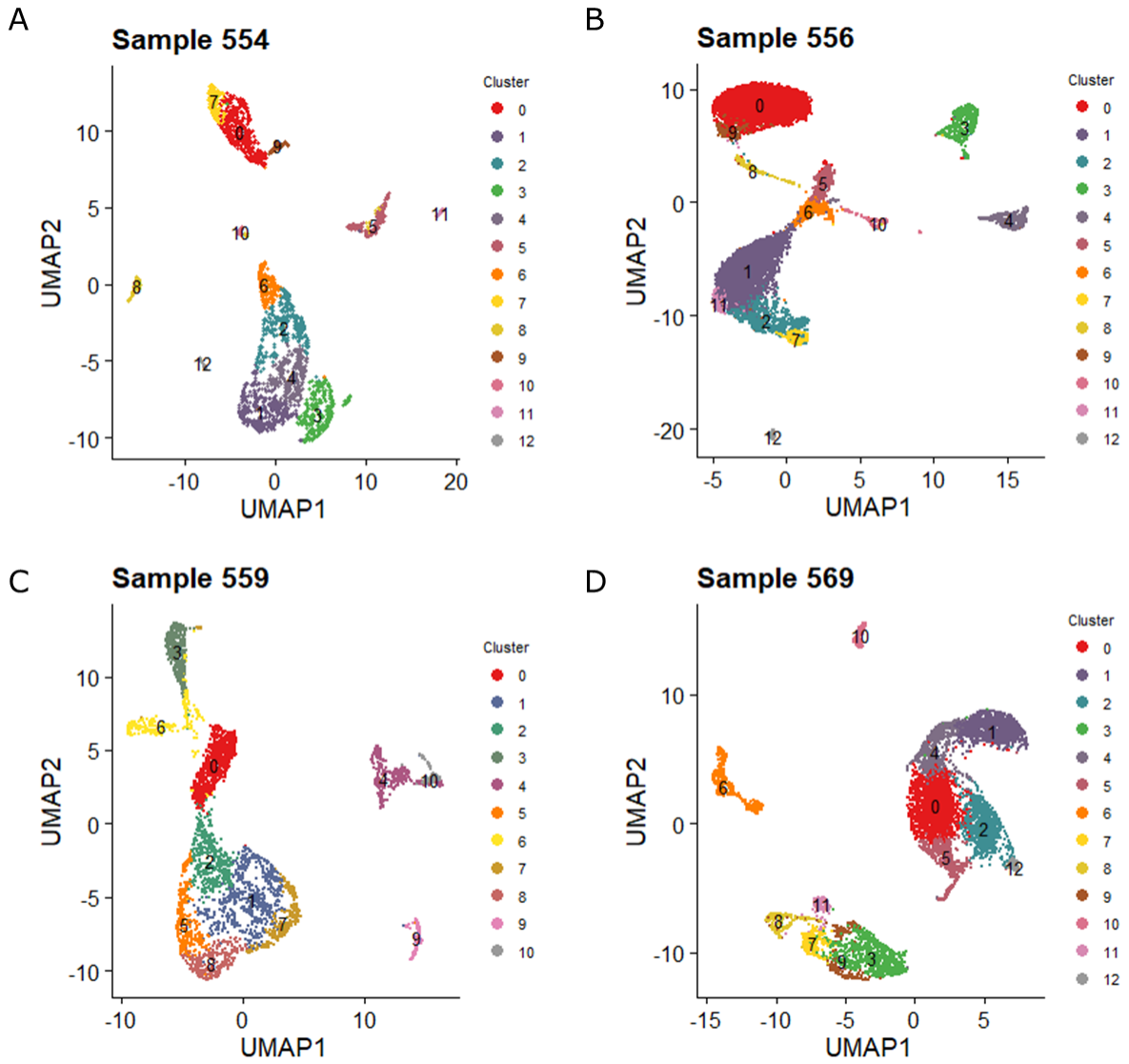


Supplementary Figure 1: Microscopy results of single cell suspensions established from the finalized protocol. The figure shows sample 554 (left), 556 (middle), and 569 (right). The microscope used was AE30 Binocular Inverted Microscope (Motic®).

Supplementary Table 1: Mean flow cytometry results of live and dead cells in single cell suspensions established from the finalized protocol. The fluorochromes calcein green, live/dead far red, and propidium iodide (PI) was used for validating cell viability, which was finally estimated to be 68.8%.

Fluorochrome	**Marks**	**Live cells [%]**	**Dead cells [%]**
Calcein green	Live cells	42.0	58.0
Live/dead far red	Dead cells	74.3	25.7
PI	Dead cells	90.1	9.9
Total mean value	-	**68.8**	**31.2**

H Supplementary section: Identified expressed genes and cell type composition in colorectal cancer tumor



Supplementary Figure 2: Non-annotated cluster graphs of each sample. (A) Sample 554 was divided into 13 clusters. (B) Sample 556 was divided into 13 clusters. (C) Sample 559 was divided into 11 clusters. (D) Sample 569 was divided into 13 clusters.

Supplementary Table 2: Top 10 expressed cell-type-specific marker genes identified for each cluster in sample 554.

#	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	5.62418482363643e-181	3.08522615872169	0.595	0.063	1.23928912588829e-176	0	LTB
2	1.49710953842519e-122	2.78940740758436	0.438	0.055	3.29888086791991e-118	0	IL7R
3	1.77221271115954e-13	2.66730023957074	0.508	0.523	3.90507070904005e-09	0	HSPA1A
4	2.12641402387697e-230	2.43466735966304	0.785	0.092	4.6855533016129e-226	0	CD2
5	1.51560395023412e-232	2.3988466921928	0.861	0.123	3.33963330434089e-228	0	PTPRC
6	1.2679082991677e-162	2.296683083677	0.602	0.083	2.79383593721603e-158	0	IKZF1
7	2.38299721469041e-156	2.24951578990374	0.658	0.105	5.25093436257031e-152	0	CD52
8	1.29171098721985e-199	2.22332493193862	0.736	0.09	2.84628516033894e-195	0	CD3D
9	2.52581092171078e-15	2.21750661298404	0.553	0.601	5.5656243659897e-11	0	JUN
10	9.64941535333054e-162	2.20829107770014	0.622	0.093	2.12624867310638e-157	0	GIMAP7
11	3.0956096411117e-229	3.12318808837633	0.935	0.197	6.82117584418963e-225	1	CEACAM7
12	3.28657458444466e-147	3.05966880721418	0.957	0.47	7.24196709682381e-143	1	FABP1
13	1.5526763337892e-196	2.75443065542098	0.997	0.392	3.42132230150451e-192	1	CEACAM5
14	2.23751893511142e-245	2.5346074901471	0.927	0.166	4.93037297351802e-241	1	KRT20
15	1.77096937334872e-176	2.47776701092213	0.992	0.419	3.9023310141739e-172	1	FXYD3
16	5.30345682663316e-111	2.3964107120927	0.846	0.317	1.16861671174862e-106	1	CKB
17	9.55963833513902e-154	2.34488506970412	0.992	0.511	2.10646630714788e-149	1	TFF1
18	3.45561924216123e-215	2.24607475326304	0.981	0.282	7.61445700010227e-211	1	TSPAN1
19	6.73321676863832e-156	2.2317740419196	1	0.703	1.48366431496945e-151	1	LGALS3
20	1.81219932990149e-153	2.20415789995523	0.978	0.432	3.99318122343793e-149	1	PHGR1
21	1.97462591904534e-79	1.30530437956179	0.859	0.421	4.3510882126164e-75	2	RNF43
22	2.4764126722612e-87	1.18277379804465	0.997	0.978	5.45677532332756e-83	2	MT-ND3
23	9.2321544863177e-85	1.16935811627262	1	0.992	2.0343052410601e-80	2	MT-ND1
24	1.8501865812552e-88	1.15536487455611	1	0.996	4.07688613179582e-84	2	MT-CO3
25	4.29066584986731e-86	1.11839347780707	0.997	0.999	9.45448220018263e-82	2	MT-CYB
26	1.75871750115141e-80	1.11312899771949	0.991	0.903	3.87533401378713e-76	2	MTRNR2L10
27	9.81694884920004e-83	1.09648203805217	0.997	0.998	2.16316467892123e-78	2	MT-CO2
28	9.99665366789205e-83	1.07905904633876	0.997	0.998	2.20276263572001e-78	2	MT-ATP6
29	5.33117066088866e-71	1.07604927227496	1	0.984	1.17472345512682e-66	2	MT-ND2
30	2.49645309455957e-64	1.07383230247887	0.89	0.464	5.500934393862e-60	2	AC103702.2
31	4.57146092515805e-238	2.55353119308453	0.807	0.076	1.00732141485858e-233	3	CENPF
32	6.49037768227184e-241	2.16758008491908	0.844	0.085	1.4301547222886e-236	3	MKI67
33	2.21051168852257e-302	1.5781163332879	0.681	0.014	4.87086250565948e-298	3	ASPM
34	6.6605712258658e-263	1.55026500853443	0.685	0.028	1.46765686961953e-258	3	TOP2A
35	8.99228596309485e-190	1.48366717318411	0.824	0.136	1.98145021196795e-185	3	CENPW
36	4.44078830379983e-106	1.47818861273248	0.824	0.227	9.78527702742292e-102	3	HMGB2
37	5.52895146748025e-143	1.40220578400333	0.705	0.107	1.21830445585927e-138	3	PTTG1
38	2.77386413674705e-91	1.37652576486215	0.949	0.509	6.11220962532212e-87	3	H2AFZ
39	1.30326614317062e-143	1.3624829743715	0.773	0.155	2.87174694647645e-139	3	CKS2
40	1.53277825517425e-95	1.33439158612204	0.986	0.875	3.37747688527646e-91	3	HMGB1
41	2.28061028057957e-100	1.51849131859421	0.83	0.26	5.02532475325708e-96	4	LEFTY1
42	1.37141561374212e-95	1.47362394018694	0.989	0.431	3.02191430488075e-91	4	GPX2
43	1.57373496839223e-70	1.42689565697899	1	0.631	3.46772500285227e-66	4	TFF3
44	1.65041623999954e-73	1.38760050675361	0.967	0.397	3.636692184839e-69	4	LCN2
45	2.40559511896971e-74	1.3705757252153	1	0.507	5.30072884464975e-70	4	PIGR
46	4.58098410564405e-71	1.19009614915994	0.985	0.439	1.00941984767867e-66	4	TSPAN8
47	2.51852747308298e-77	1.18808274327458	0.985	0.453	5.54957528693836e-73	4	FAM3D
48	3.89462207704022e-63	1.10702085880336	0.882	0.413	8.58179974675813e-59	4	SLC12A2
49	1.33963877473753e-53	1.07716636467253	0.934	0.614	2.95189404013414e-49	4	PRDX5
50	4.71845204660143e-63	1.07551155272351	0.989	0.532	1.03971090846863e-58	4	TFF1
51	5.85050632492551e-250	6.56950753995158	0.948	0.107	1.28915906869734e-245	5	CXCL14
52	0	5.53517335230791	0.99	0.045	0	5	COL1A2
53	0	5.52379628597157	1	0.041	0	5	COL3A1
54	0	5.41313732296103	0.995	0.054	0	5	COL1A1
55	0	5.09225583995372	0.943	0.02	0	5	LUM
56	0	5.00589037735879	0.967	0.019	0	5	DCN

57	0	4.09259412759228	0.986	0.023	0	5	COL6A3
58	5.00451555195467e-275	3.9412327375575	1	0.095	1.10274500187321e-270	5	CALD1
59	1.97419814123932e-279	3.91729136792391	0.714	0.023	4.35014560422084e-275	5	POSTN
60	2.31643122838028e-181	3.81287414650529	0.995	0.251	5.10425621173594e-177	5	RARRES2
61	4.49645625663014e-83	1.8350924369051	0.988	0.997	9.90794136148452e-79	6	MT-CO3
62	2.24082712589458e-76	1.79824142563134	0.988	0.98	4.9376625719087e-72	6	MT-ND3
63	7.55339166163727e-84	1.772617325482	1	0.997	1.66438985264177e-79	6	MT-ATP6
64	2.40410087727122e-80	1.76654114233971	1	0.993	5.29743628306714e-76	6	MT-ND1
65	2.98909528064277e-83	1.75797127166845	1	0.998	6.58647145089635e-79	6	MT-CO2
66	1.10597201099755e-83	1.74664005491013	1	0.999	2.4370093262331e-79	6	MT-CYB
67	3.20474004292042e-70	1.73251088337812	0.994	0.986	7.06164468457514e-66	6	MT-ND2
68	4.68633342235669e-69	1.68587871158287	1	0.996	1.0326335696163e-64	6	MTRNR2L12
69	7.2344400684334e-71	1.65682605811164	0.988	0.986	1.5941088690793e-66	6	MTRNR2L8
70	4.94064336327292e-65	1.62173432260377	0.983	0.913	1.08867076509719e-60	6	MTRNR2L1
71	3.44218740298513e-183	3.68714793793266	0.981	0.133	7.58485994247774e-179	7	CCL5
72	3.28480638286939e-191	3.46852607356743	0.938	0.104	7.23807086465271e-187	7	GZMA
73	3.46572817643686e-148	3.38413085010908	0.596	0.04	7.63673203677863e-144	7	CCL4
74	1.24781302734336e-190	3.0817255007534	0.826	0.067	2.7495560057511e-186	7	NKG7
75	7.50201739357263e-107	3.0318257245223	0.689	0.101	1.65306953267373e-102	7	TRBC1
76	1.99434263678509e-95	2.88313853692847	0.547	0.066	4.39453400015594e-91	7	KLRB1
77	8.64647295258154e-74	2.7844539930634	0.261	0.013	1.90525031510134e-69	7	CCL3
78	4.74660297398963e-155	2.70424813888789	0.919	0.13	1.04591396531862e-150	7	HCST
79	3.77838957929055e-111	2.63390737227115	0.54	0.05	8.32568143796672e-107	7	GZMB
80	1.12651791628644e-136	2.5465055022063	0.938	0.154	2.48228222853718e-132	7	CD3D
81	4.62852802903209e-168	4.19171470161794	0.769	0.05	1.01989615119722e-163	8	SPARCL1
82	5.47276988777998e-172	3.90210064994315	0.817	0.054	1.20592484477232e-167	8	MGP
83	4.48482502997953e-72	3.81102954775149	0.587	0.079	9.88231195355989e-68	8	IGFBP3
84	1.44299712785697e-98	3.691248800194	0.808	0.12	3.17964417123283e-94	8	COL4A1
85	4.71860727709083e-111	3.66654502541841	0.962	0.155	1.03974511350696e-106	8	IGFBP7
86	0	3.52752086314806	0.683	0.002	0	8	PLVAP
87	6.41148036302605e-232	3.39948944368686	0.74	0.024	1.41276969799279e-227	8	PECAM1
88	0	3.29773073242898	0.683	0.003	0	8	VWF
89	1.98512718339088e-83	3.25544433657933	0.702	0.111	4.3742277486018e-79	8	ENG
90	1.65205270255646e-69	3.10182002455143	0.76	0.176	3.64029813008316e-65	8	HSPG2
91	0.000303845751964635	3.70690514875596	0.421	0.319	1	9	LYZ
92	6.77293881195368e-15	3.61448550153361	0.719	0.458	1.49241706721399e-10	9	HLA-DRA
93	8.33335123660517e-08	3.51889209575511	0.877	0.821	0.00183625394498595	9	FTL
94	1.31692326389318e-37	3.24923589431748	0.561	0.089	2.90184041198862e-33	9	IFI30
95	1.94098069374108e-138	3.19423504351715	0.474	0.009	4.27695095865846e-134	9	AIF1
96	6.06742821435424e-71	3.11157004247905	0.561	0.041	1.33695780703296e-66	9	HLA-DQA1
97	2.06147250293563e-15	3.05514510665733	0.667	0.362	4.54245466021866e-11	9	HLA-DPB1
98	1.23014647348348e-09	2.88432677877864	0.614	0.438	2.71062775432085e-05	9	HLA-DPA1
99	2.96235404670137e-15	2.86328421733317	0.789	0.526	6.52754714190647e-11	9	HLA-DRB1
100	1.03893741532522e-121	2.83472683731534	0.404	0.007	2.28929859466913e-117	9	TYROBP
101	0	5.69101385759284	1	0.011	0	10	RGS5
102	2.52916153136969e-67	4.51566841063079	0.927	0.103	5.57300743437311e-63	10	ACTA2
103	8.93160778208433e-59	4.01997522620195	1	0.157	1.96807977478228e-54	10	CALD1
104	0	3.81083514027117	0.951	0.007	0	10	NDUFA4L2
105	6.96449333534905e-95	3.80278421032092	0.951	0.071	1.53462610644416e-90	10	MGP
106	1.01832640179329e-97	3.78139909106174	1	0.086	2.24388222635152e-93	10	COL18A1
107	6.01169806154899e-125	3.76793664787444	0.927	0.049	1.32467766786232e-120	10	CSRP2
108	8.87231840633864e-53	3.71562799479742	1	0.174	1.95501536083672e-48	10	IGFBP7
109	2.39661092192699e-56	3.63882916554077	0.902	0.115	5.28093216646613e-52	10	TAGLN
110	5.80035986471952e-62	3.635622267393	0.976	0.135	1.27810929619095e-57	10	COL4A1
111	1.07547057026724e-22	10.4236334641031	0.892	0.357	2.36979940158387e-18	11	IGKC
112	3.49920246603674e-57	9.47748061181547	0.946	0.135	7.71049263391196e-53	11	IGHA1
113	1.19860221978373e-284	8.71017607146267	0.892	0.01	2.64111999129345e-280	11	IGLC1
114	4.49454305230831e-100	8.47204270181634	1	0.075	9.90372561576136e-96	11	JCHAIN
115	6.75877729104437e-09	7.56826705705255	0.351	0.085	0.000148929657608163	11	IGLC3
116	8.33100976212644e-288	6.36557464354146	0.865	0.008	1.83573800108456e-283	11	IGHA2

117	1.07445808680706e-219	4.45083385286734	1	0.023	2.36756839427936e-215	11	MZB1
118	3.12659427484684e-40	3.6670017456938	1	0.264	6.88945048462501e-36	11	TXNDC5
119	3.76736473008 59e-152	3.16218799959687	0.324	0.001	8.30138818274428e-148	11	IGKV4-1
120	1.02472553372565e-21	2.95147064787648	0.946	0.713	2.25798271356447e-17	11	SSR4
121	0	5.03976428234347	1	0.004	0	12	SH2D6
122	5.07514222076455e-144	4.6137799672208	0.963	0.029	1.11830758834547e-139	12	LRMP
123	1.17614457627883e-14	3.31136415980819	0.889	0.454	2.5916345738304e-10	12	ANXA4
124	1.17617851691095e-27	3.2294880630769	0.741	0.108	2.59170936201328e-23	12	ALOX5AP
125	2.59558520449655e-20	3.13782145222516	0.815	0.228	5.71937199810816e-16	12	RASSF6
126	2.65031642575034e-199	3.12506498690171	0.741	0.008	5.83997224414087e-195	12	RGS13
127	6.73433105033305e-19	3.10805355807732	0.815	0.241	1.48390984694089e-14	12	PBXIP1
128	3.03083113174324e-23	2.94282203270863	0.963	0.332	6.67843639879623e-19	12	SPTLC2
129	6.00746375550399e-274	2.93521251711526	0.889	0.008	1.3237446385253e-269	12	BMX
130	1.85012062124955e-17	2.77658848293739	0.741	0.208	4.07674078892338e-13	12	AZGP1

Supplementary Table 3: Top 10 expressed cell-type-specific marker genes identified for each cluster in sample 556.

#	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	0	2.68060985034262	0.997	0.237	0	0	HLA-DRA
2	0	2.59086827466018	0.998	0.449	0	0	CD74
3	0	2.49439935282923	0.869	0.105	0	0	VPREB3
4	0	2.10928577028244	0.939	0.191	0	0	HLA-DPA1
5	0	2.08818832046786	0.853	0.097	0	0	BANK1
6	0	2.06581053616419	0.898	0.167	0	0	HLA-DPB1
7	0	2.03203870327499	0.874	0.148	0	0	CD79A
8	0	1.98825888078197	0.94	0.203	0	0	HLA-DRB1
9	0	1.95781131388115	0.863	0.113	0	0	MS4A1
10	0	1.85228138931255	0.864	0.159	0	0	HLA-DQB1
11	0	1.65735692089061	0.86	0.23	0	1	FYB1
12	0	1.45095376465311	0.763	0.189	0	1	IL7R
13	0	1.4058006557987	0.751	0.243	0	1	TRAC
14	0	1.39103771262654	0.76	0.318	0	1	TRBC2
15	0	1.38971819865136	0.831	0.207	0	1	CD3D
16	1.36219070931542e-274	1.37581520518612	0.47	0.136	3.14584322409304e-270	1	TRBC1
17	6.01072082297892e-276	1.32348524575478	0.52	0.184	1.38811586685875e-271	1	RGCC
18	0	1.30343781388006	0.604	0.156	0	1	GIMAP7
19	0	1.2877194640579	0.53	0.133	0	1	GIMAP4
20	0	1.28304378597285	0.618	0.151	0	1	CD3E
21	7.61913148713834e-219	1.92360103643632	0.586	0.145	1.75956222563973e-214	2	KLRB1
22	9.61008883470514e-196	1.79936718651185	0.823	0.347	2.21935391548681e-191	2	ID2
23	2.21152820414113e-254	1.69078699127808	0.817	0.265	5.10730323464353e-250	2	IL32
24	3.62132301649817e-186	1.6649772423685	0.714	0.262	8.36308337430087e-182	2	S100A4
25	1.12006077306469e-128	1.61409462903875	0.332	0.072	2.58666834931561e-124	2	CCL5
26	8.72058737143457e-117	1.46010478574207	0.728	0.336	2.0139324475591e-112	2	IL7R
27	3.13292286681646e-212	1.35964116843981	0.782	0.262	7.23517206862592e-208	2	CD2
28	2.05224889084261e-121	1.30594403734805	0.585	0.21	4.73946358851192e-117	2	TRBC1
29	7.88682022444861e-138	1.2626910061715	0.775	0.368	1.82138226263416e-133	2	CD3D
30	9.61538086475461e-84	1.24978943243848	0.467	0.178	2.22057605690643e-79	2	ANXA1
31	0	4.54033192782763	0.51	0.027	0	3	REG1A
32	0	4.42775764042412	0.781	0.017	0	3	PHGR1
33	0	4.14956721047083	0.702	0.009	0	3	FABP1
34	0	3.99923011833605	0.787	0.024	0	3	IFI27
35	0	3.91808084596877	0.702	0.007	0	3	ALDOB
36	0	3.7248508671425	0.721	0.006	0	3	PRAP1
37	0	3.68995183213192	0.763	0.011	0	3	PIGR
38	0	3.62151149717741	0.812	0.016	0	3	KRT8
39	0	3.54012248140992	0.692	0.014	0	3	SELENOP
40	0	3.53452817003021	0.783	0.02	0	3	LGALS4
41	4.18348121191798e-180	7.39950825690588	0.907	0.441	9.66133151080339e-176	4	IGHA1
42	1.33281695391491e-167	6.98799093513937	0.993	0.795	3.0780074733711e-163	4	IGKC
43	3.5048421267675e-45	6.94394837563628	0.516	0.214	8.09408240755687e-41	4	IGLC2
44	0	6.9272119375556	0.983	0.329	0	4	JCHAIN
45	3.00683379163263e-24	6.62089858435451	0.445	0.222	6.9439819583964e-20	4	IGLC3
46	1.45609997882967e-267	6.54109563370511	0.807	0.139	3.36271729110924e-263	4	IGLC1
47	0	5.11774248791477	0.667	0.085	0	4	IGHA2
48	2.14614891993071e-158	4.1812596037785	0.306	0.032	4.95631631568798e-154	4	IGHG3
49	0	4.15096522241146	0.956	0.075	0	4	MZB1
50	5.61874214765942e-250	3.79389750444488	0.954	0.487	1.29759231158047e-245	4	SSR4
51	1.54403809635103e-197	1.96200861391742	1	0.977	3.56580157971308e-193	5	MT-ND3
52	1.29444885642115e-202	1.93570257550762	1	0.992	2.98940018901901e-198	5	MT-ATP6
53	1.79010747216233e-185	1.91080134267029	0.998	0.98	4.13407419621169e-181	5	MTRNR2L12
54	1.79045624340099e-164	1.90837822920049	0.978	0.927	4.13487964851025e-160	5	MTRNR2L8
55	3.59118408610696e-198	1.88823522279749	1	0.99	8.29348052845542e-194	5	MT-CYB
56	2.87035156349822e-193	1.82879100772429	1	0.981	6.62878990074279e-189	5	MT-ND4

57	5.78311182853048e-197	1.82452596783699	1	0.99	1.33555184568083e-192	5	MT-CO3
58	9.94299184059862e-79	1.81041890249067	0.779	0.652	2.29623453566785e-74	5	MTRNR2L1
59	2.45682192856987e-184	1.78771559871616	0.998	0.975	5.67378456183925e-180	5	MT-ND1
60	4.432255204103e-143	1.74860937612307	0.961	0.875	1.02358501683555e-138	5	MT-ND5
61	7.77083432535668e-183	1.51119399871327	1	0.996	1.79459647909787e-178	6	MALAT1
62	2.74820022818789e-06	1.4586849439434	0.339	0.314	0.0634669360697711	6	CEMIP2
63	0.000612678308372534	1.39568990580552	0.354	0.386	1	6	TLE4
64	2.37875533745326e-05	1.36466733399857	0.437	0.503	0.549349757631456	6	NABP1
65	9.10526822951363e-46	1.33520497694787	0.809	0.812	2.10277064492388e-41	6	HSPH1
66	5.51460148484173e-08	1.30895471425237	0.488	0.587	0.00127354206690935	6	GLS
67	0.00942023026560085	1.29215172630996	0.403	0.487	1	6	SLC2A3
68	0.000432650032832348	1.27555065054038	0.256	0.232	1	6	AAK1
69	1.78008877718648e-12	1.26526105190562	0.576	0.672	4.11093702203446e-08	6	INTS6
70	5.61097595632371e-08	1.2443154131842	0.437	0.465	0.0012957987873534	6	RNF213
71	0	4.86078643481867	0.389	0.007	0	7	GNLY
72	0	4.36255245188511	0.921	0.073	0	7	CCL5
73	0	4.1811069266143	0.926	0.017	0	7	NKG7
74	0	4.05979386113184	0.861	0.02	0	7	GZMA
75	0	3.5897640457689	0.676	0.049	0	7	CCL4
76	0	3.1550319566724	0.477	0.017	0	7	IFNG
77	0	2.81845150060769	0.477	0.005	0	7	GZMB
78	0	2.65108028670218	0.481	0.003	0	7	GZMH
79	1.30885489396206e-290	2.52714061812947	0.417	0.017	3.02266949211599e-286	7	GZMK
80	0	2.32340401172507	0.505	0.005	0	7	KLRD1
81	4.1957607636471e-60	2.6159873561327	0.783	0.385	9.68968990756661e-56	8	HMGB2
82	5.35568151920067e-94	2.5558987281434	0.652	0.162	1.2368410900442e-89	8	STMN1
83	0	2.54888084848703	0.515	0.012	0	8	MKI67
84	1.41085629305477e-232	2.51579666762663	0.48	0.03	3.25823152318068e-228	8	CENPF
85	6.54014298476367e-218	2.33589715354948	0.682	0.07	1.51038062090132e-213	8	TCL1A
86	0	2.19953736694152	0.434	0.012	0	8	TOP2A
87	2.30175738689482e-115	2.13805514637708	0.727	0.166	5.3156785092949e-111	8	LRMP
88	0	2.13223416221325	0.621	0.026	0	8	MEF2B
89	5.78431430722158e-11	2.12602860825316	0.657	0.538	1.33582954610975e-06	8	HIST1H4C
90	5.52569607761345e-55	2.09399058114953	0.803	0.473	1.27610425216405e-50	8	HMGN2
91	2.18611214920907e-33	0.540282881206432	0.903	0.343	5.04860739738344e-29	9	VPREB3
92	9.4440170865305e-14	0.458896649216125	0.903	0.562	2.18100130596335e-09	9	ID3
93	2.67686592446079e-22	0.451213791941681	0.508	0.187	6.18195416594975e-18	9	LINC01781
94	1.38083180283826e-12	0.381146218313297	0.978	0.818	3.18889296547467e-08	9	RPL4
95	2.55189386961392e-11	0.380271173544415	0.908	0.629	5.89334370248638e-07	9	SNHG29
96	8.68030824889204e-16	0.377468747544686	0.984	0.64	2.00463038699913e-11	9	CD37
97	4.54110701022547e-13	0.369066586778567	0.984	0.761	1.04872325294147e-08	9	IER5
98	1.33452256255822e-12	0.366578900341903	1	0.924	3.08194640597195e-08	9	RPS23
99	9.79841695801189e-17	0.3604792284119	0.638	0.289	2.26284641228327e-12	9	KPNB1
100	3.11149406165771e-11	0.35766085959229	0.935	0.662	7.18568438599231e-07	9	KLF2
101	0	5.59591261911334	0.56	0.006	0	10	S100A9
102	0	5.49429514913931	0.547	0.014	0	10	CXCL8
103	0	5.07579096133995	0.68	0.012	0	10	LYZ
104	0	4.72581291608531	0.56	0.008	0	10	IL1B
105	0	4.63683849898389	0.433	0.003	0	10	S100A8
106	6.34356541628149e-96	4.53321905134967	0.553	0.084	1.46498299723605e-91	10	TIMP1
107	3.85415679009295e-239	4.0818754348571	0.673	0.049	8.90078969104066e-235	10	CST3
108	0	3.88885066307331	0.313	0.004	0	10	C1QA
109	0	3.77400186572498	0.293	0.002	0	10	C1QB
110	0	3.72545272727284	0.687	0.022	0	10	TYROBP
111	2.62480691558013e-13	1.10426941338143	0.483	0.236	6.06172909084075e-09	11	TRBC1
112	1.13035951000959e-23	1.09181677501155	0.797	0.45	2.61045225241614e-19	11	TRBC2
113	1.52225253652068e-22	1.03105332380286	0.762	0.395	3.51549000784085e-18	11	TRAC
114	5.17628922054574e-29	1.02734538435242	0.839	0.394	1.19541223259283e-24	11	CD3D
115	8.63248572408539e-20	1.01153788072051	0.615	0.29	1.99358625312028e-15	11	GIMAP7
116	3.85593406567688e-23	0.981101570593643	0.671	0.31	8.90489413127419e-19	11	CD3G

117	2.38655176122053e-21	0.956682488544321	0.629	0.291	5.5115026373627e-17	11	CD3E
118	7.09205509009402e-23	0.945353547798066	0.804	0.496	1.63783920250631e-18	11	LDHB
119	1.02851199813105e-16	0.932077699177703	0.608	0.299	2.37524560848385e-12	11	CD2
120	5.23871728453564e-12	0.894905912170585	0.531	0.285	1.20982936969066e-07	11	RGCC
121	0	5.92756679797814	0.933	0.011	0	12	IGFBP7
122	0	5.80471408932699	0.533	0.004	0	12	CXCL14
123	0	5.04346217151272	0.907	0.009	0	12	CALD1
124	2.35178740354583e-209	4.5964421129231	0.467	0.013	5.43121782974875e-205	12	CFD
125	0	4.57398782873211	0.773	0.001	0	12	SPARCL1
126	0	4.5401340155267	0.653	0.001	0	12	IGFBP5
127	0	4.52988390497384	0.56	0.002	0	12	DCN
128	0	4.298902674651	0.493	0.007	0	12	ADAMDEC1
129	3.0043533402879e-157	4.21871798775975	0.6	0.032	6.93825360406088e-153	12	TAGLN
130	0	4.19676169593537	0.827	0.009	0	12	A2M

Supplementary Table 4: Top 10 expressed cell-type-specific marker genes identified for each cluster in sample 559.

#	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	3.27279356045382e-291	6.60098555736931	0.661	0.109	7.17461804322686e-287	0	IGLC1
2	1.53903270170647e-96	6.39581876765836	0.64	0.329	3.37386748868093e-92	0	IGLC2
3	7.51319861984412e-92	6.15688730969689	0.574	0.264	1.64704340144223e-87	0	IGLC3
4	2.04308986321155e-248	5.72684279176455	0.907	0.524	4.47886159813237e-244	0	IGHA1
5	0	5.35800633755047	0.944	0.27	0	0	JCHAIN
6	4.96274622537544e-195	5.21818890807649	0.963	0.817	1.0879332275268e-190	0	IGKC
7	2.27924412271541e-61	5.19678596837726	0.348	0.124	4.99655896581673e-57	0	IGHM
8	6.80976654728699e-133	4.86435761016627	0.332	0.05	1.49283702249625e-128	0	IGHG1
9	3.64165383354862e-245	4.69430627231346	0.508	0.061	7.98323353390529e-241	0	IGHA2
10	1.02690615503764e-127	4.53696740747397	0.296	0.038	2.25118367307352e-123	0	IGHGP
11	8.45647133587756e-81	2.34629924612855	0.413	0.126	1.85382764625108e-76	1	SPINK4
12	3.47890859909101e-89	2.03875781504901	0.429	0.127	7.62646343092731e-85	1	MUC2
13	1.27861013130613e-45	1.72416614520468	0.52	0.272	2.80296912984929e-41	1	TFF3
14	6.85216247267218e-48	1.52403536739896	0.293	0.097	1.5021310572592e-43	1	CENPF
15	2.38516016571438e-148	1.50785944652827	0.668	0.204	5.22874811527907e-144	1	MUC5B
16	1.26337133649175e-107	1.33074000384305	0.651	0.261	2.76956264385722e-103	1	MUC4
17	1.1793643331267e-99	1.28240002270413	0.759	0.411	2.58540249108036e-95	1	PLCG2
18	6.21725087600778e-57	1.23561842956877	0.297	0.087	1.36294573703843e-52	1	MKI67
19	4.07832624576349e-31	1.20956944231921	0.357	0.175	8.94050679596272e-27	1	FCGBP
20	7.49604531409012e-109	1.1588302960149	0.801	0.359	1.64328305375484e-104	1	ELF3
21	9.76038953703872e-33	1.58387579991926	0.667	0.441	2.13967259430963e-28	2	PLCG2
22	4.28776829318216e-178	1.56907170886976	1	0.996	9.39964565231394e-174	2	MT-CO3
23	3.05905978490007e-164	1.44247910017437	1	0.997	6.70607086045794e-160	2	MT-CO1
24	1.80947175310265e-164	1.39325712941633	1	0.996	3.96672397715163e-160	2	MT-ATP6
25	2.15912529091105e-149	1.38968387095572	1	0.987	4.73323446273521e-145	2	MT-ND4
26	1.01975316983632e-143	1.37278993460958	1	0.996	2.23550289891518e-139	2	MT-CO2
27	2.51329172329212e-119	1.35093088619647	0.946	0.627	5.50963811580098e-115	2	MT-ATP8
28	4.26340916981771e-131	1.32407314348535	0.998	0.984	9.34624558207439e-127	2	MT-ND1
29	2.13656935129857e-133	1.31246270810655	0.996	0.945	4.68378733191672e-129	2	MT-ND5
30	1.02290956450975e-143	1.31030784259925	1	0.994	2.24242234731827e-139	2	MT-CYB
31	2.28210712977651e-192	3.72219478624801	0.363	0.023	5.00283524989606e-188	3	CCL5
32	0	3.55345532101006	0.782	0.02	0	3	CD3D
33	5.88216658465962e-223	3.43827976735308	0.331	0.009	1.28948855868908e-218	3	GZMA
34	0	3.42237225215721	0.667	0.038	0	3	LTB
35	0	3.36453331267556	0.937	0.102	0	3	PTPRC
36	0	3.33773172457631	0.624	0.013	0	3	TRBC1
37	0	3.22173860284988	0.701	0.059	0	3	IL7R
38	0	3.17888875758043	0.696	0.024	0	3	TRBC2
39	2.42961800987984e-269	3.17776579027863	0.88	0.23	5.32620860125859e-265	3	IL32
40	0	3.08246264658006	0.78	0.022	0	3	CD2
41	0	5.61807829336865	0.753	0.077	0	4	CXCL14
42	0	4.05593579341891	0.832	0.069	0	4	RARRES2
43	0	3.99893139830672	0.727	0.05	0	4	DCN
44	0	3.97208031697499	0.698	0.037	0	4	IGFBP5
45	0	3.95447071040266	0.693	0.05	0	4	LUM
46	0	3.83288773928033	0.897	0.066	0	4	COL3A1
47	1.12311281529621e-63	3.6295522144754	0.317	0.068	2.46208791369235e-59	4	MMP3
48	0	3.54360685381649	0.804	0.048	0	4	COL6A3
49	1.13079171931585e-257	3.51020446170978	0.544	0.042	2.47892160708421e-253	4	CFD
50	0	3.44402958453815	0.874	0.062	0	4	COL1A2
51	2.41345240012903e-176	1.89035365316428	0.518	0.054	5.29077035156286e-172	5	CLCA4
52	3.23650518205033e-224	1.86052475829163	0.842	0.129	7.09506666009074e-220	5	CEACAM7
53	8.10327455008263e-138	1.82932927131762	1	0.981	1.77639984686911e-133	5	MT-ND2
54	9.92701139152279e-127	1.76922278880512	0.912	0.355	2.17619943724963e-122	5	DST
55	5.61909803500391e-215	1.69005487627411	0.691	0.084	1.23181867123356e-210	5	SLC26A3
56	6.93257088730435e-144	1.64026963673828	1	0.978	1.51975818991486e-139	5	MT-ND3

57	2.96334654712362e-185	1.6211388097152	0.755	0.136	6.49624830060441e-181	5	MYO15B
58	1.53664887584439e-172	1.56538975120653	0.924	0.226	3.36864166562607e-168	5	FABP1
59	5.23152421052707e-134	1.55871791773043	1	0.984	1.14685473743174e-129	5	MT-ND1
60	6.83315282605976e-108	1.53267990967804	0.936	0.4	1.49796376252882e-103	5	MUC12
61	1.13955718982919e-88	4.87142393193958	0.265	0.028	2.49813727154355e-84	6	S100A9
62	4.15897952635343e-196	4.56207739431076	0.62	0.089	9.11731491767199e-192	6	HLA-DRA
63	1.45127284647628e-53	4.04816379642534	0.408	0.132	3.18148033404531e-49	6	LYZ
64	5.56271287917085e-156	3.71189759519068	0.558	0.09	1.21945791737183e-151	6	HLA-DPA1
65	1.37230823441201e-63	3.38187560546866	0.651	0.364	3.00837411147802e-59	6	CD74
66	3.16244838692481e-179	3.37785474472533	0.511	0.057	6.93271935381657e-175	6	HLA-DPB1
67	8.02017998908802e-129	3.35496175340124	0.511	0.091	1.75818385720787e-124	6	HLA-DRB1
68	1.98005621021018e-261	3.27162295307875	0.47	0.021	4.34067922402277e-257	6	TYROBP
69	5.10749662697469e-41	3.2519103315583	0.807	0.735	1.11966541056539e-36	6	FTL
70	1.06533677926459e-259	3.22547906640453	0.386	0.009	2.33543128750383e-255	6	AIF1
71	1.34607078315324e-51	3.26066418924025	0.347	0.081	2.95085637082853e-47	7	REG1A
72	6.90674068817487e-185	2.74131871344537	0.911	0.207	1.5140956936617e-180	7	AGR2
73	1.44114079261467e-109	2.4018063822148	0.571	0.114	3.15926884556988e-105	7	OLFM4
74	1.35848228067739e-72	2.35130185151091	0.782	0.274	2.97806485570097e-68	7	TFF3
75	3.28589133270457e-141	2.16886835751418	0.911	0.277	7.20333097955496e-137	7	LCN2
76	1.21051943834993e-274	2.14823648987452	0.917	0.138	2.65370071275071e-270	7	GPX2
77	4.49010790447258e-127	2.09400193585773	0.997	0.67	9.84321454818478e-123	7	RPL7
78	3.3831975735178e-140	1.99765675389333	1	0.792	7.41664572066572e-136	7	RPL8
79	2.16907647290172e-187	1.98444478203359	0.855	0.172	4.75504944389514e-183	7	TSPAN8
80	1.19252437764331e-160	1.92816349778645	0.934	0.289	2.61425194066966e-156	7	EPCAM
81	1.60082486261514e-240	3.85150087199341	0.967	0.228	3.50932826382491e-236	8	FABP1
82	9.5528133403378e-261	3.60307079838886	0.838	0.12	2.09416774046885e-256	8	TFF1
83	1.13222011243123e-307	3.3017384493785	0.927	0.128	2.48205293047174e-303	8	CEACAM7
84	4.49609740694892e-174	3.23749148034369	0.98	0.372	9.85634473551343e-170	8	PHGR1
85	3.30395778342737e-270	3.16598329789645	0.715	0.071	7.24293625282947e-266	8	CA4
86	1.17610972684121e-155	2.91421406926094	0.5	0.058	2.57826774318131e-151	8	CLCA4
87	1.29356009598658e-152	2.73938491095718	0.828	0.231	2.83574244242178e-148	8	KRT19
88	1.15945218205048e-246	2.64213815516193	0.623	0.055	2.54175107349106e-242	8	GUCA2A
89	6.62094629871755e-255	2.58919181554374	0.907	0.164	1.45144384760486e-250	8	TSPAN1
90	6.73650467784766e-168	2.5310886180799	0.99	0.431	1.47677655547776e-163	8	PIGR
91	3.76662082524655e-278	3.99105607112786	0.822	0.053	8.25718617310548e-274	9	SPARCL1
92	7.49488378070406e-197	3.94195269284317	0.941	0.122	1.64302842240594e-192	9	COL4A1
93	0	3.78115061258363	0.724	0.004	0	9	PLVAP
94	5.07109741553979e-228	3.77537747064915	0.855	0.084	1.11168597543463e-223	9	HSPG2
95	0	3.66894846056619	0.737	0.022	0	9	FLT1
96	2.97865954294688e-168	3.39402501898374	0.868	0.119	6.52981745004815e-164	9	COL4A2
97	1.42246118347612e-192	3.30711332026341	0.612	0.041	3.11831940641636e-188	9	IGFBP3
98	0	3.1600210868687	0.671	0.003	0	9	VWF
99	0	3.0709871298822	0.711	0.012	0	9	EGFL7
100	6.72798488862514e-178	3.05428228071389	0.73	0.076	1.4749088472844e-173	9	PECAM1
101	0	4.8369887204908	0.971	0.042	0	10	RGS5
102	1.20223304190756e-243	3.94498156102456	1	0.074	2.63553527446975e-239	10	THY1
103	7.71607297326423e-187	3.92397798995082	1	0.109	1.69151751719898e-182	10	SPARC
104	6.44791012083174e-118	3.88102043919942	1	0.198	1.41351085668873e-113	10	IGFBP7
105	3.6618109177621e-186	3.87018084423442	1	0.104	8.02742189391808e-182	10	COL1A1
106	2.06106071674322e-97	3.83034977684422	0.99	0.244	4.51825730324448e-93	10	TIMP1
107	1.27536469169729e-138	3.8147721484056	1	0.161	2.7958544771388e-134	10	CALD1
108	6.07379084164904e-227	3.71617687242145	0.883	0.056	1.3314964283063e-222	10	ACTA2
109	1.09765006297722e-164	3.54074131456172	0.99	0.115	2.40626846805866e-160	10	COL1A2
110	4.93691021976395e-153	3.44430296105421	1	0.13	1.08226945837665e-148	10	COL4A1

Supplementary Table 5: Top 10 expressed cell-type-specific marker genes identified for each cluster in sample 569.

#	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	3.69845546794461e-277	2.42202799453728	0.524	0.139	8.47316147706109e-273	0	CCL5
2	0	2.31485373264953	0.701	0.15	0	0	IL7R
3	0	2.13181448395676	0.646	0.143	0	0	KLRB1
4	0	1.84359661994783	0.837	0.483	0	0	ID2
5	0	1.82575965113149	0.743	0.166	0	0	CD3D
6	0	1.80940287250568	0.628	0.149	0	0	FYB1
7	0	1.74411975944706	0.583	0.123	0	0	CD3G
8	0	1.74249451836812	0.644	0.169	0	0	TRAC
9	1.10134908016706e-105	1.7409806082735	0.604	0.432	2.52319074266275e-101	0	HSPA6
10	1.77636102170315e-215	1.70411288510131	0.446	0.115	4.06964310072192e-211	0	TRBC1
11	4.28674042411181e-243	5.05291062563666	0.974	0.797	9.82092231164016e-239	1	IGKC
12	1.96429605339114e-58	4.7703442384498	0.506	0.256	4.50020225831911e-54	1	IGLC3
13	0	4.75552687656364	0.989	0.367	0	1	JCHAIN
14	1.18379004969854e-92	4.69294366427064	0.614	0.287	2.71206300385936e-88	1	IGLC2
15	0	4.65381945816148	0.82	0.114	0	1	IGLC1
16	0	3.55486639003508	0.985	0.602	0	1	IGHA1
17	0	3.47490852150996	0.785	0.126	0	1	IGHA2
18	0	3.33351405240659	0.983	0.1	0	1	MZB1
19	0	3.18951008626119	0.995	0.476	0	1	SSR4
20	0	2.68068920094942	0.91	0.059	0	1	DERL3
21	0	3.4261179680075	0.842	0.111	0	2	HLA-DRA
22	0	2.8341604274593	0.638	0.02	0	2	MS4A1
23	0	2.81865826515616	0.873	0.379	0	2	CD74
24	0	2.64119549063172	0.781	0.117	0	2	HLA-DRB1
25	0	2.52193455699643	0.771	0.143	0	2	HLA-DPA1
26	0	2.46049575504762	0.693	0.074	0	2	HLA-DPB1
27	0	2.3256137487311	0.48	0.023	0	2	VPREB3
28	0	2.306928432709	0.527	0.024	0	2	BANK1
29	0	2.30107936705159	0.594	0.047	0	2	HLA-DQA1
30	0	2.16560045820026	0.764	0.218	0	2	CD37
31	0	3.10678889513413	1	0.955	0	3	MT-ND1
32	0	2.89985313621228	1	0.988	0	3	MT-CO2
33	0	2.87869433098043	1	0.984	0	3	MT-CO1
34	0	2.86648293859409	1	0.914	0	3	MT-ND2
35	0	2.7946230480393	1	0.974	0	3	MT-CYB
36	0	2.79040784797322	1	0.987	0	3	MT-CO3
37	0	2.76034072860961	1	0.988	0	3	MT-ATP6
38	0	2.73115834632255	1	0.966	0	3	MT-ND4
39	0	2.70297477905349	1	0.937	0	3	MT-ND3
40	0	2.49103844026744	0.999	0.819	0	3	MT-ND5
41	2.61987475609e-23	1.7715503572636	0.641	0.656	6.0021330662022e-19	4	IGHA1
42	3.68507434343296e-15	1.62330889884155	0.309	0.207	8.44250532080491e-11	4	IGHA2
43	2.18728815500947e-59	1.43257394319635	0.738	0.675	5.0110771631267e-55	4	BTG2
44	3.79771689338233e-12	1.2717881933535	0.36	0.301	8.70056940273893e-08	4	XBP1
45	8.07037135999439e-19	1.21356738807732	0.641	0.664	1.84892207857471e-14	4	PPP1R15A
46	1.8905263887957e-59	1.14352542143469	0.912	0.908	4.33119595673094e-55	4	JUN
47	6.01458304388783e-06	1.13136045667182	0.455	0.522	0.13779409753547	4	SLC38A2
48	5.4257103132001e-10	1.13062558384867	0.487	0.532	1.24303023275414e-05	4	SQSTM1
49	6.98204647768284e-131	1.02147060711676	0.99	0.986	1.59958684803714e-126	4	MALAT1
50	4.03007241447224e-06	1.00423149199688	0.414	0.436	0.0923289590155589	4	GLS
51	1.8677519560451e-112	2.1178326758641	0.4	0.085	4.27901973129932e-108	5	GZMA
52	1.91267610158149e-151	1.94923877400971	0.833	0.351	4.3819409487232e-147	5	S100A4
53	7.37927289468048e-197	1.92484549985437	0.853	0.264	1.6905914201713e-192	5	CD3D
54	9.83175328113683e-128	1.81263684097843	0.723	0.269	2.25245467670845e-123	5	IL32
55	9.75941667231411e-145	1.67717932067299	0.709	0.213	2.23588235962716e-140	5	TRBC2
56	1.13408509525699e-98	1.61902649672846	0.345	0.075	2.59818895323377e-94	5	TNFRSF18

57	4.55425751515854e-162	1.61871703664122	0.747	0.216	1.04338039672282e-157	5	CD2
58	1.20560512458306e-163	1.57922454881957	0.735	0.197	2.76204134041979e-159	5	CD3G
59	3.00992574658503e-148	1.57624467322777	0.982	0.738	6.8957398854263e-144	5	ACTB
60	1.33127011693184e-172	1.55739394260556	0.55	0.104	3.04993983789085e-168	5	CD7
61	0	5.75387122284433	0.536	0.025	0	6	CXCL14
62	0	4.99771874897158	0.966	0.034	0	6	CALD1
63	0	4.73342454291607	0.834	0.01	0	6	COL3A1
64	0	4.6269424062033	0.679	0.031	0	6	TAGLN
65	0	4.59861755284853	0.957	0.067	0	6	IGFBP7
66	0	4.49985012570223	0.821	0.008	0	6	COL1A2
67	0	4.27881068908153	0.641	0.012	0	6	IGFBP5
68	0	4.168594695869	0.735	0.026	0	6	COL1A1
69	0	4.06754826915085	0.87	0.195	0	6	TIMP1
70	0	3.94561760749581	0.8	0.02	0	6	RARRES2
71	0	5.38296243668694	0.858	0.099	0	7	REG1A
72	2.12180021184324e-261	3.85408339067748	1	0.22	4.86104428533286e-257	7	PIGR
73	1.43986316137661e-303	3.56220234959523	0.815	0.087	3.2987265027138e-299	7	OLFM4
74	0	3.44428110359823	0.992	0.127	0	7	AGR2
75	0	3.39195731183029	0.937	0.093	0	7	LCN2
76	0	3.37140518945827	0.898	0.065	0	7	PLA2G2A
77	2.22727526490461e-237	3.14186911070667	0.378	0.018	5.10268763189646e-233	7	REG1B
78	1.14183283802135e-106	3.12112524273844	0.276	0.025	2.61593903190691e-102	7	SPINK4
79	0	3.09423414911327	0.866	0.042	0	7	DMBT1
80	0	3.04587383990203	0.972	0.129	0	7	TSPAN8
81	0	4.45668931023878	0.996	0.128	0	8	CEACAM5
82	0	4.38611796941067	1	0.113	0	8	TFF3
83	0	3.88089424119844	1	0.148	0	8	EPCAM
84	0	3.44096974482297	1	0.095	0	8	FXYD3
85	3.94753134263385e-192	3.24842279533981	0.902	0.168	9.04379430597416e-188	8	FABP1
86	0	3.16695972497928	0.893	0.086	0	8	OLFM4
87	6.43706786500119e-282	3.09508205133233	1	0.181	1.47473224787177e-277	8	KRT18
88	5.72457844760198e-252	3.04916592841363	1	0.196	1.31150092234561e-247	8	LGALS4
89	4.01650315288399e-179	3.03717353648918	0.996	0.338	9.20180872325721e-175	8	LGALS3
90	2.68597732600692e-279	2.97760623800907	1	0.174	6.15357405388185e-275	8	KRT8
91	5.39651830061471e-117	2.59424065556283	1	0.988	1.23634234267083e-112	9	MT-CO3
92	2.16395668114822e-171	1.96993743336948	0.711	0.109	4.95762475651058e-167	9	SOX9
93	2.06191898822449e-75	1.96790215692163	0.63	0.178	4.72385640202231e-71	9	FABP1
94	3.60887279742928e-149	1.92371656829978	0.872	0.187	8.26792757891048e-145	9	ELF3
95	4.04600123959654e-94	1.92189957690839	1	0.985	9.26938883991567e-90	9	MT-CO1
96	7.25705164403624e-120	1.91755541671925	0.553	0.092	1.6625905316487e-115	9	L1TD1
97	4.47673114048828e-114	1.82514969221868	0.779	0.183	1.02561910428586e-109	9	KRT8
98	1.20826222921696e-88	1.81505764761426	0.996	0.989	2.76812876713605e-84	9	MT-CO2
99	9.81636296268287e-90	1.80434289819684	1	0.989	2.24892875475065e-85	9	MT-ATP6
100	4.1322980297237e-45	1.73363347601427	0.511	0.165	9.46709478609699e-41	9	PHGR1
101	0	4.24670977795145	0.81	0.005	0	10	PLVAP
102	0	4.15058043929982	0.833	0.038	0	10	SPARCL1
103	0	4.11628358713694	0.856	0.05	0	10	COL4A1
104	1.37223241973028e-196	3.67746126174229	0.466	0.029	3.14378447360207e-192	10	IGFBP3
105	0	3.54163163968784	0.816	0.062	0	10	HSPG2
106	2.50582537242331e-272	3.52534535210959	0.954	0.101	5.7408459282218e-268	10	IGFBP7
107	0	3.50923783934769	0.667	0.006	0	10	FLT1
108	0	3.48910919359771	0.649	0.002	0	10	VWF
109	0	3.45165578679891	0.782	0.053	0	10	PECAM1
110	2.41793276496162e-123	3.43797766771435	0.81	0.19	5.53948396452708e-119	10	IFI27
111	5.46471627328334e-58	3.94405050485821	0.622	0.183	1.25196649820921e-53	11	FABP1
112	6.63108190337395e-118	3.78553002558854	0.776	0.163	1.51918086406297e-113	11	PHGR1
113	2.76113004195759e-143	3.73528185483757	0.929	0.208	6.32574892612484e-139	11	LGALS4
114	7.89158139094768e-120	3.71879498331139	0.564	0.075	1.80796129666611e-115	11	MT1G
115	1.77896431219096e-134	3.40795036989836	0.782	0.141	4.07560723922949e-130	11	C15orf48
116	8.85921233256028e-152	3.18694024344957	0.846	0.143	2.02964554538956e-147	11	TSPAN8

117	2.3516205798388e-82	3.17671550663108	0.878	0.349	5.3875627484107e-78	11	LGALS3
118	1.04879571986356e-98	3.08500820512489	0.628	0.118	2.40279099420742e-94	11	MT1E
119	1.55268454671998e-68	2.913783766842	0.712	0.241	3.55720029653549e-64	11	SRI
120	6.66712491811698e-93	2.88333180981766	0.724	0.171	1.5274383187406e-88	11	CD24
121	5.2050356560073e-210	5.34337961883621	0.949	0.067	1.19247366879127e-205	12	LYZ
122	3.75306759032581e-71	5.17628154808414	0.41	0.034	8.59827784943642e-67	12	APOE
123	0	4.47311654182207	0.564	0.002	0	12	C1QA
124	0	4.16734823395167	0.885	0.009	0	12	AIF1
125	3.12803924061788e-90	4.154806521134	0.962	0.19	7.16633790025557e-86	12	HLA-DRB1
126	0	4.08390995044552	0.577	0.001	0	12	C1QB
127	7.50676234851828e-90	4.0660627485096	0.974	0.192	1.71979925404554e-85	12	HLA-DRA
128	7.03776991621344e-86	3.9889734747873	0.974	0.211	1.6123530878045e-81	12	HLA-DPA1
129	4.76235878618326e-129	3.92296794622902	0.91	0.105	1.09105639791459e-124	12	HLA-DQA1
130	1.28535582733298e-26	3.84460199962382	0.692	0.28	2.94475020041985e-22	12	CTSD

Supplementary Table 6: Manually annotated clusters in sample 554 based on its top 10 cell-type-specific marker genes.

#	cluster	gene	Human Protein Atlas (HPA)	Azimuth	Panglaodb.se	Supplementary literature or other comment	Annotated cell type
1	0	LTB	NK-cells & T-cells - Immune response (mainly)				CD4+ effector memory T-cell
2	0	IL7R	Non-specific - Transcription regulation (mainly)	CD4+ effector memory T-cell (lung)			
3	0	HSPA1A	Epithelial cell types - Mixed function (mainly)				
4	0	CD2	Non-specific - Transcription regulation (mainly)	CD4+ effector memory T-cell (lung)			
5	0	PTPRC	Non-specific - Transcription regulation (mainly)	CD4+ effector memory T-cell (lung)			
6	0	IKZF1	Plasmacytoid DCs - Unknown function (mainly)				
7	0	CD52	B-cells - Immune response (mainly)				
8	0	CD3D	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (lung)			
9	0	JUN	Non-specific - Mitochondria (mainly)				
10	0	GIMAP7	NK-cells & T-cells - Immune response (mainly)				
11	1	CEACAM7	Intestinal epithelial cells - Unknown function (mainly)				Intestinal epithelial cell
12	1	FABP1	Enterocytes - Digestion (mainly)				
13	1	CEACAM5	Intestinal epithelial cells - Unknown function (mainly)				
14	1	KRT20	Intestinal epithelial cells - Unknown function (mainly)				
15	1	FXYD3	Intestinal epithelial cells - Unknown function (mainly)				
16	1	CKB	Intestinal epithelial cells - Unknown function (mainly)				
17	1	TFF1	Pancreatic endocrine cells - Mixed function (mainly)				
18	1	TSPAN1	Intestinal epithelial cells - Unknown function (mainly)				
19	1	LGALS3	Intestinal epithelial cells - Unknown function (mainly)	Intestinal epithelial cell (fetal development)			
20	1	PHGR1	Enterocytes - Digestion (mainly)				
21	2	RNF43	Intestinal epithelial cells - Unknown function (mainly)				Mitochondrial gene-expressing cell
22	2	MT-ND3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
23	2	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
24	2	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
25	2	MT-CYB	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
26	2	MTRNR2L10	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
27	2	MT-CO2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
28	2	MT-ATP6	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
29	2	MT-ND2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
30	2	AC103702.2	Not found				
31	3	CENPF	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (PBMC) CD4+ proliferating T-cell (lung)			CD4+ proliferating T-cell
32	3	MKI67	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (PBMC)			
33	3	ASPM	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (PBMC)			
34	3	TOP2A	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (PBMC) CD4+ proliferating T-cell (lung)			
35	3	CENPW	Non-specific - Cell cycle regulation (mainly)				
36	3	HMGB2	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (lung)			
37	3	PTTG1	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (PBMC)			
38	3	H2AFZ	Non-specific - Transcription regulation (mainly)	Proliferating Macrophage (lung)			
39	3	CKS2	Non-specific - Mitochondria (mainly)				
40	3	HMGB1	Non-specific - Cell cycle regulation (mainly)				
41	4	LEFTY1	Intestinal epithelial cells - Unknown function (mainly)				Intestinal epithelial cell
42	4	GPX2	Intestinal epithelial cells - Unknown function (mainly)				

43	4	TFF3	Mucus-secreting cells - Mucin production (mainly)			Classical marker of goblet cell (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
44	4	LCN2	Respiratory epithelial cells - Mucosal defense (mainly)	Mucous cell or goblet cell (lung)			
45	4	PIGR	Intestinal epithelial cells - Unknown function (mainly)				
46	4	TSPAN8	Enterocytes - Digestion (mainly)				
47	4	FAM3D	Intestinal epithelial cells - Unknown function (mainly)	Goblet cell (lung)			
48	4	SLC12A2	Mucus-secreting cells - Mucin production (mainly)			Significantly expressed in human intestinal stem cells (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
49	4	PRDX5	Respiratory epithelial cells - Mucosal defense (mainly)				
50	4	TFF1	Pancreatic endocrine cells - Mixed function (mainly)				
51	5	CXCL14	Fibroblasts - ECM organization (mainly)	Intestine-Stromal cells (fetal development)			Fibroblast
52	5	COL1A2	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Myofibroblast (lung)			
53	5	COL3A1	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Smooth muscle cells (fetal development) Myofibroblast (lung)			
54	5	COL1A1	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Smooth muscle cells (fetal development) Myofibroblast (lung)			
55	5	LUM	Fibroblasts - ECM organization (mainly)	Stormal cells (bone marrow) Fibroblast (lung)			
56	5	DCN	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Fibroblast (lung)			
57	5	COL6A3	Fibroblasts - ECM organization (mainly)	Smooth muscle cells (fetal development) Myofibroblast (lung)			
58	5	CALD1	Smooth muscle cells - ECM organization (mainly)	Smooth muscle cells (fetal development) Smooth muscle cells (lung) Myofibroblast (lung)			
59	5	POSTN	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Stromal cells (fetal development) Myofibroblast (lung)			
60	5	RARRES2	Hepatocytes - Metabolism (mainly)	Alveolar fibroblast (lung)			
61	6	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	Mitochondrial gene-expressing cell
62	6	MT-ND3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
63	6	MT-ATP6	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
64	6	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
65	6	MT-CO2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
66	6	MT-CYB	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
67	6	MT-ND2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
68	6	MTRNR2L12	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
69	6	MTRNR2L8	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
70	6	MTRNR2L1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
71	7	CCL5	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ effector memory T-cell (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ central memory t-cell (PBMC) NK-cell (lung)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	CD8+ effector memory T-cell
72	7	GZMA	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ T-cell (lung)			
73	7	CCL4	NK-cells & T-cells - Immune response (mainly)	CD8+ effector T-cell (bone marrow) NK proliferating cells (bone marrow)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	

74	7	NKG7	NK-cells & T-cells - Immune response (mainly)	CD4+ cytotoxic T-cell (PBMC) CD8+ effector memory T-cell (PBMC) NK-cell (PBMC) NK-cell (lung) CD8+ T-cell (lung) CD8+ effector memory T-cell (lung)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	
75	7	TRBC1	NK-cells & T-cells - Immune response (mainly)				
76	7	KLRB1	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ central memory T-cell (PBMC) NK-cell (PBMC) CD4+ effector T-cell (bone marrow) CD8+ effector T-cell (bone marrow)			
77	7	CCL3	Macrophages - Innate immune response (mainly)	CD8+ effector T-cell (bone marrow)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	
78	7	HCST	NK-cells & T-cells - Immune response (mainly)	CD8+ T-cell (PBMC)			
79	7	GZMB	NK-cells & T-cells - Immune response (mainly)	NK-cell (PBMC) CD8+ effector memory T-cell (PBMC) NK-cell (lung) CD8+ T-cell (lung)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	
80	7	CD3D	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ proliferating T-cell (PBMC) NK-cell (PBMC) CD4+ effector memory T-cell (lung) CD8+ T-cell (lung)			
81	8	SPARCL1	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Bronchial vessel (lung) Stromal cell (bone marrow)			Vascular endothelial cell
82	8	MGP	Glandular cells - Unknown function (mainly)				
83	8	IGFBP3	Fibroblasts - ECM organization (mainly)	Stellate cells (fetal development)			
84	8	COL4A1	Smooth muscle cells - ECM organization (mainly)	Vascular endothelial cell (fetal development) Pericyte (lung)			
85	8	IGFBP7	Smooth muscle cells - ECM organization (mainly)	Vascular endothelial cell (fetal development) Lymphatic vessel (lung)			
86	8	PLVAP	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Intestine-Vascular endothelial cells (fetal development)			
87	8	PECAM1	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Capillary endothelial cell (lung)			
88	8	VWF	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Lymphatic or vascular endothelial cells (fetal development) Bronchial vessel endothelial cell (lung)			
89	8	ENG	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Endocardial cells (fetal development)			
90	8	HSPG2	Granulosa cells - Unknown function (mainly)	Vascular endothelial cells (fetal development) Endocardial cells (fetal development)			
91	9	LYZ	Monocytes & Neutrophils - Innate immune response (mainly)	Monocyte (PBMC) CD14+ monocyte (PBMC) Stomach-Goblet cells (fetal development)			Myeloid cell
92	9	HLA-DRA	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development) Myeloid dendritic cell (lung)			
93	9	FTL	Macrophages - Innate immune response (mainly)	Myeloid cell (fetal development) Neutrophil (kidney)			
94	9	IFI30	Monocytes & Neutrophils - Innate immune response (mainly)	Dendritic cell (bone marrow)			
95	9	AIF1	Monocytes & Neutrophils - Innate immune response (mainly)	Monocyte (PBMC) CD16+ monocyte (PBMC) CD16+ monocyte (lung)			

96	9	HLA-DQA1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) B-cell (PBMC) Stomach-Myeloid cells (fetal development) Dendritic cell (lung) Myeloid dendritic cell (lung)			
97	9	HLA-DPB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development) Myeloid dendritic cell (lung)			
98	9	HLA-DPA1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Myeloid dendritic cell (lung)			
99	9	HLA-DRB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development)			
100	9	TYROBP	Macrophages - Innate immune response (mainly)	Monocyte (PBMC) NK-cell (PBMC) Dendritic cell (lung)			
101	10	RGS5	Smooth muscle cells - ECM organization (mainly)				Smooth muscle cell
102	10	ACTA2	Smooth muscle cells - ECM organization (mainly)				
103	10	CALD1	Smooth muscle cells - ECM organization (mainly)				
104	10	NDUFA4L2	Smooth muscle cells - ECM organization (mainly)				
105	10	MGP	Glandular cells - Unknown function (mainly)				
106	10	COL18A1	Smooth muscle cells - ECM organization (mainly)				
107	10	CSRP2	Fibroblasts - ECM organization (mainly)				
108	10	IGFBP7	Smooth muscle cells - ECM organization (mainly)				
109	10	TAGLN	Smooth muscle cells - ECM organization (mainly)				
110	10	COL4A1	Smooth muscle cells - ECM organization (mainly)				
111	11	IGKC	Plasma cells - Humoral immune response (mainly)				Plasma B-cell
112	11	IGHA1	Plasma cells - Humoral immune response (mainly)				
113	11	IGLC1	Alveolar cells - Smell perception (mainly)				
114	11	JCHAIN	Plasma cells - Humoral immune response (mainly)				
115	11	IGLC3	Plasma cells - Humoral immune response (mainly)				
116	11	IGHA2	Plasma cells - Humoral immune response (mainly)				
117	11	MZB1	Plasma cells - Humoral immune response (mainly)				
118	11	TXNDC5	Plasma cells - Humoral immune response (mainly)				
119	11	IGKV4-1	Plasma cells - Humoral immune response (mainly)				
120	11	SSR4	Plasma cells - Humoral immune response (mainly)				
121	12	SH2D6	Proximal tubular cells - Tubular reabsorption (mainly)		Cholangiocytes - Intestinal epithelial cell		Intestinal epithelial cell
122	12	LRMP	Non-specific - Transcription regulation (mainly)		Cholangiocytes - Intestinal epithelial cell		
123	12	ANXA4	Pancreas - Digestion (mainly)		Cholangiocytes - Intestinal epithelial cell		
124	12	ALOX5AP	Macrophages - Innate immune response (mainly)		Cholangiocytes or dendritic cells		
125	12	RASSF6	Intestinal epithelial cells - Unknown function (mainly)		Cholangiocytes - Intestinal epithelial cell		
126	12	RGS13	Granulocytes - Receptor signaling (mainly)	Myeloid cell (lung)			
127	12	PBXIP1	Non-specific - Transcription regulation (mainly)				
128	12	SPTLC2	Smooth muscle cells - Unknown function (mainly)		Cholangiocytes - Intestinal epithelial cell		
129	12	BMX	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Artery (lung)	Cholangiocytes - Intestinal epithelial cell		
130	12	AZGP1	Glandular cells - Unknown function (mainly)				

Supplementary Table 7: Manually annotated clusters in sample 556 based on its top 10 cell-type-specific marker genes.

#	cluster	gene	Human Protein Atlas (HPA)	Azimuth	Panglaodb.se	Supplementary literature or other comment	Annotated cell type
1	0	HLA-DRA	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (PBMC) Antigen-presenting cells (fetal development) Myeloid dendritic cell (lung)	Dendritic cell B-cell		Dendritic cell or B-cell
2	0	CD74	Macrophages - Immune response (mainly)	B-cell (PBMC) Dendritic cell (PBMC) Antigen-presenting cells (fetal development)			
3	0	VPREB3	B-cells - Immune response (mainly)	B-cell (lung)	B-cell		
4	0	HLA-DPA1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Myeloid dendritic cell (lung)	Dendritic cell B-cell		
5	0	BANK1	B-cells - Immune response (mainly)	B-cell (PBMC) B-cell (lung) B-cell (kidney)	B-cell		
6	0	HLA-DPB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development) Myeloid dendritic cell (lung)	Dendritic cell B-cell		
7	0	CD79A	B-cells - Immune response (mainly)	B-cell (PBMC) B-cell (lung)	B-cell		
8	0	HLA-DRB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development)	Dendritic cell B-cell		
9	0	MS4A1	B-cells - Immune response (mainly)	B-cell (PBMC B-cell (lung)	B-cell		
10	0	HLA-DQB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid dendritic cell (lung)	Dendritic cell B-cell		
11	1	FYB1	Non-specific - Mitochondria (mainly)	CD4+ effector memory T-cell (PBMC)			CD4+ effector memory T-cell
12	1	IL7R	Non-specific - Transcription regulation (mainly)	CD8+ central memory T-cell (PBMC) CD4+ effector memory T-cell (lung)			
13	1	TRAC	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ effector memory T-cell (PBMC)			
14	1	TRBC2	Non-specific - Mitochondria (mainly)				
15	1	CD3D	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ proliferating T-cell (PBMC)			
16	1	TRBC1	NK-cells & T-cells - Immune response (mainly)				
17	1	RGCC	Non-specific - Mitochondria (mainly)	Epithelial cell (pancreas)			
18	1	GIMAP7	NK-cells & T-cells - Immune response (mainly)				
19	1	GIMAP4	NK-cells & T-cells - Immune response (mainly)			Membrane-expressed CD4+ T-helper cells (https://doi.org/10.1155/2010/268589)	
20	1	CD3E	T-cells - T-cell receptor (mainly)	CD4+ effector memory T-cell (PBMC) CD4+ effector memory T-cell (lung) CD8+ T-cell (PBMC) CD8+ T-cell (lung) NK-cell (PBMC) NK-cell (lung)			
21	2	KLRB1	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ central memory T-cell (PBMC) NK-cell (PBMC) CD4+ effector T-cell (bone marrow) CD8+ effector T-cell (bone marrow)			CD8+ effector memory T-cell

22	2	ID2	Monocytes - Immune response regulation (mainly)	CD8+ naive T-cell (PBMC)			
23	2	IL32	Pancreas - Digestion (mainly)	CD4+ T-cell (PBMC) CD8+ effector memory T-cell (PBMC) CD8+ effector memory cell (lung)			
24	2	S100A4	Monocytes & Neutrophils - Innate immune response (mainly)	CD4+ cytotoxic T-cell (PBMC) CD4+ central memory T-cell (PBMC) CD4+ effector memory T-cell (PBMC) NK-cell (PBMC)			
25	2	CCL5	NK-cells & T-cells - Immune response (mainly)	CD4+ T-cell (PBMC) CD8+ effector memory T-cell (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ central memory T-cell (PBMC) NK-cell (lung)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	
26	2	IL7R	Non-specific - Transcription regulation (mainly)	CD8+ central memory T-cell (PBMC) CD4+ effector memory T-cell (lung)			
27	2	CD2	Non-specific - Transcription regulation (mainly)	CD8+ effector memory T-cell (PBMC) CD4+ effector memory T-cell (lung) CD8+ T-cell (lung) NK-cell (lung)			
28	2	TRBC1	NK-cells & T-cells - Immune response (mainly)				
29	2	CD3D	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ proliferating T-cell (PBMC)			
30	2	ANXA1	Epithelial cell types - Mixed function (mainly)	CD8+ central memory T-cell (PBMC)			
31	3	REG1A	Pancreas - Digestion (mainly)				
32	3	PHGR1	Enterocytes - Digestion (mainly)		Epithelial cells		
33	3	FABP1	Enterocytes - Digestion (mainly)		Epithelial cells		
34	3	IFI27	Pancreatic endocrine cells - Mixed function (mainly)	Endothelial cell (motor cortex)	Endothelial cells Epithelial cells		
35	3	ALDOB	Proximal tubular cells - Tubular reabsorption (mainly)		Hepatocytes		
36	3	PRAP1	Enterocytes - Digestion (mainly)		Epithelial cells Enterocytes		Intestinal epithelial cell
37	3	PIGR	Intestinal epithelial cells - Unknown function (mainly)		Epithelial cells		
38	3	KRT8	Pancreatic endocrine cells - Mixed function (mainly)		Epithelial cells		
39	3	SELENOP	Macrophages - Innate immune response (mainly)		Epithelial cells Hepatocytes		
40	3	LGALS4	Intestinal epithelial cells - Unknown function (mainly)		Epithelial cells		
41	4	IGHA1	Plasma cells - Humoral immune response (mainly)				
42	4	IGKC	Plasma cells - Humoral immune response (mainly)				
43	4	IGLC2	Plasma cells - Humoral immune response (mainly)				
44	4	JCHAIN	Plasma cells - Humoral immune response (mainly)				
45	4	IGLC3	Plasma cells - Humoral immune response (mainly)				
46	4	IGLC1	Alveolar cells - Smell perception (mainly)				Plasma B-cell
47	4	IGHA2	Plasma cells - Humoral immune response (mainly)				
48	4	IGHG3	Plasma cells - Humoral immune response (mainly)				
49	4	MZB1	Plasma cells - Humoral immune response (mainly)				
50	4	SSR4	Plasma cells - Humoral immune response (mainly)				
51	5	MT-ND3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
52	5	MT-ATP6	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
53	5	MTRNR2L12	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	Mitochondrial gene-expressing cell
54	5	MTRNR2L8	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
55	5	MT-CYB	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	

56	5	MT-ND4	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
57	5	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
58	5	MTRNR2L1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
59	5	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
60	5	MT-ND5	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
61	6	MALAT1			Unknown		Unknown cell type
62	6	CEMIP2	Non-specific - Transcription regulation (mainly)				
63	6	TLE4	Spermatids - Unknown function (mainly)		Germ cells		
64	6	NABP1	Alveolar cells - Smell perception (mainly)		Germ cells		
65	6	HSPH1	Epithelial cell types - Mixed function (mainly)		Germ cells		
66	6	GLS	Proximal tubular cells - Tubular reabsorption (mainly)				
67	6	SLC2A3	Adipocytes & Endothelial cells - Angiogenesis (mainly)		Endothelial cells Fibroblasts		
68	6	AAK1	Photoreceptor cells - Phototransduction (mainly)		T-cells NK-cells		
69	6	INTS6	Non-specific - Mitochondria (mainly)		Germ cells Unknown		
70	6	RNF213	NK-cells & T-cells - Immune response (mainly)		Unknown T-cells NK-cells		
71	7	GNLY	NK-cells & T-cells - Immune response (mainly)	NK-cells (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ effector memory cell (PBMC) NK-cell (lung)			CD8+ effector memory T-cell
72	7	CCL5	NK-cells & T-cells - Immune response (mainly)	CD4+ T-cell (PBMC) CD8+ effector memory T-cell (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ central memory T-cell (PBMC) NK-cell (Azimuth lung)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	
73	7	NKG7	NK-cells & T-cells - Immune response (mainly)	CD4+ cytotoxic T-cell (PBMC) CD8+ effector memory T-cell (PBMC) NK-cell (PBMC) NK-cell (lung) CD8+ T-cell (lung) CD8+ effector memory T-cell (lung)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	
74	7	GZMA	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ T-cell (lung)			
75	7	CCL4	NK-cells & T-cells - Immune response (mainly)	CD8+ effector T-cell (bone marrow) NK proliferating cells (bone marrow)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	
76	7	IFNG	Non-specific - Mitochondria (mainly)	CD8+ effector T-cell (bone marrow)			
77	7	GZMB	NK-cells & T-cells - Immune response (mainly)	NK-cell (PBMC) CD8+ effector memory T-cell (PBMC) NK-cell (lung) CD8+ T-cell (lung)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	
78	7	GZMH	NK-cells & T-cells - Immune response (mainly)	CD8+ effector memory T-cell (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ T-cell (lung)			
79	7	GZMK	Non-specific - Transcription regulation (mainly)	CD4+ T-cell (PBMC) CD8+ effector memory T-cell (PBMC) NK-cell (PBMC)			
80	7	KLRD1	NK-cells & T-cells - Immune response (mainly)	NK-cell (PBMC) CD8+ effector memory T-cell (PBMC) NK-cell (lung)			

81	8	HMGB2	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (lung)			CD4+ proliferating T-cell
82	8	STMN1	Non-specific - Cell cycle regulation (mainly)	NK proliferating cell (PBMC) Proliferating macrophage (lung)			
83	8	MKI67	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (PBMC)			
84	8	CENPF	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (PBMC CD4+ proligerating T-cell (lung)			
85	8	TCL1A	B-cells - Immune response (mainly)	Naive B-cell (PBMC)			
86	8	TOP2A	Non-specific - Cell cycle regulation (mainly)	CD4+ proliferating T-cell (PBMC) CD4+ proliferating T-cell (lung)			
87	8	LRMP	Non-specific - Transcription regulation (mainly)		Cholangiocytes - Intestinal epithelial cell		
88	8	MEF2B	Plasma cells - Humoral immune response (mainly)				
89	8	HIST1H4C	Non-specific - Transcription regulation (mainly)				
90	8	HMGN2	Non-specific - Cell cycle regulation (mainly)				
91	9	VPREB3	B-cells - Immune response (mainly)	B-cell (lung)			B-cell
92	9	ID3	Adipocytes & Endothelial cells - Angiogenesis (mainly)				
93	9	LINC01781	Not found	B memory cell (PBMC)	B-cells		
94	9	RPL4	Non-specific - Translation (mainly)				
95	9	SNHG29	Not found				
96	9	CD37	B-cells - Immune response (mainly)	Naive B-cell (PBMC)			
97	9	IER5	Non-specific - Mitochondria (mainly)				
98	9	RPS23	Non-specific - Translation (mainly)				
99	9	KPNB1	Plasma cells - Humoral immune response (mainly)				
100	9	KLF2	Fibroblasts - ECM organization (mainly)				
101	10	S100A9	Monocytes & Neutrophils - Innate immune response (mainly)	CD14+ monocyte (PBMC CD14+ monocyte (lung) Neutrophil (kidney)			Monocyte
102	10	CXCL8	Epithelial cell types - Mixed function (mainly)	Classical monocyte (lung)			
103	10	LYZ	Monocytes & Neutrophils - Innate immune response (mainly)	Monocyte (PBMC) CD14+ monocyte (PBMC) Stomach-Goblet cells (fetal development)			
104	10	IL1B	Macrophages - Immune response (mainly)	CD14+ monocyte (PBMC) Classical monocyte (lung)			
105	10	S100A8	Monocytes & Neutrophils - Innate immune response (mainly)	CD14+ monocyte (PBMC CD14+ monocyte (lung) Classical monocyte (lung)			
106	10	TIMP1	Fibroblasts - ECM organization (mainly)				
107	10	CST3	Macrophages - Immune response (mainly)				
108	10	C1QA	Macrophages - Innate immune response (mainly)	Macrophage (lung)			
109	10	C1QB	Macrophages - Innate immune response (mainly)	Macrophage (lung)			
110	10	TYROBP	Macrophages - Innate immune response (mainly)	Monocyte (PBMC) NK-cell (PBMC) Dendritic cell (lung)			
111	11	TRBC1	NK-cells & T-cells - Immune response (mainly)				T-cell
112	11	TRBC2	Non-specific - Mitochondria (mainly)				
113	11	TRAC	NK-cells & T-cells - Immune response (mainly)	CD4+ T-cell (PBMC) CD8+ central memory T-cell (PBMC) CD8+ effector memoryT-cell (PBMC) CD8+ proliferating T-cell (PBMC)			
114	11	CD3D	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD4+ effector memory T-cell (lung)			

				CD8+ proliferating T-cell (PBMC CD8+ proliferating T-cell (lung)			
115	11	GIMAP7	NK-cells & T-cells - Immune response (mainly)				
116	11	CD3G	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory cell (PBMC) CD8+ proliferating T-cell (PBMC)			
117	11	CD3E	T-cells - T-cell receptor (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ T-cell (PBMC) NK-cell (PBMC) CD4+ effector memory T-cell (lung) CD8+ effector memory T-cell (lung)			
118	11	LDHB	Non-specific - Basic cellular processes (mainly)	CD4+ central memory T-cell (PBMC) CD8+ central memory T-cell (PBMC)			
119	11	CD2	Non-specific - Transcription regulation (mainly)	CD8+ effector memory T-cell (PBMC) CD4+ effector memory T-cell (lung) CD8+ T-cell (lung) NK-cell (lung)			
120	11	RGCC	Non-specific - Mitochondria (mainly)	Endothelial cell (pancreas) Endothelial cell (lung)			
121	12	IGFBP7	Smooth muscle cells - ECM organization (mainly)	Lymphatic or vascular endothelial cells (fetal development) Lymphatic or vascular endothelial cells (lung)		Expressed in cancer-associated fibroblasts and tumor vessels (https://doi.org/10.1038/onc.2014.18)	Fibroblast
122	12	CXCL14	Fibroblasts - ECM organization (mainly)	Intestine-Stromal cells (fetal development)			
123	12	CALD1	Smooth muscle cells - ECM organization (mainly)	Smooth muscle cell (fetal development) Myofibroblast (lung) Pericyte (lung) Smooth muscle cell (lung)			
124	12	CFD	Fibroblasts - ECM organization (mainly)	Fibroblast (Azimuth lung)			
125	12	SPARCL1	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Bronchial epithelial vessel (lung)			
126	12	IGFBP5	Smooth muscle cells - ECM organization (mainly)				
127	12	DCN	Fibroblasts - ECM organization (mainly)	Stromal cell (fetal development) Fibroblast (lung)			
128	12	ADAMDEC1	Macrophages - Innate immune response (mainly)				
129	12	TAGLN	Smooth muscle cells - ECM organization (mainly)	Intestine-Smooth muscle cells (fetal development) Vascular smooth muscle cell (lung)			
130	12	A2M	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Fibroblast (lung)			

Supplementary Table 8: Manually annotated clusters in sample 559 based on its top 10 cell-type-specific marker genes.

#	cluster	gene	Human Protein Atlas (HPA)	Azimuth	Panglaodb.se	Supplementary literature or other comment	Annotated cell type
1	0	IGLC1	Alveolar cells - Smell perception (mainly)				Plasma B-cell
2	0	IGLC2	Plasma cells - Humoral immune response (mainly)				
3	0	IGLC3	Plasma cells - Humoral immune response (mainly)				
4	0	IGHA1	Plasma cells - Humoral immune response (mainly)				
5	0	JCHAIN	Plasma cells - Humoral immune response (mainly)				
6	0	IGKC	Plasma cells - Humoral immune response (mainly)				
7	0	IGHM	Plasma cells - Humoral immune response (mainly)				
8	0	IGHG1	Plasma cells - Humoral immune response (mainly)				
9	0	IGHA2	Plasma cells - Humoral immune response (mainly)				
10	0	IGHGP	Not found	Plasma cells (kidney)			
11	1	SPINK4	Mucus-secreting cells - Mucin production (mainly)				Intestinal goblet cell
12	1	MUC2	Mucus-secreting cells - Mucin production (mainly)			Classical marker of goblet cell (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
13	1	TFF3	Mucus-secreting cells - Mucin production (mainly)			Classical marker of goblet cell (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
14	1	CENPF	Non-specific - Cell cycle regulation (mainly)				
15	1	MUC5B	Intestinal epithelial cells - Unknown function (mainly)			Classical marker of goblet cell (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
16	1	MUC4	Respiratory epithelial cells - Mucosal defense (mainly)			Classical marker of goblet cell (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
17	1	PLCG2	Non-specific - Mitochondria (mainly)				
18	1	MKI67	Non-specific - Cell cycle regulation (mainly)				
19	1	FCGBP	Mucus-secreting cells - Mucin production (mainly)				
20	1	ELF3	Respiratory epithelial cells - Mucosal defense (mainly)				
21	2	PLCG2	Non-specific - Mitochondria (mainly)			Mitochondrial gene	Mitochondrial gene-expressing cell
22	2	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
23	2	MT-CO1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
24	2	MT-ATP6	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
25	2	MT-ND4	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
26	2	MT-CO2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
27	2	MT-ATP8	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
28	2	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
29	2	MT-ND5	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
30	2	MT-CYB	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
31	3	CCL5	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ effector memory T-cell (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ central memory T-cell (PBMC) NK-cell (lung)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/onc.2014.18)	CD4+ effector memory T-cell
32	3	CD3D	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ proliferating T-cell (PBMC)			
33	3	GZMA	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ T-cell (lung)			
34	3	LTB	NK-cells & T-cells - Immune response (mainly)	CD4+ central T-cell (PBMC) CD4+ effector memory T-cell (PBMC) CD8+ central memory T-cell (PBMC)			

35	3	PTPRC	Non-specific - Transcription regulation (mainly)	CD4+ effector memory T-cell (lung)			
36	3	TRBC1	NK-cells & T-cells - Immune response (mainly)				
37	3	IL7R	Non-specific - Transcription regulation (mainly)	CD8+ central memory T-cell (PBMC) CD4+ effector memory T-cell (lung)			
38	3	TRBC2	Non-specific - Mitochondria (mainly)				
39	3	IL32	Pancreas - Digestion (mainly)	CD4+ cytotoxic T-cell (PBMC) CD4+ central memory T-cell (PBMC) CD4+ effector memory T-cell (PBMC) CD8+ effector memory T-cell (PBMC)			
40	3	CD2	Non-specific - Transcription regulation (mainly)	CD8+ effector memory T-cell (PBMC) CD4+ effector memory T-cell (lung) CD8+ T-cell (lung) NK-cell (lung)			
41	4	CXCL14	Fibroblasts - ECM organization (mainly)	Intestine-Stromal cells (fetal development)			Fibroblast
42	4	RARRES2	Hepatocytes - Metabolism (mainly)	Alveolar fibroblast (lung)			
43	4	DCN	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Fibroblast (lung)			
44	4	IGFBP5	Smooth muscle cells - ECM organization (mainly)				
45	4	LUM	Fibroblasts - ECM organization (mainly)	Stromal cells (bone marrow) Fibroblast (lung)			
46	4	COL3A1	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Smooth muscle cells (fetal development) Myofibroblast (lung)			
47	4	MMP3	Fibroblasts - ECM organization (mainly)				
48	4	COL6A3	Fibroblasts - ECM organization (mainly)	Smooth muscle cells (fetal development) Myofibroblast (lung)			
49	4	CFD	Fibroblasts - ECM organization (mainly)	Fibroblast (lung)			
50	4	COL1A2	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Myofibroblast (lung)			
51	5	CLCA4	Intestinal epithelial cells - Unknown function (mainly)				Intestinal epithelial cell
52	5	CEACAM7	Intestinal epithelial cells - Unknown function (mainly)				
53	5	MT-ND2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
54	5	DST	Epithelial cell types - Mixed function (mainly)				
55	5	SLC26A3	Intestinal epithelial cells - Unknown function (mainly)				
56	5	MT-ND3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
57	5	MYO15B	Intestinal epithelial cells - Unknown function (mainly)				
58	5	FABP1	Enterocytes - Digestion (mainly)			Marker of colon enterocyte (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
59	5	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
60	5	MUC12	Intestinal epithelial cells - Unknown function (mainly)				
61	6	S100A9	Monocytes & Neutrophils - Innate immune response (mainly)	CD14+ monocyte (PBMC) CD14+ monocyte (lung) Neutrophil (kidney)			Myeloid cell
62	6	HLA-DRA	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development) Myeloid dendritic cell (lung)			
63	6	LYZ	Monocytes & Neutrophils - Innate immune response (mainly)	Monocyte (PBMC) CD14+ monocyte (PBMC) Stomach-Goblet cells (fetal development)			

64	6	HLA-DPA1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Myeloid dendritic cell (lung)			
65	6	CD74	Macrophages - Immune response (mainly)	B-cell (PBMC) Dendritic cell (PBMC) Antigen-presenting cells (fetal development)			
66	6	HLA-DPB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development) Myeloid dendritic cell (lung)			
67	6	HLA-DRB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development)			
68	6	TYROBP	Macrophages - Innate immune response (mainly)	Monocyte (PBMC) NK-cell (PBMC) Dendritic cell (lung)			
69	6	FTL	Macrophages - Innate immune response (mainly)	Myeloid cell (fetal development) Neutrophil (kidney)			
70	6	AIF1	Monocytes & Neutrophils - Innate immune response (mainly)	Monocyte (PBMC) CD16+ monocyte (PBMC) CD16+ monocyte (lung)			
71	7	REG1A	Pancreas - Digestion (mainly)				Intestinal epithelial cell
72	7	AGR2	Mucus-secreting cells - Mucin production (mainly)			Marker of crypt-resident goblet cells (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
73	7	OLFM4	Intestinal epithelial cells - Unknown function (mainly)			Marker of intestinal stem cells (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
74	7	TFF3	Mucus-secreting cells - Mucin production (mainly)			Classical marker of goblet cell (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
75	7	LCN2	Respiratory epithelial cells - Mucosal defense (mainly)	Mucous cell (lung) Goblet cell (lung)			
76	7	GPX2	Intestinal epithelial cells - Unknown function (mainly)				
77	7	RPL7	Non-specific - Translation (mainly)				
78	7	RPL8	Non-specific - Translation (mainly)				
79	7	TSPAN8	Enterocytes - Digestion (mainly)				
80	7	EPCAM	Intestinal epithelial cells - Unknown function (mainly)				
81	8	FABP1	Enterocytes - Digestion (mainly)			Marker of colon enterocyte (https://doi.org/10.1016/j.jcmgh.2022.02.007)	Intestinal enterocyte cell
82	8	TFF1	Pancreatic endocrine cells - Mixed function (mainly)				
83	8	CEACAM7	Intestinal epithelial cells - Unknown function (mainly)				
84	8	PHGR1	Enterocytes - Digestion (mainly)				
85	8	CA4	Intestinal epithelial cells - Unknown function (mainly)				
86	8	CLCA4	Intestinal epithelial cells - Unknown function (mainly)	Basal cell (lung)		Marker of colon enterocyte (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
87	8	KRT19	Respiratory epithelial cells - Mucosal defense (mainly)	Stomach-Squamos epithelial cells (fetal development) Basal cells (lung)			
88	8	GUCA2A	Enterocytes - Digestion (mainly)				
89	8	TSPAN1	Intestinal epithelial cells - Unknown function (mainly)				
90	8	PIGR	Intestinal epithelial cells - Unknown function (mainly)				
91	9	SPARCL1	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Bronchial epithelial vessel (lung)			Vascular endothelial cell
92	9	COL4A1	Smooth muscle cells - ECM organization (mainly)	Vascular endothelial cell (fetal development) Pericyte (lung)			
93	9	PLVAP	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Intestine-Vascular endothelial cells (fetal development)			

94	9	HSPG2	Granulosa cells - Unknown function (mainly)	Vascular endothelial cells (fetal development) Endocardial cells (fetal development)			
95	9	FLT1	Syncytiotrophoblasts - Pregnancy hormone signaling (mainly)	Vascular endothelial cell (fetal development)			
96	9	COL4A2	Smooth muscle cells - ECM organization (mainly)	Vascular endothelial cell (fetal development) Pericyte (lung)			
97	9	IGFBP3	Fibroblasts - ECM organization (mainly)	Stellate cells (fetal development)			
98	9	VWF	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Lymphatic or vascular endothelial cells (fetal development) Bronchial vessel endothelial cell (lung)			
99	9	EGFL7	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Hematopoietic stem and progenitor cell (PBMC) Peritubular capillary endothelial cell (kidney)			
100	9	PECAM1	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Capillary endothelial cell (lung)			
101	10	RGS5	Smooth muscle cells - ECM organization (mainly)	Sympathoblasts (fetal development) Chromaffin cells (fetal development) Smooth muscle cells (fetal development) Vascular smooth muscle cell (kidney)			Vascular smooth muscle cell
102	10	THY1	Fibroblasts - ECM organization (mainly)	Vascular smooth muscle cell (Azimuth lung)			
103	10	SPARC	Fibroblasts - ECM organization (mainly)				
104	10	IGFBP7	Smooth muscle cells - ECM organization (mainly)	Vascular or lymphatic endothelial cells (fetal development) Lymphatic vessel cells (lung)			
105	10	COL1A1	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Smooth muscle cells (fetal development) Myofibroblast (lung)			
106	10	TIMP1	Fibroblasts - ECM organization (mainly)				
107	10	CALD1	Smooth muscle cells - ECM organization (mainly)	Smooth muscle cells (fetal development Smooth muscle cells (lung) Myofibroblast (lung)			
108	10	ACTA2	Smooth muscle cells - ECM organization (mainly)	Intestine-Smooth muscle cell (fetal development) Smooth muscle cell (lung) Vascular associated smooth muscle cell (lung)			
109	10	COL1A2	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Myofibroblast (lung)			
110	10	COL4A1	Smooth muscle cells - ECM organization (mainly)	Vascular endothelial cell (fetal development) Pericyte (lung)			

Supplementary Table 9: Manually annotated clusters in sample 569 based on its top 10 cell-type-specific marker genes.

#	cluster	gene	Human Protein Atlas (HPA)	Azimuth	Panglaodb.se	Supplementary literature or other comment	Annotated cell type
1	0	CCL5	NK-cells & T-cells - Immune response (mainly)	CD4+ T-cell (PBMC) CD8+ effector memory T-cell (PBMC) CD4+ cytotoxic T-cell (PBMC) CD8+ central memory T-cell (PBMC) NK-cell (lung)		Differentiates CD8+ from CD4+ (https://doi.org/10.1038/s41467-019-12464-3)	CD8+ effector memory T-cell
2	0	IL7R	Non-specific - Transcription regulation (mainly)	CD8+ central memory T-cell (PBMC) CD4+ effector memory T-cell (lung)			
3	0	KLRB1	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ central memory T-cell (PBMC) NK-cell (PBMC) CD4+ T-cell (bone marrow) CD8+ effector T-cell (bone marrow)			
4	0	ID2	Monocytes - Immune response regulation (mainly)	CD8+ naive T-cell (PBMC)			
5	0	CD3D	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ proliferating T-cell (PBMC)			
6	0	FYB1	Non-specific - Mitochondria (mainly)	CD4+ effector memory T-cell (PBMC)			
7	0	CD3G	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory cell (PBMC) CD8+ proliferating T-cell (PBMC)			
8	0	TRAC	NK-cells & T-cells - Immune response (mainly)	CD4+ central memory T-cell (PBMC) CD4+ effector memory T-cell (PBMC) CD4+ proliferating T-cell (PBMC) CD8+ central memory T-cell (PBMC) CD8+ effector memory T-cell (PBMC) CD8+ proliferating T-cell (PBMC)			
9	0	HSPA6	Epithelial cell types - Mixed function (mainly)				
10	0	TRBC1	NK-cells & T-cells - Immune response (mainly)				
11	1	IGKC	Plasma cells - Humoral immune response (mainly)				Plasma B-cell
12	1	IGLC3	Plasma cells - Humoral immune response (mainly)				
13	1	JCHAIN	Plasma cells - Humoral immune response (mainly)				
14	1	IGLC2	Plasma cells - Humoral immune response (mainly)				
15	1	IGLC1	Alveolar cells - Smell perception (mainly)				
16	1	IGHA1	Plasma cells - Humoral immune response (mainly)				
17	1	IGHA2	Plasma cells - Humoral immune response (mainly)				
18	1	MZB1	Plasma cells - Humoral immune response (mainly)				
19	1	SSR4	Plasma cells - Humoral immune response (mainly)				
20	1	DERL3	Plasma cells - Humoral immune response (mainly)				
21	2	HLA-DRA	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells Antigen-presenting cells (fetal development) Myeloid dendritic cell (lung)			Dendritic cell or B-cell
22	2	MS4A1	B-cells - Immune response (mainly)	B-cell (PBMC B-cell (lung)	B-cell		
23	2	CD74	Macrophages - Immune response (mainly)	B-cell (PBMC) Dendritic cell (PBMC) Antigen-presenting cells (fetal development)			
24	2	HLA-DRB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development)	Dendritic cell B-cell		
25	2	HLA-DPA1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Myeloid dendritic cell (lung)	Dendritic cell B-cell		
26	2	HLA-DPB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development) Myeloid dendritic cell (lung)	Dendritic cell B-cell		
27	2	VPREB3	B-cells - Immune response (mainly)	B-cell (lung)	B-cell		
28	2	BANK1	B-cells - Immune response (mainly)	B-cell (PBMC) B-cell (lung) B-cell (kidney)	B-cell		

29	2	HLA-DQA1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) B-cell (PBMC) Stomach-Myeloid cells (fetal development) Dendritic cell (lung) Myeloid dendritic cell (lung)			
30	2	CD37	B-cells - Immune response (mainly)	Naive B-cell (PBMC)			
31	3	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	Mitochondrial gene-expressing cell
32	3	MT-CO2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
33	3	MT-CO1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
34	3	MT-ND2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
35	3	MT-CYB	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
36	3	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
37	3	MT-ATP6	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
38	3	MT-ND4	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
39	3	MT-ND3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
40	3	MT-ND5	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
41	4	IGHA1	Plasma cells - Humoral immune response (mainly)	Plasma cell (PBMC) Plasma cell (kidney)			Unknown cell type
42	4	IGHA2	Plasma cells - Humoral immune response (mainly)	Plasmablast (PBMC) Memory B-cell (PBMC)			
43	4	BTG2	Non-specific - Mitochondria (mainly)				
44	4	XBP1	Glandular cells - Unknown function (mainly)	Plasma cells (lung)			
45	4	PPP1R15A	Non-specific - Mitochondria (mainly)				
46	4	JUN	Non-specific - Mitochondria (mainly)	Ductal cells (fetal development)			
47	4	SLC38A2	Non-specific - Mitochondria (mainly)				
48	4	SQSTM1	Epithelial cell types - Mixed function (mainly)				
49	4	MALAT1			Unknown		
50	4	GLS	Proximal tubular cells - Tubular reabsorption (mainly)				
51	5	GZMA	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell CD4+ cytotoxic T-cell (PBMC) CD8+ T-cell (lung)			T-cell
52	5	S100A4	Monocytes & Neutrophils - Innate immune response (mainly)	CD4+ cytotoxic (PBMC) CD4+ central memory (PBMC) CD4+ effector memory T-cell (PBMC) NK-cell (PBMC)			
53	5	CD3D	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory T-cell (PBMC) CD8+ proliferating T-cell (PBMC)			
54	5	IL32	Pancreas - Digestion (mainly)	CD4+ T-cell (PBMC) CD8+ effector memory T-cell (PBMC) CD8+ effector memory cell (lung)			
55	5	TRBC2	Non-specific - Mitochondria (mainly)				
56	5	TNFRSF18	Non-specific - Transcription regulation (mainly)				
57	5	CD2	Non-specific - Transcription regulation (mainly)	CD4+ effector memory T-cell (lung)			
58	5	CD3G	NK-cells & T-cells - Immune response (mainly)	CD4+ effector memory cell (PBMC) CD8+ proliferating T-cell (PBMC)			
59	5	ACTB	Monocytes - Immune response regulation (mainly)				
60	5	CD7	NK-cells & T-cells - Immune response (mainly)				
61	6	CXCL14	Fibroblasts - ECM organization (mainly)	Intestine-Stromal cells (fetal development)			Vascular smooth muscle cell
62	6	CALD1	Smooth muscle cells - ECM organization (mainly)	Smooth muscle cells (fetal development) Smooth muscle cells (lung) Myofibroblast (lung)			
63	6	COL3A1	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Smooth muscle cells (fetal development) Myofibroblast (lung)			
64	6	TAGLN	Smooth muscle cells - ECM organization (mainly)	Intestine-Smooth muscle cells (fetal development) Vascular smooth muscle cell (lung)			
65	6	IGFBP7	Smooth muscle cells - ECM organization (mainly)	Lymphatic or vascular endothelial cells (fetal development) Lymphatic or vascular endothelial cells (lung)		Is expressed in cancer-associated fibroblasts and tumor vessels (https://doi.org/10.1038/onc.2014.18)	
66	6	COL1A2	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Myofibroblast (lung)			

67	6	IGFBP5	Smooth muscle cells - ECM organization (mainly)				
68	6	COL1A1	Fibroblasts - ECM organization (mainly)	Stromal cells (fetal development) Smooth muscle cells (fetal development) Myofibroblast (lung)			
69	6	TIMP1	Fibroblasts - ECM organization (mainly)				
70	6	RARRES2	Hepatocytes - Metabolism (mainly)	Alveolar fibroblast (lung)			
71	7	REG1A	Pancreas - Digestion (mainly)				
72	7	PIGR	Intestinal epithelial cells - Unknown function (mainly)		Cholangiocytes or epithelial cells		
73	7	OLFM4	Intestinal epithelial cells - Unknown function (mainly)			Marker of intestinal stem cells (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
74	7	AGR2	Mucus-secreting cells - Mucin production (mainly)	Goblet cells (fetal development Goblet cells (lung)		Marker of crypt-resident goblet cells (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
75	7	LCN2	Respiratory epithelial cells - Mucosal defense (mainly)	Mucous cell (lung) Goblet cell (lung)			Intestinal epithelial cell
76	7	PLA2G2A	Fibroblasts - ECM organization (mainly)	Epicardial fat cells (fetal development)			
77	7	REG1B	Pancreas - Digestion (mainly)				
78	7	SPINK4	Mucus-secreting cells - Mucin production (mainly)				
79	7	DMBT1	Enterocytes - Digestion (mainly)	Stomach-MUC13/DMBT1 positive cells (fetal development)			
80	7	TSPAN8	Enterocytes - Digestion (mainly)	Basal cell (lung)			
81	8	CEACAM5	Intestinal epithelial cells - Unknown function (mainly)				
82	8	TFF3	Mucus-secreting cells - Mucin production (mainly)	Lymphatic vessel (lung)		Classical marker of goblet cell (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
83	8	EPCAM	Intestinal epithelial cells - Unknown function (mainly)	Hematopoeitic stem and progenitor cell (bone marrow) Late erythroid (bone marrow)			
84	8	FXYD3	Intestinal epithelial cells - Unknown function (mainly)				
85	8	FABP1	Enterocytes - Digestion (mainly)		Cholangiocytes or epithelial cells		Intestinal epithelial cell
86	8	OLFM4	Intestinal epithelial cells - Unknown function (mainly)			Marker of intestinal stem cells (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
87	8	KRT18	Pancreas - Digestion (mainly)				
88	8	LGALS4	Intestinal epithelial cells - Unknown function (mainly)		Cholangiocytes or epithelial cells		
89	8	LGALS3	Intestinal epithelial cells - Unknown function (mainly)	Intestinal epithelial cell (fetal development)			
90	8	KRT8	Pancreatic endocrine cells - Mixed function (mainly)		Cholangiocytes or epithelial cells		
91	9	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
92	9	SOX9	Squamous epithelial cells - Cornification (mainly)				
93	9	FABP1	Enterocytes - Digestion (mainly)		Cholangiocytes or epithelial cells		
94	9	ELF3	Respiratory epithelial cells - Mucosal defense (mainly)	Stomach-Goblet cells (fetal development) Squamous epithelial cells (fetal development)			
95	9	MT-CO1	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	Intestinal epithelial cell
96	9	L1TD1	Cytotrophoblasts - Unknown function (mainly)				
97	9	KRT8	Pancreatic endocrine cells - Mixed function (mainly)		Cholangiocytes or epithelial cells		
98	9	MT-CO2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
99	9	MT-ATP6	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial gene	
100	9	PHGR1	Enterocytes - Digestion (mainly)	Stomach-MUC13/DMBT1 positive cells (fetal development)			
101	10	PLVAP	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Intestine-Vascular endothelial cells (fetal development)			
102	10	SPARCL1	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Bronchial epithelial vessel (lung)			
103	10	COL4A1	Smooth muscle cells - ECM organization (mainly)	Vascular endothelial cell (fetal development) Pericyte (lung)			
104	10	IGFBP3	Fibroblasts - ECM organization (mainly)	Stellate cells (fetal development)			
105	10	HSPG2	Granulosa cells - Unknown function (mainly)	Vascular endothelial cells (fetal development) Endocardial cells (fetal development)			
106	10	IGFBP7	Smooth muscle cells - ECM organization (mainly)	Vascular endothelial cell (fetal development) Lymphatic vessel (lung)			Vascular endothelial cell
107	10	FLT1	Syncytiotrophoblasts - Pregnancy hormone signaling (mainly)	Vascular endothelial cell (fetal development)			
108	10	VWF	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Lymphatic or vascular endothelial cells (fetal development) Bronchial vessel endothelial cell (lung)			
109	10	PECAM1	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Capillary endothelial cell (lung)			
110	10	IFI27	Pancreatic endocrine cells - Mixed function (mainly)	Endothelial cell (motor cortex)	Endothelial cells, choangiocytes or epithelial cells		

111	11	FABP1	Enterocytes - Digestion (mainly)			Marker of colon enterocyte (https://doi.org/10.1016/j.jcmgh.2022.02.007)	
112	11	PHGR1	Enterocytes - Digestion (mainly)	Stomach-MUC13/DMBT1 positive cells (fetal development)	Cholangiocytes or epithelial cells		
113	11	LGALS4	Intestinal epithelial cells - Unknown function (mainly)		Cholangiocytes or epithelial cells		
114	11	MT1G	Proximal tubular cells - Tubular reabsorption (mainly)				
115	11	C15orf48	Pancreatic endocrine cells - Mixed function (mainly)	Classical monocyte (lung)			Intestinal epithelial cell
116	11	TSPAN8	Enterocytes - Digestion (mainly)	Basal cell (lung)			
117	11	LGALS3	Intestinal epithelial cells - Unknown function (mainly)	Intestinal epithelial cell (fetal development)			
118	11	MT1E	Proximal tubular cells - Tubular reabsorption (mainly)				
119	11	SRI	Intestinal epithelial cells - Unknown function (mainly)				
120	11	CD24	Mammary glandular cells - Lactation (mainly)				
121	12	LYZ	Monocytes & Neutrophils - Innate immune response (mainly)	Monocyte (PBMC) CD14+ monocyte (PBMC) Stomach-Goblet cells (fetal development)	Dendritic cells or monocytes		Myeloid cell
122	12	APOE	Smooth muscle cells - Unknown function (mainly)				
123	12	C1QA	Macrophages - Innate immune response (mainly)	Macrophage (lung)	Dendritic cells		
124	12	AIF1	Monocytes & Neutrophils - Innate immune response (mainly)	Monocyte (PBMC) CD16+ monocyte (PBMC) CD16+ monocyte (lung)	Dendritic cells		
125	12	HLA-DRB1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development)	Dendritic cells		
126	12	C1QB	Macrophages - Innate immune response (mainly)	Macrophage (lung)	Dendritic cells		
127	12	HLA-DRA	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Antigen-presenting cells (fetal development) Myeloid dendritic cell (lung)	Dendritic cells		
128	12	HLA-DPA1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) Myeloid cells (fetal development) Myeloid dendritic cell (lung)	Dendritic cells		
129	12	HLA-DQA1	Macrophages - Immune response (mainly)	Dendritic cell (PBMC) B-cell (PBMC) Stomach-Myeloid cells (fetal development) Dendritic cell (lung) Myeloid dendritic cell (lung)	Dendritic cells		
130	12	CTSD	Macrophages - Innate immune response (mainly)	CD14+ monocyte (PBMC)	Dendritic cells		

Supplementary Table 10: Top 10 expressed cell-type-specific marker genes identified for each cluster in the integrated dataset.

#	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	5.5577360104809e-35	0.996435946114834	0.935	0.68	1.11154720209618e-31	B-cell	KLF2
2	3.94592222935048e-37	0.900613230895783	0.865	0.387	7.89184445870095e-34	B-cell	MS4A1
3	1.59403104172259e-50	0.846955564355128	0.978	0.569	3.18806208344518e-47	B-cell	HLA-DPA1
4	2.29881789388497e-31	0.839810267465021	0.87	0.46	4.59763578776993e-28	B-cell	CD79A
5	2.4398546266429e-31	0.795144627645023	0.973	0.803	4.8797092532858e-28	B-cell	HSPD1
6	1.24198572779814e-29	0.764971051050715	0.973	0.847	2.48397145559629e-26	B-cell	DUSP1
7	2.33488277991728e-22	0.759813039730757	0.93	0.763	4.66976555983456e-19	B-cell	ID3
8	1.53848022757692e-46	0.711566478652121	0.897	0.375	3.07696045515384e-43	B-cell	VPREB3
9	3.49179055420791e-26	0.666711641239115	0.946	0.652	6.98358110841582e-23	B-cell	CD52
10	9.13132644853588e-58	0.65021593469735	0.93	0.436	1.82626528970718e-54	B-cell	PHACTR1
11	0	1.86554296924995	0.97	0.571	0	CD4_EM_T	IL7R
12	0	1.60693656955151	0.998	0.61	0	CD4_EM_T	KLRB1
13	0	1.5469605959572	0.97	0.575	0	CD4_EM_T	CD3D
14	0	1.53820521554124	0.938	0.602	0	CD4_EM_T	TRAC
15	0	1.53807753905134	0.834	0.527	0	CD4_EM_T	TRBC2
16	0	1.43139737021653	0.945	0.585	0	CD4_EM_T	CD3G
17	0	1.38952867874797	0.796	0.515	0	CD4_EM_T	TRBC1
18	0	1.36904420934468	0.983	0.812	0	CD4_EM_T	FOS
19	0	1.31234992150082	0.868	0.52	0	CD4_EM_T	GIMAP7
20	0	1.21752943051763	0.958	0.598	0	CD4_EM_T	SRGN
21	9.23352809196374e-123	1.94246350623607	0.822	0.42	1.84670561839275e-119	CD4_P_T	MKI67
22	3.08990918033407e-119	1.69569388873018	0.824	0.341	6.17981836066813e-116	CD4_P_T	CENPF
23	3.08504797459573e-157	1.65618019054618	0.923	0.439	6.17009594919147e-154	CD4_P_T	STMN1
24	2.23146206882805e-126	1.64365747659448	0.824	0.413	4.46292413765611e-123	CD4_P_T	TOP2A
25	6.56534944553024e-105	1.62429166436307	0.899	0.504	1.31306988910605e-101	CD4_P_T	HMGB2
26	1.27163261130236e-124	1.6177781455562	0.978	0.636	2.54326522260473e-121	CD4_P_T	CEACAM5
27	3.0182258352702e-109	1.49707698762823	0.955	0.614	6.0364516705404e-106	CD4_P_T	EPCAM
28	9.80078366842919e-105	1.44766234361236	0.963	0.697	1.96015673368584e-101	CD4_P_T	H2AFZ
29	1.59614031540028e-109	1.36218758933709	0.917	0.523	3.19228063080057e-106	CD4_P_T	TUBB
30	6.11735704237928e-69	1.34288107462466	0.724	0.201	1.22347140847586e-65	CD4_P_T	ASPM
31	0	3.02467482007889	0.71	0.636	0	CD8_EM_T	CCL5
32	4.18342004100914e-86	2.18834440214203	0.48	0.509	8.36684008201828e-83	CD8_EM_T	CCL4
33	0	1.88007394355982	0.775	0.659	0	CD8_EM_T	KLRB1
34	3.04423053193186e-34	1.81313503152468	0.483	0.568	6.08846106386372e-31	CD8_EM_T	IFNG
35	9.69446919488298e-97	1.81256422906357	0.508	0.574	1.9388938389766e-93	CD8_EM_T	GZMA
36	0	1.68187260683061	0.837	0.612	0	CD8_EM_T	CD3D
37	0	1.63096354325159	0.783	0.616	0	CD8_EM_T	IL7R
38	0	1.59497756047185	0.741	0.642	0	CD8_EM_T	CD2
39	4.04868489608353e-197	1.58672660403432	0.609	0.554	8.09736979216705e-194	CD8_EM_T	TRBC1
40	0	1.5084744510053	0.726	0.632	0	CD8_EM_T	CD3G
41	0	2.96587367445765	0.963	0.424	0	Dendritic or B-cell	HLA-DRA
42	0	2.55904014181327	0.847	0.299	0	Dendritic or B-cell	MS4A1
43	0	2.36036407812434	0.932	0.395	0	Dendritic or B-cell	HLA-DRB1
44	0	2.11824812593022	0.92	0.501	0	Dendritic or B-cell	HLA-DPA1
45	0	1.99826190387037	0.872	0.405	0	Dendritic or B-cell	HLA-DPB1
46	0	1.95469265649303	0.877	0.643	0	Dendritic or B-cell	KLF2
47	0	1.91370144609718	0.778	0.299	0	Dendritic or B-cell	VPREB3
48	0	1.88075266096785	0.813	0.393	0	Dendritic or B-cell	CD79A
49	0	1.79379022320489	0.824	0.462	0	Dendritic or B-cell	HLA-DQA1
50	0	1.40188337482887	0.783	0.479	0	Dendritic or B-cell	HLA-DQB1
51	0	4.81070367924788	0.939	0.44	0	Fibroblast	CXCL14
52	0	4.10986701227108	0.958	0.385	0	Fibroblast	COL3A1
53	0	4.03359393452223	0.987	0.39	0	Fibroblast	CALD1
54	0	3.9866866862772	0.952	0.326	0	Fibroblast	COL1A2
55	0	3.77688833034442	0.97	0.344	0	Fibroblast	RARRES2
56	3.40929491818435e-166	3.74973490569558	0.814	0.386	6.8185898363687e-163	Fibroblast	LUM
57	1.08076907765694e-237	3.74262073607613	0.877	0.281	2.16153815531387e-234	Fibroblast	IGFBP5
58	0	3.71753604642614	0.982	0.439	0	Fibroblast	IGFBP7
59	0	3.69808689840023	0.935	0.405	0	Fibroblast	COL1A1
60	5.22326635295981e-182	3.58495819755064	0.826	0.242	1.04465327059196e-178	Fibroblast	DCN
61	2.53300143300428e-166	3.77650588053881	0.987	0.661	5.06600286600856e-163	Intestinal enterocyte	FABP1
62	3.63883798273329e-158	3.31323429058953	0.98	0.516	7.27767596546659e-155	Intestinal enterocyte	PHGR1

63	2.02504351040697e-160	3.03464242314781	0.99	0.604	4.05008702081394e-157	Intestinal enterocyte	KRT19
64	6.86087540058614e-135	2.76845599546363	0.934	0.406	1.37217508011723e-131	Intestinal enterocyte	KRT20
65	3.43681404888244e-156	2.73274323337295	0.997	0.586	6.87362809776488e-153	Intestinal enterocyte	KRT8
66	1.0535378941605e-148	2.65529878833842	0.99	0.715	2.10707578832099e-145	Intestinal enterocyte	LGALS3
67	6.09913100367696e-149	2.63552470785492	0.997	0.803	1.21982620073539e-145	Intestinal enterocyte	S100A6
68	1.53837780650731e-140	2.62433210146743	0.98	0.639	3.07675561301461e-137	Intestinal enterocyte	CEACAM5
69	1.2226847729111e-135	2.52424888438844	0.94	0.43	2.44536954582219e-132	Intestinal enterocyte	TSPAN1
70	1.14426947563516e-141	2.46323011091727	0.99	0.63	2.28853895127032e-138	Intestinal enterocyte	PIGR
71	0	3.25171560953615	0.959	0.592	0	Intestinal epithelial	PIGR
72	0	3.12872452734832	0.863	0.639	0	Intestinal epithelial	FABP1
73	0	3.12727062412888	0.918	0.583	0	Intestinal epithelial	EPCAM
74	0	3.12698366920457	0.958	0.628	0	Intestinal epithelial	LGALS4
75	0	3.11231934739894	0.897	0.61	0	Intestinal epithelial	CEACAM5
76	0	2.92841028015244	0.916	0.577	0	Intestinal epithelial	AGR2
77	0	2.92835908338668	0.812	0.567	0	Intestinal epithelial	TFF3
78	0	2.90616051766875	0.911	0.573	0	Intestinal epithelial	TSPAN8
79	0	2.85687136420903	0.78	0.601	0	Intestinal epithelial	OLFM4
80	0	2.72537733788794	0.932	0.547	0	Intestinal epithelial	KRT8
81	1.92027650602407e-12	2.15912661010369	0.411	0.285	3.84055301204814e-09	Intestinal goblet	SPINK4
82	1.55438695347356e-08	2.03190509407752	0.432	0.336	3.10877390694711e-05	Intestinal goblet	MUC2
83	2.73329355554356e-208	1.73800881438483	0.958	0.555	5.46658711108712e-205	Intestinal goblet	ELF3
84	7.42199490669518e-219	1.71996386090657	1	0.996	1.48439898133904e-215	Intestinal goblet	MT-CO3
85	1.77645019139611e-76	1.62104982012098	0.798	0.59	3.55290038279221e-73	Intestinal goblet	TFF3
86	4.2806002304405e-170	1.53027916430263	0.941	0.625	8.56120046088099e-167	Intestinal goblet	PIGR
87	2.51064108980983e-194	1.52359063520726	0.995	0.989	5.02128217961967e-191	Intestinal goblet	MT-ND4
88	1.90470307802877e-192	1.48407528992806	0.998	0.995	3.80940615605754e-189	Intestinal goblet	MT-CO1
89	1.54840556343501e-133	1.45933781743717	0.876	0.469	3.09681112687001e-130	Intestinal goblet	KRT18
90	2.58027783195161e-191	1.44655503905425	1	0.996	5.16055566390321e-188	Intestinal goblet	MT-CO2
91	3.12671880277615e-52	4.72139276757755	0.86	0.334	6.2534376055523e-49	Monocyte	LYZ
92	3.95569297438669e-41	4.59600743635357	0.86	0.332	7.91138594877338e-38	Monocyte	APOE
93	4.79407846051166e-11	3.95565702626364	0.647	0.283	9.58815692102333e-08	Monocyte	CXCL8
94	1.29921161171567e-53	3.73631733237577	0.86	0.264	2.59842322343133e-50	Monocyte	C1QA
95	6.04628513278047e-36	3.67174241672702	0.747	0.194	1.20925702655609e-32	Monocyte	C1QB
96	1.43728788722754e-45	3.656286261915	0.833	0.249	2.87457577445507e-42	Monocyte	S100A9
97	3.87667076028938e-32	3.49808609166163	0.753	0.289	7.75334152057876e-29	Monocyte	AIF1
98	1.74845042461976e-63	3.35601407672422	0.927	0.436	3.49690084923952e-60	Monocyte	FCER1G
99	4.07497227159806e-33	3.31047714030176	0.787	0.381	8.14994454319611e-30	Monocyte	TYROBP
100	8.28642436859483e-13	3.28488744880362	0.653	0.094	1.65728487371897e-09	Monocyte	IL1B
101	0	2.51017142649539	1	0.983	0	MT	MT-ND1
102	0	2.37905182409657	1	0.964	0	MT	MT-ND2
103	0	2.36631932872781	1	0.996	0	MT	MT-CO3
104	0	2.32967929128179	1	0.996	0	MT	MT-CO2
105	0	2.24431378621784	1	0.994	0	MT	MT-CO1
106	0	2.23838812901709	1	0.988	0	MT	MT-ND4
107	0	1.31239728064642	0.907	0.532	0	MT	ELF3
108	0	1.02170731566387	0.776	0.463	0	MT	AC103702.2
109	1.1124259993099e-265	0.955162509827864	0.882	0.725	2.2248519986198e-262	MT	SOX4
110	6.63227240337852e-265	0.916514899390364	0.702	0.445	1.3264544806757e-261	MT	MUC4
111	6.65714406040975e-21	4.06358417034459	0.588	0.332	1.33142881208195e-17	Myeloid	LYZ
112	1.23005474645413e-104	3.32840417688972	0.868	0.516	2.46010949290827e-101	Myeloid	HLA-DQA1
113	0.000183988940907665	3.17104672653129	0.336	0.161	0.36797788181533	Myeloid	SPP1
114	9.4182742397084e-80	3.16112095507646	0.82	0.479	1.88365484794168e-76	Myeloid	HLA-DRB1
115	1.11973374698121e-28	3.15568810157338	0.485	0.263	2.23946749396242e-25	Myeloid	C1QA
116	9.79684744656628e-101	3.10560878330259	0.899	0.565	1.95936948931326e-97	Myeloid	HLA-DPA1
117	1.04448525623566e-69	3.09173341320345	0.781	0.478	2.08897051247131e-66	Myeloid	HLA-DPB1
118	2.43160033857019e-106	3.06347365799398	0.886	0.507	4.86320067714037e-103	Myeloid	HLA-DRA
119	4.46216142433095e-34	2.90825243240973	0.594	0.286	8.9243228486619e-31	Myeloid	AIF1
120	1.38277895809276e-26	2.80626503581838	0.816	0.645	2.76555791618553e-23	Myeloid	CTSD
121	1.00318725714639e-83	5.75579055775936	0.68	0.668	2.00637451429279e-80	Plasma B-cell	IGLC2
122	0	5.72994893687575	0.985	0.913	0	Plasma B-cell	IGKC
123	0	5.58935838228394	0.994	0.741	0	Plasma B-cell	JCHAIN
124	1.63962384895331e-31	5.5885964720522	0.612	0.639	3.27924769790662e-28	Plasma B-cell	IGLC3
125	0	5.53911875377749	0.873	0.441	0	Plasma B-cell	IGLC1
126	0	5.15754195743015	0.992	0.86	0	Plasma B-cell	IGHA1
127	0	4.69422498199663	0.903	0.653	0	Plasma B-cell	IGHA2
128	1.20519118349 63e-09	4.20036478284454	0.42	0.461	2.41038236699261e-06	Plasma B-cell	IGHG1
129	1.30731774944494e-13	3.72056069358559	0.279	0.414	2.61463549888988e-10	Plasma B-cell	IGHGP

130	0	3.70205587188969	0.988	0.567	0	Plasma B-cell	MZB1
131	1.10109967700586e-29	4.96278550685082	1	0.153	2.20219935401172e-26	Smooth muscle	RGS5
132	1.10881968002976e-27	4.58097586995537	1	0.454	2.21763936005952e-24	Smooth muscle	IGFBP7
133	1.45233716862105e-27	4.20535612449817	1	0.406	2.9046743372421e-24	Smooth muscle	CALD1
134	6.15108326762987e-28	4.113067374549	1	0.495	1.23021665352597e-24	Smooth muscle	MGP
135	1.43379378720324e-27	3.98226937487447	1	0.483	2.86758757440649e-24	Smooth muscle	ACTA2
136	2.19335176592187e-24	3.81648588889668	0.976	0.415	4.38670353184374e-21	Smooth muscle	TAGLN
137	8.60933184885635e-28	3.63403251452961	1	0.353	1.72186636977127e-24	Smooth muscle	MYL9
138	7.72554472538298e-25	3.52574963901268	0.976	0.342	1.54510894507666e-21	Smooth muscle	COL4A1
139	2.36033680817845e-27	3.51807275446319	1	0.45	4.72067361635691e-24	Smooth muscle	C11orf96
140	8.55974475598098e-28	3.46886134563415	1	0.49	1.7119489511962e-24	Smooth muscle	ADIRF
141	2.86482610300301e-11	1.89150971834611	0.49	0.568	5.72965220600602e-08	T-cell	GZMA
142	9.03783569719231e-161	1.83671762253201	0.875	0.633	1.80756713943846e-157	T-cell	CD3D
143	5.60133470888849e-139	1.78316248498142	0.871	0.646	1.1202669417777e-135	T-cell	S100A4
144	0.00936650785605324	1.69136526079055	0.331	0.511	1	T-cell	IL17A
145	5.91249239387696e-106	1.62445771664907	0.786	0.65	1.18249847877539e-102	T-cell	CD2
146	3.27227561899048e-94	1.56329796523075	0.746	0.572	6.54455123798097e-91	T-cell	TRBC2
147	2.25704161563916e-50	1.54858505560029	0.621	0.559	4.51408323127833e-47	T-cell	TRBC1
148	3.28644920878662e-97	1.52735109331089	0.777	0.64	6.57289841757324e-94	T-cell	CD3G
149	1.64089069133169e-127	1.52695191763789	0.847	0.651	3.28178138266339e-124	T-cell	TRAC
150	2.36560378611126e-42	1.45961094129109	0.548	0.467	4.73120757222251e-39	T-cell	CD7
151	7.22113849916778e-09	1.54176251722609	0.748	0.879	1.44422769983356e-05	Unknown	IGHA1
152	3.10199510073198e-22	1.11727945558532	0.39	0.401	6.20399020146397e-19	Unknown	IGHGP
153	9.44256258531742e-05	1.05755302892704	0.475	0.55	0.188851251706348	Unknown	XBP1
154	5.85030369611448e-31	1.00060462049001	0.31	0.604	1.1700607392229e-27	Unknown	CCDC144A
155	5.86683417080393e-07	0.976482825361853	0.251	0.353	0.00117336683416079	Unknown	IGHG2
156	7.02518907634328e-10	0.962968539061084	0.416	0.459	1.40503781526866e-06	Unknown	IGHG1
157	3.52718732273558e-07	0.909380778469833	0.377	0.476	0.000705437464547115	Unknown	IGHG3
158	0.00449281262785531	0.886989091864964	0.539	0.631	1	Unknown	RRBP1
159	1.82185636678404e-06	0.82817188374692	0.451	0.526	0.00364371273356808	Unknown	JSRP1
160	2.84547334999246e-17	0.787724399446066	0.657	0.895	5.69094669998492e-14	Unknown	HSPB1
161	9.64115859732082e-194	4.16386100316324	0.847	0.343	1.92823171946416e-190	Vascular endothelial	PLVAP
162	2.004295783419e-162	4.15117393549922	0.863	0.545	4.008591566838e-159	Vascular endothelial	SPARCL1
163	9.0303541465262e-211	3.94719843306475	0.914	0.333	1.80607082930524e-207	Vascular endothelial	COL4A1
164	1.9973155887113e-179	3.4730216300525	0.874	0.481	3.9946311774226e-176	Vascular endothelial	HSPG2
165	1.25257296522927e-218	3.47050241861554	0.96	0.445	2.50514593045855e-215	Vascular endothelial	IGFBP7
166	8.93194549802638e-175	3.43531574404722	0.907	0.471	1.78638909960528e-171	Vascular endothelial	IFI27
167	1.1255750166463e-157	3.37011818035176	0.833	0.365	2.2511500332926e-154	Vascular endothelial	PECAM1
168	1.69604897539334e-92	3.32881246672951	0.73	0.454	3.39209795078668e-89	Vascular endothelial	FLT1
169	2.9491165293433e-51	3.21562377102103	0.635	0.454	5.8982330586866e-48	Vascular endothelial	IGFBP3
170	1.00744085557873e-152	3.15986049890305	0.83	0.395	2.01488171115746e-149	Vascular endothelial	COL4A2
171	1.24252136329454e-305	4.28938264763799	0.965	0.442	2.48504272658907e-302	Vascular smooth muscle	IGFBP7
172	0	4.15907804071485	0.973	0.394	0	Vascular smooth muscle	CALD1
173	5.24760633009876e-39	4.03559430052531	0.508	0.453	1.04952126601975e-35	Vascular smooth muscle	CXCL14
174	8.13806220366876e-173	4.01212213622026	0.738	0.409	1.62761244073375e-169	Vascular smooth muscle	TAGLN
175	5.59767825187436e-146	3.86281701093029	0.459	0.148	1.11953565037487e-142	Vascular smooth muscle	RGS5
176	1.00439123949978e-244	3.79838953809464	0.863	0.39	2.00878247899957e-241	Vascular smooth muscle	COL3A1
177	2.89515603565704e-248	3.67212507168441	0.852	0.331	5.79031207131408e-245	Vascular smooth muscle	COL1A2
178	1.0076248639121e-180	3.63403927743572	0.765	0.443	2.01524972782419e-177	Vascular smooth muscle	C11orf96
179	4.02832550606806e-119	3.61895486818913	0.67	0.479	8.05665101213612e-116	Vascular smooth muscle	ACTA2
180	8.45234049797074e-228	3.60975390539386	0.893	0.589	1.69046809959415e-224	Vascular smooth muscle	TIMP1

Supplementary Table 11: Top 10 expressed cell-type-specific marker genes identified for each cluster in the stromal cell type group of the integrated dataset.

#	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	8.24803802898859e-73	3.31482708954221	0.848	0.485	1.64960760579772e-69	0	RGS5
2	4.38889999365778e-07	2.8464152149164	0.548	0.418	0.000877779998731557	0	CCL19
3	2.09353139288206e-75	2.79817651513196	0.805	0.399	4.18706278576411e-72	0	NDUFA4L2
4	1.69364750164853e-95	2.61630961262554	0.977	0.704	3.38729500329705e-92	0	MGP
5	4.2393236575153e-80	2.55551952750451	0.921	0.58	8.4786473150306e-77	0	ADIRF
6	7.13089747362669e-72	2.46731309353597	0.894	0.62	1.42617949472534e-68	0	CRIP1
7	6.79086925489374e-87	2.26081487646343	0.901	0.613	1.35817385097875e-83	0	MCAM
8	3.78375767697376e-93	2.17725319162662	0.884	0.405	7.56751535394751e-90	0	NOTCH3
9	1.03518691977802e-61	2.03851448076243	0.858	0.584	2.07037383955605e-58	0	CSRP2
10	2.30846157851043e-71	1.94961832784958	0.947	0.686	4.61692315702086e-68	0	SOD3
11	3.99199667522722e-16	2.2871838165971	0.667	0.417	7.98399335045444e-13	1	MMP3
12	2.90552210604396e-89	2.26640404605111	0.915	0.483	5.81104421208792e-86	1	CTHRC1
13	8.48503677074342e-80	1.9263087545181	0.982	0.513	1.69700735414868e-76	1	LUM
14	8.92098198396217e-43	1.85535021438247	0.73	0.327	1.78419639679243e-39	1	APOD
15	5.33022261161277e-35	1.85479897001601	0.812	0.491	1.06604452232255e-31	1	MMP11
16	3.80235678490056e-34	1.77221930603538	0.975	0.841	7.60471356980113e-31	1	COL1A1
17	2.55502455085869e-75	1.57688486170167	0.837	0.311	5.11004910171737e-72	1	GREM1
18	7.80705044427652e-70	1.49077005413075	0.989	0.58	1.5614100888553e-66	1	DCN
19	2.32492066177109e-40	1.46933248744092	0.993	0.885	4.64984132354218e-37	1	COL1A2
20	9.53799433213864e-46	1.36920055332138	0.734	0.314	1.90759886642773e-42	1	C3
21	2.44108102989048e-09	4.386779664977	0.76	0.914	4.88216205978096e-06	2	IGKC
22	1.14756456738752e-05	4.30181978568145	0.432	0.672	0.0229512913477503	2	IGLC3
23	5.724359342951e-07	2.71666648028644	0.384	0.295	0.0011448718685902	2	IGHG1
24	0.00293997626760181	2.66269977452309	0.568	0.617	1	2	IGHA2
25	1.1186633911676e-11	2.30683639775677	0.157	0.396	2.23732678233519e-08	2	IGHM
26	4.6705856000254e-07	2.19683079340044	0.341	0.603	0.000934117120005079	2	IGLC2
27	5.54549325285951e-31	1.78427279490718	0.996	0.993	1.1090986505719e-27	2	MT-ND1
28	1.4642581837222e-32	1.73814433260441	1	1	2.92851636744439e-29	2	MT-CO3
29	7.72895791902292e-39	1.70281242005084	0.991	0.982	1.54579158380458e-35	2	MT-ND2
30	4.97355737855063e-06	1.68156009316208	0.41	0.374	0.00994711475710126	2	PIGR
31	4.69472538318075e-89	2.70048817148626	0.972	0.538	9.38945076636151e-86	3	F3
32	3.81811809219812e-80	2.62358950521226	0.897	0.477	7.63623618439624e-77	3	ALKAL2
33	1.66302412754921e-75	2.43178662915539	1	0.7	3.32604825509842e-72	3	CXCL14
34	6.20059572585592e-72	2.39197801650368	0.85	0.345	1.24011914517118e-68	3	HSD17B2
35	7.13825706416449e-68	2.379020468481	0.874	0.376	1.4276514128329e-64	3	NRG1
36	6.29746068994379e-78	2.37288167636241	0.776	0.203	1.25949213798876e-74	3	PAPPA2
37	5.97459049529668e-94	2.35964885225114	0.991	0.509	1.19491809905934e-90	3	PDGFRA
38	1.29504162203197e-72	2.28504804151887	0.958	0.621	2.59008324406395e-69	3	PLAT
39	3.84921301724526e-73	2.16044324358139	0.85	0.403	7.69842603449052e-70	3	PDGFD
40	8.54577896555905e-69	2.11865651112858	0.864	0.485	1.70915579311181e-65	3	BMP5
41	1.15430340422757e-75	4.0928043964449	0.969	0.357	2.30860680845514e-72	4	HHIP
42	1.77427189963444e-61	3.08879490088283	1	0.574	3.54854379926888e-58	4	MYH11
43	4.11975553756243e-57	2.54164046946809	0.923	0.5	8.23951107512486e-54	4	NPNT
44	6.31439668153984e-43	2.36288171671209	0.885	0.534	1.26287933630797e-39	4	ACTG2
45	6.82491170144655e-53	2.19524237727431	0.992	0.715	1.36498234028931e-49	4	MYLK
46	5.057568022086e-52	2.03837229780226	1	0.817	1.0115136044172e-48	4	LPP
47	1.92269611550291e-28	2.01016177391155	0.738	0.381	3.84539223100583e-25	4	MFAP5
48	1.58760463472037e-07	1.94843666113894	0.215	0.368	0.000317520926944075	4	IGHM
49	5.20649677477352e-52	1.92486022756019	0.915	0.513	1.0412993549547e-48	4	PLN
50	4.10002778748364e-34	1.80668172771251	0.962	0.779	8.20005557496727e-31	4	FLNA
51	2.97327947719934e-51	4.7260691453686	0.914	0.312	5.94655895439868e-48	5	CCL13
52	2.78314997222092e-30	4.13667213281186	0.8	0.277	5.56629994444185e-27	5	CCL11
53	1.06896458903868e-52	3.71775659430517	0.962	0.409	2.13792917807736e-49	5	TFPI2
54	6.95086028949104e-53	3.14315381421059	0.981	0.527	1.39017205789821e-49	5	ADAMDEC1
55	5.59932523781442e-49	2.71784928206614	0.99	0.581	1.11986504756288e-45	5	CFD
56	7.80306974745108e-42	2.52142201940211	0.905	0.368	1.56061394949022e-38	5	ADH1B

57	5.69419402919265e-32	2.38298309999147	0.895	0.524	1.13883880583853e-28	5	CCL2
58	3.9211177682823e-31	2.25612901054906	0.886	0.357	7.8422355365646e-28	5	PTGDS
59	5.2991906253375e-24	2.18449263524503	0.886	0.683	1.0598381250675e-20	5	APOE
60	5.46485381120368e-49	2.16385381981828	0.905	0.324	1.09297076224074e-45	5	HAPLN1

Supplementary Table 12: Top 10 expressed cell-type-specific marker genes identified for each cluster in the endothelial cell type group of the integrated dataset.

#	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	0.0014196053140049	2.57011544191584	0.819	0.894	1	0	IGKC
2	3.79772087555401e-07	1.6020391631515	0.331	0.158	0.00075954417511080 2	0	TPM2
3	4.57784921855912e-27	1.58797825513627	1	0.993	9.15569843711823e-24	0	MT-CO3
4	3.4917551980183e-22	1.50486685153108	1	0.99	6.9835103960366e-19	0	MT-CO1
5	1.50303146193707e-25	1.50010858833375	0.992	0.983	3.00606292387413e-22	0	MT-ND1
6	9.93674790605183e-24	1.4357276977966	1	0.993	1.98734958121037e-20	0	MT-CO2
7	2.34688138338737e-20	1.34623234052102	1	0.98	4.69376276677474e-17	0	MT-ND4
8	3.1920127068034e-16	1.3252460259707	0.976	0.974	6.38402541360679e-13	0	MT-ND2
9	6.00763917357249e-10	1.19034877786501	0.236	0.541	1.2015278347145e-06	0	DCN
10	3.12074340358068e-05	1.18418490811233	0.362	0.248	0.0624148680716137	0	CARMN
11	3.30116408377947e-29	2.54979142764621	0.922	0.694	6.60232816755893e-26	1	CLDN5
12	3.13608849863858e-17	2.36207964763267	0.825	0.557	6.27217699727717e-14	1	CCN2
13	3.0816992637859e-20	2.31117876770238	0.699	0.407	6.1633985275718e-17	1	CPE
14	6.0339611300138e-29	2.18721636727916	0.913	0.682	1.20679222600276e-25	1	HLA-DRB1
15	1.09165212188921e-14	2.13069619255594	0.893	0.761	2.18330424377842e-11	1	FABP5
16	1.19773995409232e-22	1.89886896694449	0.845	0.654	2.39547990818464e-19	1	HLA-DRA
17	4.42456730788519e-16	1.88007327142296	0.728	0.587	8.84913461577039e-13	1	CLU
18	1.1004938926674e-28	1.86981914632694	0.903	0.618	2.20098778533479e-25	1	HLA-DPA1
19	1.59832476844682e-15	1.80405869544043	0.854	0.682	3.19664953689364e-12	1	ENPP2
20	8.88302700922494e-12	1.76319491103381	0.689	0.535	1.77660540184499e-08	1	CCL14
21	2.07552895665929e-27	1.56671253847362	1	0.777	4.15105791331858e-24	2	SPARC
22	1.83279695258219e-18	1.37123613437634	0.864	0.468	3.66559390516439e-15	2	H19
23	1.20585384714544e-20	1.28005956982624	0.981	0.716	2.41170769429087e-17	2	PODXL
24	5.56203844684617e-21	1.22347122797722	1	0.798	1.11240768936923e-17	2	PLVAP
25	1.62954755235938e-20	1.19324632935	0.981	0.645	3.25909510471876e-17	2	CD34
26	3.07482546772592e-17	1.18668728070343	0.903	0.627	6.14965093545185e-14	2	CCND1
27	1.4311291708734e-05	1.14585941116866	0.505	0.196	0.028622583417468	2	SPP1
28	1.19598368232448e-06	1.14013903554302	0.592	0.306	0.00239196736464896	2	ANGPT2
29	1.25624020177924e-15	1.1192786780391	0.932	0.59	2.51248040355848e-12	2	PLPP3
30	3.78984662899739e-11	1.09171510583024	0.796	0.446	7.57969325799477e-08	2	IGFBP5
31	1.24496850153544e-06	0.644570614152526	0.937	0.909	0.00248993700307087	3	COL4A1
32	3.99189446113472e-05	0.611158070234288	0.962	0.954	0.0798378892226945	3	NEAT1
33	0.000946668834510617	0.604916980257373	0.62	0.467	1	3	F2RL3
34	0.000411001516878506	0.572960663720036	0.911	0.866	0.822003033757012	3	HSPG2
35	0.000224025122069471	0.526802944189715	0.506	0.365	0.448050244138943	3	HOXA9
36	3.55514744410875e-06	0.502240517151131	0.582	0.399	0.0071102948882175	3	RBMS3
37	0.00436643282568136	0.47039056683852	0.532	0.376	1	3	EBF1
38	1.68301002038838e-06	0.468691872493644	0.595	0.416	0.00336602004077677	3	GJC1
39	0.00973325412797188	0.46705903772647	0.671	0.57	1	3	JAG1
40	0.00214442133347931	0.455225389604451	0.582	0.487	1	3	HES4
41	1.37462702501357e-09	3.23142320925507	0.889	0.214	2.74925405002715e-06	4	CENPF
42	1.39872499943461e-11	3.00202371451517	0.944	0.311	2.79744999886921e-08	4	MKI67
43	3.63260790135205e-10	2.70749112679098	0.889	0.114	7.26521580270411e-07	4	NUSAP1
44	2.23549213482038e-09	2.51046768504307	0.889	0.51	4.47098426964076e-06	4	ASPM
45	7.19867172539703e-09	2.20054934337488	0.944	0.561	1.43973434507941e-05	4	HIST1H4C
46	1.21345622234018e-10	2.18401453211948	0.944	0.575	2.42691244468037e-07	4	PRC1
47	1.34492706218981e-08	2.16088544007969	0.944	0.566	2.68985412437962e-05	4	HMGB2
48	1.31984364782801e-08	2.14658657437917	1	0.49	2.63968729565602e-05	4	STMN1
49	3.27815060569296e-08	2.06966206062393	1	0.621	6.55630121138591e-05	4	TUBA1B
50	1.79108472354077e-11	2.05153678574389	0.944	0.456	3.58216944708153e-08	4	PCLAF

Supplementary Table 13: Top 10 expressed cell-type-specific marker genes identified for each cluster in the intestinal epithelial cell type group of the integrated dataset.

#	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	7.0140748725894e-202	1.16013661975799	0.999	0.962	1.40281497451788e-198	0	MT-ND2
2	2.98792218981081e-220	1.12723407494992	0.999	0.972	5.97584437962161e-217	0	MT-ND1
3	6.33077037011465e-238	0.99323573004407	1	0.979	1.26615407402293e-234	0	MT-CO1
4	4.09645838695878e-230	0.972587637010815	1	0.983	8.19291677391755e-227	0	MT-CO2
5	3.9260671748685e-50	0.825672148559514	0.798	0.738	7.85213434973701e-47	0	CDHR5
6	6.11007774568966e-175	0.800362690489388	0.999	0.968	1.22201554913793e-171	0	MT-ND4
7	2.16908075390735e-144	0.720239384762326	1	0.986	4.33816150781471e-141	0	MT-CO3
8	5.33364669285221e-43	0.692516838469648	0.798	0.67	1.06672933857044e-39	0	DST
9	2.31355752469972e-34	0.671187747695864	0.655	0.374	4.62711504939943e-31	0	SLC26A3
10	3.4614135920201e-34	0.631242861919084	0.949	0.845	6.92282718404019e-31	0	SLC26A2
11	4.0243992417063e-05	1.06680605201054	0.414	0.506	0.0804879848341259	1	CENPF
12	1.45443126642983e-98	0.800212672546545	1	0.987	2.90886253285967e-95	1	MT-CO3
13	4.54712226150745e-18	0.723080859891552	0.924	0.917	9.09424452301491e-15	1	SOX4
14	7.81834223689281e-68	0.64061856375736	0.971	0.888	1.56366844737856e-64	1	L1TD1
15	0.000254807499942834	0.62566287373202	0.623	0.69	0.509614999885668	1	KCNQ1OT1
16	1.69399679302418e-72	0.619162985723849	0.996	0.972	3.38799358604836e-69	1	MT-ND4
17	1.09671474896562e-22	0.569305865427715	0.837	0.773	2.19342949793124e-19	1	EPHB3
18	1.34480241279498e-50	0.52635080365432	0.997	0.975	2.68960482558997e-47	1	MT-ND1
19	7.6232314443144e-23	0.524044636896557	0.363	0.557	1.52464628886288e-19	1	HELLS
20	1.82820188278875e-08	0.485095448199464	0.453	0.65	3.6564037655775e-05	1	FCGBP
21	9.27174453599014e-173	2.32658447108361	0.994	0.774	1.85434890719803e-169	2	TFF3
22	2.57681907605976e-242	2.13879947821619	0.965	0.613	5.15363815211952e-239	2	LYZ
23	1.1523470320181e-256	1.9029778393954	0.944	0.624	2.3046940640362e-253	2	RETNLB
24	6.99572308461009e-250	1.78633933153877	0.989	0.885	1.39914461692202e-246	2	L1TD1
25	1.01974862015834e-153	1.66299261882368	0.956	0.796	2.03949724031669e-150	2	OLFM4
26	1.66943935236559e-227	1.58439711683888	0.982	0.739	3.33887870473118e-224	2	IFITM3
27	1.79359361372699e-159	1.55827859958669	0.868	0.496	3.58718722745398e-156	2	WFDC2
28	7.97374148802964e-198	1.50917421499884	0.991	0.736	1.59474829760593e-194	2	SLC12A2
29	2.25835259359832e-189	1.50079630273694	0.991	0.885	4.51670518719664e-186	2	PRDX5
30	2.62991379503629e-190	1.4813479163233	0.989	0.786	5.25982759007258e-187	2	H2AFZ
31	1.95441224610809e-193	2.23094443598597	0.984	0.678	3.90882449221617e-190	3	PLA2G2A
32	7.41456293699114e-210	2.23002698339765	0.98	0.687	1.48291258739823e-206	3	DMBT1
33	2.13664455133617e-236	1.71736360214932	0.995	0.791	4.27328910267235e-233	3	ADH1C
34	1.14086793236769e-138	1.71349747637607	0.978	0.753	2.28173586473538e-135	3	LCN2
35	1.48962870585832e-183	1.59269832563763	0.998	0.808	2.97925741171664e-180	3	C15orf48
36	4.46241154928361e-161	1.56386953264866	0.947	0.716	8.92482309856722e-158	3	LEFTY1
37	3.59637076668718e-189	1.55022344327148	1	0.951	7.19274153337437e-186	3	PIGR
38	3.23303004740047e-172	1.45524219368941	0.998	0.814	6.46606009480095e-169	3	FABP5
39	4.17773989561861e-71	1.4297285299865	0.947	0.803	8.35547979123722e-68	3	OLFM4
40	1.94343572644078e-143	1.35417516040092	0.998	0.917	3.88687145288157e-140	3	TSPAN8
41	5.99452989698332e-160	2.13798915492581	0.992	0.783	1.19890597939666e-156	4	TSPAN1
42	5.72451637141937e-142	2.13524654640551	0.984	0.615	1.14490327428387e-138	4	CEACAM6
43	6.94086625849962e-106	2.0500263713898	0.979	0.87	1.38817325169992e-102	4	FABP1
44	1.13659133376146e-141	2.02841007801822	0.995	0.608	2.27318266752292e-138	4	KRT20
45	5.43006798864358e-96	1.82632590515942	0.989	0.851	1.08601359772872e-92	4	KRT19
46	3.26662412900609e-122	1.78112709002261	1	0.885	6.53324825801218e-119	4	CEACAM5
47	1.67201884137235e-99	1.75850628961413	0.902	0.445	3.34403768274469e-96	4	TFF1
48	4.10400396451828e-100	1.71225815592373	0.989	0.857	8.20800792903655e-97	4	FXYD3
49	1.23436737068251e-92	1.68945258109313	0.887	0.399	2.46873474136501e-89	4	SLC26A3
50	1.41969081321586e-111	1.62018960133643	1	0.944	2.83938162643171e-108	4	LGALS3
51	1.86453138659332e-27	5.08158264095771	0.964	0.916	3.72906277318665e-24	5	IGKC
52	6.84287548413577e-18	4.80873591960323	0.929	0.919	1.36857509682715e-14	5	IGHA1
53	2.60430919611414e-17	4.78334379992724	0.798	0.747	5.20861839222828e-14	5	JCHAIN
54	2.7481022545345e-09	3.87111392328699	0.69	0.679	5.49620450906899e-06	5	IGHA2
55	0.008961341547791	2.55073075447726	0.411	0.416	1	5	IGHGP
56	1.01021580003044e-45	2.37984623296974	0.774	0.444	2.02043160006089e-42	5	VIM

57	0.000381594177580489	2.08712436961251	0.494	0.463	0.763188355160978	5	MZB1
58	1.03774134684106e-44	1.94508891807715	0.923	0.777	2.07548269368212e-41	5	FOSB
59	9.0901375639629e-05	1.89707822132773	0.625	0.683	0.181802751279258	5	CCL5
60	3.40376571896291e-18	1.76927383411058	0.857	0.84	6.80753143792581e-15	5	DUSP1
61	6.05787510117936e-10	3.03444885573123	0.604	0.741	1.21157502023587e-06	6	MT1G
62	3.27245016415381e-13	2.60969602677009	0.672	0.83	6.54490032830762e-10	6	MT2A
63	1.22501436773456e-11	2.39797039260907	0.679	0.859	2.45002873546912e-08	6	MT1E
64	1.63822429600196e-27	2.26955757275624	0.791	0.844	3.27644859200392e-24	6	FABP5
65	5.48540046655935e-22	2.17670813290563	0.746	0.855	1.09708009331187e-18	6	SRI
66	3.46984928301415e-68	2.16243694344893	1	0.967	6.9396985660283e-65	6	FTH1
67	3.12353701397008e-11	2.15100995776168	0.672	0.827	6.24707402794015e-08	6	ADH1C
68	9.49929288920015e-31	2.10823967867565	0.813	0.838	1.89985857784003e-27	6	C15orf48
69	5.00783436057358e-60	1.95635043931564	0.985	0.963	1.00156687211472e-56	6	FTL
70	2.41591009508127e-25	1.89310594977768	0.836	0.877	4.83182019016254e-22	6	PHGR1
71	1.39483510092813e-35	5.19152999923548	1	0.305	2.78967020185626e-32	7	SH2D6
72	4.78610321045835e-38	4.70267513880045	0.977	0.323	9.5722064209167e-35	7	HPGDS
73	2.53398938158483e-27	4.2597431443216	0.955	0.461	5.06797876316966e-24	7	LRMP
74	1.05748702907353e-26	3.52912193226618	0.932	0.357	2.11497405814705e-23	7	PSTPIP2
75	2.74976893194468e-20	3.09169572988634	0.909	0.616	5.49953786388936e-17	7	AZGP1
76	8.86219162416844e-21	3.06992644100085	0.886	0.538	1.77243832483369e-17	7	SPIB
77	2.29662423271188e-20	2.75611357823398	0.795	0.162	4.59324846542376e-17	7	HCK
78	4.44586708515955e-23	2.73477646716412	0.886	0.306	8.89173417031911e-20	7	ANXA13
79	4.56911676881362e-23	2.60797312979722	0.955	0.623	9.13823353762723e-20	7	CRIP1
80	6.80222190310728e-12	2.50056307739035	0.773	0.573	1.36044438062146e-08	7	RASSF6

Supplementary Table 14: Manually annotated clusters in stromal cell type group based on its top 10 cell-type-specific marker genes.

#	cluster	gene	Human Protein Atlas (HPA)	Panglaodb.se	Supplementary literature or other comment	Annotated cell type
1	0	RGS5	Smooth muscle cells - ECM organization (mainly)	Canonical marker for smooth muscle cell Canonical marker for pericyte	Marker gene for pericyte (https://doi.org/10.1038/s41586-021-03852-1) Signature molecule of tumor-associated pericytes (https://doi.org/10.1038/s41418-021-00801-3)	Pericyte
2	0	CCL19	Smooth muscle cells - ECM organization (mainly)			
3	0	NDUFA4L2	Smooth muscle cells - ECM organization (mainly)	Marker for pericyte		
4	0	MGP	Glandular cells - Unknown function (mainly)			
5	0	ADIRF	Enterocytes - Digestion (mainly)			
6	0	CRIP1	Alveolar cells - Smell perception (mainly)			
7	0	MCAM	Smooth muscle cells - ECM organization (mainly)	Canonical marker for pericyte	Marker gene for pericyte (https://doi.org/10.1038/s41586-021-03852-1) Increased expression in fibroblasts and pericytes during tumorigenesis, and confirmed as a prognostic factor to poor overall survival (https://doi.org/10.1038/s41575-021-00573-8)	
8	0	NOTCH3	Smooth muscle cells - ECM organization (mainly)	Canonical marker for smooth muscle cell Canonical marker for pericyte	Marker gene for pericyte (https://doi.org/10.1038/s41586-021-03852-1) Expressed in pericytes (https://doi.org/10.1152/ajpcell.00320.2021)	
9	0	CSRP2	Fibroblasts - ECM organization (mainly)			
10	0	SOD3	Fibroblasts - ECM organization (mainly)	Canonical marker for smooth muscle cell		
11	1	MMP3	Fibroblasts - ECM organization (mainly)	Canonical marker for fibroblast	Marker gene for cancer-associated fibroblasts (https://doi.org/10.1093/hmg/ddaa130) (https://doi.org/10.1101/2020.01.10.901579)	Cancer-associated fibroblast (CAF)
12	1	CTHRC1	Fibroblasts - ECM organization (mainly)	Canonical marker for fibroblast		
13	1	LUM	Fibroblasts - ECM organization (mainly)	Canonical marker for fibroblast		
14	1	APOD	Fibroblasts - ECM organization (mainly)			
15	1	MMP11	Stromal cells - Cell proliferation (mainly)		Marker gene for cancer-associated fibroblasts (https://doi.org/10.1093/hmg/ddaa130) (https://doi.org/10.1101/2020.01.10.901579)	
16	1	COL1A1	Fibroblasts - ECM organization (mainly)	Canonical marker for fibroblast	Marker gene for cancer-associated fibroblasts (https://doi.org/10.1101/2020.01.10.901579)	
17	1	GREM1	Granulosa cells - Unknown function (mainly)	Canonical marker for fibroblast		
18	1	DCN	Fibroblasts - ECM organization (mainly)	Canonical marker for fibroblast		
19	1	COL1A2	Fibroblasts - ECM organization (mainly)	Canonical marker for fibroblast	Marker gene for cancer-associated fibroblasts (https://doi.org/10.1101/2020.01.10.901579)	
20	1	C3	Hepatocytes - Hemostasis (mainly)			
21	2	IGKC	Plasma cells - Humoral immune response (mainly)			Plasma B-cell
22	2	IGLC3	Plasma cells - Humoral immune response (mainly)	Canonical marker for plasma B-cell		
23	2	IGHG1	Plasma cells - Humoral immune response (mainly)	Canonical marker for plasma B-cell		
24	2	IGHA2	Plasma cells - Humoral immune response (mainly)	Canonical marker for plasma B-cell		
25	2	IGHM	Plasma cells - Humoral immune response (mainly)	Canonical marker for plasma B-cell		
26	2	IGLC2	Plasma cells - Humoral immune response (mainly)	Canonical marker for plasma B-cell		

27	2	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)		Mitochondrial derived	
28	2	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)		Mitochondrial derived	
29	2	MT-ND2	Cardiomyocytes - Muscle contraction (mainly)		Mitochondrial derived	
30	2	PIGR	Intestinal epithelial cells - Unknown function (mainly)			
31	3	F3	Pancreatic endocrine cells - Mixed function (mainly)			Crypt-top fibroblast (CTF)
32	3	ALKAL2	Granulosa cells - Unknown function (mainly)			
33	3	CXCL14	Fibroblasts - ECM organization (mainly)	Canonical marker for fibroblast		
34	3	HSD17B2	Enterocytes - Digestion (mainly)			
35	3	NRG1	Neurons & Oligodendrocytes - Neuronal signaling (mainly)			
36	3	PAPPA2	Intestinal endocrine cells - Hormone signaling (mainly)			
37	3	PDGFRA	Stromal cells - Cell proliferation (mainly)	Canonical marker for fibroblast	Marker of cancer-associated fibroblasts (https://doi.org/10.1038/s41417-021-00318-4) Marker of crypt-associated colonic fibroblast-population (https://doi.org/10.1371/journal.pbio.3001032)	
38	3	PLAT	Epithelial cell types - Mixed function (mainly)			
39	3	PDGFD	Neurons & Oligodendrocytes - Synaptic function (mainly)	Canonical marker for smooth muscle cell		
40	3	BMP5	Fibroblasts - ECM organization (mainly)		Secreted by crypt-top fibroblasts (https://doi.org/10.1371/journal.pbio.3001032)	
41	4	HHIP	Oligodendrocytes - Myelin sheath organization (mainly)	Canonical marker for smooth muscle cell Canonical marker for fibroblast	Marker gene for myofibroblast (https://doi.org/10.1038/s41586-021-03852-1)	Myofibroblast
42	4	MYH11	Smooth muscle cells - ECM organization (mainly)	Canonical marker for smooth muscle cell	Marker gene for myofibroblast (https://doi.org/10.1093/hmg/ddaa130)	
43	4	NPNT	Alveolar cells - Smell perception (mainly)		Marker gene for myofibroblast (https://doi.org/10.1038/s41586-021-03852-1)	
44	4	ACTG2	Smooth muscle cells - ECM organization (mainly)	Canonical marker for smooth muscle cell	Marker gene for myofibroblast (https://doi.org/10.1093/hmg/ddaa130)	
45	4	MYLK	Smooth muscle cells - ECM organization (mainly)	Canonical marker for smooth muscle cell		
46	4	LPP	Spermatids - Spermatogenesis (mainly)			
47	4	MFAP5	Fibroblasts - ECM organization (mainly)	Canonical marker for smooth muscle cell Marker for fibroblast		
48	4	IGHM	Plasma cells - Humoral immune response (mainly)			
49	4	PLN	Cardiomyocytes - Muscle contraction (mainly)	Canonical marker for smooth muscle cell		
50	4	FLNA	Smooth muscle cells - ECM organization (mainly)			
51	5	CCL13	Macrophages - Immune response (mainly)		Marker gene for lamina propria fibroblast (https://doi.org/10.1093/hmg/ddaa130)	Lamina propria fibroblast (LPF)
52	5	CCL11	Fibroblasts - ECM organization (mainly)	Canonical marker for fibroblast	Marker gene for lamina propria fibroblast (https://doi.org/10.1093/hmg/ddaa130)	
53	5	TFPI2	Syncytiotrophoblasts - Pregnancy hormone signaling (mainly)			
54	5	ADAMDEC1	Macrophages - Innate immune response (mainly)	Canonical marker for smooth muscle cell	Marker gene for lamina propria fibroblast (https://doi.org/10.1093/hmg/ddaa130)	
55	5	CFD	Fibroblasts - ECM organization (mainly)			
56	5	ADH1B	Hepatocytes - Metabolism (mainly)			
57	5	CCL2	Smooth muscle cells - ECM organization (mainly)	Canonical marker for macrophage	Marker gene for lamina propria fibroblast (https://doi.org/10.1093/hmg/ddaa130)	

58	5	PTGDS	Non-specific - Transcription (mainly)			
59	5	APOE	Smooth muscle cells - Unknown function (mainly)		Marker gene for lamina propria fibrolast (https://doi.org/10.1093/hmg/ddaa130)	
60	5	HAPLN1	Fibroblasts - ECM organization (mainly)			

Supplementary Table 15: Manually annotated clusters in endothelial cell type group based on its top 10 cell-type-specific marker genes.

#	cluster	gene	Human Protein Atlas (HPA)	Panglaodb.se	Supplementary literature or other comment	Annotated cell type
1	0	IGKC	Plasma cells - Humoral immune response (mainly)			Mitochondrial gene-expressing cell
2	0	TPM2	Smooth muscle cells - ECM organization (mainly)			
3	0	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)		Mitochondrial derived	
4	0	MT-CO1	Cardiomyocytes - Muscle contraction (mainly)		Mitochondrial derived	
5	0	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)		Mitochondrial derived	
6	0	MT-CO2	Cardiomyocytes - Muscle contraction (mainly)		Mitochondrial derived	
7	0	MT-ND4	Cardiomyocytes - Muscle contraction (mainly)		Mitochondrial derived	
8	0	MT-ND2	Cardiomyocytes - Muscle contraction (mainly)		Mitochondrial derived	
9	0	DCN	Fibroblasts - ECM organization (mainly)			
10	0	CARMN	Not found			
11	1	CLDN5	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Canonical marker for endothelial cell	Tip-like endothelial cell marker gene (https://doi.org/10.1007/s12079-019-00511-z)	Activated TEC
12	1	CCN2	Fibroblasts - ECM organization (mainly)		Expressed in endothelial cells, and it increases vascular angiogenesis (https://doi.org/10.1038/cddis.2014.453)	
13	1	CPE	Smooth muscle cells - ECM organization (mainly)		Expressed by activated postcapillary vein tumor-associated endothelial cells (https://doi.org/10.1016/j.ccell.2019.12.001)	
14	1	HLA-DRB1	Macrophages - Immune response (mainly)		Expressed by capillary endothelial cells, gene involved in antigen presentation (https://doi.org/10.1016/j.ccell.2019.12.001)	
15	1	FABP5	Squamous epithelial cells - Cornification (mainly)			
16	1	HLA-DRA	Macrophages - Immune response (mainly)		Expressed by capillary endothelial cells, gene involved in antigen presentation (https://doi.org/10.1016/j.ccell.2019.12.001)	
17	1	CLU	Erythroid cells - Oxygen transport (mainly)		Expressed by activated postcapillary vein tumor-associated endothelial cells (https://doi.org/10.1016/j.ccell.2019.12.001)	
18	1	HLA-DPA1	Macrophages - Immune response (mainly)		Expressed by capillary endothelial cells, gene involved in antigen presentation (https://doi.org/10.1016/j.ccell.2019.12.001)	
19	1	ENPP2	Oligodendrocytes - Myelin sheath organization (mainly)			
20	1	CCL14	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Canonical marker for endothelial cell	Expressed by activated postcapillary vein tumor-associated endothelial cells (https://doi.org/10.1016/j.ccell.2019.12.001)	
21	2	SPARC	Fibroblasts - ECM organization (mainly)	Canonical marker for endothelial cell	Up-regulated in colorectal cancer endothelial cells (https://doi.org/10.3934/mbe.2021360)	Tip TEC

					Marker gene of tip endothelial cells (https://doi.org/10.1016/j.ccell.2019.12.001)	
22	2	H19	Not found			
23	2	PODXL	Endometrium - Transcription (mainly)	Canonical marker for endothelial cell		
24	2	PLVAP	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Canonical marker for endothelial cell	Marker gene of immature endothelial cell (https://doi.org/10.1016/j.ccell.2019.12.001)	
25	2	CD34	Adipocytes & Endothelial cells - Angiogenesis (mainly)	Canonical marker for endothelial cell	Up-regulated in tumor endothelial cell (https://doi.org/10.3389/fcell.2020.00766) Marker of endothelial tip cell (https://doi.org/10.1007/s10456-011-9251-z)	
26	2	CCND1	Pancreas - Digestion (mainly)			
27	2	SPP1	Macrophages - Innate immune response (mainly)			
28	2	ANGPT2	Smooth muscle cells - ECM organization (mainly)		Marker gene of tumor-associated endothelial cells (https://doi.org/10.3390/ijms19051272) Marker gene of tip EC (https://doi.org/10.1016/j.ccell.2019.12.001) (https://doi.org/10.1016/j.devcel.2021.06.021)	
29	2	PLPP3	Fibroblasts - ECM organization (mainly)			
30	2	IGFBP5	Smooth muscle cells - ECM organization (mainly)		Up-regulated in colorectal cancer endothelial cells (https://doi.org/10.3934/mbe.2021360) Marker gene of immature tumor-associated endothelial cell (https://doi.org/10.1038/s41467-021-21346-6)	
31	3	COL4A1	Smooth muscle cells - ECM organization (mainly)		Hub gene of colon tumor-associated endothelial cell (https://doi.org/10.3934/mbe.2021360) Marker gene of tip endothelial cell (https://doi.org/10.1016/j.ccell.2019.12.001) Marker gene of tip-like endothelial cell (https://doi.org/10.1007/s12079-019-00511-z)	Immature TEC
32	3	NEAT1	Not found			
33	3	F2RL3	Myeloid cells - Hemostasis (mainly)			
34	3	HSPG2	Granulosa cells - Unknown function (mainly)	Canonical marker for endothelial cell	Marker gene of immature endothelial cell (https://doi.org/10.1016/j.ccell.2019.12.001) Marker gene of tip-like endothelial cell (https://doi.org/10.1007/s12079-019-00511-z)	
35	3	HOXA9	Proximal tubular cells - Tubular reabsorption (mainly)			
36	3	RBMS3	Neurons & Oligodendrocytes - Synaptic function (mainly)			
37	3	EBF1	Fibroblasts - ECM organization (mainly)			
38	3	GJC1	Smooth muscle cells - ECM organization (mainly)			
39	3	JAG1	Squamous epithelial cells - Cornification (mainly)		Expressed by tumor-associated endothelial cells (https://doi.org/10.3389/fcell.2020.00766)	

					Marker gene of immature endothelial cell (https://doi.org/10.1016/j.ccell.2019.12.001)	
40	3	HES4	Smooth muscle cells - ECM organization (mainly)			
41	4	CENPF	Non-specific - Cell cycle regulation (mainly)			Proliferative EC
42	4	MKI67	Non-specific - Cell cycle regulation (mainly)		Marker of proliferative cancer cell (https://doi.org/10.1038/s41418-021-00823-x)	
43	4	NUSAP1	Non-specific - Cell cycle regulation (mainly)		Promotes cell-proliferation (https://doi.org/10.1016/j.yexcr.2018.03.039)	
44	4	ASPM	Non-specific - Cell cycle regulation (mainly)			
45	4	HIST1H4C	Non-specific - Transcription regulation (mainly)			
46	4	PRC1	Non-specific - Cell cycle regulation (mainly)			
47	4	HMGB2	Non-specific - Cell cycle regulation (mainly)		Marker gene of proliferative endothelial cell in liver and spleen (https://doi.org/10.1016/j.cell.2020.01.015)	
48	4	STMN1	Non-specific - Cell cycle regulation (mainly)		Marker gene of proliferative endothelial cell in liver and spleen (https://doi.org/10.1016/j.cell.2020.01.015)	
49	4	TUBA1B	Non-specific - Cell cycle regulation (mainly)		Marker gene of proliferative endothelial cell in liver and spleen (https://doi.org/10.1016/j.cell.2020.01.015)	
50	4	PCLAF	Non-specific - Cell cycle regulation (mainly)			

Supplementary Table 16: Manually annotated clusters in intestinal epithelial cell type group based on its top 10 cell-type-specific marker genes.

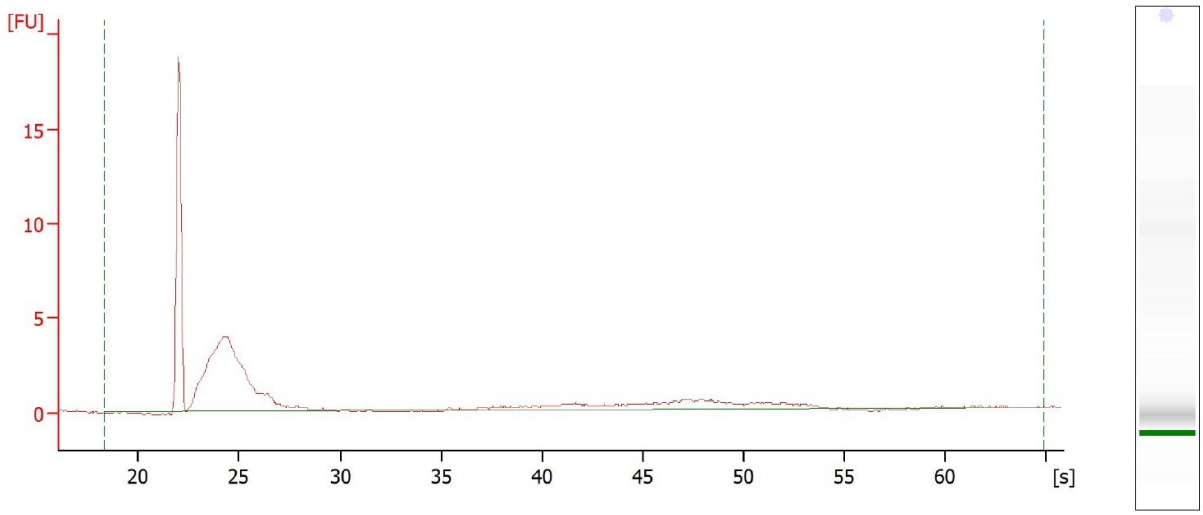
#	cluster	gene	Human Protein Atlas	Panglaodb.se	Supplementary literature or other comment	Annotated cell type
1	0	MT-ND2	Cardiomyocytes - Muscle contraction (mainly)			Mitochondrial derived
2	0	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)			
3	0	MT-CO1	Cardiomyocytes - Muscle contraction (mainly)			
4	0	MT-CO2	Cardiomyocytes - Muscle contraction (mainly)			
5	0	CDHR5	Enterocytes - Digestion (mainly)			
6	0	MT-ND4	Cardiomyocytes - Muscle contraction (mainly)			
7	0	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)			
8	0	DST	Epithelial cell types - Mixed function (mainly)			
9	0	SLC26A3	Intestinal epithelial cells - Unknown function (mainly)			
10	0	SLC26A2	Intestinal epithelial cells - Unknown function (mainly)			
11	1	CENPF	Non-specific - Cell cycle regulation (mainly)		Associated with stem cell characteristics (https://www.doi.org/10.4240/wjgs.v12.i11.442)	Secretory progenitor 1
12	1	MT-CO3	Cardiomyocytes - Muscle contraction (mainly)			
13	1	SOX4	Granulosa cells - Unknown function (mainly)		Expressed in stem cell and enterocyte progenitors (https://doi.org/10.1016/j.celrep.2021.109484) Promotes secretory progenitor differentiation (https://www.doi.org/10.1053/j.gastro.2018.07.023)	
14	1	L1TD1	Cytotrophoblasts - Unknown function (mainly)		Colorectal cancer stem cell-related gene (https://doi.org/10.3390/biomedicines9020179) Embryonic stem cell factor (https://doi.org/10.1186/s12885-019-5952-2)	
15	1	KCNQ1OT1	Not found		Expressed in tumor cells, promoting colorectal cancer development (https://doi.org/10.3389/fcell.2021.653808)	
16	1	MT-ND4	Cardiomyocytes - Muscle contraction (mainly)			
17	1	EPHB3	Intestinal epithelial cells - Unknown function (mainly)		Expressed in cells at the crypt base, associated with other intestinal stem cell markers (https://www.doi.org/10.3390/biom10040602) Expressed in stem cells in the crypt base of small intestine (https://www.doi.org/10.1111/j.1469-7580.2008.00925.x) Expressed by paneth cell progenitors (https://www.doi.org/10.15252/embr.201540188)	
18	1	MT-ND1	Cardiomyocytes - Muscle contraction (mainly)			
19	1	HELLS	Non-specific - Cell cycle regulation (mainly)		Upregulated in colorectal cancer (https://www.doi.org/10.2147/OTT.S223668)	
20	1	FCGBP	Mucus-secreting cells - Mucin production (mainly)		Expressed in colonic stem cell's transition to the progenitor stage (https://doi.org/10.1038/s42003-020-01181-z)	
21	2	TFF3	Mucus-secreting cells - Mucin production (mainly)	Canonical marker for goblet cell		Crypt base columnar cell (CBC) and paneth cell
22	2	LYZ	Monocytes & Neutrophils - Innate immune response (mainly)	Canonical marker for paneth	Marker gene of paneth cell (https://doi.org/10.1016/j.celrep.2016.08.054)	
23	2	RETNLB	Mucus-secreting cells - Mucin production (mainly)	Marker for paneth cell		

24	2	L1TD1	Cytotrophoblasts - Unknown function (mainly)		Colorectal cancer stem cell-related gene (https://doi.org/10.3390/biomedicines9020179) Embryonic stem cell factor (https://doi.org/10.1186/s12885-019-5952-2)	
25	2	OLFM4	Intestinal epithelial cells - Unknown function (mainly)	Canonical marker for crypt cell	Expressed in colorectal crypt base cells (https://doi.org/10.1053/j.gastro.2009.05.035)	
26	2	IFITM3	Fibroblasts - ECM organization (mainly)			
27	2	WFDC2	Respiratory epithelial cells - Mucosal defense (mainly)			
28	2	SLC12A2	Mucus-secreting cells - Mucin production (mainly)	Canonical marker for crypt cell		
29	2	PRDX5	Respiratory epithelial cells - Mucosal defense (mainly)			
30	2	H2AFZ	Non-specific - Transcription regulation (mainly)			
31	3	PLA2G2A	Fibroblasts - ECM organization (mainly)		Marker gene for transit amplifying cells, and marker gene for Paneth cells (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596) Expressed in paneth-like cells in the colon (https://doi.org/10.1016/j.stem.2016.05.023)	Secretory progenitor 2
32	3	DMBT1	Enterocytes - Digestion (mainly)		Enteroendocrine progenitor marker gene (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596)	
33	3	ADH1C	Pancreas - Digestion (mainly)			
34	3	LCN2	Respiratory epithelial cells - Mucosal defense (mainly)			
35	3	C15orf48	Pancreatic endocrine cells - Mixed function (mainly)		Expressed in stem cell, enterocyte, tuft cell, goblet cell, and enteroendocrine cell (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596)	
36	3	LEFTY1	Intestinal epithelial cells - Unknown function (mainly)		Expressed in stem cell (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596)	
37	3	PIGR	Intestinal epithelial cells - Unknown function (mainly)		Expressed in enteroendocrine progenitor, enterocyte, and goblet cell (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596)	
38	3	FABP5	Squamous epithelial cells - Cornification (mainly)		Expressed in enteroendocrine (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596)	
39	3	OLFM4	Intestinal epithelial cells - Unknown function (mainly)	Canonical marker of crypt cell	Marker gene for intestinal crypt base cell (https://doi.org/10.1053/j.gastro.2009.05.035) Canonical stem cell marker gene (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596)	
40	3	TSPAN8	Enterocytes - Digestion (mainly)		Expressed in enterocyte, stem cell, and goblet cell (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596)	
41	4	TSPAN1	Intestinal epithelial cells - Unknown function (mainly)			Enterocyte
42	4	CEACAM6	Intestinal epithelial cells - Unknown function (mainly)			
43	4	FABP1	Enterocytes - Digestion (mainly)	Canonical marker for enterocyte		
44	4	KRT20	Intestinal epithelial cells - Unknown function (mainly)	Canonical marker for enterocyte		
45	4	KRT19	Respiratory epithelial cells - Mucosal defense (mainly)			
46	4	CEACAM5	Intestinal epithelial cells - Unknown function (mainly)			
47	4	TFF1	Pancreatic endocrine cells - Mixed function (mainly)			
48	4	FXYD3	Intestinal epithelial cells - Unknown function (mainly)			
49	4	SLC26A3	Intestinal epithelial cells - Unknown function (mainly)	Canonical marker for enterocyte		
50	4	LGALS3	Intestinal epithelial cells - Unknown function (mainly)			

51	5	IGKC	Plasma cells - Humoral immune response (mainly)			Plasma B-cell
52	5	IGHA1	Plasma cells - Humoral immune response (mainly)			
53	5	JCHAIN	Plasma cells - Humoral immune response (mainly)			
54	5	IGHA2	Plasma cells - Humoral immune response (mainly)			
55	5	IGHGP	Not found			
56	5	VIM	Macrophages - Immune response (mainly)			
57	5	MZB1	Plasma cells - Humoral immune response (mainly)			
58	5	FOSB	Non-specific - Mitochondria (mainly)			
59	5	CCL5	NK-cells & T-cells - Immune response (mainly)			
60	5	DUSP1	Macrophages - Immune response (mainly)			
61	6	MT1G	Proximal tubular cells - Tubular reabsorption (mainly)		Metallothionein are expressed during oxidative stress responses in enteroendocrine cells and endothelial cells (https://doi.org/10.1101/721662)	Iron-storing epithelial cell
62	6	MT2A	Adipocytes & Endothelial cells - Angiogenesis (mainly)		Metallothionein are expressed during oxidative stress responses in enteroendocrine cells and endothelial cells (https://doi.org/10.1101/721662)	
63	6	MT1E	Proximal tubular cells - Tubular reabsorption (mainly)		Metallothionein are expressed during oxidative stress responses in enteroendocrine cells and endothelial cells (https://doi.org/10.1101/721662)	
64	6	FABP5	Squamous epithelial cells - Cornification (mainly)	Canonical marker for enteroendocrine cell Canonical marker for M cells	Marker for enteroendocrine cells (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596) Expressed in enteroendocrine cells (https://www.doi.org/10.1210/me.2014-1194)	
65	6	SRI	Intestinal epithelial cells - Unknown function (mainly)			
66	6	FTH1	Monocytes - Immune response regulation (mainly)		One of two subunits of the iron storage ferritin protein, which is expressed by colorectal epithelial cells (https://doi.org/10.1038/s41419-020-2299-1) (https://doi.org/10.1038/s41419-021-03559-1)	
67	6	ADH1C	Pancreas - Digestion (mainly)			
68	6	C15orf48	Pancreatic endocrine cells - Mixed function (mainly)		Expressed in stem cell, enterocyte, tuft cell, goblet cell, and enteroendocrine (https://patentscope.wipo.int/search/en/detail.jsf?docId=US317630596)	
69	6	FTL	Macrophages - Innate immune response (mainly)		One of two subunits of the iron storage ferritin protein, which is expressed by colorectal epithelial cells (https://doi.org/10.1038/s41419-020-2299-1) (https://doi.org/10.1038/s41419-021-03559-1)	
70	6	PHGR1	Enterocytes - Digestion (mainly)	Marker for enterocyte		
71	7	SH2D6	Proximal tubular cells - Tubular reabsorption (mainly)		Signature marker for CD45+ tuft-2 cells (https://doi.org/10.1016/j.immuni.2022.03.001)	Tuft-2
72	7	HPGDS	Granulocytes - Receptor signaling (mainly)		Mainly expressed in CD45+ tuft-2 cells (https://doi.org/10.1016/j.immuni.2022.03.001)	
73	7	LRMP	Non-specific - Transcription regulation (mainly)	Canonical marker for tuft cell	Marker gene for tuft cell (https://www.doi.org/10.1038/s41598-019-52049-0)	
74	7	PSTPIP2	Myeloid cells - Hemostasis (mainly)			
75	7	AZGP1	Glandular cells - Unknown function (mainly)			

76	7	SPIB	Plasmacytoid DCs - Unknown function (mainly)		Involved in tuft-2 cell development (https://doi.org/10.1016/j.immuni.2022.03.001) Required for M cell differentiation (https://doi.org/10.1038/mi.2016.68) (https://www.doi.org/10.1038/ni.2352)	
77	7	HCK	Monocytes & Neutrophils - Innate immune response (mainly)			
78	7	ANXA13	Mucus-secreting cells - Mucin production (mainly)			
79	7	CRIP1	Alveolar cells - Smell perception (mainly)		Associated with cancer (https://doi.org/10.1016/j.lfs.2018.05.054)	
80	7	RASSF6	Intestinal epithelial cells - Unknown function (mainly)		Associated with colorectal cancer (https://www.doi.org/10.18632/oncotarget.7852)	

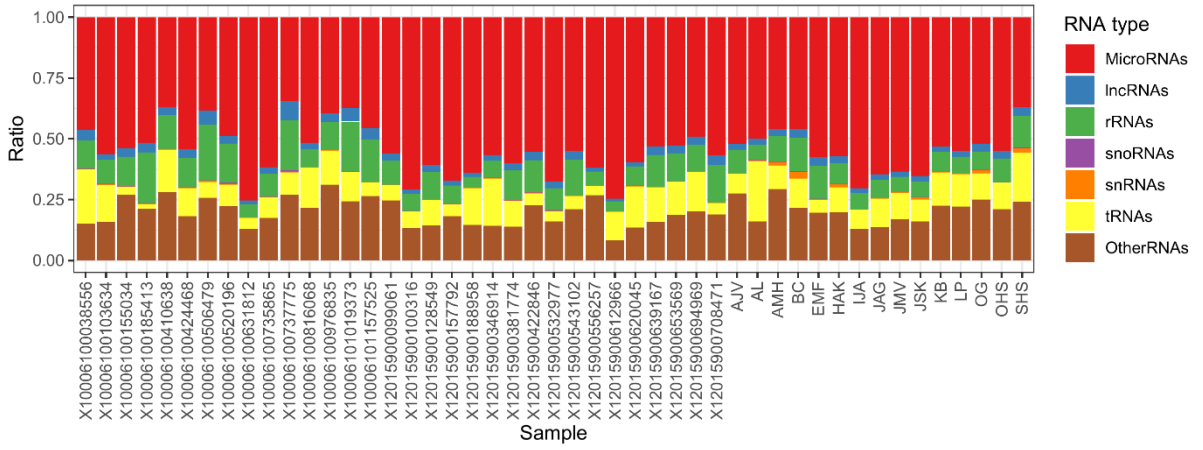
I Supplementary section: Identified differentially expressed circulating miRNAs between colorectal cancer patient groups



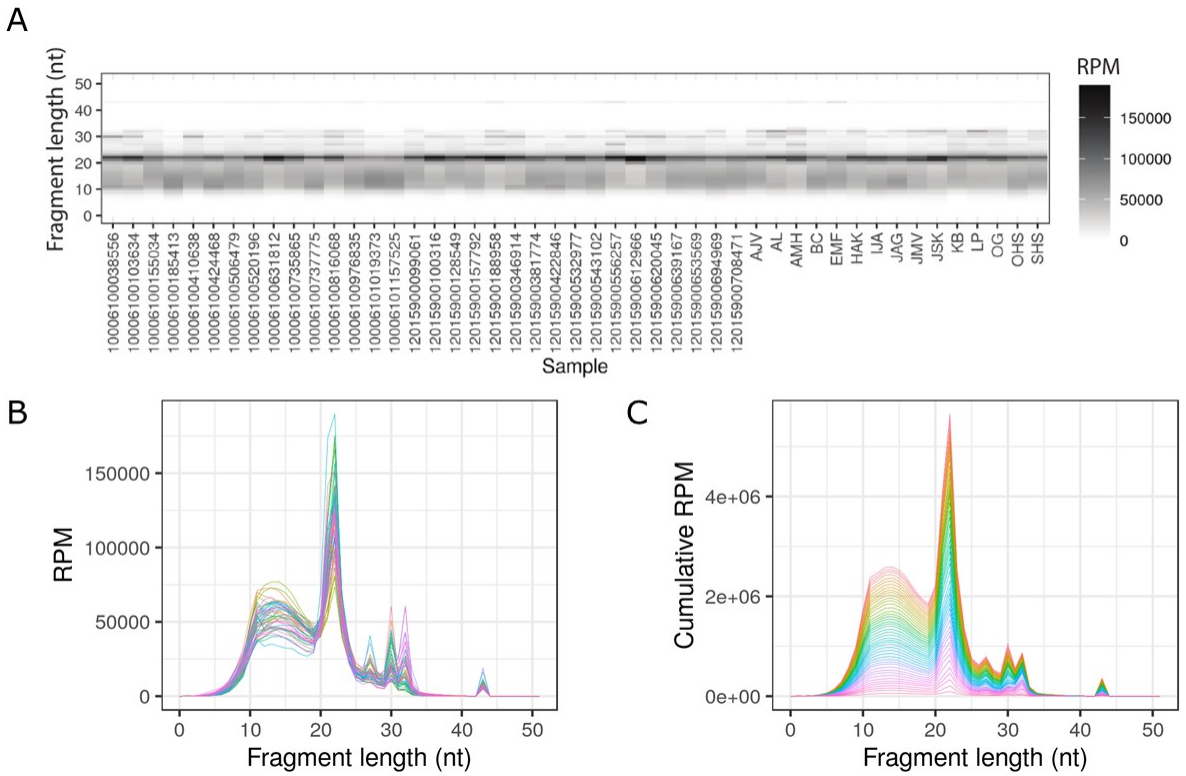
Overall Results for sample 5 : SHS

RNA Area:	48.8	RNA Integrity Number (RIN):	2.1 (B.02.10)
RNA Concentration:	106 pg/µl	Result Flagging Color:	
rRNA Ratio [28s / 18s]:	0.0	Result Flagging Label:	RIN: 2.10

Supplementary Figure 3: Sequencing quality statistics. Assessment of RNA quality and relative size were conducted by measuring the samples on a 2100 Bioanalyzer. The plot is representative of all 47 samples.



Supplementary Figure 4: RNA distribution plot of the sequencing libraries. The plot shows the relative abundance of the major classes of RNAs detected in the sequencing libraries.



Supplementary Figure 5: Length distribution of the sequence reads after adapter and quality trimming in 47 samples. (A) Each tile represents a read, where the intensity of each tile depicts the number of reads. Each sample is listed on the x-axis, while the y-axis shows the fragment lengths. **(B)** Each line represents a sample, where the fragment length is shown on the x-axis and RPM is shown on the y-axis. **(C)** Cumulative graphs of mapped reads. Fragment length is depicted on the x-axis and the cumulative RPM is shown on the y-axis. Abbreviations: nt, nucleotides; RPM, reads per million mapped reads.

Supplementary Table 17: Differentially expressed miRNAs for true positive CRC patients with localized disease versus false positive CRC patients (healthy).

miRNA	logFC	AveExpr	t	P.Value	adj.P.Val	B
hsa-miR-30a-5p	-1,596596089	11,10190101	-7,114034271	3,34278E-09	8,75691E-07	10,92319798
hsa-miR-484	1,068278498	10,95168084	7,015345406	4,7983E-09	8,75691E-07	10,57766389
hsa-miR-126-5p	-1,225022842	9,046003843	-5,701198121	5,76678E-07	6,78168E-05	6,01384864
hsa-miR-122-5p	-2,621624982	14,61508189	-5,630650163	7,43197E-07	6,78168E-05	5,668734635
hsa-miR-342-3p	-1,063749702	10,05995758	-5,414044673	1,61317E-06	9,72317E-05	5,024543288
hsa-miR-125a-5p	-1,176987769	11,98978333	-5,38504196	1,78869E-06	9,72317E-05	4,882770939
hsa-miR-150-5p	-1,492751876	11,26420134	-5,373343165	1,86472E-06	9,72317E-05	4,859149111
hsa-miR-142-5p	-0,856045718	12,28061237	-5,304409873	2,38204E-06	0,000107034	4,593590174
hsa-let-7c-5p	-1,199612809	8,985287488	-5,275486604	2,63919E-06	0,000107034	4,583531
hsa-miR-10b-5p	-1,250397023	12,21465319	-5,159747814	3,97215E-06	0,000144983	4,10954825
hsa-miR-29a-3p	-1,288731592	10,73818744	-5,076970651	5,31361E-06	0,000176315	3,880644448
hsa-miR-101-3p	-0,847234268	10,32161997	-4,970046631	7,723E-06	0,000234908	3,53103366
hsa-miR-146b-5p	-1,025483113	9,80802919	-4,872208473	1,0852E-05	0,000304689	3,223467216
hsa-miR-27b-3p	-1,03647021	11,33471254	-4,731416003	1,76401E-05	0,000459902	2,717712077
hsa-miR-375-3p	-1,940138677	10,26476405	-4,626807812	2,52329E-05	0,000597395	2,443485858
hsa-miR-143-3p	-1,066141986	12,26428624	-4,615915175	2,61872E-05	0,000597395	2,301150124
hsa-miR-30a-3p	-2,0261264	5,809094095	-4,468319717	4,31736E-05	0,000926962	1,770840195
hsa-miR-584-5p	1,057063145	8,711311846	4,441422236	4,72622E-05	0,000958372	1,88111541

hsa-miR-29c-3p	-1,156042368	7,015478138	-4,359652268	6,2149E-05	0,001193916	1,673017613
hsa-miR-1249-3p	-1,973526992	6,168566659	-4,239705325	9,2542E-05	0,001673925	1,14247178
hsa-miR-16-5p	-0,569929532	14,7414419	-4,227614634	9,6308E-05	0,001673925	0,961969718
hsa-miR-483-3p	-3,00079154	6,404134527	-4,160463887	0,000120092	0,001992428	0,838468371
hsa-miR-1-3p	-2,023105077	7,913374618	-4,09713723	0,000147682	0,002265995	0,884610663
hsa-miR-192-5p	-1,371054703	9,180609279	-4,094412789	0,000148997	0,002265995	0,819528046
hsa-miR-1228-3p	-3,182280296	3,990660622	-4,081643889	0,000155315	0,002267594	0,116973658
hsa-miR-99a-5p	-1,017695046	10,04454709	-3,975803616	0,000218664	0,003069702	0,395275774
hsa-miR-30e-5p	-0,511201191	13,20882657	-3,923128228	0,00025887	0,003499541	0,090876175
hsa-miR-423-5p	0,555770493	14,42809973	3,890535415	0,000287225	0,003744185	-0,060018329
hsa-miR-125b-5p	-0,939298896	10,17125742	-3,801555336	0,000380707	0,004791655	-0,13051478
hsa-miR-885-3p	-3,449441685	2,102083348	-3,688778295	0,000541746	0,006557052	-0,930124647
hsa-miR-95-3p	-2,256857886	4,115791539	-3,679890272	0,0005569	0,006557052	-0,717431463
hsa-miR-206	-3,634517428	3,112465394	-3,573067846	0,000773907	0,008827371	-0,971006823
hsa-miR-629-5p	0,795141253	8,08871325	3,470264678	0,00105745	0,011696042	-0,937671203
hsa-miR-148a-3p	-0,650294823	11,41034787	-3,442956113	0,001147993	0,012274673	-1,218108591
hsa-miR-3613-5p	-0,605191386	9,305116332	-3,434633015	0,001177023	0,012274673	-1,134041029
hsa-miR-433-3p	3,062761522	3,79921006	3,413058322	0,001255559	0,012729974	-1,213731751
hsa-miR-126-3p	-0,719133622	13,16483187	-3,384921227	0,001365482	0,013470296	-1,465107525
hsa-miR-215-5p	-2,035513872	5,858001738	-3,367038283	0,001440028	0,01383185	-1,128812377
hsa-miR-16-2-3p	-0,46322482	10,61792793	-3,261123654	0,001966921	0,018408364	-1,685462316
hsa-miR-125a-3p	2,414850684	3,778931583	3,185843066	0,002447041	0,022329245	-1,726240222
hsa-miR-6803-3p	-1,176893196	7,137953939	-3,16776175	0,002577787	0,022948591	-1,612672787
hsa-miR-106b-5p	-0,821860666	7,649102505	-3,120732613	0,002949297	0,025630792	-1,817682849
hsa-miR-410-3p	2,374426856	2,532897633	3,102436142	0,003106977	0,026373178	-1,955262239
hsa-miR-150-3p	-1,391743708	4,918311469	-3,092655807	0,003194476	0,026499626	-1,804308784
hsa-miR-589-5p	1,611823171	5,002666887	3,042056353	0,003685249	0,029891461	-1,893878327
hsa-miR-144-3p	-0,816187261	8,670425702	-3,023991315	0,003876942	0,030762688	-2,171291847
hsa-miR-323a-3p	1,709939127	5,848623727	3,001976861	0,00412314	0,032020126	-1,965197921
hsa-miR-122-3p	-2,766496698	3,072619853	-2,993108999	0,004226377	0,032138078	-2,197255307
hsa-miR-185-3p	2,311229079	2,352847101	2,949384996	0,004771607	0,035543606	-2,272734764
hsa-miR-1284	2,296867895	3,148564437	2,933093534	0,004991006	0,036434343	-2,265042655
hsa-miR-134-5p	2,577295914	2,872928941	2,906936407	0,005362989	0,038382175	-2,324453501
hsa-miR-323b-3p	1,823149249	6,189608397	2,860883521	0,006081159	0,042685061	-2,317137658
hsa-miR-5010-5p	1,981828426	3,631422281	2,848840671	0,006283147	0,04327073	-2,398088028

Supplementary Table 18: Differentially expressed miRNAs for true positive CRC patients with metastatic disease versus false positive CRC patients (healthy).

miRNA	logFC	AveExpr	t	P.Value	adj.P.Val
hsa-miR-142-5p	-0,758121654	12,28061237	-4,647375395	2,35229E-05	0,008585875
hsa-miR-16-5p	-0,527298842	14,7414419	-3,867055797	0,000309483	0,04489124
hsa-miR-143-3p	-0,861242486	12,26428624	-3,704878692	0,000515296	0,04489124
hsa-miR-10a-5p	1,301945804	12,2467303	3,674827669	0,000565713	0,04489124
hsa-miR-126-5p	-0,768854748	9,046003843	-3,592101585	0,000730086	0,04489124
hsa-miR-16-2-3p	-0,519897748	10,61792793	-3,588612082	0,000737938	0,04489124
hsa-miR-92b-3p	0,930361924	7,797164534	3,532367658	0,000876186	0,045686867

Supplementary Table 19: Differentially expressed miRNAs for true positive CRC patients with metastatic disease versus localized disease.

miRNA	logFC	AveExpr	t	P.Value	adj.P.Val	B
hsa-miR-375-3p	2,86503989	10,26476405	6,749822412	1,26906E-08	4,63207E-06	9,607767602
hsa-miR-484	-0,968992088	10,95168084	-6,06716186	1,5344E-07	2,80029E-05	7,265515729

hsa-miR-10a-5p	2,081024981	12,2467303	5,528308441	1,07266E-06	0,000130507	5,396306017
hsa-miR-1228-3p	4,051003354	3,990660622	5,334588305	2,14011E-06	0,000156845	2,937167011
hsa-miR-192-5p	1,830449954	9,180609279	5,333477988	2,14856E-06	0,000156845	4,777281961
hsa-miR-122-5p	2,55578042	14,61508189	5,281995364	2,57904E-06	0,000156891	4,513796403
hsa-miR-483-3p	3,414206586	6,404134527	4,695920592	1,99243E-05	0,001038908	2,172741694
hsa-miR-29a-3p	1,206293588	10,73818744	4,551485239	3,2597E-05	0,001487237	2,202567045
hsa-miR-200a-3p	3,465997337	4,866574171	4,494627868	3,95093E-05	0,001530252	1,331509038
hsa-miR-95-3p	2,808628418	4,115791539	4,477032544	4,19247E-05	0,001530252	0,901745275
hsa-miR-141-3p	4,208288498	2,224073379	4,430977061	4,89497E-05	0,001624242	0,551131259
hsa-miR-200b-5p	4,150171764	1,738451866	4,34535686	6,51837E-05	0,001946062	0,31774572
hsa-miR-1249-3p	2,094002549	6,168566659	4,326917505	6,93118E-05	0,001946062	1,278252597
hsa-miR-194-5p	1,405051065	8,569272577	4,245047167	9,09241E-05	0,002370521	1,31682598
hsa-miR-30a-5p	1,001619429	11,10190101	4,20739491	0,000102941	0,002504901	1,116525756
hsa-miR-210-3p	1,800614636	5,295307845	4,184081001	0,000111141	0,002535394	0,72339533
hsa-miR-6803-3p	1,537335453	7,137953939	3,993748352	0,000206399	0,004033432	0,587617442
hsa-miR-885-3p	3,817229233	2,102083348	3,985694286	0,000211819	0,004033432	-0,498003928
hsa-miR-27b-3p	0,913507616	11,33471254	3,982325318	0,000214127	0,004033432	0,41578504
hsa-let-7c-5p	0,960124114	8,985287488	3,972481989	0,00022101	0,004033432	0,493589685
hsa-miR-92b-3p	1,109425308	7,797164534	3,954870876	0,000233862	0,004064738	0,479599265
hsa-miR-125a-5p	0,888810795	11,98978333	3,868876534	0,000307699	0,005105009	0,053046362
hsa-miR-429	3,691249769	2,83812068	3,840085619	0,000337101	0,005349652	-0,602891449
hsa-miR-3605-5p	-2,749120112	2,243448026	-3,300179381	0,001754349	0,026680728	-1,756628496
hsa-miR-410-3p	-2,638457106	2,532897633	-3,27286162	0,001900603	0,027748801	-1,744064304
hsa-miR-1228-5p	2,279605731	4,037448717	3,214985921	0,00224937	0,031577696	-1,743031044
hsa-miR-1306-5p	0,917897482	7,163214442	3,173312791	0,002536964	0,033480689	-1,590939879
hsa-miR-21-5p	0,664671736	15,09252068	3,169033287	0,002568382	0,033480689	-2,06321988
hsa-miR-185-3p	-2,555249814	2,352847101	-3,103911577	0,003093974	0,0389414	-2,090584402
hsa-miR-99a-5p	0,82263466	10,04454709	3,053824647	0,003565181	0,043376369	-2,108862202

Hanna Sofie Kjemperud

Investigation of single cells in tumor tissue and microRNAs in serum of colorectal cancer patients

NTNU
Norwegian University of
Science and Technology